

Universidade Nova de Lisboa - ISEGI
Lisboa, Portugal

SAKWeb[©]

SPATIAL AUTOCORRELATION AND KRIGING WEB
A W3 GEOCOMPUTATION PERSPECTIVE

A thesis submitted in partial satisfaction
of the requirements for the degree of
Doctor of Philosophy in Information Systems
by
João Garrott Marques Negreiros

Chairperson: Marco Painho, Ph.D.

2004

Copyright © by
João Garrott Marques Negreiros
All Rights Reserved
2004

ABSTRACT

SAKWeb[®]

by

João Garrott Marques Negreiros

The present research focuses on the first software to offer spatial autocorrelation and association measures, spatial exploratory tools, variography and Ordinary Kriging (OK) in the World Wide Web. Exploiting IE[®], ASP[®], PHP[®] and IIS[®] capabilities, SAKWeb[®] version 2.0 was designed in an attractive and straightforward way for everyone's use. Major commercial off-the-shelf spatial software is reviewed to confirm SAKWeb[®] improvements. As confirmed by Green and Bossomaier [2002], "it is essential for geographical studies to provide basic spatial autocorrelation indices and significance test computer programs for personal computers".

Spatial autocorrelation can provide an Exploratory Spatial Data Analysis (ESDA) tool for atypical locations, trends, transition regions and spatial patterns. The weight matrix (**W**) is assessed on the basis of the covariogram contiguity. In accordance with various distance ranges, SAKWeb[®] selects the one with the highest global Moran I to setup the conventional and the Moran location scatterplot. In addition, the Moran I correlogram may lead to a re-estimation process of the best variogram range, sill, model and nugget-effect factors. Some of the possibilities are dropping outliers, tracing patterns of quadrants I and III, highlighting transition samples or maintaining the original dataset. IDW is also stressed as a poor method for **W** specification. Nearest neighborhood analysis and brushing operations are added features for enhanced dataset familiarity.

SAKWeb[®] presents the Moran variance scatterplot because the conventional one is not sensitive to neighborhood variance. The 2D view stresses outliers and stable and unpredictable regions. It can also be used with trend residuals to reveal heteroskedasticity.

The Web driven interface using hypermedia help with sound and movement is also provided in an e-Learning structure that includes an IRC, Email and News service, to close the knowledge gap between experts and common GIS users. This requirement results from two spatial autocorrelation and Kriging knowledge surveys which were carried out in several prestigious Portuguese institutions, and revealed a vast lack of awareness of both issues.

The SAKWeb[®] approach regarding the nugget-effect is significant. Four strategies are presented:

- $\gamma(0)=0$;
- $\gamma(0)=0$ but including C_0 for distances greater than zero;
- $\gamma(0)$ division into two scales: $\gamma_1(0)=0$ and $\gamma_1(0 < \mathbf{h}_1 \leq \text{Shortest Sampling Interval}) = \text{the Spherical, Exponential or Gaussian model whose sill equals the second long range structure, } \gamma_2(\mathbf{h})$, at SSI lag distance;
- If measurement error is given then this factor will be included in the Kriging system but it will lead to a non-exact Kriging version.

“The variogram sill does not reflect the true global variance, which is generally lower, because the variogram function poorly describes the short scale variability” [Isaaks, Srivastava, 1989]. Thus, if the estimated global mean (EGM) and the estimated global variance (EGV) indices are assessed on the basis of the nearest neighbor analysis then a sill rescaling operation is possible for OK variance and region confidence interval enhancement (both local and global). A new classical local spatial variance index that is both data and geometry dependent is introduced. The assessment of combined Kriging estimates weighted by the OK variance is also accomplished. The Colorado (USA) grasshopper datasets of 1993, 1994, 1995 and 1997 are also analyzed and presented by SAKWeb[®] with remarkable conclusions between the Moran I correlogram and the variogram parameters.

RESUMO DA DISSERTAÇÃO

SAKWeb[®]

por

João Garrott Marques Negreiros

A presente investigação introduz o primeiro geosoftware Web relativo a medidas de autocorrelação espacial e associação, ferramentas de análise exploratória de dados, variografia e Krigagem ordinária. Através do IE[®], ASP[®], PHP[®] e IIS[®], o SAKWeb[®] versão 2.0 foi desenvolvido de um modo atractivo e directo para todo o tipo de utilizadores. Assim, a maior parte dos geosoftware comerciais foram re-analisados de modo a confirmar a sua inovação e presente desenvolvimento. Confirmado por Green e Bossomaier [2002], “é essencial nos estudos geográficos providenciar software relacionado com os índices básicos de autocorrelação espacial”.

A autocorrelação espacial permite fornecer um conjunto de ferramentas exploratórias para dados espaciais sobre a localização de observações atípicas, a tendencia geral da distribuição espacial, as regiões de transição e os padrões espaciais. A matriz de pesos (**W**) é calculado com base na contiguidade do covariograma. De acordo com várias distancias, o SAKWeb[®] seleciona a amplitude baseado no valor mais alto indicado pelo Moran I e que servirá de base ao Moran Scatterplot convencional e de localização. Uma outra vantagem deste método iterativo é a sua capacidade de indicar pistas sobre os parâmetros principais do variograma (modelo, efeito de pepita, amplitude e patamar) num processo de re-estimação. Rejeitar outliers, detectar os padrões do I e III quadrante, enfatizar as observações localizadas em regiões de transição ou manter o conjunto das amostras iniciais são algumas das possibilidades. O método IDW é indicado como um método pobre e ineficaz na especificação da matriz W. Igualmente, a análise de índices de vizinhança mais próxima e operações de sub-selecção de amostras em tempo real são características enaltecidas pelo SAKWeb[®].

SAKWeb[®] estabelece uma nova ferramenta, a variância do Moran Scatterplot, pois a tradicional não é sensível ao segundo momento de uma distribuição. A sua visão 2D sublinha ‘outliers’, regiões estacionárias e áreas instáveis podendo também este ser utilizado para testemunhar a variância dos resíduos resultantes de uma operação de “detrend”.

A interface Web usando hipermédia no seu menu de Help numa estrutura de e-Learning incluindo IRC e serviço de News é também importante. De facto, os resultados dos inquéritos sobre geoestatística realizado em várias instituições prestigiadas Portuguesas revelou uma grande falta de conhecimentos e sensibilidade associado a este assunto.

A metodologia associado ao SAKWeb[®] relativo ao efeito de pepita é significativa. Quatro estratégias são apresentadas:

- $\gamma(0)=0$;
- $\gamma(0)=0$ mas incluindo o efeito de pepita para distâncias maiores que zero;
- $\gamma(0)$ é dividido em duas estruturas embricadas: $\gamma_1(0)=0$ e $\gamma_1(0 < \mathbf{h}_1 \leq \text{Distância Mínima Amostral}) = \text{Esférico, Exponencial ou Gaussiano cujo patamar é igual à segunda macro-estrutura, } \gamma_2(\mathbf{h})$, na distância DMA;
- Se o erro de amostragem for considerado então este factor é incluído no sistema linear de Krigagem mas originando uma versão de interpolador não exacta.

“O patamar do variograma não reflecte a variância global das observações, geralmente mais baixa, porque o variograma descreve pobremente a variabilidade local” [Isaaks e Srivastava, 1989]. Assim, a média e a variância global estimada pela análise da vizinhança mais próxima é calculado. Numa operação de re-escalamento do patamar é possível aperfeiçoar a variância da Krigagem ordinária e do intervalo da região de confiança. Um novo índice de incerteza, the spatial local classical variance, baseado na dependência geométrica e valores das amostras é desenvolvido. Finalmente, quatro conjunto de observações de pragas de gafanhotos no estado do Colorado, USA, é analisado pelo SAKWeb[®] com algumas conclusões notáveis.

RÉSUMÉ ANALYTIQUE

SAKWeb[®]

par

João Garrott Marques Negreiros

La recherche présentée ici traite du premier logiciel Web appliqué à la géographie qui a pour objet les mesures d'autocorrélation spatiale et d'association, les outils d'analyse exploratoire de données, la variographie et le Kriging ordinaire (OK). Au moyen de l'IE[®], ASP[®], PHP[®] and IIS[®], la version 2.0 du SAKWeb[®] a été élaborée d'une manière attrayante et directe pour tout type d'utilisateurs. Ainsi, la majeure partie des logiciels destinés aux géographes et disponibles dans le commerce a été ré-examinée afin de pouvoir confirmer le caractère innovateur du logiciel SAKWeb[®] et son développement actuel. Comme cela a été confirmé par Green and Bossomaier [2002], "il est indispensable pour les études géographiques de disposer d'un programme traitant des indices de base d'autocorrélation spatiale".

L'autocorrélation spatiale permet de fournir un ensemble d'outils d'investigation de données spatiales au sujet de la localisation d'observations atypiques, des tendances générales de la distribution spatiale, des régions de transition et des modèles spatiaux. La matrice de poids (**W**) est évaluée sur base de la contiguïté du co-variogramme. En fonction des diverses distances, le SAKWeb[®] sélectionne l'amplitude en se basant sur la valeur la plus haute indiquée par le Moran I et qui servira de base au Scatterplot Moran conventionnel ainsi que de localisation. Un autre avantage de cette méthode répétitive est sa capacité d'indiquer des pistes au sujet des paramètres principaux du variogramme (modèle, effet de pépité, amplitude et seuil) dans un processus de ré-évaluation. Le rejet des observations aberrantes (outliers), la détection des modèles de quadrant I et III, la mise en évidence des observations localisées dans des régions de transition ou le maintien de l'ensemble des échantillons initiaux sont quelques-unes de possibilités qu'offre le SAKWeb[®]. La méthode IDW est considérée comme étant limitée et inefficace pour la spécification de la matrice W. De même, l'analyse des indices de voisinage

le plus proche et les opérations de sous-sélection des échantillons en temps réel sont des caractéristiques supplémentaires du SAKWeb[®].

Le SAKWeb[®] crée un nouvel outil, la variance du scatterplot Moran parce que le scatterplot conventionnel ne réagit pas aux variations de voisinage au second moment d'une distribution. La vision 2D du SAKWeb[®] met en évidence les observations aberrantes (outliers) ainsi que les régions stationnaires et les aires instables. Elle peut aussi être utilisée pour prouver la variation des éléments résiduels résultant d'une opération de 'detrend'.

L'interface de réseau (web) est aussi importante. Elle utilise l'hypermédia dans son menu Aide incorporé dans une structure d'apprentissage électronique (e-Learning) comprenant un IRC et un service de News. En fait, les résultats des enquêtes sur la géostatistique réalisées dans diverses institutions portugaises prestigieuses ont révélé un important manque de connaissances et de sensibilisation à ce sujet.

La méthodologie associée au SAKWeb[®] en ce qui concerne l'effet de pépite est importante. Quatre stratégies sont présentées:

- $\gamma(0)=0$;
- $\gamma(0)=0$ mais comprenant l'effet de pépite pour des distances supérieures à zéro;
- $\gamma(0)$ est divisé en deux échelles reliées entre elles: $\gamma_1(0)=0$ et $\gamma_1(0 < \mathbf{h}_1 \leq \text{distance d'échantillonnage la plus courte, SSI}) = \text{Sphérique, Exponentielle ou du modèle de Gauss dont le seuil est égal à la seconde macrostructure, } \gamma_2(\mathbf{h})$, à distance SSI (valeur de la force du signal);
- Si l'erreur d'échantillonnage est prise en compte, alors ce facteur est inclus dans le système linéaire du Kriging mais cela créera une version inexacte d'interpolateur.

“Le seuil du variogramme ne reflète pas la véritable variance globale des observations, qui est généralement inférieure, parce que le variogramme décrit mal la variabilité locale” [Isaaks, Srivastava, 1989]. Ainsi sont calculées la moyenne globale estimée et la variance globale estimée sur base de l'analyse de voisinage la

plus proche. Par une opération de ré-échelonnement du seuil, il est possible de perfectionner la variance de l'opération de Kriging ordinaire et de l'intervalle de la région de confiance. Ainsi est développé un nouvel indice d'incertitude, la variance spatiale locale classique, qui est basé sur la dépendance géométrique et les valeurs des échantillons. Enfin, quatre ensembles d'observations d'invasions de sauterelles dans l'état du Colorado (États-Unis) sont analysés par le SAKWeb[®] et ont conduit à quelques conclusions remarquables.

ACKNOWLEDGMENTS

This document represents the culmination of a process which required a tremendous amount of dedication, not only from the author but also from a number of people who assisted in producing it in countless ways. In recognition of his unflagging enthusiasm and constant encouragement, I would like to thank Professor Marco Painho, Ph.D., chairperson of my thesis committee, and my advisor, for his knowledgeable guidance during the course of my doctoral studies at ISEGI, Universidade Nova de Lisboa. I would like to extend my appreciation to Professor Jorge de Sousa, Ph.D., from IST, Universidade Técnica de Lisboa, for leading me to this motivating research field and to give special thanks to Professor Pedro Coelho, Ph.D., for his helpful statistical suggestions. The kind support provided by ISEGI, in particular Dr. Francisco Marques, and CIIST computer center and Professor Pierre Goovaerts, Professor Ana Costa and Professor Fernando Bação's criticism are gratefully acknowledged. Thanks to Professor Archie Collins for the English review. I am also grateful to my parents who believed in my ability to get a Ph.D. long before I did. Finally, I would like to thank my eternal irreplaceable friend Paulo Pereira for his valuable Web graphics help and my wife Claudia for clearing my brain of the dust accumulated by spending too many hours behind a computer screen. Certainly, I could not forget my son Daniel for waking me up every night for a diaper change while Dad finished school. This endeavor is dedicated to them.

The research presented in this thesis was partially funded and supported by *Fundação Ciência e Tecnologia* (Praxis XXI/BD/11233/97), Lisboa, Portugal.

“The emphasis is on where, the jurisdiction of spatial statistics.”

J. Berry

“What makes spatial statistics different from statistics? Lots and lots of computer programming.”

Unknown author

I am sorry for this lengthy thesis but I did not have the time to write a smaller one.

Adaptation from Padre António Vieira

TABLE OF CONTENTS

1. Introduction	27
1.1. Statement of the Problem	27
1.2. Research Objectives	32
1.3. Constraints and Limitations	34
1.4. Hypothesis	35
1.5. Methodology	35
1.5.1. Spatial Analysis: The Missing Data Issue	37
1.5.2. Spatial Analysis: Interpolation, Implementation Review and Software Evaluation	38
1.5.3. Spatial Analysis: User Know-how	40
1.5.4. SAKWeb [®]	40
1.6. Relevance	42
1.7. Organization	47
2. Spatial Analysis: The Missing Data Issue	49
2.1. Geography and Geographical Data	49
2.2. GIS and Spatial Analysis: An Overview	52
2.3. Spatial Autocorrelation	56
2.4. Visualizing What Is Not Observed	62
2.4.1. The Variogram Cloud and the Cressie Box Plot	63
2.4.2. The Spatial Lag Scatterplot	64
2.4.3. The Spatial Lag Pies	64
2.4.4. The Moran I	65
2.4.5. The Geary C	66
2.4.6. The Local $G_i(d)$ and the Global $G(d)$	67
2.4.7. The Local Moran	69
2.4.8. The Moran Scatterplot	69
2.5. Spatial Autoregression Models	70
2.5.1. Spatial Autoregression: Residuals	72
2.5.1.1. LADS	75

2.5.1.2.	Join-Count Statistics	76
2.6.	Spatial Autocorrelation, Autoregression and Kriging: The Missing Data Issue	77
2.7.	W Matrix: IDW versus Binary Contiguity	80
2.7.1.	Case Study 1: Appendix I	81
2.7.2.	Case Study 2: Appendix J	84
2.7.3.	Discussion	86
2.8.	Conclusions	89
3.	Spatial Analysis: Interpolation	91
3.1.	Spatial Deterministic Interpolators	91
3.1.1.	Drawing Boundaries and TIN	91
3.1.2.	Global Polynomial Trend	92
3.1.3.	Fourier	94
3.1.4.	B-Spline	95
3.1.5.	Inverse Distance Weighting	95
3.2.	The Nature of Regionalized Data: Introduction	97
3.3.	Exploratory (A)Spatial Data Analysis	104
3.3.1.	Distribution Standardization	106
3.3.2.	Sampling Considerations	110
3.4.	The Variogram	112
3.4.1.	SAKWeb [®] Isotropic Variogram Models	119
3.4.2.	Computational Variogram Issues	121
3.5.	Ordinary Kriging	124
3.5.1.	The Cross-Validation Procedure	126
3.6.	Assessing Uncertainty	127
3.6.1.	The Classical Local Spatial Variance	131
3.7.	Conclusions	135
4.	Spatial Analysis: Implementation Review	137
4.1.	Statistical Packages	137
4.2.	GIS Modules	140
4.3.	Independent GeoSoftware	142
4.4.	Spatial Analysis Versus Topology Linkage: A Need	144

4.5.	e-Learning and Interfaces	146
4.6.	Conclusions	149
5.	Spatial Analysis: Software Evaluation	151
5.1.	Surface III[®]	151
5.2.	VarioWin[®]	152
5.3.	GS+[®]	153
5.4.	GeoMS[®]	154
5.5.	REGARD[®]	155
5.6.	GMS[®]	157
5.7.	GSLib[®] and GsTL[®]	158
5.8.	Kriging For Berkley Unix[®]	159
5.9.	Uncert[®]	160
5.10.	GEO-EAS[®]	161
5.11.	Surfer[®]	162
5.12.	EcoSSe[®]	163
5.13.	GeoPack[®]	164
5.14.	ArcGIS Geostatistical Analyst[®] Module	164
5.15.	Kriging ArcScripts[®]	166
5.16.	SpaceStat[®]	167
5.17.	Conclusions	171
6.	Spatial Analysis: User Know-how	173
6.1.	Modest Expertise in Sophisticated Spatial Analysis	173
6.2.	Statistical and Geographical Knowledge: The Perfect Marriage	178
6.3.	Spatial Autocorrelation and Kriging: The Web Solution	180
6.4.	Conclusions	188
7.	SAKWeb[®]	189
7.1.	SAKWeb[®] Technologies	192
7.1.1.	Internet Information Server[®]	192

7.1.2.	From CGI to ASP [®] and PHP [®]	195
7.1.3.	JavaScript [®]	200
7.1.4.	ActiveX [®] and Component One WebChart [®]	202
7.1.5.	Java Applets [®]	208
7.1.6.	DHTML	211
7.2.	SAKWeb[®] Developments	212
7.2.1.	The SAKWeb [®] Variogram - Moran I - Moran Scatterplot Optimal ESDA	212
7.2.1.1.	Case Study: The San Diego Dataset	213
7.2.1.2.	Case Study: The Africa Dataset	214
7.2.1.3.	Log Transformation Impact on Moran I	215
7.2.1.4.	Moran I as a Scale Pattern Factor for the Moran Scatterplot	216
7.2.1.5.	The Issue of True and False Outliers	219
7.2.1.6.	The Undercover Uncertainty Region of the Conventional Moran Scatterplot	221
7.2.1.7.	The SAKWeb [®] Moran Location Scatterplot	223
7.2.2.	The SAKWeb [®] Moran Variance Scatterplot ESDA	228
7.2.2.1.	Case Study: The San Diego Dataset	231
7.2.2.2.	Case Study: The GeoEAS [®] Lead Dataset	232
7.2.2.3.	Case Study: The Organic Dataset for the Soil in Nebraska	233
7.2.3.	SAKWeb [®] Variogram Sill Rescaling	235
7.2.3.1.	Case Study: The San Diego Dataset	238
7.2.3.2.	Case Study: The Africa Dataset	238
7.2.3.3.	SAKWeb [®] Global Confidence Interval	239
7.2.3.4.	SAKWeb [®] Local Confidence Interval	241
7.2.4.	SAKWeb [®] Nugget-effect	243
7.2.4.1.	Case Study: The GS+ [®] Pb Dataset	252
7.2.4.2.	Case Study: The San Diego Dataset	254
7.3.	SAKWeb[®] Functionalities	255
7.3.1.	Data Input and Exploring View	257
7.3.2.	Variography and Spatial Autocorrelation	264
7.3.3.	SAKWeb [®] (Ordinary) Kriging	268
7.3.4.	SAKWeb [®] Help	275
7.4.	SAKWeb[®] End-User Evaluation	282
7.5.	Conclusion	287
8.	SAKWeb[®]: Spatial Analysis of the Grasshopper in Colorado, USA	289
8.1.	The Grasshopper Ecology	290

8.2.	The Grasshopper Prediction Issue	293
8.3.	The Grasshopper Study Region	294
8.4.	SAKWeb [®] Exploratory Data Analysis	296
8.5.	SAKWeb [®] Spatial Autocorrelation	300
8.6.	SAKWeb [®] Variography	303
8.7.	The SAKWeb [®] Moran Location Scatterplot and Moran Variance Scatterplot	307
8.8.	SAKWeb [®] Ordinary Kriging	315
8.9.	SAKWeb [®] Validation Procedures	321
8.10.	Conclusion	328
9.	Summary and General Conclusion	331
9.1.	Summary	332
9.2.	Discussion of the Hyphoteses	334
9.3.	Limitations and Future Research	337
	Literature Cited	340
	Appendix A	367
	Appendix B	373
	Appendix C	376
	Appendix D	379
	Appendix E	380
	Appendix F	401
	Appendix G	403
	Appendix H	406
	Appendix I	409
	Appendix J	415
	Appendix K	419
	Appendix L	421

Appendix M	423

Appendix N	426

Appendix O	428

Acronyms and Mathematical Symbols	430

LIST OF FIGURES

Figure 1 – The problems in specifying parameters for area objects. _____	30
Figure 2 – The SAKWeb [®] overall framework. _____	36
Figure 3 – The missing data issue and spatial interpolation of SAKWeb [®] . _____	38
Figure 4 – The SAKWeb [®] organization of spatial analysis implementation review and software evaluation. _____	39
Figure 5 – The overall SAKWeb [®] development and technical implementation framework. ____	41
Figure 6 – The SAKWeb [®] technical implementation view. _____	42
Figure 7 – The SAKWeb [®] development view. _____	42
Figure 8 – Population gains and losses on the Portuguese electoral roll at town level. _____	51
Figure 9 – The spatial lag pies of a Columbus Ohio dataset [Anselin, 1998], where the red bars indicate neighborhood crime average and the blue ones represent crime in a particular ward. ____	65
Figure 10 – The Moran I and Geary C measures inferences. _____	67
Figure 11 – Residual plots where the x-axis represents the predicted values while the y-axis is the associated error [Clark and Hosking, 1986]. _____	74
Figure 12 – The Moran scatterplot exchanged observation position according to Moran I. ____	83
Figure 13 – The linear relationships of Moran scatterplots regarding case one and case six. ____	83
Figure 14 – A comparison of 1/d ^x contiguities for the impact of weight samples. _____	84
Figure 15 – The linear and quadratic polynomial matrices. _____	93
Figure 16 – The variogram and the covariogram function with and without the nugget-effect. _	102
Figure 17 – The long trend versus short range stationary [Soares, 2000]. _____	104
Figure 18 – The variogram. _____	113
Figure 19 - The four type of behavior near the origin. _____	114
Figure 20 – Geometric (left) and geometric with zonal (right) anisotropy. _____	118
Figure 21 – An example of the proportional effect [Sousa, 1998]. _____	119
Figure 22 – The spherical model. _____	120
Figure 23 – The exponential model. _____	120
Figure 24 – The gaussian model. _____	120
Figure 25 – The setup factors for the variogram construction with regard to the sector binning method. _____	122
Figure 26 – The Kriging system. _____	125
Figure 27 – The absolute true error against OK variance (left) and local variance (right). ____	133
Figure 28 – The absolute true error against classical local spatial variance (left) and OK variance (right). _____	134
Figure 29 – The number of users that use Kriging within GIS modules revealed by the first survey. _____	141
Figure 30 – The Kriging software awareness revealed by the September 2001 survey. ____	142
Figure 31 – The XGobi [®] menu within ArcView [®] . _____	144

Figure 32 – The train access time from Lisboa [Figueira, 2000].	145
Figure 33 – The Surface III [®] data postings.	151
Figure 34 – The Surface III [®] search neighborhood.	152
Figure 35 – The Regard [®] brushing and linking features.	156
Figure 36 – The Regard [®] brushing and linking features within the scatterplot view.	157
Figure 37 – The Kriging for Berkley Unix [®] OK interpolation and variance.	160
Figure 38 – The GEO-EAS [®] Vario options.	162
Figure 39 – The Kriging ArcScript [®] window.	167
Figure 40 – The SpaceStat [®] Regress module framework.	169
Figure 41 – The SpaceStat [®] Extension for ArcView [®] .	170
Figure 42 – Portuguese GIS users' acquaintance with spatial interpolation.	175
Figure 43 – Portuguese GIS users' knowledge of spatial autocorrelation software.	176
Figure 44 – The high-school teacher's questionnaire: final answers after fifty hours of five ArcView [®] modules.	178
Figure 45 – User agreement (%) concerning the linkage of spatial autocorrelation and GIS software.	182
Figure 46 – The client-side technology.	190
Figure 47 – The server-side technology.	190
Figure 48 – The home and authentication page of SAKWeb [®] that includes the keywords meta tag for search engines robots.	194
Figure 49 – An example of an SSI that returns the date and time of the Web server.	196
Figure 50 – The DOS Commands selection of SAKWeb [®] including the PHP [®] code.	199
Figure 51 – Partial view of PHP [®] configuration.	200
Figure 52 – The JavaScript [®] clock output.	202
Figure 53 – The SAKWeb [®] main menu.	205
Figure 54 – The SAKWeb [®] trend surface analysis of the GS [®] + Pb dataset.	208
Figure 55 – The independent structure of Java [®] language.	209
Figure 56 – The general covariogram contiguity plot.	213
Figure 57 – The Moran I correlogram for the San Diego dataset.	214
Figure 58 – The Pb dataset results with the Variogram - Moran I - Moran scatterplot optimal technique.	220
Figure 59 – The location samples of the transition regime.	222
Figure 60 – The Pb location samples of the transition regime.	222
Figure 61 – The optimal Moran scatterplot flowing chart of SAKWeb [®] .	226
Figure 62 – The SAKWeb [®] view concerning the quadrant I and II of the San Diego dataset.	227
Figure 63 – The SAKWeb [®] view concerning the quadrant III and IV of the San Diego dataset.	228
Figure 64 – The heteroskedasticity question, a problem not resolved by the Moran scatterplot.	229
Figure 65 – The 2D view of the Moran scatterplot extension of Figure 64.	231
Figure 66 – The Moran variance scatterplot for the San Diego dataset.	232
Figure 67 – The Moran variance scatterplot for the GeoEAS [®] lead dataset.	233

Figure 68 – The Moran variance scatterplot of the first polynomial residuals for the Nebraska soil organic dataset. _____	235
Figure 69 – The weight assignment methodologies. _____	236
Figure 70 – The Global Region Confidence Interval option of SAKWeb [®] , a three step process. _____	240
Figure 71 – The three step process of the Local Region Confidence Interval option. _____	243
Figure 72 – The Atkinson [1997] internal and external error view. _____	245
Figure 73 – The Kriging interpolation of a Swedish glacier mass with and without the nugget-effect. _____	245
Figure 74 – The three exact nugget-effect approaches of SAKWeb [®] . _____	247
Figure 75 – OK comparisons between the exact version and the one with all measurement error (1 and 2 represents data locations). _____	248
Figure 76 – The OK system of the working example. _____	249
Figure 77 – The OK system of the working example, considering with and without measurement error. _____	249
Figure 78 – Comparison between the exact version of SAKWeb [®] OK and the one with 30% measurement error. _____	251
Figure 79 – The spatial location of the grasshopper densities in 1993 where the dark dots represent the sites whose OK estimates with a 30% measurement error are lower than the original samples. _____	252
Figure 80 – The SAKWeb [®] overall structure. _____	257
Figure 81 – The MS-Excel [®] data input selection. In addition, a movie was added for beginner's users. _____	258
Figure 82 – The visitor distribution graphic of SAKWeb [®] . _____	260
Figure 83 – The local interactive statistics option of SAKWeb [®] . _____	261
Figure 84 – The 3D scatter view of SAKWeb [®] . _____	263
Figure 85 – The SAKWeb [®] indicator mapping in a two step process. _____	264
Figure 86 – The SAKWeb [®] experimental variogram setup. _____	266
Figure 87 – A partial view of the SAKWeb [®] variogram fitness process. _____	267
Figure 88 – SAKWeb [®] OK Calculus option. _____	269
Figure 89 – Combination of Kriging estimates from OK with C0 (model 1), OK without C0 (model 2) and with two structures (model 3) for the grasshopper 1993 dataset. This includes differences with other SAKWeb [®] OK results. Notice that the OK variance difference among the four approaches: the combined version rounds 66% less when compared with the other three. _____	271
Figure 90 – Partial view of SAKWeb [®] OK assessment based on GSLib [®] routines. _____	272
Figure 91 – Partial view of SAKWeb [®] Cross-Validation option with its true-false/positive-negative analysis. _____	274
Figure 92 – Notice that the black line in the left image is not drawn by SAKWeb [®] . _____	275
Figure 93 – The SAKWeb [®] presentation movie in terms of its construction (left) and final output (right). _____	277
Figure 94 – The SAKWeb [®] Help. _____	279

Figure 95 – The message administration option of SAKWeb [®] .	281
Figure 96 – The SAKWeb [®] News option.	282
Figure 97 – The average classification for the impact of SAKWeb [®] features.	283
Figure 98 – The ranking for the Web solution.	285
Figure 99 – Map of the cities of Colorado, USA, as shown in the SAKWeb [®] Image Mapping option.	296
Figure 100 – Topographic map of Colorado, USA.	299
Figure 101 – The map of Colorado, USA.	307
Figure 102 – The SAKWeb [®] Moran location scatterplot of Colorado in 1993.	309
Figure 103 – The SAKWeb [®] Moran location scatterplot of Colorado for 1994.	311
Figure 104 – The SAKWeb [®] Moran location scatterplot of Colorado in 1995.	313
Figure 105 – The SAKWeb [®] Moran location scatterplot of Colorado for 1997.	315
Figure 106 – The SAKWeb [®] OK estimation and OK variance for a 100 km range.	322

LIST OF TABLES

Table 1 – Internet technology can radically reduce transaction costs in many industries [Spotlight, 1999].	43
Table 2 – The OLS regression of ANOVA and MLE spatial lag regression of Spatial ANOVA (to emphasize that the treatments are spatial in the sense of being subregions of the dataset) for a study of 49 wards in Columbus, Ohio [adapted from Anselin, 1992].	59
Table 3 – The theoretical expected degrees of freedom for the modified t-Student test for two autoregressive lattice processes in relation to the number of locations and autocorrelation parameters, A_{px} (x axis direction) and A_{py} (y axis direction) [Dutilleul, 1994].	59
Table 4 – The standard global G and I indices.	68
Table 5 – The SAR benchmark by Li [1996].	71
Table 6 – The average weights for the first four distance decay contiguities and the binary one.	86
Table 7 – The Moran I results according to distance decay contiguities.	87
Table 8– The number of direct neighbors according to the binary contiguity of some San Diego districts.	89
Table 9– The two geographical arrangements for two datasets.	99
Table 10– Univariate statistics.	105
Table 11– Bivariate statistics.	105
Table 12– Univariate spatial descriptors.	106
Table 13 – A global behavior comparison of previous variogram models [Geostatistics FAQ, 1997].	121
Table 14 – The major differences between Kriging and other interpolation methods.	135
Table 15 – Spatial data characteristics by Openshaw [1994].	139
Table 16– Internet economic indicators: The evidence.	149
Table 17 – Analysis of geostatistical software.	171
Table 18– The Kriging strategies for the same Nebraska (USA) dataset.	178
Table 19 – Moran I according to various covariograms ranges for the Africa GDP dataset.	217
Table 20 – The Moran variance scatterplot results for the five cases above.	230
Table 21 – The SAKWeb [®] OK estimations for the Pb dataset.	254
Table 22 – The SAKWeb [®] OK estimations for the San Diego housing costs.	255
Table 23 – The ASP [®] (server-side) code that produces the VBScript [®] (client-side) one.	262
Table 24 – Mortality among grasshoppers exposed to insecticide in laboratory and field trials [Onsager, 1978].	292
Table 25 – Costs and returns associated with control of grasshoppers (forage was valued at \$44 per metric ton and grasshopper density was 10 to 15 per m ²) [Onsager, 1978].	293
Table 26 – SAKWeb [®] aspatial descriptive results for the four grasshopper datasets.	297
Table 27 – Aspatial descriptive results according to each year and region.	298
Table 28 – SAKWeb [®] nearest neighborhood analysis of the four grasshoppers datasets.	300

Table 29 – SAKWeb [®] Moran’s I correlogram for the grasshopper dataset. _____	301
Table 30 – SAKWeb [®] variogram models for the four datasets. _____	304
Table 31 – The SAKWeb [®] overall variogram for the years of 1993, 1994, 1995 and 1997. ____	306
Table 32 – Main trends in the Moran I correlogram and the variogram parameters. _____	312
Table 33 – Comparison of EGV, classical variance and variogram sill results for the four years. 316	
Table 34 – SAKWeb [®] OK and OK variance with measurement error (30%) of grasshopper Colorado (1993). _____	317
Table 35 – SAKWeb [®] OK and OK variance with measurement error (20% with a rescaled variogram sill factor of 0.7602) for the grasshopper dataset, Colorado (1994). _____	318
Table 36 – SAKWeb [®] OK and OK variance with measurement error (20%) for the grasshopper, Colorado (1995). _____	319
Table 37 – SAKWeb [®] OK and OK variance with measurement error (20%) for the grasshopper, Colorado (1997). _____	320
Table 38 – According to the two variogram ranges, the error comparison for the 1993 and 1994 grasshopper datasets clearly demonstrates the appropriate hints of the Moran I correlogram. ____	322
Table 39 – The grasshopper cross-validation for 1993, 1994 (upper row), 1995 and 1997 (lower row). _____	323
Table 40 – The global confidence interval for a level of 80% and a cutoff value of 9 with an OK model without C0. _____	325
Table 41 – The local confidence interval for a level of 80% and a cutoff value of 9 with an OK model without C0. _____	326
Table 42 – The grasshopper analysis of cross-validation for 1993, 1994, 1995 and 1997 (OK model with C0). _____	328

LIST OF EQUATIONS

Equation 1 – The Chi-Square of statistical interdependency among regions and groups.	57
Equation 2 – The Cressie [1993] resistant variogram estimator.	63
Equation 3 – The spatial lag pies.	64
Equation 4 – The Moran I.	65
Equation 5 – The Geary C.	66
Equation 6 – The local $G_i(d)$.	68
Equation 7 – The global $G(d)$.	68
Equation 8 – The local Moran.	69
Equation 9 – The SAR model.	70
Equation 10 – The variance of SAR model.	70
Equation 11 – The alternative SAR model.	71
Equation 12 – The alternative SAR variance model.	71
Equation 13 – The Columbus, Ohio, OLS model by Anselin [1992].	72
Equation 14 – The MLE spatial lag model of Columbus, Ohio [Anselin, 1992].	72
Equation 15 – The error assessment of the multiple regression model.	73
Equation 16 – The OLS regression model of the Brown [1996] Rwanda case.	74
Equation 17 – The MLE spatial lag model of the Brown [1996] Rwanda case.	75
Equation 18 – The regression model of the US government expenditure.	75
Equation 19 – The new regression model of the US government expenditure.	76
Equation 20 – The standard negative residuals contiguity of the binary approach.	76
Equation 21 – The standard positive residuals contiguity of the binary approach.	76
Equation 22 – The standard negative-positive residuals contiguity of binary approach.	77
Equation 23 – The linear estimator.	78
Equation 24 – The distance decay model by Upton [1992].	80
Equation 25 – The clockwise rotation angle of the standard deviational ellipse.	85
Equations 26 and 27 – The shorter and the longer axes lengths of the standard deviational ellipse.	85
Equation 28 – The pyritic sulfur prediction formula.	97
Equations 29 and 30 – The stationary assumptions.	99
Equations 31 and 32 – The intrinsic hypothesis.	100
Equation 33 – The variogram definition.	100
Equation 34 – The variogram definition for intrinsic variables.	100
Equation 35 – The variogram estimator.	101
Equation 36 – The interpolation structure component.	103
Equation 37 – The Chi-square Normal distribution test.	107
Equation 38 – The Sichel optimal estimator for the lognormal arithmetic mean.	108
Equation 39 – The exponential OLS linear for the regression back (bias) transformation.	108
Equation 40 – The exponential OLS linear regression for the back (unbiased)	

transformation.	108
Equation 41 – The back (unbiased) transformation of LK.	108
Equation 42 –The G goodness of prediction index.	109
Equation 43 – The lognormal linear correlation coefficient between elevation and zinc.	110
Equation 44 – The relative variogram.	115
Equation 45 – The general relative variogram.	116
Equation 46 – The pairwise relative variogram.	116
Equation 47 – The geometric anisotropic variogram.	117
Equation 48 – The zonal anisotropic variogram.	118
Equation 49 – The spherical variogram.	120
Equation 50 – The exponential variogram.	120
Equation 51 – The Gaussian variogram.	120
Equation 52 – The search radius formula for irregular grid samples.	122
Equation 53 – The Cressie [1993] automatic fitting formula.	123
Equations 54 and 55 – The variogram OK system.	124
Equations 56 and 57 – The covariogram OK system.	124
Equation 58 – The Matheron OK system.	125
Equation 59 – The Universal Soil Loss Equation for the rainfall-rainoff impact.	128
Equation 60 – The estimation of the true population standard deviation.	129
Equation 61 – The population standard deviation confidence interval.	130
Equation 62 – The mean confidence interval.	130
Equation 63 – Estimation of the probability for a specific cutoff value.	130
Equation 64 – The general Kriging variance.	131
Equation 65 – The concise OK variance in terms of the variogram.	131
Equation 66 – The classical local spatial variance.	132
Equation 67 – The exponential variogram for the San Diego dataset.	213
Equations 68 and 69 – The neighborhood linear regressions for the San Diego dataset.	214
Equation 70 – The spherical variogram for the Africa population density dataset.	216
Equation 71 – The spherical variogram for the Africa GDP dataset.	216
Equations 72 and 73 – The SAKWeb [®] Moran variance scatterplot.	229
Equation 74 – The Geostatistical Analyst [®] factor to add concerning the outer rectangle boundary.	236
Equation 75 – The estimated global mean formula.	238
Equation 76 – The estimated global variance formula.	238
Equation 77 - The exponential variogram for the San Diego dataset.	238
Equation 78 – The new exponential variogram for the San Diego dataset.	238
Equation 79 – The new spherical variogram for the Africa GDP dataset.	239
Equation 80 – The new spherical variogram for the Africa population density dataset.	239
Equation 81 – The non-convexity property.	243
Equation 82 – The OK variance including measurement error.	248

SAKWeb [®] , a W3 Software of Spatial Autocorrelation and Kriging.....	26

Equation 83 – The combined OK estimates for three datasets.	270

Equation 84 – The combined OK variance estimates for three datasets.	270

1.Introduction

1.1. STATEMENT OF THE PROBLEM

The problem of statistical spatial analysis encompasses an expanding range of methods which address different spatial problems, from image enhancement and pattern recognition to spatial interpolation and socio-economic trend modeling. Each of these methods focuses on a particular aspect but what emerges is something that is clearly identifiable as spatial statistics, geographically raw data correlated by statistical methods.

Bação [1997], Anselin [1998] and Longley *et al.* [2001] argue that within specialized software there is a lack of spatial analysis techniques in general and of spatial statistics in particular. This may be a major impediment to GIS, which is a ‘data rich, theory poor’ environment. These authors put an emphasis on creativity, comparing spatial analysis to art and not to science. Applied geostatistics is a talent science because its results may be widely diverse depending on several factors such as the variogram modeling, the Kriging approach, the sampling design or the scientist’s previous experience.

“Very few existing GIS packages have truly analytical functionality” [Levine, 1997, Wong and Lee, 2001]. There has been little progress in incorporating advanced spatial techniques into commercial GIS products in comparison to simple analytical functions. “Most of the functionalities should be more properly regarded as data manipulation rather than analysis” [Openshaw, 1998]. These functions are quite deterministic and include routing, site location, data summarization or vector-raster conversion. Future GIS issues should be viewed as questions of probabilistic thinking, e.g. the Modifiable Areal Unit Problem and Monte Carlo error propagation during an overlay operation.

“At present, most of the literature on statistics for geography should be seen for what it really is: a series of introductions to statistical methods that are of geographical interest only because they have been applied to spatial datasets” [Openshaw 1994, Christensen, 1999]. Such applications are aspatial since they ignore the special characteristics of spatial data such as spatial dependence (autocorrelation) and spatial heterogeneity (association). Hence, it is analysis of spatial data, not spatial analysis of data. It is essential to implement spatial analysis tools that are simultaneously relevant to GIS users and already available in research.

Every year, dozens of research articles are printed with new spatial analysis methods, involving ever more fields such as stochastic simulation, multivariate geostatistic and spatial correlation. Regrettably, the majority of those ideas end up on the bookshelf without any practical consequence for the spatial analysis user. Algorithms, physical realities and thoughts exist for every person. A theory only succeeds if it leads to a practical purpose. “The spirit of geocomputation is fundamentally about matching the philosophy of science with practice, geometry with application, analysis with local context and technology with environment” [Longley, 1998]. To materialize research findings to the real world is a major dilemma nowadays since some of these methods are relatively crude. They have theoretical and implementation problems that remain to be solved. The validation of such findings by peers is a difficulty as well.

Geographers and major GIS users do not think in spatial terms and do not ask questions about spatial pattern and spatial autocorrelation. The goal of most users is to handle large volumes of spatial data and produce beautiful maps rather than carrying out sophisticated statistical spatial analysis. This leads to a corresponding lack of pressure for such methods to be made available.

Since the space aspect of Tobler’s Law¹ is the rule for Mother Nature rather than the exception [Clark and Harper, 2000], the classical and convenient assumption of data independency is in the line of fire and the most common statistical procedures must be rethought. The strength of temporal interaction follows the same principle

too (the time aspect of Tobler's Law). Activities performed on a succeeding day present a higher autocorrelation than yearly activities. Hence, models that involve statistical dependence are often more realistic.

Because spatial autocorrelation is often present, there is a need for spatial analysis. The inexistence of the assumption of spatial independence would create a chaotic basis for natural phenomena. "Hell might be a world without spatial dependence since it would be impossible to live there in any practical and meaningful way" [Longley *et al.*, 2001]. However, random spatial autocorrelation is difficult to detect visually because patches display similar information while others contain fragmented locational information. Most real world patterns are somewhere between a random and a clustered pattern. "Very rarely would we find a spatial pattern that is extremely clustered, extremely dispersed or purely random" [Wong and Lee, 2001].

Some measures of spatial autocorrelation such as the Moran I index try to summarize a complete spatial distribution in a single number. This can lead to a serious flaw: with large spatial datasets, this global measure of spatial autocorrelation can hide spatial heterogeneity or local patterns of spatial clustering.

Spatial residual autocorrelation measurement is crucial although it might not be easy to carry out because most available techniques rely on topological concepts and topology normally stored in a GIS. The common approach consists of using an intermediate application to transfer topological information from the GIS to the statistics application. Naturally, in this situation it is necessary to switch back and forth between the GIS and the analytical tool. On the other hand, "it has been accepted that in the presence of a great number of data points the computation is very intensive and it is probably better handled with specialized software than by integration into the GIS" [Cassetari, 1993].

Neighborhood definition can be difficult for the computation of spatial autocorrelation indices. Mayhew [1999] presents two divergent views: "...where inhabitants recognize each other by sight" versus "others deny the existence of

identifiable subculture areas...neighborhoods do not really exist”. Moran I, Geary C, Lisa, the spatial lag pies and the Moran scatterplot embrace consequences of this constraint. In the same way, spatial autocorrelation and association indices reduce vicinity to a single mean (first moment) but this they do not consider the variance (second moment) regarding the neighborhood distribution. Time, cost and convenience represent different aspects to be considered as well.

Concerning software implementation of spatial statistics, several approaches have been made, e.g., specialized software, commercial GIS modules or direct linkage between GIS and statistical packages. From the programmer’s perspective, the most important potential benefit between the latter two approaches involves direct access to spatial query and spatial objects while, at the same time, deriving spatial properties and relationships such as area, distance, adjacency or distance along a network. When spatial objects are areas, spatial properties are essential to compute spatial indices (namely the **W** weight matrix). However, Figure 1 reveals three problematic situations: the area A centroid is a non-neighbor of area C although its distance is closer than d while area E has no neighbors. The U-shaped region R has its centroid on polygon S. Finally, the four corners of the United States raise neighborhood questions such as whether Utah is a natural neighbor of New Mexico.

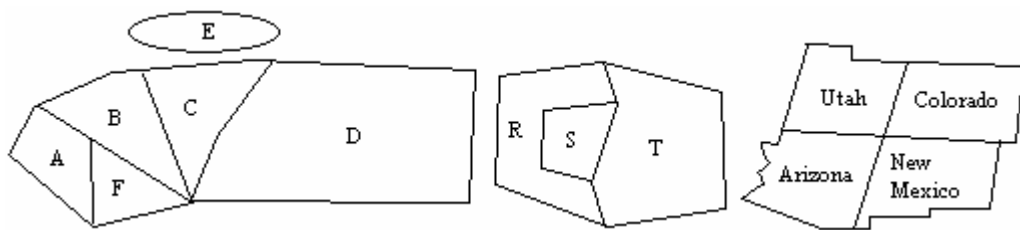


Figure 1 – The problems in specifying parameters for area objects.

Interpolation quality is a result of the number and distribution of the known points used and how well the mathematical function correctly models the phenomena. Thus, the choice of the appropriate model is essential if reasonable results are to be obtained. For example, with a low inverse distance exponent or high radius, each

local mean moves toward the overall sample mean, leading to a smoother density surface. In contrast, the density surface becomes relatively rough.

Major physical functions and chemical laws are deterministic and based on mathematical expressions. On the contrary, spatial interpolation, an intelligent guesswork process, must include error, autocorrelation, anisotropy and uncertainty constraints where observations are the result of a random process with spatial dependence.

Coined by Matheron at the Institute for Mathematical Morphology in Fontainebleau, Kriging takes into account the trend of each vicinity point for prediction but emphasizes the apparently contradictory aspects of regionalized variables: a random local process versus a structured large-scale one. “For this author, randomness is a fluctuation, not an error of the phenomena with its own structure” [Armstrong, 1998]. Once the structural analysis is made, the estimation can be achieved more accurately using the rate at which the variance between points changes over space. This distinctive feature is expressed by the variogram. However, it assumes that a surface holds a constant mean, a situation almost never true [Nalder and Wein, 1998]. Spatial autocorrelation exists and samples follow a Gaussian distribution. The more severely these assumptions are violated, the less accurate interpolation results become.

The variogram estimation includes a set of parameters² to be setup by the expert user for implementation. According to Kelly [1996], “there is little to be gained by making Kriging and spatial autocorrelation a main tool of spatial analysis unless the GIS user sophistication is sufficient for this information to be preferred”. Major Kriging software offers a wide range of options but its interpretation requires a thorough background. Certainly, an essential foundation of GIScience is the availability of spatially-aware professionals. Finding the best variogram fitness procedure (interactive or automatic) can also be a problem.

Because estimation error cannot be avoided, Ordinary Kriging variance error is essential for assigning confidence intervals. There is a discrepancy between the

predicted error variance and the real one since in most cases the variogram sill overestimates global sample variance. However, cell declustering and polygonal methods present difficulties in weighting the sample contribution, particularly for the samples on the edge. Outlier negative influence on the variogram and non-sensitivity to data of Ordinary Kriging variance are also real.

Another concern is the assumption of precise and accurate sample values by considering the zero nugget-effect. It is quite common to have a discontinuity at the variogram origin although major Kriging software forces it to be zero at zero distance. On the other hand, if the variogram for zero distance equals total nugget-effect then the estimation Kriging variance becomes lower, leading to a nonsense situation.

1.2. RESEARCH OBJECTIVES

The research thesis presented here develops the first non-commercial application service provider dealing with spatial autocorrelation measurement, Ordinary Kriging with the Internet Explorer[®] interface and hypermedia help in an e-Learning context.

The research objectives proposed are:

- To discuss spatial statistical analysis implementation in software and classical statistical limitation with regard to spatial data.
- To investigate and compare IDW and binary approaches regarding **W** matrix construction on Moran I and the Moran scatterplot.
- To be aware of variography and theoretical Kriging concepts.
- To analyze nugget-effect, range, model and sill rescaling factors in Kriging in conjunction with spatial autocorrelation measures.
- To evaluate Kriging software features at present and to explore Web programming technologies available nowadays.

- To create the first spatial autocorrelation and Ordinary Kriging Web software with the following major characteristics:
 - To build a Web interface where simplicity is fundamental.
 - To offer a hypermedia help facility in an e-Learning framework.
 - To investigate spatial distribution measures like the standard distance deviation and the nearest neighbor index.
 - To include brushing with updated local statistical results and indicator maps (based on the conventional and nearest weighted cumulative histogram) for the original samples.
 - To compute the first trend surface analysis³ with the join count statistics of its residuals. This includes the location discrimination of positive and negative residuals.
 - To plot the Moran I correlogram.
 - In accordance with the highest Moran I and variogram contiguity⁴ function, to create the conventional and Moran location scatterplot in a dynamic mapping structure and dataset renewal. The Moran scatterplot correlogram may also lead to a re-estimated process of the variogram parameters. The IDW deterministic interpolation is also based on the same parameters.
 - To plot the Moran variance scatterplot on the basis of the mean variance among neighbors and between the center location and its neighbors.
 - To monitor interactive variogram fitness including the azimuth angle factor, anisotropy map, major and minor range, main direction, correlogram, madogram, general relative variogram, pairwise relative variogram, spatial structure proportion and variogram fitness indicator.
 - To handle the nugget-effect of four available strategies including differences in results.
 - To combine the Kriged estimates by their estimation variances on the basis of the three exact OK models.
 - To rescale the variogram function for a higher Ordinary Kriging variance and global and local assessment of area confidence interval.

- To map validation with an extra dataset and cross-validation. This includes the location of true-false positive and negative samples based on the latter analysis.
- To layout an image file with the site mapping in an additional window.
- To delineate interactively the profile of any OK 3D surface to 2D.
- To launch a new menu for other forms of Kriging such as SK, UK, IK and CK.
- To analyze the Colorado, USA, grasshopper case with SAKWeb[®].

1.3. CONSTRAINTS AND LIMITATIONS

The present research will be based on the next constraints and limitations:

- Samples hold the same support (the affine and the indirect lognormal correction are not taken into account).
- Observation values are not negative.
- Only one variable at a time is estimated.
- GLS, WLS, Bayesian and stepwise regression, logbit models and the minimum Chi-square method for the Bernoulli and Poisson probability functions, time-series residual correlation, non-parametric tests for data sample comparison such as the Durbin-Watson test, Kruskal-Wallis H, Chi-Square and U Mann-Whitney, non-stationary regionalized variables and Bayesian interpolation, Grubb's outlier test and geostatistical simulation models are not included in the scope of the research.
- Any type of artificial intelligence (expert systems, genetic algorithms, fuzzy logic, neural networks and hybrid systems) and temporal space analysis are not covered.

1.4. HYPOTHESIS

The research will be based on the following premises:

- There is a lack of knowledge regarding spatial autocorrelation and Kriging issues among Portuguese GIS users.
- Spatial autocorrelation and Kriging software within W3 with hypermedia help in an e-Learning context contributes to the educational sharing and effortless accessing of spatial analysis tools among common users.
- IDW is not a good process for the **W** specification because it is not sensitive to its natural neighbours such as binary or standard deviational ellipse.
- It is possible to create an uncertainty measure that is both data and geometry dependent.
- A spatial neighborhood pattern can be found regarding which nugget-effect approach fits best. It is expected that in stable and transition regions, a low and a high C_0 , respectively, will fit best.
- By implementing the variogram sill rescaling based on nearest neighborhood analysis enhances the region confidence interval centered on the Ordinary Kriging variance.
- The Variogram - Moran I - Moran scatterplot optimal SAKWeb[®] approach helps users as an ESDA tool. In addition, it can help the estimation of variogram parameters in a continuous process.
- The Moran variance scatterplot uncovers observations whose variability is quite different. It can also be used with residuals to locate heteroskedasticity in a detrend Kriging process and particularly significant to validate the stationary assumptions.

1.5. METHODOLOGY

The methodology used is presented in Figures 2, 3, 4, 5, 6 and 7. They give an overview of the methodology, including five major topics:

- 1) Spatial analysis: The missing data issue;
- 2) Spatial analysis: Interpolation;
- 3) Spatial analysis: Implementation review and software evaluation;
- 4) Spatial analysis: User know-how;
- 5) SAKWeb[®] structure and functionalities, including an evaluation survey, enhancements and a spatial data analysis of the Colorado (USA) grasshopper dataset.

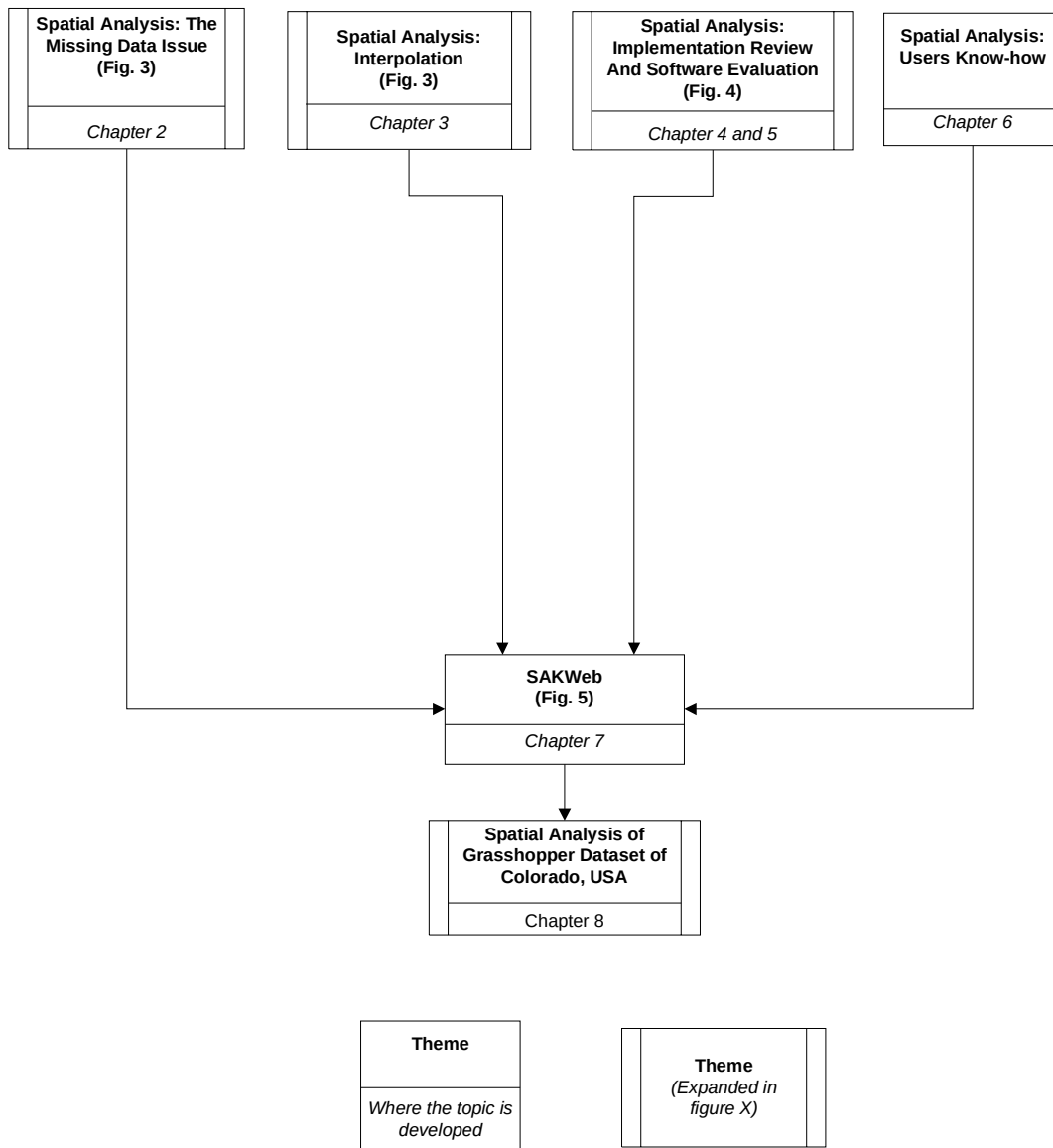


Figure 2 – The SAKWeb[®] overall framework.

1.5.1. Spatial Analysis: The Missing Data Issue

According to Rogerson, Fotheringham [1996] and Walter [2001], “there is an increased demand for systems that do more than display and organize data”. The set of potential applications for spatial analysis is enormous, for example, accident patterns, victim profiles within a residential population and spread rates for pollution levels. “Techniques are needed to let spatial data speak for themselves” [Griffith and Layne, 1999]. Thus, spatial statistics must hold a specific spatial framework to apply quantitative and statistical methods for a better understanding of spatial relationships. If the finding of spatial structures is fundamental then autocorrelation, interpolation and autoregressive models⁵ are three major spatial methods that fulfill this constraint. These methods are reviewed here. Although these fields have been developed autonomously, Figure 3 gives the details of the missing data issue that involves the estimation of missing georeferenced data when undertaking spatial autoregression and Kriging. However, spatial autocorrelation can also be exploited for spatial prediction purposes such as patterns and outliers. Spatial autocorrelation measures can also indicate if there is an equally likely chance of predicting neighboring values while the degree of redundancy indicates how much information is free to vary. As expected, certain issues emerged during the review of existing literature, such as:

- The dependence of the OK uncertainty measure solely on geometry;
- Definition of the relationships that exist among the above three concepts;
- **W** matrix codification;
- Heavy spatial analysis computation;
- Standardization of observation distribution and the impact of sampling approaches.

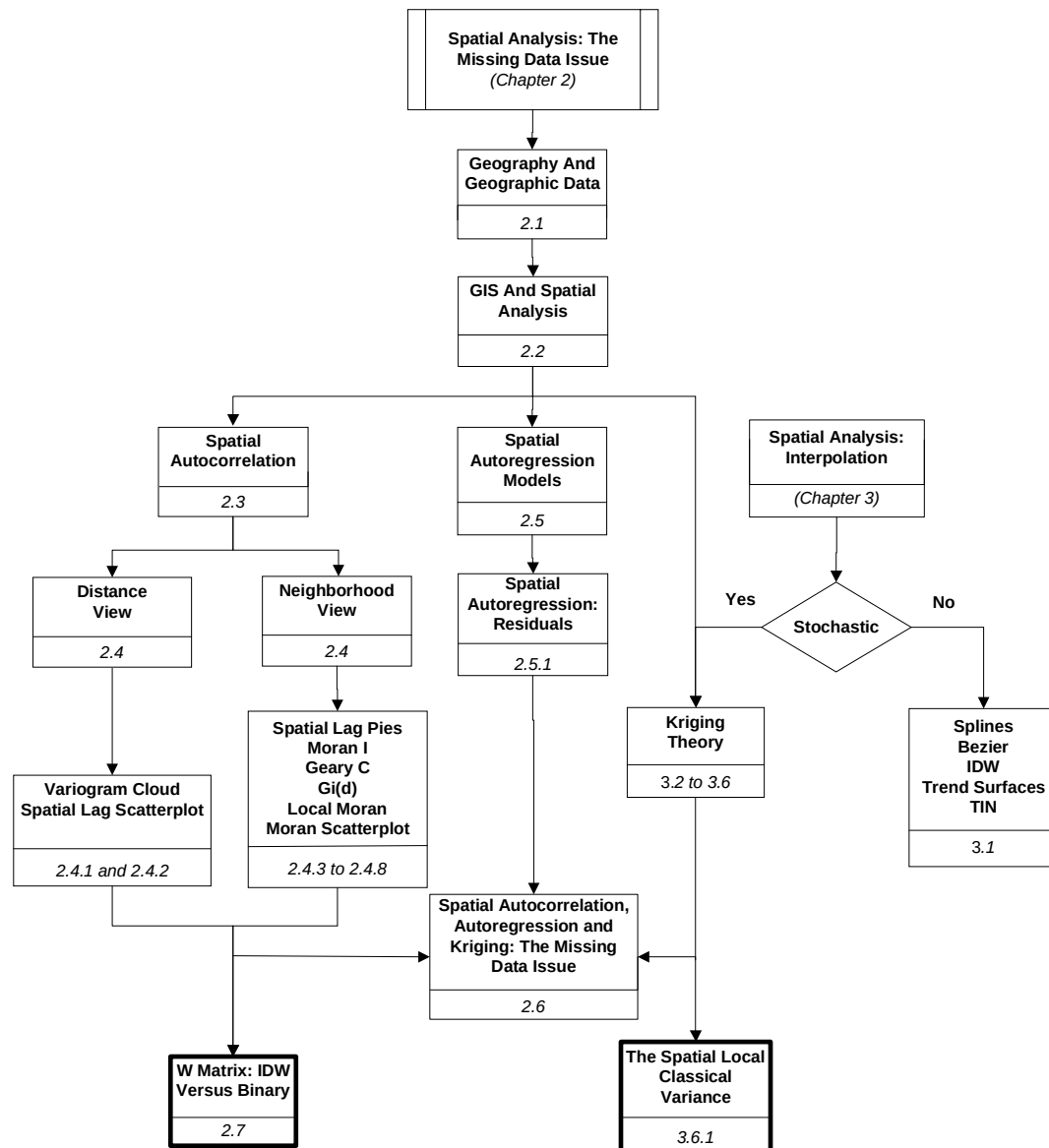


Figure 3 – The missing data issue and spatial interpolation of SAKWeb®.

1.5.2. Spatial Analysis: Interpolation, Implementation Review and Software Evaluation

Clearly, two methods emerge regarding spatial interpolation: the deterministic and the stochastic view. If the former has no capability for assessing uncertainty, the latter must be viewed as a process instead of a single function with an increase in computer power demand, particularly with Kriging. In terms of software development, highly regarded literature such as Geocomputation, A Primer [1998]

refers to three major implementations: independent geosoftware, GIS modules and statistical packages with spatial extensions (see Figure 4). Data exchange among systems has also become a requirement. In fact, spatial analysis and GIS are well-established technologies that operate in different spheres at present although their aims are similar: to describe and to infer the nature of spatial variation. Hence, some kind of systems integration would benefit both groups of users: within GIS, by increasing the scope of their spatial analysis and, within spatial analysis, by providing topology access. Yet, the lack of spatial autocorrelation measures in the software is huge. In fact, these measures may also help the variogram fitness interactive procedure. To date, the WWW implementation environment has never been considered. The nugget-effect variogram factor in software including e-Learning tools is seldom well developed. The rescaling of OK variance for true spatial sampling variance is not considered at all.

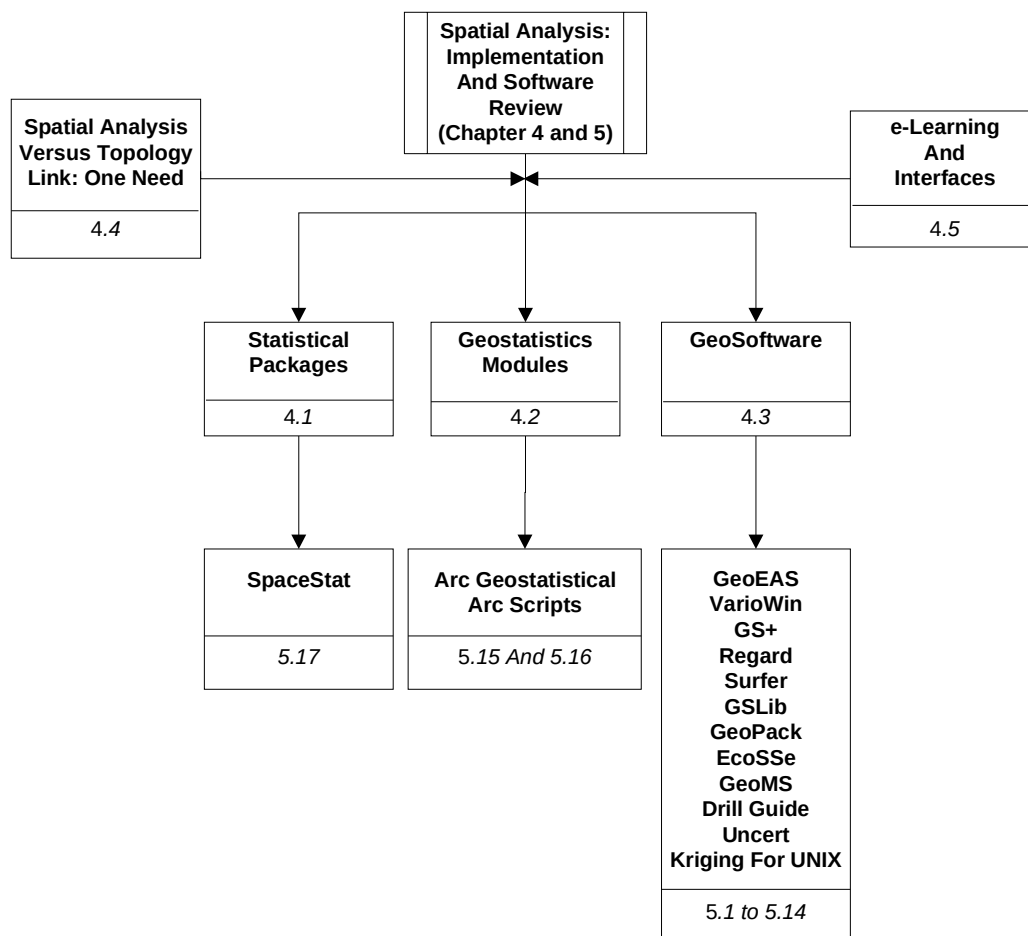


Figure 4 – The SAKWeb[®] organization of spatial analysis implementation review and software evaluation.

1.5.3. Spatial Analysis: User Know-how

With regard to the implementation of new spatial analysis tools, the future does not seem very bright on account of the undeclared divorce between the research community and the software developers. There is a market for spatial analysis but, at the moment, most spatial analysis users are oriented towards simple and deterministic rather than complex spatial tasks. The surveys carried out among GIS users in several Portuguese universities and research institutions confirm this situation (see Figure 2, above):

- Spatial autocorrelation measures and Kriging interpolation are two unknown issues;
- The need to update the geography and GIS curriculum as regards stochastic issues is urgent;
- Any initial spatial problem may result in different Kriging solutions;
- The best solution relies on the best justification model;
- Both statistical and geographical background is needed in order to obtain better results;
- Quite often, the software access for the appropriate spatial analysis can be an issue for common users.

1.5.4. SAKWeb[®]

When common users begin to handle and understand the background concepts of spatial autocorrelation and Kriging, the difficulty of getting specialized software becomes all too clear⁶. It is important to forget installation procedures and operating system requirements, to embrace a direct, friendly and known GUI⁷ with mobile access, to incorporate standard exchange for data input, to validate input parameters including suggestions and hints on-the-fly⁸, to add new features available and to provide the capability of discussing these issues on an on-line basis with good hypermedia help. The final major SAKWeb[®] specifications are, thus, setup (see Figures 5, 6 and 7). Major theoretical improvements are introduced

and, because software should be evaluated by end-users, a third survey has been carried out, giving new insights for a future version and some adjustments. Finally, the Colorado grasshopper dataset is analyzed by SAKWeb[®] in Chapter 8, which also includes an evaluation of the previous theory and a practical application.

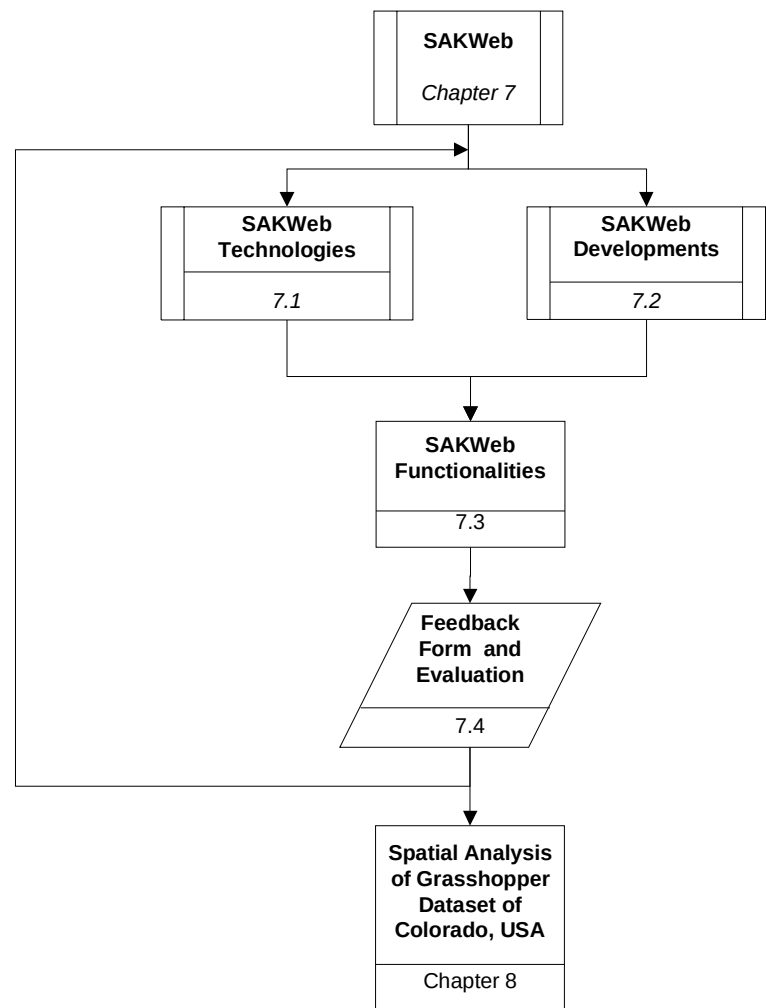


Figure 5 – The overall SAKWeb[®] development and technical implementation framework.

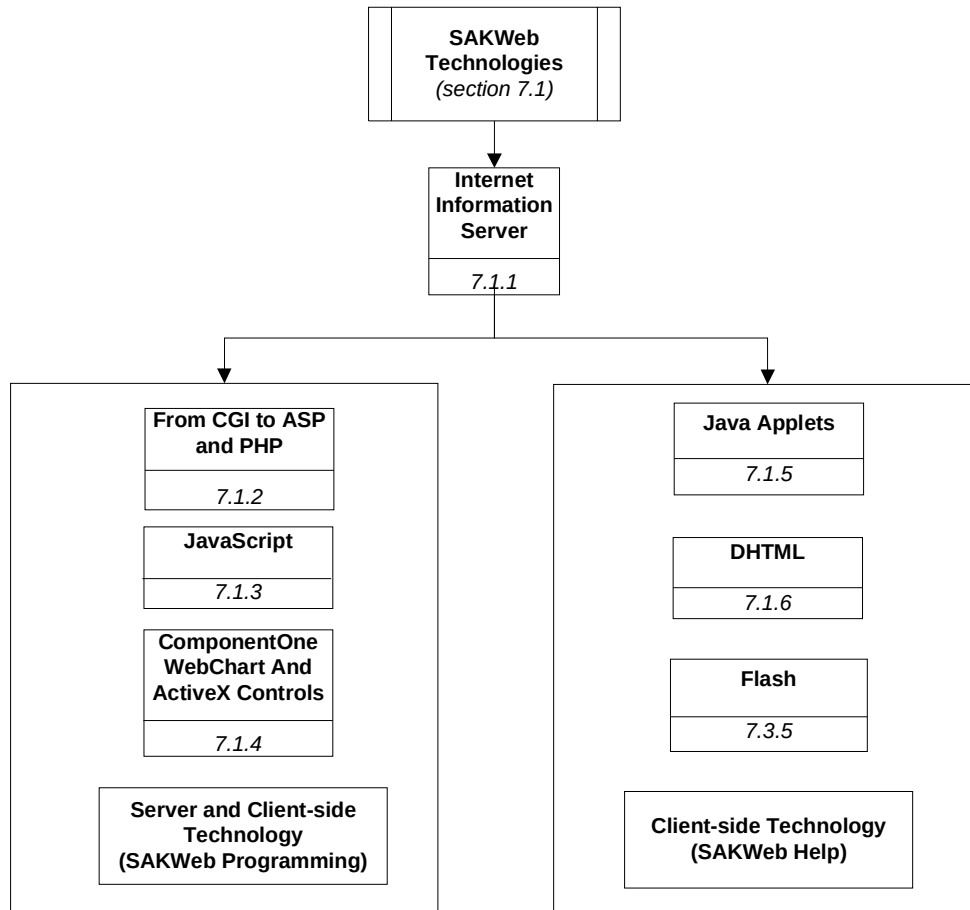


Figure 6 – The SAKWeb[®] technical implementation view.

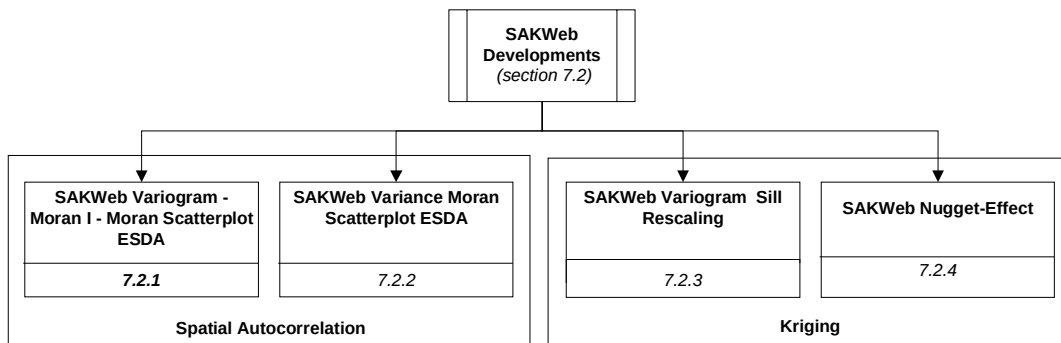


Figure 7 – The SAKWeb[®] development view.

1.6. RELEVANCE

Spatial analysis technology is now beginning to reach the stage where advanced users are struggling with new methods of spatial analysis relating to possible explanations for any spatial phenomena. From an explanation of the unitary to a

description of the collective, all views of space are correct if the where question is answered, what spatial correlations exist and why spatial distribution is structured in that way are the central questions. Hence, this thesis mainly discusses methods that address the inherent stochastic nature of patterns and relationships based on measured point properties⁹ rather than aspatial forms and deterministic ones such as 3D slopes or isomorphic buffer zones.

The questionnaires applied during this research clearly demonstrate that the W3 environment emerges as the final solution. “From a technological standpoint, cooperative work, computation distribution and networking¹⁰ contribute to the widespread dissemination of geographical knowledge with more intuitive, more analytical and more diverse embedding technologies” [Negreiros, 1999]. From the commercial perspective, the Internet can be viewed as an extension of the traditional competitive marketplace. The Internet is capable of dramatically lowering the transaction and agency costs facing most organizations (see Table 1).

E-GOODS	Traditional System	Internet	Percent Savings
Airline Tickets	\$8	\$1	87%
Banking	\$1.08	\$0.13	89%
Bill Payment	\$2.22 to \$3.32	\$0.65 to \$1.1	71% to 67%
Term Life Insurance Policy	\$400 to \$700	\$200 to \$350	50%
Software	\$15	\$0.2 to \$0.5	97% to 99%

Table 1 – Internet technology can radically reduce transaction costs in many industries [Spotlight, 1999].

Equally important, Internet technology is providing the infrastructure for electronic business because its technology can also be used to make information flow seamlessly from one part of the organization to another. For instance, Internet standards can be used to link disparate systems, such as ordering and logistics tracking that previously could not communicate with each other. Moreover, the cost of comparison shopping was very high, because people had to physically travel from store to store. Internet has changed this relationship. “Once everyone is connected electronically, information about products and services can flow on its own directly and instantly to consumers” [Laudon, Laudon, 2002] although this situation might create a channel conflict with the firm’s traditional channels. Thus, Internet shrinks information asymmetry. In the same way, using the Internet and

Web multimedia capabilities, companies can quickly and inexpensively provide detailed product information specific to each customer to very large numbers of people simultaneously [Evans, Wurster, 2000], the richness-reach concept.

Similarly, Internet help B2C companies to create and capture profit in new ways by providing the foundation for new products and services such as people can make on-line bids for items like rock concert tickets or antiques and collectibles, the pure-play business model. This includes m-commerce, electronic payment systems and customer support via e-mail. Another current trend is Web personalization, the capability to present to each customer a modified Web page based on that person's purchase history.

B2B e-commerce generates efficiencies by enabling companies to electronically locate suppliers, solicit bids, place orders, and track shipments in transit. That is, trading partners can directly communicate with each other, bypassing intermediaries and inefficient multilayered procedures. For instance, "vendors of some large retailers can access retailers' internal Web sites directly for up-to-the-minute sales information and initiate replenishment orders instantly" [Laudon, Laudon, 2002]. This means simpler business processes, fewer employees, and much flatter organizations than in the past in a redesigned and more closely integrated framework.

From the science organization point of view, the synergy effect of the online educational material¹¹ in which the overall information may be greater than the sum of its parts is a conviction. Under the GIS view, the greatest W3 impact is to close the data access gap among users¹² with free and direct retrieval of spatial analysis tools. Hence, to fill this present spatial analysis software vacuum is a major aim of this thesis.

The ambition of SAKWeb[®] version 2.0 is to develop the first spatial autocorrelation and Ordinary Kriging approach in a Web environment for people with data observation problems that can be handled by both methodologies. SAKWeb[®] tries to reconcile theory with practice and concepts with application.

Regarding programming concerns, due to the large number of variables, SAKWeb[®] makes extensive use of the dynamic memory allocation concept as RAM memory is a key concern in the Von Neumann computer framework. For instance, “the latest ESRI[®] AML language development concerns the garbage collection concept, the capability of regaining memory that is no longer being used by the application through the mark and sweep algorithm” [ArcUser, 2000].

With regard to **W** matrix construction, inverse distance weighting (IDW) is indicated as a poor process because it is not sensitive to its direct neighbors¹³. Hence, SAKWeb[®] presents the covariogram contiguity for which neighbors' boundaries are setup by the variogram range. “Any other weight can be applied provided it means something to the researcher and can be justified theoretically” [Ebdon, 1998]. In accordance with various lag distances, Moran I is presented as a good search factor of the neighbor radius. Based on the range of the highest Moran I, the weight matrix is defined while the Moran location scatterplot is created. The IDW deterministic interpolation is also based on the same procedure. The mapping computing framework allows the study of spatial patterns, outliers, changeover areas and trends. As the Colorado grasshopper case will also illustrate in Chapter 8, the Moran I correlogram may lead to a re-estimation process of the variogram sill, model, nugget-effect and range. In addition, the Moran variance scatterplot is presented as a valuable extension of the traditional one. To reproduce a new spatial dataset is an option. To contribute successfully to an ESDA tool before Ordinary Kriging is a major goal.

The discontinuity of the variogram origin can be viewed as a non-sense situation if spatial reality is considered a continuous surface. Among all variogram factors, the nugget-effect is also the most unpredictable because of the lack of close samples. So, SAKWeb[®] offers four strategies to handle the nugget-effect¹⁴:

- $\gamma(0) = 0$ (model 1);
- $\gamma(0) = 0$ but including C_0 for superior lags (model 2);
- $\gamma(0)$ with micro-scale, $\gamma_1(\mathbf{h})$, and long-range, $\gamma_2(\mathbf{h})$, assessment (model 3);
- If measurement error is given¹⁵, this attribute will be incorporated within the Kriging system (model 4).

The combination of model 1, 2 and 3 estimates “requires weighted averaging of the Kriged estimates by their estimation variances in order to give a more accurate sight of zones” [Bryan, 1994]. This approach is also undertaken by SAKWeb[®].

Contrary to hypothesis number five, it is proved that there is no evidence of spatial neighborhood patterns as regards which approach fits best. In addition, “it is assumed that the sill is a good global population variance estimate, a condition that most of the times is not true” [Soares, 2000]. Just as the arithmetic sample mean is a poor global mean estimate, preferential sampling with proportional effect leads to a false variability value. It is crucial to readjust the variogram sill to reflect the true global variance. Using the nearest neighborhood analysis for the samples weights, it is possible to find the relationship between the Estimated Global Variance (EGV) and the original variogram sill. It also avoids the difficulties of weighting samples at the edge. If this ratio multiplies the variogram model, a rescaling operation, then the Kriging classical variance and region confidence interval, both local and global, can be enhanced. In addition, a new spatial local variance index is presented as a good uncertainty measure.

The definition of the variogram model is a time-consuming and somewhat subjective operation. Moreover, it requires theoretical knowledge to setup its seven input parameters¹⁶. For this reason, input validation with an attractive Web interface with hypermedia help in an e-Learning context¹⁷ was undertaken to help to publicize spatial autocorrelation and Kriging methodologies because neither of the processes is well known among GIS users and classical statisticians. Quoting Soares [2000], “the dissemination of geostatistics is only possible with good books and first-class application software”.

1.7. ORGANIZATION

The present thesis is organized into nine chapters.

Chapter II is introductory but provides background to some of the issues addressed in the research. It presents spatial analysis concepts and discusses various subjects, for example, aspatial versus spatial statistical methods. Global and local spatial autocorrelation and association measurements are fully dissected while IDW is pointed out as a bad methodology for **W** construction.

Chapter III focuses on a review of Kriging. This debate starts with a spatial deterministic interpolator overview, e.g., B-Splines, Fourier, IDW and polynomial trend surfaces. Afterwards, the regionalized data nature and variogram models are stated: local, general relative, pairwise relative, covariance and correlogram. Spatial and aspatial exploratory data analysis is considered including data standardization and sampling considerations. Cressie's [1993] automatic fitness is mentioned. Some statistical and computational Kriging aspects are examined. Uncertainty confidence enhancement using a new uncertainty index, the spatial classical variance, is stated.

Chapters IV and V review the three major implementation techniques available at present. As expected, the foremost software features of spatial analysis are described. E-Learning and the close link between topology information and spatial analysis needs are particularly highlighted.

Chapter VI reveals the lack of spatial analysis development by the research community for the production of GIS commercial packages and, in particular, the modest knowledge of its users as regards spatial interpolation and autocorrelation. The WWW environment for SAKWeb[®] final specification is emphasized.

Chapter VII focuses on the creation of SAKWeb[®] version 2.0. Major Active Server Pages[®], Personal Home Pages[®], Internet Information Server[®], JavaScript[®], ActiveX[®] and Flash[®] features are reviewed. Particular features are examined on the basis of a technical debate. Major innovations, specifications and interfaces are fully analyzed. A third feedback survey is presented in conjunction with its evaluation results. Further SAKWeb[®] investigation and programming needs are pointed out, though presenting algorithm resolution in theory is quite different from solving it in practice¹⁸. “The major cost to be borne is the implementation and development of the software” [Upton and Fingleton, 1995].

Chapter VIII uses SAKWeb[®] to explore the grasshopper dataset from Colorado, USA, for the years 1993, 1994, 1995 and 1997 in order to appraise its contribution to geostatistical software. This includes descriptive and nearest neighborhood analysis, variography fitness, SAKWeb[®] Moran scatterplot (location and variance), OK surfaces with validation test and cross-validation analysis. A research summary with concluding remarks is presented as well.

Chapter IX presents a general conclusion that summarizes the contribution of this thesis to the body of geocomputational knowledge. The initial hypotheses are analyzed. Some other issues are also reviewed while further investigation and programming needed for SAKWeb[®] is point out.

2. Spatial Analysis: The Missing Data Issue

Chapter 2 is divided into eight major sections. Section 2.1 focuses on the role of space attributes in geography while the second one discusses spatial analysis within a GIS environment. Section 2.3 centers on the impact of spatial autocorrelation in spatial data. Section 2.4 presents spatial autocorrelation tests while the following one reviews spatial autoregressive methods. Their residuals are mentioned in Subsection 2.5.1. Section 2.6 establishes a close and important link among Kriging, spatial autocorrelation and spatial autoregressive models: the missing data issue. The last section with its two subsections presents the relationship between Moran I and the Moran scatterplot and stresses IDW contiguity as a poor **W** specification model when compared with the binary cutoff distance. The conclusion section summarizes major inferences throughout this chapter including a linkage with Chapter 7 that discusses the SAKWeb[®] vision of spatial autocorrelation.

2.1. GEOGRAPHY AND GEOGRAPHICAL DATA

For Bação [1997], “geography does not hold its own methodology because it does not have convictions towards its own study object, space”. He argues that there is a serious absence of laws and techniques on how space rules the Third Rock from the Sun phenomena. It seems that geography is not a spatial science but a reflection on the universe, which suggests that it cooks up a global explanation from others sciences even when reality is the same. Hence, the geographer becomes an intruder into the physical and social matters. The Central Place Theory by Walter Christaller [Dictionary of Geography, 1997] is a good example of how geography and economics intersect, giving an insight into the nature of town development.

It would be absurd to forget the permanent social and movement of society within space and believe that quantitative methods could explain all Earth phenomena. However, African decolonization or the Gulf war should be studied in a spatial context only within geography, a fragmented point of view. For Mayhew [1999], “this quantification is attacked for being unrealistic and bloodless, turning humans into automata, for being too deterministic and for ignoring the importance of subjective experience”.

For a long time, geography did not hold a specific mathematical approach for any interpretation of space and this was the major reason why geography degrees cover a wide variety of subjects like demography, geology, topography, hydrology, history, biology, morphology, climatology, sociology and economy to fill the curriculum. Yet, geography is finally creating its own research agenda to meet the needs of a spatial location framework because what really matters for geographers is the spatial attributes, while the non-spatial ones belong to other fields. From this viewpoint, “geography becomes a widow of space” [Santos, 1978], although space has a close brother called time: geography for space and history for time. Definitely, ‘Where is...?’, ‘Is there a general spatial pattern?’, ‘What are the anomalies?’ and ‘Why do these phenomena pursue certain spatial distributions?’ are the added value given by geographers to other sciences [Ferreira and Simões, 1993, Longley *et al.*, 2001].

It is known that most spatial variables are the result of a vast number of processes that cannot be described quantitatively. However, the earth holds some spatial continuity patterns leading geography to a new effort to create its own general space rules instead of begging methodologies from other sciences such as history or economics. “Geography is the science of the differentiation of space based on its own laws“ [Longley *et al.*, 2001], to detect urban hierarchy, future industrial localizations, military targets accessibility, a hospital’s regional influence or agriculture patterns, for instance. Certainly, this narrow view of geography emphasizes the major feature of its uniqueness in relation to other research fields.

For instance, the darker areas in Figure 8 represent population losses between 1993 and 1997 from the Portuguese electoral roll at town level (the Lisboa and Setubal regions) while the lighter ones represent the resident gains for the same time period. As expected, a strong positive spatial autocorrelation is to be found, a trouble-free detective job in this particular case.

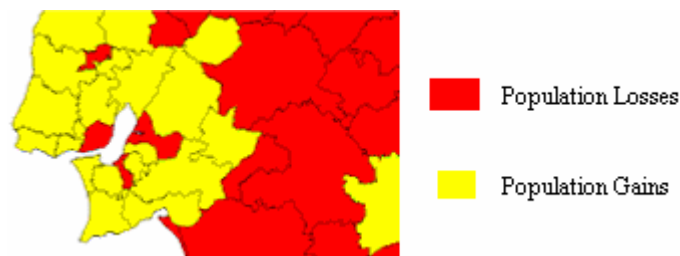


Figure 8 – Population gains and losses on the Portuguese electoral roll at town level.

It is undoubtedly a geographer's job to interpolate the number of inhabitants for a particular *freguesia*¹⁹, to extrapolate the socialist voters for Odemira *concelho*²⁰, to check the relationship and variability abstentions among *freguesias*, *concelhos* and *distritos*²¹, to look for local pockets of similar political party trends, to generate the new Odivelas *concelho* boundaries with Loures, to design a package delivery route between Estoril and Sesimbra, to route the nearest ambulance to an accident taken place at Vila Expo, considering the traffic jam, to find how many customers live near Avenida Luisa Tody, Setubal, to find the nearest drug store available at 3 am or to transform the present *freguesia* dataset to the post code in a spatial overlay operation. As Bação [1997] rightly points out, “each discipline should explain any phenomena based on its own laws, thus creating complementary conclusions”. If the dynamic spatial relationships between places reflect social, physical and economic processes, to reject all traditional approaches would be as ridiculous as to reject new and complementary conceptualizations. Therefore, discussing spatial autocorrelation and spatial interpolation processes²² is central to this thesis, where absolute and relative localization is the core of finding space rules.

Spatial analysis should be centered on the search for patterns within spatial phenomena involving understanding, prediction and simulation, where location, time and geometry are significant. The difference between geography and other

sciences is locations, spatial structures and spatial processes, because no other study field concerns itself with the distribution of spatial phenomena, for example, interpolating gold in South African mines, checking spatial relationships among religions, looking for local pockets of elephants in Bangladesh, extrapolating oil production in Alaska from the Russia Kamchatka peninsula, simulating the Egyptian Nile basin dam and the spread of infectious hepatitis C, calculating the shortest distance between Moscow and Berlin, specifying the range area influence of a new police station or allocating voters in establishing new electoral boundaries. If astronomers or meteorologists, for instance, apply geographical methodologies with their own objects then they become geographers and geography becomes identifiable. It is with the explanation of each discipline but based on their own rules in a puzzle framework that any physical or social phenomena should be decoded.

2.2. GIS AND SPATIAL ANALYSIS: AN OVERVIEW

Several classical statements relating to the definition of GIS can be found in specialized literature such as the GIS International Journal, expressing the idea that spatial analysis can somehow be useful. “GIS can be seen as a spatial analysis engine and its main end relies on the ability to predict outcomes and, above all, to understand those” [ArcNews, 2000]. “GIS is simultaneously the telescope, the microscope, the computer and the Xerox machine of regional analysis and the synthesis of spatial data” [Abler, 1988]. “GIS is a system of hardware, software and liveware implemented with the aim of storing, processing, visualizing and analyzing data of a spatial nature. Other definitions are also possible” [Painho, 1999]. “GIS is a tool for revealing what is otherwise invisible in geographical information” [Longley *et al.*, 2001]. Certainly, GIS is not a graphic database.

According to Unwin [1998] and Bailey [1996], the spatial analysis concept has nothing to do with the general ability to describe spatial data, generally incorporated in all commercial GIS but it presents the challenge, given a spatial pattern²³, to explore it with an appropriate model and represent it with a graphical

display. Certainly, to describe is not to explore. As Murteira [1993] confirms, exploring is a detective work, a search for clues and evidence, while describing is a job of judgment, a job of analyzing and evaluating clues.

The geographical view is essentially cartographical driven regarding the recognition and description of spatial patterns involving simple statistics and the direct use of visualization. Anselin [1993, 1998] wrote that spatial data analysis could be defined as the study of statistical phenomena that manifest themselves in space. As a result, location, area, topology²⁴, spatial arrangement, distance and interaction become the focus of attention. As confirmed by Longley *et al.* [2001], “spatial analysis is a set of methods whose results change when object location being analyzed changes too”. Hence, spatial analysis is more widespread than statistical analysis of non-spatial information because it requires access not only to attributes but also to location and topological knowledge. It could be described as part of the process of transforming spatial data into geographical knowledge. Quoting Longley *et al.* [2001], “it can make what is implicit explicit”.

Rossiter [1999] also includes spatial flow modeling and deterministic processes like groundwater movement and environmental quality management based on economic criteria such as land use and transportation. It seems that geographical analysis comprises GIS (an applied computer-science view), spatial statistics including uncertainty issues (spatial autocorrelation, spatial autoregression, Kriging, stochastic simulation, morphologic geostatistics and space-time processes), classical aspatial statistics, remote detection and deterministic spatial analysis such as optimization routing, B-Splines, overlay, buffering and DEM operations (cartographic modeling).

Clark and Hosking [1986] see spatial analysis as spatial modeling of a decision support such as GADS²⁵ for solid waste spatial planning. In conjunction with the network and spatial analysis of GIS modules, the Decision Support System Location Planner[®] analyzes market saturation, retail facilities accessibility, population mobility and demand-supply prediction based on demographic and socio-economic attributes, warehouse locations, distance or travel time between

sites and expenditure flows between demand and supply chains [Arentze *et al.*, 1998]. However, according to Openshaw [1994, 1998], “an emphasis of DSS is a convenient distraction to hide a lack of the relevant GIS technology”.

Ignoring technical matters and human technology resistance²⁶, spatial DSS, education and W3 services create political pressure on governments regarding public policy decisions in such fields as environmental protection, natural resource management, hazardous waste location and regional development (‘government of the people, by the people and for the people’). “This millennium will be about slowing the rate of deforestation, improving water quality, restoring wildlife habitats and understanding the earth’s limitations” [ArcNews, 2000]. Thus, the handling of spatial data should generate definitive answers. McMaster [1993] confirms this belief in his hazardous materials modeling for Santa Monica, CA. By identifying the explosives, flammable gases, solids and liquids, radioactive materials, corrosives and poisonous materials present within the community of 100,000 people, the evacuation plan by the Santa Monica Police and Fire Department became very clear.

Yu [1998] substantiate the difficulty of reaching a good balance between acceptable landscape planning in the Red Stone National Park, South China, and conflicts with ecologically protective boundaries for endangered medium-sized mammals and amphibian species, tourist preferences and newly reclaimed agriculture land. It is not easy to find the perfect solution. For Vale [1994], resolution of this conflict for the same space involves three solutions: hierarchical dominance (certain matters are more important than others), multiple use (the same space may have several uses) and trade-off (certain issues are chosen to the detriment of others).

Another reasonable spatial analysis goal is temporal trigger scheduling. Quoting Albert Einstein, “space-time is indivisible” [Schwartz and McGuinness, 1982]. Historical and rollback information systems are essential for contracts, death and inheritance, legal proceedings and earthquakes situations. Moreover, the space-time point pattern recognition, e.g., with forests fires or mail order deliveries can

be a complicated issue because space clustering may have a weak time interaction and vice-versa.

Recently, Cressie and Wikle [1999] reported new advances in spatio-temporal prediction for large datasets using Kalman filter modelling to reduce the time dimension in an empirical method of moments²⁷ implementation, while Hopkins *et al.* [1999] mentioned Kriging use for (x,y,t) ozone concentration estimation with GMS^{®28}. Another classical time example is shoreline limits, which are particularly useful for the study of global warming effects on coastal cities. Which shoreline should be adopted? Should the geographical database be dynamic and capable of tracking fluctuations? One possibility is to consider the shore slope classification, the sine angle and the time sea level above the height of the lowest tide. Since tides follow a deterministic formula, it is possible to calculate the exact shoreline location for a given time *t*. It is implicit, therefore, that spatial analysis is a GIS component to support decision-making²⁹ for solving problems with a spatial component.

Another central component of GIS is technology management. “There is a world market for maybe five computers” is the famous statement by the IBM[®] chairman in 1943. He happened to pass on a few years later and IBM[®] was saved. Ken Olson from Digital[®] Corporation, the second largest computer manufacturer in the seventies, did not survive (“There is no reason why anyone would want a computer in their home”). When he woke up, Open VMS[®] was already dead and DEC[®] had been bought by Compaq[®] Corporation. Bill Gates recovered from the “640k is enough for everyone” statement on time.

Decision-making is not a planning sequence and policy-making exercise in a SWOT strategy. There are quantitative decisions such as what to buy, for how much and how many and there are qualitative decisions such as who should do the job. “The most important factor in the whole decision-making process is the human component” [Negreiros, 1998]. This is the most unpredictable and unstable one because high-tech often modifies people’s professions. Human management and how to match human resource with the organization’s goals present the

greatest challenge to the decision-making structure and not the technology itself. “The future focus will be the I² concept: Integration Implementation of people, procedures, data, technology and communication networks that result in the solution of business needs” [ArcNews, 2000].

2.3. SPATIAL AUTOCORRELATION

The role of location³⁰ holds two major implications for the way statistical analysis should be carried out. Location leads to spatial dependence (correlation or variation that each neighbor holds in relation to a particular point) and spatial heterogeneity (clustering, concentration or proportion of neighborhood average in relation to a specific point) established by Tobler’s First Law of Geography³¹. Since regional differentiation respects the intrinsic uniqueness of each location, spatial autocorrelation can be viewed, hence, as a map pattern descriptor.

It is significant to emphasize a particular dilemma associated with the definition of spatial autocorrelation and spatial association. For Goodchild [1987], “spatial autocorrelation can be interpreted as a descriptive index, measuring aspects of the way things are distributed in space (clustering) but, at the same time, it can be seen as a casual process measuring the degree of influence exerted by something over its neighbors (correlation)”. Yet, according to Levine [1997], while the first of them describes the relationship among different locations for a single variable, indicating the degree of concentration or dispersion³², spatial association describes the correlation or association between variables distributed over space³³. Several considerations can be made. To define spatial association with the word association is a gray situation. According to Amrhein and Reynolds [1998], the term correlation is connected with the definition of autocorrelation, another gloomy position. In mathematical terms, the number of variables by clustering and dependency indices is only one. For Ebdon [1998], “autocorrelation is concerned with the relationship between successive residuals values along a regression line”. It seems that there is an assorted exactness in the description of association, correlation, dispersion, concentration and the number of variables.

Classical statistics offer a wide range of inferential methods based on restricted assumptions. The 95% confidence level of the true unknown population mean³⁴ and standard deviation³⁵ is only valid if samples are independent and homoskedastic with uncorrelated error. Another classical possibility is the chi-square (χ^2) contingency test, if the independent random sample of size n is sufficiently large, in order to find particular relationships between two variables where each variable consists of a number of classes.

$$\chi^2 = \sum_{i=1}^N \sum_{j=1}^Z \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad [1]$$

where:

O_{ij} - Observed number of i region and j group;

E_{ij} - Expected number of i region and j group;

Z – Number of groups;

N – Number of regions.

If the χ^2 of equation one³⁶ is greater than the expected χ^2 for a certain probability p then the initial hypothesis (H_0) is rejected and, therefore, a statistical interdependency between regions and groups has been found. To confirm the difference mean of two subpopulations³⁷ is an alternative based on the independent random samples and Normal distribution assumptions. After the standard error of the difference between sample means³⁸ is assessed, if the t-Student for the test of both means³⁹ is rejected for a particular α level of confidence level then both dataset means are dissimilar. Furthermore, linear and multiple regression models including the global F^{40} and individual t-Student⁴¹ tests assume linearity, independency of random variables, homoskedasticity (the weak assumptions) and distribution normality (the strong assumption). “If these assumptions are correct then the Gauss-Markov theorem proves that the least squares yield the best linear unbiased estimators” [Pinto and Curto, 1999].

Yet, spatial data does not follow statistician’s rules and spatial patterns must be taken into account. As Griffith and Layne [1999] points out, “spatial autocorrelation is quite often positive with a correlation range between 0.3 and

0.5". Spatial ANOVA of SpaceStat[®] regards spatial differentiation as a regression model whose spatial **W** matrix can be included and explanatory variables are converted to categorical ones such as North/South sub-regions. "This assumption of indicator variables in the form of a dummy variable that measures the difference between the sub-region mean and the overall one" [Anselin, 1992]. A positive and high t-Student indicates a strong discrepancy among sub-regions. Furthermore, "if spatial residual dependency exists then biased estimates can be assumed and a spatial lag autocorrelation regression⁴² can be applied" [Anselin, 1992]. Based on the crime dataset for the 49 wards of Columbus, Ohio, for instance, the center/periphery (CP) sub-regions crime discrepancy (CP=0/1) was tested using standard OLS regression and MLE spatial lag. This dummy variable that correspond to subregions in the dataset measure the difference between the mean for the subregion and the overall mean. In effect and according to the same author, the positive and highly significant value for its coefficient (24.9) indicates a strong discrepancy between the mean crime rates in the two subregions. As expected, the later method led to an R^2 improvement with a significant CP that backs up strong spatial clustering in crime rates. According to Table 2, non-colinearity, errors normality, homoskedasticity was found with OLS. Nevertheless, spatial error dependency is a reality and, therefore, these estimates may be biased. Performing a spatial lag estimation for the ANOVA model, the W_crime lag variable (regional effect) became positive and highly significant with MLE. The diagnostics for this model do not reveal any further problems and confirm the significance of the spatial autoregressive term.

OLS Regression of ANOVA			MLE Spatial Lag Regression of Spatial ANOVA		
$R^2=0.56$	LIK=-186.67	AIC ⁴³ =377.35	$R^2=0.59$	LIK=-184.48	AIC=374.97
SC=381.13	F=61.05	RSS=5845.19	SC=380.64		
Variable	Coefficient	Standard Deviation	Variable	Coefficient	Standard Deviation
			W_crime	0.31	0.14
Constant	22.93	2.23	Constant	14.60	4.45
CP_1	24.90	3.18	CP_1	19.45	3.7
Multicollinearity Condition Number ⁴⁴ : 2.37					
Kieffer-Salmon Errors Normality ⁴⁵ : 0.04 (χ^2 , df=2)			Likelihood Ratio significance for W_crime: 4.38 (χ^2 , df=1)		

Breush-Pagan Heteroskedasticity: 0.008 (χ^2 , df=1)	Breush-Pagan Heteroskedasticity: 0.26 (χ^2 , df=1)
Moran I Error: 0.18 and LaGrange Multiplier Test	LaGrange Multiplier Test On Spatial Error
On Spatial Error Dependence: 3.5 (χ^2 , df=1)	Dependence: 0.071 (χ^2 , df=1)

Table 2 – The OLS regression of ANOVA and MLE spatial lag regression of Spatial ANOVA (to emphasize that the treatments are spatial in the sense of being subregions of the dataset) for a study of 49 wards in Columbus, Ohio [adapted from Anselin, 1992].

“Autocorrelation damages the ability to perform hypothesis standard statistical hypothesis tests because the confidence interval estimated by the classical Pearson product moment is narrow inducing, thus, biased conclusions” [Nass and Garfinkle, 1992, Legendre, 1994]. “The estimator standard errors are not be minimized and regression coefficients from least squares are unbiased while their variances are underestimated” [Clark and Hosking, 1986, Ebdon, 1998]. This occurs because new observations, under the lack of independence, do not each bring one full degree of freedom since the observer holds some prior knowledge at new locations reflecting information loss⁴⁶. As spatial autocorrelation approaches 1, the effective degrees of freedom approach 0⁴⁷. As spatial autocorrelation approaches 0, the effective degrees of freedom tend to the total sampling number. As spatial autocorrelation approaches -1, the effective degrees of freedom increase beyond the number of observations (see Table 3).

Number of Locations	$A_{px} = A_{py} = A$				$A_{px} = -A_{py} = A$			
	$A=.05$		$A=.10$		$A=.05$		$A=.10$	
	df ₁	df ₂	df ₁	df ₂	df ₁	df ₂	df ₁	df ₂
4	2.68	1.92	1.80	1.79	3.35	2.11	4.78	2.51
16	14.12	13.33	11.41	11.35	15.90	14.71	19.24	17.25
25	22.71	21.92	18.67	18.62	25.32	24.13	30.16	28.20
36	33.22	32.42	27.59	27.57	36.82	35.63	43.46	41.51
49	45.64	44.85	38.18	38.17	50.40	49.22	59.15	57.21
64	59.99	59.2	50.43	50.44	66.06	64.88	77.22	75.28
100	94.44	93.65	79.94	79.94	103.65	102.47	120.51	118.58
144	136.57	135.77	116.03	115.98	149.58	148.39	173.37	171.40
256	243.87	243.05	208.24	208.21	266.44	265.25	307.69	305.63
400	381.94	381.12	326.98	326.97	416.68	415.46	480.47	478.30

Table 3 – The theoretical expected degrees of freedom for the modified t-Student test for two autoregressive lattice processes in relation to the number of locations and autocorrelation parameters, A_{px} (x axis direction) and A_{py} (y axis direction) [Dutilleul, 1994].

Fewer samples are therefore needed as positive spatial autocorrelation increases, a possible reduction solution for the huge spatial data computation issue⁴⁸. Under this viewpoint, there is a better than equal chance of predicting neighboring values if this information is already available. As spatial autocorrelation also decreases, error prediction increases⁴⁹. Spatial autocorrelation is the spine of spatial interpolation. If someone records downtown air pollution in Madrid, Spain, monitoring devices will report the same levels adding almost no new information to the previous samples. Information redundancy is the logical consequence with autocorrelated georeferenced data because samples collected for two juxtaposed points in space tend to represent essentially the same information. Thus, spatial autocorrelation may be defined as a measure of the true but masked information content in spatial data.

“If spatial autocorrelation exists, the standard error of the linear estimators will not be minimized, the dependent variable estimation will be inefficient, the confidence intervals will be larger, R^2 , t and F tests will no longer be valid and the residuals may be highly correlated” [Dutilleul, 1994]. The independent zero spatial autocorrelation assumption conflicts with Tobler’s Law: if strong residual autocorrelation exists then OLS inference breaks down because near things hold the same weight as distance ones.

With independent data, the two-sided 95% confidence interval for the mean μ is $[\bar{x} - 1.96\sigma/\sqrt{n}, \bar{x} + 1.96\sigma/\sqrt{n}]$ but with positive n correlated samples⁵⁰, Cressie [1993] proved that the previous interval⁵¹ equals $[\bar{x} - 2.485\sigma/\sqrt{n}, \bar{x} + 2.485\sigma/\sqrt{n}]$ which leads to a wider confidence interval. Moreover, the estimated variance of the mean with a population of n samples is not σ^2/n but σ^2/n' where $n' = n/[1 + 2\{\rho/(1-\rho)\}\{1 - (1/n)\} - 2\{\rho/(1-\rho)\}^2(1-\rho^{n-1})/n]$, which can be interpreted as the equivalent independent observations number. If $n=10$ and spatial autocorrelation equals 0.26 then $n'=6.2$, that is, 6 independent observations achieve approximately the same accuracy as 10 correlated samples. Hence, spatial autocorrelation works like a variance inflator factor: by decreasing spatial

dependency or increasing distance between sample points, prediction intervals move from a platykurtic to a leptokurtic profile.

The researcher's interest in spatial autocorrelation derives from the inability of the traditional estimator inability to capture redundant locational information. The same grid may hold different spatial distributions in spite of the same mean and standard deviation. Regarding the journey-to-work decay factors for Canadian urban areas, for instance, Griffith [1996] reports that 40% of the observed variation was due to areal unit configuration. Location information composes a non-marginal amount of the total information content of the georeferenced data. A parallel situation occurs in classical statistics when p variables, in a multivariate dataset, are correlated among them although principal component analysis tries to summarize them into q variables, with $q \ll p$. In this respect, if spatial autocorrelation is among observations for a single variable, in a multicollinearity situation, correlation is among variables.

To identify outliers is another aim of most spatial autocorrelation measure. If the outlier is located on the center of the map then only smaller variogram lags will contain it, leading to a variogram decrease for greater distances. However, if the outlier is located in the corner of the field then the variogram will present no difference from the one in which the outlier does not exist. Certainly, if heteroskedasticity exists then variability will increase with distance lags. Quite commonly, as a rule of thumb, a convex variogram at higher distance lags reveals a situation for outlier detection and spatial patterns.

Spatial investigators are held hostage by unsuitable conventional statistical processes. Particularly, if dependency among samples emerges then Kriging weights depend on the covariance between samples and estimations. Nonetheless, Clark, Hosking [1986] and Ebdon [1998] present three reasons for classical statistics and OLS use. First, some of the statistical analysis executed by geographers, the spatial component is of little or no significance. This means spatial data are aspatial. Second, summary information can always be compared to different locations for different times since the interest is not the variability study

over space. Third, score variability includes a random error component and, therefore, statistical models can be applied. Additionally, the potential gains with simple computing and estimation efficiency are attractive, mostly as a diagnostic tool at an exploratory stage and for residuals analysis. Finally, “with the absence of spatial dependency, the best interpolator for constructing a generalized map is a coordinate OLS function” [Clark and Hosking, 1986].

Classical statistics (see Section 3.3 for a further review) constitute one method in the whole field of spatial analysis. The classical inferential statistical method used in this process is just another analysis extension of spatial data. Its role should be centered on the description phase, on the score analysis variability over space and on the check in the phenomena over time. Non-spatial operations can often lead to significant spatial pattern discovery. Griffith and Layne [1999] recommend standard OLS multiregression for all spatial statistics toolboxes. Thus, this type of spatial autocorrelation may be viewed as a self-correlation of attribute values.

2.4. VISUALIZING WHAT IS NOT OBSERVED

As in the datamining process of finding attribute anomalies, spatial autocorrelation measurements look for patterns and relationships within vast spatial digital archives. Spatial autocorrelation indices are categorized into two groups:

- Distance view, an approach based on Kriging where spatial interaction is conceptualized as a continuous distance metric function (the variogram)⁵². It focuses spatial prediction. “Geostatistics uses the data twice: first to estimate spatial autocorrelation and second to make predictions” [ESRI, 2001]. Thus, the variogram represents the two faces of the same coin: spatial description and spatial prediction.
- Neighboring view, where the interaction is concerned with phenomena aggregations of discrete regions or correlation among adjacent observations. Neighbors, for instance, are defined as spatial units with a common boundary or within a given critical distance (no interaction occurs further away). Commonly, the **W** matrix contents are $W_{ij}=1$, if *i* and *j* regions are adjacent

to or below the cutoff distance, and zero, otherwise. W_{ii} also equals zero. This view enhances statistical description. If extended with autoregressive models then spatial interpolation becomes real.

2.4.1. The Variogram Cloud and the Cressie Box Plot

The variogram cloud is the squared difference between two locations, x_i and x_j , against distance d that separates them in a certain direction. Outliers hold high values in the variogram cloud no matter what the distance is because it affects computation independently of being a true outlier, skewness datum or wrong observation. Closer distances characterize local outliers. Global ones are considered by two distinct high point groups for small and larger distances. Thus, Cressie [1993] introduces a more resistant estimator (see Equation 2) by down weighting the contaminated data and relying on the square-root difference cloud instead of the traditional square difference. According to this author, this residual variogram produces lower values than the classical one because it down weights for atypical observations.

$$2\gamma(h) = \left\{ \frac{1}{|N(h)|} \sum_{i=1}^{N(h)} (x_i - x_j)^{1/2} \right\}^4 \bigg/ \left(0.457 + \frac{0.494}{|N(h)|} \right) \quad [2]$$

where

$N(h)$ indicates the number of pairs available for a specific lag distance and direction;

x_i and x_j are two observations that are separated by distance h .

If the cloud sill represents the sample variability and the cloud range characterizes spatial continuity then six other major inferences emerge:

- A lower sill signifies lower variability whose range indicates its size.
- A higher sill suggests spatial unpredictability.
- Location pairs beyond the sill are considered to be uncorrelated.
- If point pairs produce a horizontal straight line then no spatial autocorrelation situation emerges (the Moran I affinity is 0).

- The variogram cloud surface can illustrate anisotropy (see Section 3.4 for review). This feature is confirmed by close points in one direction and spread out points in another.
- To avoid a point tempest, Cressie [1993] presents the box plot associated with each distance band that yields the median, average, outliers, maximum and minimum and the inter-quartile range.

2.4.2. The Spatial Lag Scatterplot

This scatterplot is distinguished by a graphic where the horizontal axis corresponds to the value at each site while the vertical one shows the corresponding value at locations within a specific distance d apart, representing different contiguity orders along different distance bands. Hence, looking for local outliers⁵³ becomes obvious while a linear relationship along the diagonal indicates a strong association.

2.4.3. The Spatial Lag Pies

Spatial lag pies consist of pie charts located on observations coordinates comparing the magnitude of sample values relative to their spatial lag⁵⁴ (see Figure 9). More precisely, if the observation value at location i is x_i then its spatial lag becomes:

$$\text{spatial\_lag}_i = \sum_j w_{ij} x_j \quad [3]$$

where:

x_j represents i 's neighbor;

w_{ij} is the weight assigned between site i and neighbor j samples.

Similar bar heights suggest positive spatial autocorrelation while unequal ones tend to indicate negative spatial autocorrelation or a transition between regimes.

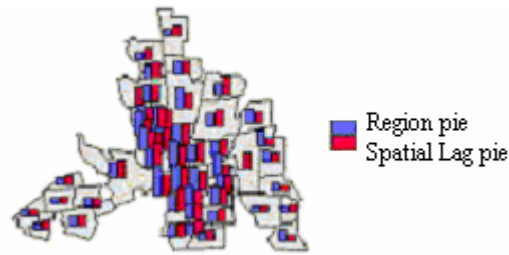


Figure 9 – The spatial lag pies of a Columbus Ohio dataset [Anselin, 1998], where the red bars indicate neighborhood crime average and the blue ones represent crime in a particular ward.

2.4.4. The Moran I

Under the spatial independency null hypothesis, the global Moran I (see Equation 4) is evaluated by measuring the covariance between attributes at each place and near sites towards the overall mean. If both neighboring values are above or below the mean (similar high-high or low-low values), the product is positive, reflecting the presence of a similar spatial autocorrelation. Otherwise, the product of the two mean deviations will be negative (unrelated high-low and low-high values), indicating a non-positive one.

$$I = \frac{n \sum_{i=1}^n \sum_{j=1(i < j)}^n w_{ij} (x_i - \bar{x})(x_j - \bar{x})}{S_0 \sum_{i=1}^n (x_i - \bar{x})^2} \quad [4]$$

where:

- $\bar{x}$ symbolizes the overall sampling mean;
- w_{ij} is the weight of matrix $\mathbf{W}$ between i^{th} and j^{th} ;
- n represents the total number of observations;
- x_i and x_j are i^{th} and j^{th} sample value;
- S_0 equals $\sum_i \sum_j w_{ij}$ (the sum of the spatial weights).

For Hart [1999], this index is the modified Pearson coefficient for one single variable including the $\mathbf{W}$ matrix. Its output domain varies between +1.0 (similar patterns were found) and –1.0 (nearby areas tend to be dissimilar) although a zero outcome for all neighborhood distances denotes a pure nugget-effect situation for

Kriging (see Section 3.4 for a further review). If there is no distance relationship, how can anyone weight by a distance function?

For Wong and Lee [2001], the mean of Moran I equals $E(I) = -(n-1)^{-1}$ although this value will tend to zero as the sample size increases. “Because Moran I is calculated under all $n!$ ⁵⁶ random permutations, one does not have to know the true distribution since it becomes asymptotically Normal for datasets greater than 25 observations” [Skinner, 1999]. If the standard value of Moran I⁵⁷ is less than the Gaussian significance value for a certain confidence level⁵⁸ then the null hypothesis of the random observed arrangement must be accepted. Therefore, the location within the research region can be regarded as a random value drawn from a Gaussian distribution and not due to a systematic process. In addition, the Moran I cumulative peak⁵⁹ should be analogous with the range for the sample variogram and it can be expected to have a similar value. “If these ranges do not match then larger-scale patterns were not modeled for larger lags by the variogram” [Skinner, 1999].

2.4.5. The Geary C

Wong and Lee [2001] recommend a variant from the previous coefficient for positive values only, consisting of the sum within neighbor locations for a specific range divided by the sum of all observations (see Equation 5).

$$C = \frac{(n-1) \sum_i \sum_j w_{ij} (x_i - \bar{x})(x_j - \bar{x})}{\sum_i \sum_j w_{ij} (x_i - \bar{x})^2 + \sum_j \sum_i w_{ji} (x_j - \bar{x})^2} \quad [5]$$

where:

$\bar{x}$ symbolizes the overall mean;

w_{ij} is the weight of matrix **W** between i^{th} and j^{th} ;

n represents the total number of observations;

x_i and x_j are i^{th} and j^{th} sample value.

It should be noted that the standard deviation denominator of Moran I and Geary C are almost the same, although the latter is more sensitive to extreme values. Moreover, areal units with more connections are being weighted more heavily with Moran I than Geary C. Under the random map pattern for a Gaussian distribution, the expected value of Geary C equals 1. For positive autocorrelated data, Geary C becomes less than one while a dissimilar cluster tendency reveals a Geary C greater than one (see Figure 10).

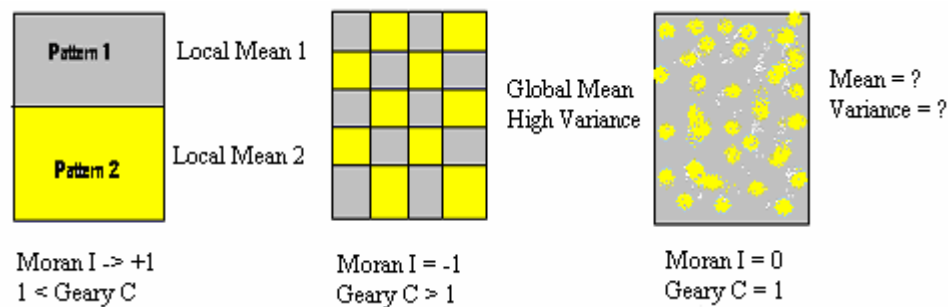


Figure 10 – The Moran I and Geary C measures inferences.

2.4.6. The Local $G_i(d)$ and the Global $G(d)$

As an alternative for identifying local patterns not revealed by the global ones, a spatial scale problem, Getis and Ord [1992] introduced the local $G_i(d)$ statistic for positive variables. This index measures (see Equation 6) the degree of clustering that result from the spatial concentration, not correlation, of the sum of points within distance d of the point i^{th} . Thus, $G_i(d)$ is the sum of all $x_i x_j$ values within distance d of point i . A large standard negative index⁶¹ implies small values of x_j ⁶² within location i while a large positive one involves larger ones⁶³. Accordingly, a spatial pattern has been found. This statistic is particularly suitable for detecting spatial concentration in one sub-region. "If one notes changes in $G_i(d)$ as d increases, trends in the data can be identified" [Getis, 1996]. Furthermore, " $G_i(d)$ is scale-invariant⁶⁴ but not location-invariant⁶⁵, inappropriate with regression residuals because it is centered on positive variables and sensitive to transformations such as $x_i = \log(x_i)$ " [Ord, Getis, 1992].

$$G_i(d) = \frac{\sum_{j=1}^n w_{ij}(d)x_j}{\sum_{j=1}^n x_j} \quad [6]$$

$$G(d) = \frac{\sum_i \sum_j w_{ij}(d)x_i x_j}{\sum_i \sum_j x_i x_j} \quad [7]$$

where:

$w_{ij}(d)$ is the weight of matrix **W** between i^{th} and j^{th} according to cutoff distance d ;

n represents the number of observations;

x_j is the j^{th} neighborhood sample value of the i^{th} site.

Table 3 presents the $G(d)^{66}$ standard normal variation that measures the overall concentration or lack of it under differing circumstances. Moran $I(d)^{67}$ has a peculiar weakness because it does not discriminate patterns that hold high or low dominant values while $G(d)$ does not recognize clearly high-low regions. For this reason, there are advantages in using both statistics to explore spatial processes. Yet, both coefficients have difficulty in differentiating between a random pattern and one with little spatial variation.

Values Situation	Z(G)	Z(I)
High-High	++	++
High-Medium	+	+
Medium-Medium	0	0
Random	0	0
High-Low	-	--
Medium-Low	-	-
Low-Low	--	++

Table 4 – The standard global G and I indices.

According to Ord and Getis [1992], Griffith and Layne [1999], $I(d)$ and $G(d)$ expected no statistical spatial correlation for the Sudden Infant Death Syndrome dataset in North Carolina 100 counties between 1974 and 1978. But their values clearly show a tendency toward a positive spatial autocorrelation for a distance of between 30 and 40 miles. The standard value of $G_i(33)$ was computed and four highly spatially-clustered counties in the south were statistically found for the

social investigator, epidemiologist or other regional scientist to explain the subtle pattern.

2.4.7. The Local Moran

“Local Moran spatial association focuses on hot spots and on diagnosing local instabilities within the overall spatial association” [Anselin, 1996, Wong and Lee, 2001]. Based on the null hypothesis of no local clustering, the LISA coefficient (see Equation 8) combines the value resemblance with Moran I. In addition, this statistic is similar to the local G_i with a smaller difference: if a positive G_i value indicates high-value clustering, a positive L_i represents spatial clustering of similar values (either high or low) while a negative one describes a dissimilar one (high values surround by lower ones or vice-versa).

$$L_i = \frac{x_i}{m_2} \sum_j w_{ij} x_j \quad [8]$$

where:

x_j represents the neighbor values of the i^{th} location;

$$m_2 = \sum_i x_i^2 / n ;$$

x_i is the sample value.

2.4.8. The Moran Scatterplot

Anselin [1996, 1998] indicates the Moran scatterplot to visualize and to identify the degree of spatial instability based on the bivariate regression coefficient of the spatial lagged variable ⁶⁸($\mathbf{Wx}$) against the original variable (x). The four quadrants centered on the global mean are composed of the x-axis, deviations from the original variable mean, and the y-axis, neighborhood weight average mean deviations. As expected, this scatterplot is divided into four association types: the upper right quadrant (high values above the mean surrounded by high values), the lower left (high values surrounded by low values), the upper left (low values

surrounded by high values) and the lower right (high values surrounded by low ones). Peculiarly, $G_i(d)$ captures the first and third quadrants. However, a positive L_i groups the first and third quadrants as well as the second and fourth with a negative L_i .

2.5. SPATIAL AUTOREGRESSION MODELS

Can [1997] presents the spatial autoregressive model (SAR). Given an n -by-1 vector of spatially correlated $\mathbf{Y}$ -dependent observations, the SAR model and its variance with several independent explanatory variables has the form:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\xi} = \mathbf{X}\boldsymbol{\beta} + \rho(\mathbf{I}_n - \rho\mathbf{W})^{-1}\mathbf{W}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \quad [9]$$

$$\text{SAR}(\sigma^2) = \sigma^2[(\mathbf{I}_n - \rho\mathbf{W})'(\mathbf{I}_n - \rho\mathbf{W})]^{-1} \quad [10]$$

where:

$\mathbf{I}_n$ is the identity matrix;

$\mathbf{W}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = \boldsymbol{\xi}$ is the n -by-1 residuals vector;

$\mathbf{W}$ is the n -by- n matrix connectivity row;

$\mathbf{X}$ is the n -by- p explanatory observations matrix;

$\mathbf{Y}$ is the n -by-1 independent observations vector;

ρ is the spatial autoregressive coefficient between units i and j ;

σ^2 equals sampling variance;

$\boldsymbol{\beta}$ is the 1 -by- p vector to be estimated using MLE.

“Although no estimator is uniformly superior under all conditions, most research studies indicate that MLE⁶⁹ is superior to the conditional least squares estimators for small samples” [SAS Institute, 2001]. In terms of results, this spatial MLE error model⁷⁰ fits best with the highest Information Criterion (IC), log likelihood (LIK) and pseudo- R^2 or with the lowest Akaike Information Criterion (AIC) and Schwartz Criterion (SC). The significance hypothesis test of the model coefficients is based on the standard Normal distribution instead of the traditional t -Student. Furthermore, the χ^2 heteroskedasticity and χ^2 spatial lag dependence tests for the correlation coefficient are confirmed by the Breusch-Pagan⁷¹ and the Likelihood Ratio⁷² tests, respectively. For Li [1996], the best estimator of the SAR model and variance are:

$$\mathbf{Y}=[(\mathbf{X}-\rho\mathbf{WX})'(\mathbf{X}-\rho\mathbf{WX})]^{-1}(\mathbf{X}-\rho\mathbf{WX})'(\mathbf{Y}-\rho\mathbf{WY}) \quad [11]$$

$$\text{SAR}(\sigma^2)=1/n(\mathbf{Y}-\rho\mathbf{WY}-\mathbf{X}\beta+\rho\mathbf{WX}\beta) \quad [12]$$

where:

$\mathbf{W}$ is the n-by-n matrix connectivity row;

$\mathbf{X}$ is the n-by-p exploratory observations matrix;

$\mathbf{Y}$ is the n-by-1 independent observations vector;

ρ is the spatial autoregressive coefficient between units i and j ;

σ^2 equals sampling variance;

n is the number of observations;

β is the 1-by-p vector to be estimated using MLE.

Another two versions presented by Griffith and Layne [1999] are the conditional autoregressive model (CAN), frequently used in remote sensing, and the autoregressive model (AR) used by spatial econometricians. For instance, Long [1996] discusses a 34 cultivar block study in Montana, USA, using conventional ANOVA and SAR⁷³. The R^2 increased from 69.9% to 74.6% with SAR and a reduction in MSE⁷⁴ (15.4%) and MSS⁷⁵ (7.2%). The spatial autocorrelation factor accounts for 9.8% of total variation in grain yield. According to Griffith and Layne [1999], “incorporating spatial autocorrelation results in a 5-10% improvement in the statistical description for a given georeferenced dataset”. Because underlying it is a mathematical design with massive computation, unless the brute-force of parallel computers is used⁷⁶, it is unlikely to be computed with very large spatial-temporal datasets. For instance, Table 5 presents the SAR benchmark comparison between the sequential IBM[®] RS6000 and parallel computers. The great performance gap between both approaches is well-known. However, for matrices smaller than 100-by-100, the RS6000 calculates faster than the CM5.

Matrix Size	RS6000 (second)	CM5 (second)	Ratio
400	38.19	3.6	10.6
2500	8566.5	48	178.47

Table 5 – The SAR benchmark by Li [1996].

Confirming previous conclusions, Anselin [1988] produced a good model by introducing the spatially lagged autocorrelation factor⁷⁷ into the Columbus Ohio

dataset that relates crime (y) with income and housing⁷⁸. With robust OLS (see Equation 13), the R^2 equals 53.29% while the White heteroskedasticity index equals 19.94⁷⁹, a highly significant value.

$$y=68.61-1.59\times\text{income}-0.27\times\text{housing} \quad [13]$$

Since spatial error dependence is a reality⁸⁰, the MLE spatial lag model (see equation 14) improves the previous one:

- The LIK index changed from -197.4 to 380.3;
- The AIC and SC indices decreased from 380.8 to 372.8 and 386.4 to 380.3, respectively;
- The R^2 improved to 61.51%;
- No spatial error dependence is found.

$$y=45.07+0.43\times w\_crime-1.03\times\text{income}-0.26\times\text{housing} \quad [14]$$

Under the general linear mixed model for spatial variability, Coelho and Pinheiro [1998] describe a research case concerning the study of a consumption function using 248 families distributed among 64 regions. The case exhibits a spatial correlation between observations groups. Using the Restricted Maximum Likelihood (AIC=-460.9, SBC=-466.1), family consumption per region equals 9.10 plus 0.61 of their income. Comparing this with the traditional linear model⁸¹, both authors conclude that the former approach offers not only a way of characterizing spatial correlation structures but also improves description efficiency.

2.5.1. Spatial Autoregression: Residuals

Residuals are the only available data that can be considered as a reasonable reality-approximation test. The chance to verify existing patterns among residuals provides a key information source on possible assumption violations, variable transformations, outliers, trends surface and inappropriate formulations for raw

data [Clark and Hosking, 1986, Haining, 1996]. For instance, “the estimator variance of OLS regression depends on residual variance” [Pinto and Curto, 1999].

In theory, errors assumptions are:

- It is expected that the sum of the residuals equals zero;
- Errors display a constant variance;
- Residuals have a Gaussian distribution;
- Errors are independent of each other.

With linear autocorrelation residuals for multiple regression models, according to Ebdon [1998], error assessment equals:

$$\frac{\sum_{a,b} \frac{ab}{n} - \frac{a}{n} \frac{b}{n}}{S_a S_b} \quad [15]$$

where:

$$S_a = \sqrt{\frac{\sum a^2}{n} - \frac{a^2}{n}} \quad \text{and} \quad S_b = \sqrt{\frac{\sum b^2}{n} - \frac{b^2}{n}};$$

a equals the first $n-1$ residuals;

b is the last $n-1$ residuals;

ab represents the product between two consecutive or adjacent ones.

The complete absence of autocorrelation among residuals would give zero while a value close to +1.0 or –1.0 suggests a directly or inversely close relationship presence between successive residuals. “If so, the inference based on the usual t and F statistics will be misleading and R^2 goodness-of-fit will be wrong” [Pinto and Curto, 1999].

Graphically, with regard to predicted values against the associated error (see Figure 11) in a spatial linear framework⁸², situation (a) is desirable with homogeneous variance. The prediction intervals associated with interpolation are correct. However, (b) involves a WLS approach to model the continuous spatial drift. Situation (c) needs a linear structure while case (d) requires a quadratic model or a transformation prior to regression.

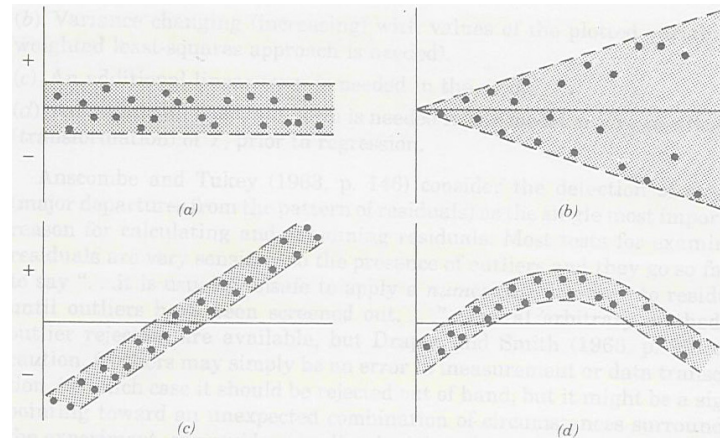


Figure 11 – Residual plots where the x-axis represents the predicted values while the y-axis is the associated error [Clark and Hosking, 1986].

“If map residuals show some systematic pattern distribution then squeeze the residuals until they talk” [Griffith and Layne, 1999]. In other words, spatial autocorrelation among residuals can also be viewed as a diagnostic process tool and a correction factor. Hence, spatial autocorrelation error can be defined as a specification error artifact for spatial modeling. Under this post-ESDA view, a large negative or positive residuals show spatial regimes because of the lack of model explanatory variables. With the net migration Rwanda [Brown, 1996], for instance, spatial residuals lead to the discovery of unknown influences using an OLS regression model ($R^2=0.417$):

$$mi=1.362-2.4e-05 \times pd+0.004 \times sl+2.7e-07 \times in+4.9e-05 \times de-0.0003 \times ra-0.012 \times sc \quad [16]$$

where:

mi is migration;
 pd is population density;
 sl is spare land;
 in is income;
 de is development;
 ra is rainfall;
 sc is soil capability.

However, the immigration underestimation by the previous model within the southern region is due to the swampy area⁸³ while northern region overestimation is because of the former Gishwati Forest, a recently opened settlement. Since these two extra factors were not taken into account, a spatial lag variable was added to represent the neighborhood effect of the net migration rate into the autoregressive

model using MLE ($R^2=0.472$). As expected, the sum confidence of the residual squares improves from 10.38 to 7.74 and AIC decreased while the standard errors of the estimates become lower.

$$mi=0.648+0.402\times\text{lagged\_migration}+0.003\times sl+1.71e-07\times in+5.15e-05\times de \quad [17]$$

where:

mi is migration;
sl is spare land;
in is income;
de is development.

2.5.1.1. LADS

“The localized autocorrelation diagnostic⁸⁴ test is a further diagnosis tool for spatial pattern residuals and for omitted variable identification, distinct regimes and error heteroskedasticity within spatial multivariate regression models” [Nass and Garfinkle, 1992]. The aim of LADS($N\times N,E,C$) is to determine whether the presence of N geographical entities with C contiguous blocks among the most E extreme positive and negative residuals is worth further investigation. Therefore, if the localized autocorrelation level is smaller than the established critical level then the identified units can suggest a new variable to improve the current model. According to both authors, the regression model of the US government expenditure is ($R^2=0.65$):

$$\text{USA\_Expenditure}=-072+0.38\times in-0.93\times sc-0.14\times pd+116.3\times 1/po \quad [18]$$

where:

in is income;
sc is school;
pd is population density;
po is population.

Given the seven most negative residuals of ten, the LADS($50\times 50,10,7$) computation resulted in 0.02 and, for a critical level of 5%, the contiguity pattern of the seven states (Maryland, Michigan, New Jersey, North Carolina, Ohio, Pennsylvania and Virginia) was worthy of further investigation. Urbanization

population, an outside independent variable with mean=53.1⁸⁵ were added to the model leading to a new one ($R^2=0.7$):

$$\text{USA\_Expenditure} = -0.31 + 0.30 \times \text{in} - 0.68 \times \text{sc} - 0.14 \times \text{pd} + 53.0 \times 1/\text{po} - 0.23 \times \text{ur} \quad [19]$$

where:

in is income;
 sc is school;
 pd is population density;
 po is population;
 ur is urbanization.

Remark that the omitted variable, urbanization, came to the attention of the researcher only after the initial model had been estimated. Thus, LADS can be a useful tool for the diagnosis of localized spatial pattern residuals and for the identification of omitted variables or regimes in regression models.

2.5.1.2. Join-Count Statistics

Residuals can also be tested by the binary presence-absence of the join-count statistic⁸⁶. These dual state variables are based on the contiguity test of negative (N) and positive (P) residuals patterns, a join measurement between neighboring points with non-zero boundaries. A further downgrading possibility to this binary form is cut-off level. So, unit joins can be PN, NN or PP, following the ‘bell’ curve within the binary map but based on the random spatial arrangement null hypothesis⁸⁷ in a randomization sampling test⁸⁸. Equations 20, 21 and 22 give the three standard indices that represent the joins number between negative, positive and negative-positive units. If any of these indices are greater or less than ± 1.96 for a 95% confidence level then the respective null hypothesis is rejected and residual contiguity emerges.

$$Z(\text{NN}) = \frac{J(\text{NN}) - \mu(\text{NN})}{\sigma(\text{NN})} \quad [20]$$

$$Z(\text{PP}) = \frac{J(\text{PP}) - \mu(\text{PP})}{\sigma(\text{PP})} \quad [21]$$

$$Z(PN) = \frac{J(PN) - \mu(PN)}{\sigma(PN)} \quad [22]$$

where:

$J(NN)$ is the number of links between negative neighbor's pairs;

$J(PP)$ is the number of links between positive neighbor's pairs;

$J(PN)$ is the number of links between positive and negative neighbor's pairs;

$\sigma(NN)$ is the variance of negative neighbor's pairs;

$\sigma(PP)$ is the variance of positive neighbor's pairs;

$\sigma(PN)$ is the variance of between positive and negative neighbor's pairs;

$\mu(PP) = 1/2 \sum \sum w_{ij} x_i x_j$ (the expected link average between positive neighbor's pairs);

$\mu(NN) = 1/2 \sum \sum w_{ij} (1-x_i)(1-x_j)$ (the expected link average between negative neighbor's pairs);

$\mu(PN) = 1/2 \sum \sum w_{ij} (x_i x_j)^2$ (the expected link average between positive and negative neighbors pairs).

2.6. SPATIAL AUTOCORRELATION, AUTOREGRESSION AND KRIGING: THE MISSING DATA ISSUE

Spatial autocorrelation and statistical heterogeneity hangovers hold the ability to compare two regions and to characterize texture differences. Quite often, distance location pairs are less similar (competitive spatial processes) than closer ones (cooperative spatial processes). Probably, some landscapes can exhibit extremely irregular shapes. As a consequence, indices of spatial autocorrelation calculated globally and locally are valuable for descriptive purposes because they provide a measure of how similar objects are to their spatial neighbors. This spatial dependence impact is also crucial on spatial inference interpolation such as Kriging.

“The word Kriging is synonymous with the optimal prediction of unknown values from observed data at known locations” [Journel, Huijbregts, 1978, Aunón, Hernandez, 2000]. After the variogram has been defined, the algebraic relationship between values at individual distances including clustering and direction is used to estimate Kriging weights. Like any linear estimator, the estimation equals:

$$x_0 = \sum_{i=1}^n w_i x_i \quad [23]$$

where:

w_i is the weight given to each available sample;

x_0 is the estimated point;

x_i represents the i^{th} sample.

Mostly, four factors are taken into account in assigning weights:

- Closeness to the location being estimated;
- Redundancy between data values;
- Anisotropic continuity;
- Magnitude of continuity.

“Kriging is B.L.U.P. (best linear unbiased predictor) whether or not data are normally distributed” [ESRI, 2001]. It is linear since estimations are weighted linear combinations of the available data. It is unbiased because the error mean is zero (no over or under-estimates). It is best since its goal is to minimize error variance. Conceptually, the variogram curve has minimal error variance at any known data location⁸⁹ and maximum value at some specified range away from that point. Also, “all locations outside the range are considered to be unaffected by known data although those weights may not equal zero due to their contribution to the trend component estimation at that particular location” [Geostatistics FAQ, 1997]. As expected, the predictable maximum error Kriging deviation occurs in areas where there are no points.

Once again, it is the notion that variation is spatially autocorrelated in some predictable fashion that creates a fundamental conflict between the goals of classical statistics and its assumptions because classical estimators tend to be inconsistent as the degree of spatial dependence increases. As Levine [1997] and Limaïem and Elmaraghy [1999] point out, “this may come as a surprise to ecologists and geographers who have been trained in the belief that Mother Nature follows the assumptions of classical statistics”.

The non-linear regression predictor plus spatial autocorrelation provides another linkage with Kriging, the missing data issue. Aspects of the missing data issue require the estimation of omitted spatial data when undertaking spatial autoregressive methods and Kriging. Yet, spatial autocorrelation can also deal with spatial prediction. For instance, a high degree of spatial autocorrelation suggests an equally likely chance of predicting neighboring values. Also, a low value reveals a low level of spatial data redundancy.

Eight relationships among these three concepts emerge:

- Spatial autocorrelation is the progenitor of Kriging and autoregression models.
- Spatial autocorrelation itself seeks description and diagnosis while spatial regression and Kriging seeks prediction.
- The variance-covariance matrix is included within spatial regression and Kriging.
- Once a variogram is fitted to the sample data, Kriging can be used for locations at which data are not sampled.
- “With spatial regression models, the missing data can be regarded as an interactive re-estimation solution fashioned with updated variable imputations based on R^2 and IC decision parameters in ML, OLS and bootstrap procedures” [Griffith and Layne, 1999].
- Kriging is primarily concerned with more or less continuous attributes while spatial autoregression involves aggregations of phenomena into discrete regions such as areal units.
- If autoregressive and trend surface methods assume that samples follow an underlying trend plus random residuals, Kriging assumes the exact reverse, the values of which are interrelated with (Universal Kriging) or without (Ordinary Kriging) background trend.
- Since Kriging honors data at sampled locations, autoregressive residuals are not precisely similar to Kriging estimation errors.

2.7. W MATRIX: IDW VERSUS BINARY CONTIGUITY

“Understanding spatial patterns allows geographers to recognize phenomena distribution and how they can be compared with others by determining the underlying environmental factors that are changing them” [Wong and Lee, 2001]. To emphasize them or to correct those changes becomes, thus, the issue. By taking into account both location and attribute data, spatial autocorrelation measures can become a useful tool for describing their forms and detecting changes over time.

Major literature such as Practical Handbook of Spatial Statistics recognizes that previous autocorrelation measures only concern polygon distribution where spatial **W** quantification is based on the IDW, rook, bishop and queen binary⁹⁰ matrix. Yet, if spatial objects are points, the vicinity boundaries become unclear, reflected in **W** matrix contents and in spatial autocorrelation and clustering indices results. “The choice of geographical weights for spatial statistical modeling is not a clear cut issue although it is mostly done in an *ad hoc* manner” [Wong and Lee, 2001]. In fact, it is governed, primarily, by convenience or convention if the rook and queen is chosen: it is symmetric, its diagonal is zero and row sum indicates total neighbor number for each area unit, while its division by two equals the total number of shared joins.

Upton [1992] has already explored a singular distance decay model:

$$(\text{population})^{\alpha_1} \times (\text{area})^{\alpha_2} / d_{ij}^5 \quad [24]$$

where:

- α_1 and α_2 are the binary exponents;
- d_{ij} is the distance between areal centroids i and j ;
- Area and population factors regard one particular region;
- The diagonal weights are given by $\text{area}_j / d_{ij}^{-3}$.

Additionally, Florax and Rey [1996] have studied the statistical power impact of **W** misspecification within spatial econometric models:

- Both under⁹¹ and over specification⁹² produce an increase in the mean squared error;
- Over and under-specification within autoregressive models results in an inflated variance estimation of 10%;
- Moran I seems to retain better statistical power than any other spatial autocorrelation test in the presence of **W** misspecification;
- Smaller sample size magnifies differences created by geographical **W** matrices;
- It is better to conceive some reasonable geographical **W** specification than to assume zero entries⁹³;
- For a spatial autocorrelation of +0.2, the mean estimator loss efficiency for autoregressive models due to **W** misspecification is roughly equivalent to a 10% reduction in sample size.
- With weak positive spatial autocorrelation presence, over and under-specification makes little difference.
- With a moderate one, under-specification appears to be more seriously affected while, in the presence of strong spatial autocorrelation, complications become worse with over-specification.

Three **W** matrix criteria will be covered here with regard to their impact on Moran I and Moran scatterplots:

- The Euclidean cutoff distances where the connectivity matrix is constructed with 0s and 1s (binary contiguity);
- $w_{ij}=1/d_{ij}$ (distance decay contiguity);
- $w_{ij}=1/d_{ij}^2$ (double distance decay contiguity).

2.7.1. Case Study 1: Appendix I

“There is no reason why we should not interpret Tobler’s Law in terms of a gradual, incremental, attenuating distance effect as we traverse successive cells” [Longley *et al.*, 2001]. The chosen sampling layout involves six areas presented in

Appendix I where all distances between centroids are 2 units, with the exception of a distance of 1.4 units between points D-E and D-B. It is clear that different weights lead to an assorted Moran I and Moran scatterplots:

- The **W** matrix is not influenced by sample values.
- With the first case, as Moran I gets positively stronger, the [A,B,C] region patterns (quadrant I) become apparent although IDW enhances it in a clear way. Area [E] is considered an outlier (quadrant IV) because two out of its three neighbors hold very low values. Analogously, low region [D] with three high neighbors and one low leads to a vague position between quadrants II and III. As Moran I increases, [B] shifts to quadrant I.
- With the second case, as Moran I gets less negative, region [B] moves towards quadrant I. On the contrary, low areas [A,D,F] do not layout within quadrant III because of the connectivity lacking between areas [A] and [D,F] and strong connections with high values of [B,C,E]. If low area [A] is exchanged with high [F], the [A,D,F] pattern will move to quadrant III. Analogously, high regions [B,E] are enclosed by low ones and, therefore, they are displayed on quadrant IV.
- With case four, negative Moran I emphasizes outlier [F] of quadrant II and regions [E,D] of IV. Case five corroborates with the sense that the IDW matrix never holds the extreme Moran I indices.
- By choosing different neighbor limits for the same layout, observations can only exchange positions between quadrants I and IV or II and III in a vertical movement whose amplitude is, in average, 15% of the overall mean. Positive Moran I stresses local pockets of dependence (quadrant I and III) and a negative one highlights outliers (quadrant II and IV). Figure 12 also highlights the region of the shift or transition area between quadrants due to a **W** misspecification change (I-IV or II-III) or mean cutoff value (I-II or III-IV)⁹⁴.

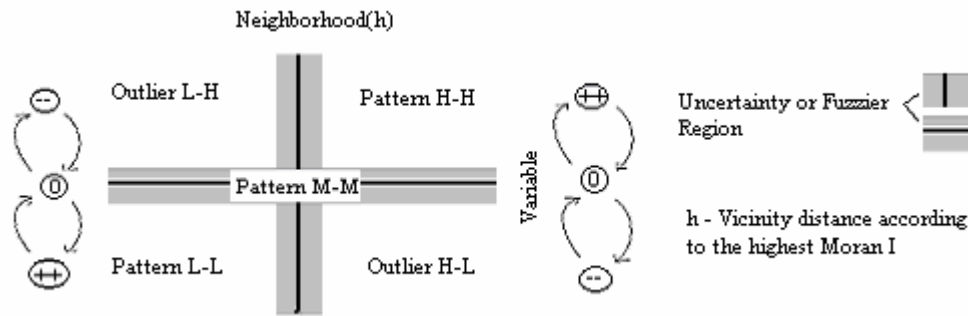


Figure 12 – The Moran scatterplot exchanged observation position according to Moran I.

- Case six proves that zero Moran I produces a random and dispersed outcome along all quadrants. By plotting a neighborhood average (Wx) against the sample value (x) regression line (see Figure 13), it is possible to verify what type of spatial autocorrelation exists within the dataset.

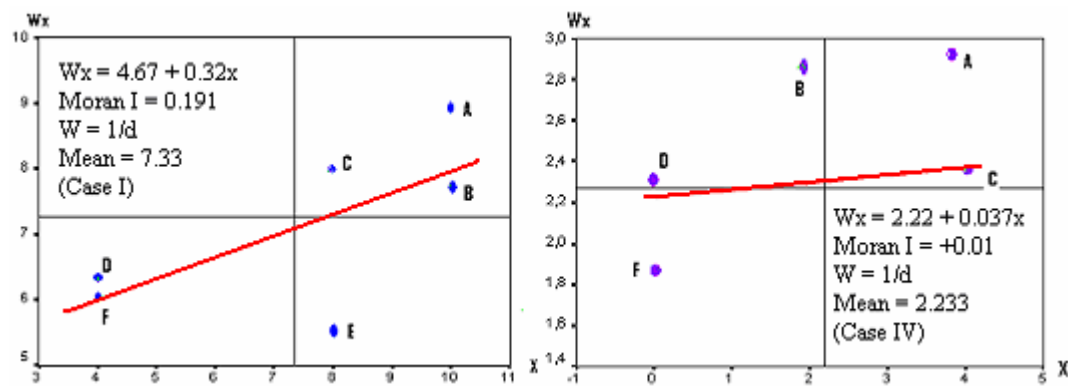


Figure 13 – The linear relationships of Moran scatterplots regarding case one and case six.

- In Figure 14, the contribution of $1/d^5$ and $1/d^4$ to each sample regarding point K holds a massive power for the closest two samples. If binary contiguity were considered, the same weight would be given to its natural neighbors (1, 2, 3 and 4, for instance) but leading to a more smooth weight situation among neighbors.



Figure 14 – A comparison of $1/d^x$ contiguities for the impact of weight samples.

2.7.2. Case Study 2: Appendix J

The city of San Diego, California, is not a homogeneous region where reasonable small sections of the metropolitan area follow a distinct spatial autocorrelation tendency like major cities in this world. “Because Rancho Santa Fe is the richest suburban area with house prices nearly three times higher than the next highest district, La Jolla, its value was removed because of the lack of robustness of spatial autocorrelation statistics” [Getis, Ord, 1992]. This situation might be regarded as a restrictive sampling issue. The total housing costs mean equals 192.81 while variance, skewness and kurtosis are 5523, 0.83 and 0.20, respectively. The Moran I of +0.2783 clearly indicates there is a significant spatial autocorrelation for the zero and one **W** specification although its value decreased to +0.1106 when $1/d^4$ distance decay contiguity was applied. As expected, the total outlier number increased. The Coronado (43) district does not hold any neighbor according to the binary contiguity approach, confirming this difficulty regarding island regions (see Figure 15).

From Appendix J, a certain refinement of knowledge can be gained:

- Binary contiguity and $1/d^4$ highlights similar patterns and outliers. However, the outlier layout of quadrant II is quite different.
- IDW holds the capability of limiting points around the Moran scatterplot center (44 districts), confirmed by the weak Moran I of +0.0742.
- $1/d^2$ and $1/d^3$ **W** matrices recognize patterns better than $1/d$ and this is reinforced by their Moran I: +0.1630 and +0.1809.

- Binary contiguity puts more emphasis on the rich northern districts (a long, thin pattern) of Encinitas (1), Cardiff (2), Del Mar (4), Lake Hodges (5), Rancho Penasquitos (6), Rancho Bernardo (7), La Jolla (11), Beaches (14), Bay Park (15) and Kearny Mesa (16).
- The $1/d^4$ distance decay contiguity stresses the San Diego center region (quadrant III in a circular pattern context) of Scripps Ranch (9), Mission Village (17), Tierrasanta (18), Del Cerro (19), San Carlos (20), Santee (21), Lakeside (22), El Cajon East (23), El Cajon West (24), La Mesa (25), Mission Valley (28), Lemon Grove (32), Spring Valley (33), National City (37), Logan Heights (38), East San Diego (39) and Center San Diego (48).

“In geostatistics, one type of neighborhood structure is often used: the elliptical⁹⁵” [Remy, 2001]. So, the Moran scatterplot is readjusted in order to include sampling anisotropy whose sample vicinity limits are defined by the standard deviational ellipse⁹⁶. Centered on the mean center⁹⁷, the standard deviational ellipse reveals directional process bias of sampling. Computationally, three steps follow:

- 1) Coordinate system transposition of all samples to the mean center ($x_i' = \text{coord\_}x_i - \text{average\_coords\_x}$, $y_i' = \text{coord\_}y_i - \text{average\_coords\_y}$);
- 2) Clockwise rotation angle (θ).

$$\text{tangent}(\theta) = \frac{(\sum x_i'^2 - \sum y_i'^2) + \sqrt{(\sum x_i'^2 - \sum y_i'^2)^2 + 4(\sum x_i'y_i')^2}}{2\sum x_i'y_i'} \quad [25]$$

where:

x_i' is the transformed coordinate x of the i^{th} sample;

y_i' is the transformed coordinate y of the i^{th} sample;

θ represents the clockwise angle in radians.

- 3) Shorter and longer axis length assessment.

$$\text{Length}_{xx} = \sqrt{(\sum x_i'^2)\cos^2\theta - 2(\sum x_i'y_i')\sin\theta\cos\theta + (\sum y_i'^2)\sin^2\theta} / n \quad [26]$$

$$\text{Length}_{yy} = \sqrt{(\sum x_i'^2)\sin^2\theta - 2(\sum x_i'y_i')\sin\theta\cos\theta + (\sum y_i'^2)\cos^2\theta} / n \quad [27]$$

The rotation angle equals 74° while major and minor axes equal 9 and 6 units, respectively. Comparing its Moran scatterplot with its binary contiguity counterpart, few differences are of importance: 29 districts hold the same position while 21 shift slightly its position, particularly in the lower central and higher northern districts on account of the presence of the strongest spatial autocorrelation.

2.7.3. Discussion

These results justify some clarification. When a local pattern respects a long but narrow arrangement like the rich northern area of San Diego, binary contiguity fits best because it avoids the smoothing effect of inverse distance decay contiguity and, thus, it places those points farther away from the Moran scatterplot center. With binary contiguity, each neighbor's weight is inversely proportional to the total number of neighbors and, therefore, it becomes more force resistant to the gravity force in relation to small patterns or long but narrow ones than with inverse distance contiguity. Cardiff (2), for instance, is the only Encinitas (1) neighbor receiving, thus, 100% of the total weight. With the $1/d$ approach, Cardiff (2), Solana Beach (3) and Del Mar (4) receives 13%, 7% and 5%, a more even situation, while with $1/d^4$ methodology, those three regions' contribution reaches 85%, 11% and 1%, respectively. As the distance difference among neighbors gets larger, more weight is given to the closest regions (see Table 6).

Distance Decay Contiguity Weights					Binary
Nearest Neighbor	$1/d^4$	$1/d^3$	$1/d^2$	$1/d$	
1 st	80%	65%	40%	32.5%	100% divided by the total direct number neighbors
2 nd	15%	17.5%	25%	27.5%	
3 rd	5%	7.5%	12.5%	20%	
4 th		10%	10%	12.5%	
5 th			12.5%	5%	
Others				2.5%	

Table 6 – The average weights for the first four distance decay contiguities and the binary one.

However, the eight neighbors of Center of San Diego (48) receive 12.5% each, if binary contiguity is applied⁹⁸. With $1/d$ methodology, the neighborhood weights range between 2.9% and 17.4% leading to a smoother neighborhood average. With $1/d^4$, this scope fluctuates between 0.8% and 52%. This trend is confirmed by the following domains: [218.67, 168.87] of $1/d$ against [283.44, 114.58] of $1/d^4$. The major concern of IDW specification is the smoothing effect by of concentrating their positions close to the Moran scatterplot center and, thus, a smaller Moran I. On the other hand, as $1/d^z$ decreases when $z \rightarrow \infty$, the neighborhood average is reduced to the closest sample unless the distance difference between the closest and the next ones are very small. With $1/d^5$, the closest neighbor holds, on average, 90% of all the weights. This gives the suggestion that the more $1/d^z$ decreases, the wider the neighborhood range turn out to be and, thus, the more anti-gravitational, the Moran scatterplot layout becomes. What, then, is the best distance decay contiguity? No rules can be setup.

If from the statistical viewpoint this contiguity process is correct, it does not make sense to reduce vicinity to the closest sample in spite of the fact that neighborhood concept is a fuzzy one. Another major concern is the dissimilarity found in the Moran I regarding binary contiguity (+0.27) against distance decay contiguity. With $1/d^6$, for instance, this index equals -0.21 (see Table 7). However and looking closely at the distribution map, the city of San Diego pursues a distinct spatial autocorrelation pattern and a positive Moran I is expected like major spatial Mother Nature processes.

Distance Decay Contiguity	Moran I	Distance Decay Contiguity	Moran I
$1/d$	+0.07	$1/d^6$	-0.21
$1/d^2$	+0.16	$1/d^7$	-0.42
$1/d^3$	+0.18	$1/d^8$	-0.62
$1/d^4$	+0.11	$1/d^9$	-0.81
$1/d^5$	-0.02	$1/d^{10}$	-0.98

Table 7 – The Moran I results according to distance decay contiguities.

So, the key question regarding the IDW view (see Section 3.1.5 for a further discussion) is the regular number of neighbors' numbers considered by this

methodology. With $1/d$, the nearest five neighbors represent 97.5%, in average. With $1/d^2$, the nearest four neighbors hold 87.5%. With $1/d^3$, the nearest three neighbors stand for 90%. With $1/d^4$, the nearest two neighbors own 95%. With $1/d^5$, the nearest one retains around 97%. Therefore, inverse distance decay is not sensitive to the total number of natural neighbors that any region might hold (see Table 8). For instance, with $1/d$, the nearest five neighbors are the ones that counted regardless of the true number of direct neighbors. With $1/d^4$, the two nearest neighbors are the ones that make a difference despite vicinity districts. However, according to the standard deviational ellipse approach (Appendix J), twenty-one counties have only one or two natural neighbors, fifteen have three, four or five while eleven counties have between six and eight.

According to Anselin, [1992], “the neighborhood first order contiguity structure for West Virginia counties averages five links per county, although their composition varies quite considerably”. This statement supports the San Diego argument: one county has eight neighbors, thirteen counties enclose two or three connections while thirty four register between four and seven neighbors. There are no islands and anisotropy is not being taken into account. Distance decay contiguity does not respect the real direct neighbors that any location might hold. Finally, distances of less than one may create problems regarding IDW computation although distance matrix rescaling such that the smallest distance equals one can be made.

The fact is IDW by itself is not a good process for the **W** specification because it is blind and insensitive to its obvious natural neighbours, unlike the binary approach. Contrary to major literature, IDW must be used in combination with the binary approach to recognize direct neighbors whose standard weights are closely proportionate to the central-neighbors’ distances. Returning to the eight neighbors of San Diego Center (48), for instance, their weights become 17% (Spring Valley), 8.3% (Paradise Hills), 12.6% (National City), 12.6% (Logan Heights), 12.6% (West San Diego), 11.4% (East San Diego), 16% (La Mesa) and 9.5% (Lemon Grove) with this mix approach leading, therefore, to a more prudent situation.

Number of Natural Neighbors	Districts Examples
0	Coronado (43)
1	Encinitas (1)
2	Otay Base (2)
3	Jamuk (34)
4	National City (37)
5	College (26)
6	Rancho Penasquitos (6)
7	East San Diego (31)
8	Kearny Mesa (16)

Table 8– The number of direct neighbors according to the binary contiguity of some San Diego districts.

2.8. CONCLUSIONS

Much work in statistics is inferential. Still, conventional statistical models must be reformulated to properly account for spatial correlation and spatial heterogeneity within georeferenced data. In emphasizing classical statistical assumptions about spatial data, five major problems emerge:

- Non-Gaussian distribution of variables: trend removal and Box-Cox power transformations are the corrective actions before OK, SK and UK can be applied [ESRI, 2001].
- Correlated random error with non-zero mean and non-Normal distribution⁹⁹.
- Non-homoskedasticity of error distribution: it is easier to handle nonstationary time series with changing means than nonstationary time series with changing variances [SAS, 2001].
- Error measurement only involves the dependent factors.
- Non-multicollinearity leads to a larger estimated variance of the regression coefficients (According to Anselin [1992], “the combination of a high R^2 with a very low t statistic is often a good indicator that something is wrong in terms of colinearity. Another indication is when estimates vary considerably by adding or dropping a single observation”).).

Yet, spatial autocorrelation is a physical reality and is also a necessity for interpolating nearby values. By ignoring this assumption, the right ingredients to compromise an inferential basis may result in a false sense of security. Thus, this chapter dissects these concepts from spatial autocorrelation to its implications towards spatial autoregressive models. Their spatial residuals impacts were not forgotten.

Classical statistics is commonly regarded as an aspatial diagnostic tool for variography and Kriging. According to the literature reviewed here, spatial autocorrelation, particularly the neighbor approach, can provide a valuable ESDA tool for the recognition of atypical locations, trends, transitional regions and spatial patterns. Then again, **W** definition is primarily a problem of ordinary observations. But even more important, if IDW is not used in conjunction with the binary approach, IDW is a bad option for constructing the weight matrix.

The spatial autocorrelation perspective of SAKWeb[®] is presented in Chapter 7.

3. Spatial Analysis: Interpolation

Chapter 3 is divided into seven major sections. Section 1 focuses on deterministic interpolators such as the Thiessen, trend surface, Fourier, IDW and B-Splines. The nature of the regionalized data justifies a special attention on the next section. Section 3.3 reports a detailed exploratory (a)spatial and spatial data analysis that geostatistics software should hold, including nearest neighborhood analysis. Its first subsection regards distribution standardization and subsection 3.3.2 concerns sampling considerations and major implications for Kriging. Section 3.4 clarifies variogram models and their features such as anisotropy, sill and behavior near the origin. Each subsection concerns itself with the formal variogram models. Subsection 3.4.2 highlights the automatic Cressie fit statistic. Section 3.5 discusses the theoretical structure of OK and the cross-validation procedure while the following one confronts several uncertainty techniques available at present. A new uncertainty index is presented and compared with the traditional OK variance. The last section reconsiders major issues presented in this chapter including a close linkage with the SAKWeb[®] sill rescaling and SAKWeb[®] nugget-effect approaches that are undertaken in Chapter 7.

3.1. SPATIAL DETERMINISTIC INTERPOLATORS

3.1.1. Drawing Boundaries and TIN

Boundary drawing uses external landscape features to delineate land units, assuming variations only at borders, whereas variation is null within the limits. Spatial autocorrelation is ignored between categories. The removal of systematic differences before continuous interpolation and the quick assessment with sparse data are, thus, two intrinsic features of drawing boundaries. With Voronoi or Thiessen vector polygons, the best information about an unvisited point is the

nearest data prediction point, a weighted linear combination approach where all weights are given to the closest sample. Thus, polygon output depends on the sample layout, a point-in-polygon computation problem resolved by the semi-line algorithm¹⁰⁰. There are never any null regions while its size varies inversely with the point density.

The ESRI Geostatistical Analyst[®] extension refers to this tool to analyze stationary and local outliers on the basis of the cluster assignment rule. After cells are placed into five classes, if the class interval is different from each of its neighbors then the cell will be colored gray to distinguish it from its neighbors. Griffith and Layne [1999] mention the use of Thiessen polygon surface to construct the geographical connectivity matrix. With their Canada boreal forest research, Nalder and Wein [1998] ranked Voronoi third best for temperature and fourth for precipitation out of seven methods¹⁰¹, lending support to the principle that complex methods are not inevitably more accurate.

According to Goovaerts [1999], TIN has been applied successfully to environmental modeling, categorical data¹⁰² and elevation-erosive¹⁰³ estimation although the assignment of quantities can be a major difficulty. Nested tessellation is another possibility. Tobler's pycnophylactic interpolation also addressed this problem by creating a continuous interpolator that removes the changes of abrupt boundaries on the basis of the mass-preserving reallocation approach from primary data. This view assures that the attribute volume within the spatial entity remains the same. In conjunction with ancillary data, both methods can be useful for areal interpolation.

3.1.2. Global Polynomial Trend

A polynomial trend surface¹⁰⁴ is used as a fitting regression procedure of a global surface for smoothing, filtering and data interpolation by separating the study variable into two components¹⁰⁵: large-scale variation and random error. Developed by Whitten in 1957, the unknown b_i coefficients of the coordinates'

polynomial regression are found by solving a set of simultaneous non-linear equations of cross products sums of the (x,y) coordinates and y_i values. If the n terms presented in **X** matrix correspond to the coordinates' data points and the n terms of vector **Y** equal the correspondent values then b_0 of vector **B** represents the constant term while the remain are the regression coefficients of the polynomial (see Figure 15)¹⁰⁶. Thus, the value of a particular (x_i, y_i) location equals $b_0 + b_1x_i + b_2y_i$, if a linear polynomial is considered, or $b_0 + b_1x_i + b_2y_i + b_3x_i^2 + b_4x_iy_i + b_5y_i^2$, if a quadratic one.

$$\begin{array}{c}
 \mathbf{Y} \\
 \begin{bmatrix} y_1 \\ \dots \\ y_n \end{bmatrix}
 \end{array}
 =
 \begin{array}{c}
 \mathbf{X} \\
 \begin{bmatrix} 1 & x_1 & y_1 \\ \dots & \dots & \dots \\ 1 & x_n & y_n \end{bmatrix}
 \end{array}
 *
 \begin{array}{c}
 \mathbf{B} \\
 \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix}
 \end{array}$$

$$\begin{array}{c}
 \mathbf{Y} \\
 \begin{bmatrix} y_1 \\ \dots \\ y_n \end{bmatrix}
 \end{array}
 =
 \begin{array}{c}
 \mathbf{X} \\
 \begin{bmatrix} 1 & x_1 & y_1 & x_1x_1 & x_1y_1 & y_1y_1 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & x_n & y_n & x_nx_n & x_ny_n & y_ny_n \end{bmatrix}
 \end{array}
 *
 \begin{array}{c}
 \mathbf{B} \\
 \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \end{bmatrix}
 \end{array}$$

Figure 15 – The linear¹⁰⁷ and quadratic¹⁰⁸ polynomial matrices.

It is this simplicity that makes this approach worth using instead of Kriging. “If with Kriging there is a price to pay with the estimation of variogram weights, the quadratic trend surface predictor uses up to six factors to describe the mean structure but it only needs one for the error structure” [Sarkozy, 1999]. “A high multicollinearity diagnosis is, however, a trend surface characteristic due to the strong functional relationship among the polynomial terms” [Anselin, 1992]. So, the indication of significant t-Student and R^2 indices should be suspicious. For instance, Anselin and O’Loughlin [1991] report the use of a quadratic trend surface regression to assess space importance as an explanatory variable for total conflict in Africa. In spite of the random white noise, a high multicollinearity index and a significant t test were found among the residuals.

In terms of software, EcoSse[®] develops this least square methodology quite well. As Appendix N shows, the organic matter in the soil of Nebraska, USA, does not

reveal any spatial autocorrelation among the data. The Moran I correlogram is zero for all distances, the isotropic variogram is a pure nugget-effect and its cross-validation has a negative slope. Based on Anova variance, a cubic trend surface was applied for detrend operation while the variogram, OK and cross-validation estimation were estimated regarding the residuals. This example illustrates the close link between trend surface analysis and Kriging: the detrend operation.

Like any OLS approach, this approach is highly susceptible to edge effects and rarely passes exactly through the original data points. Wang and Zhang [1999] report a tendency to accelerate without limit in areas like map edges where there is no point's control, an extrapolation issue. Another limitation regards uncertainty assessment. Finally, negative values can also be achieved depicting, thus, a serious distortion of reality.

3.1.3. Fourier

“The Fourier series is a mathematics-based technique for resolving any time domain function into a frequency spectrum much like a prism splitting light into a spectrum of colors” [Cross, 2000]. It is based on the discovery that it is possible to take any periodic time function and resolve it into an equivalent infinite summation of sine and cosine waves with frequencies that start at zero along multiples base frequencies. “In a clear way, the Fourier analysis transforms the data value using a series of sine's and cosines” [Isaaks and Srivastava, 1989]. “This global deterministic interpolator fits best with spatial datasets that exhibit marked periodicity such as ocean waves” [Walter, 2001]. Levine [1997] mentions this technique within the SASP[®] spatial pattern for Unix[®] platforms and exemplifies it with traffic flow shift over time for major city arterials. Cross [2000] reports this methodology for bird and insect type identification via seismographic information. Kratky [1981] describes its use in photogrammetry¹⁰⁹ while Billingsley [1983] uses it with image raster analysis to enhance both short and long range pattern components. Curiously, Yao [1999] describes a new automatic cross-

covariance matrix methodology between primary and secondary variables for CK estimation using FFT and the Bochner theorem.

3.1.4. B-Spline

“This local spatial inference model uses a radial basis polynomial to fit to the available data point with minimum curvature in a continuous surface” [Walter, 2001]. Often used for smoothing digitized lines like soil maps, B-Splines predictions can reach above the maximum or below the minimum available.

Burrough and McDonnell [1998] approve this method for arc or chain representation within a vector model for saving data storage space at processing time expense. ESRI [2001] recommends it for DEMs and labeling contour features while Miller [1997] reports its use with nitrogen concentration interpolation over space and time at Chesapeake Bay. “Yet, this adjustment may interpolate values whose nature has no relation with the original data but, rather, with the shape of how a polynomial adjusts to them” [Matos, 2001]. Concerning computational programming, B-Spline surfaces are defined in terms of blending functions that generate surface points as weighted averages of the point coordinates.

3.1.5. Inverse Distance Weighting

This exact deterministic method is a weighted average value at some known points. Like all distance weight averages, its common concern is the Krige regression effect¹¹⁰: under-estimation of high values and over-estimation of low values¹¹¹. The higher the function power, the more weight will be given to closer samples in conformity with Tobler’s Law where the best guess is the measured value at closest observations. Also, as more samples are included into the weighted linear combination, the resulting estimates become less variable, which leads to fewer extremes. Essentially, interpolation is smoothing.

Moving averages with a narrow window emphasize best spatial variation although, according to Goovaerts [1998], simulation maps reproduce best the spatial variability of the original data¹¹². For Zimmerman [1999], the recommended number of neighbors is around 6, while Declercq [1996] suggests the same for smooth surfaces but 20 for abruptly changing ones. Kravchenko and Bullock [1999] found no correlation between the number of neighbors to be considered and classical statistics parameters. “Usually, this is indicative of a stable situation, suitable for the application of weighted average estimation” [Clark and Harper, 2000]. Yet, with skew spatial datasets, the estimations were best accomplished with a power of 4 while low ones prefer a power of 1.

The layout map also depends on data clustering¹¹³ and outlier presence because this estimation is strongly influenced by the closest samples. Also, the reduction of data points should be mandatory in software for screened data. Anisotropy, an extra factor to consider, can be resolved with the quadrant and the octant search.

Regarding programming, the resulting map first sets a grid over the area while the estimated value at each node is calculated. However, in some of those grid nodes, there are already samples implying an infinity value since distance equals zero. Therefore, a special case for this calculation should be made¹¹⁴ as with the moon hole-effect: zero estimated value due to samples lacking within its neighborhood¹¹⁵. Categorically, both factors may contribute to a bizarre choropleth map.

Another dilemma is the uniform population distribution assumption within zones and how to choose the best weights for the weight function. A statistical search possibility for the optimal IDW is based on the cross-validation procedure. By using the original samples and their predictions, the best weight parameter is the one that holds the lowest root-mean-square prediction error. A more analytical approach is Kriging whose weights are optimized at each interpolation to produce a surface that satisfies minimum error variance. However, Dalthorp *et al.* [1999] present an estimation comparison study of the local mean for insect densities in

golf courses using Kriging against IDW. On the basis of cross-validation errors, Kriging had a substantially higher MSE¹¹⁶ effectiveness than IDW on a particular golf course because of low sampling density. “Being complex does not imply better results” [Goovaerts, 2002].

3.2. THE NATURE OF REGIONALIZED DATA: INTRODUCTION

“The existing knowledge of spatial behavior and the ability to estimate resources have been translated into three models” [Armstrong, 1998]: The genesis of phenomena is based on geological processes with extremely complicated parameters to represent them. An additional model was the creation of a polynomial deterministic term plus a random error component that is uncorrelated from one place to another (see equation 28¹¹⁷). Certainly, a persistent uncorrelated error between several distances, assuming error randomness, instead of translating this factor as a function that varies from place to place, creates a difficult problem to resolve. “Although major earth science processes do not behave as most physical and chemical laws, spatial data are not a pure random process, either” [GSLib, 2002].

$$\begin{aligned} \text{Pyritic Sulfur} = & 14.0548 - 5.97910 \times \text{AS} / 1.35753 \times \text{SU} + 1.34232 \times \text{SU} \times (\text{AS} - \text{SU}) - 0.419448 \times \text{AS} \times \text{SU} \times (\text{AS} - \text{SU}) \\ & + 4.95307 \times 10^{-3} [\text{AS} \times \text{SU} \times (\text{AS} - \text{SU})]^2 - 2.69728 \times 10^{-4} [\text{AS} \times \text{SU} \times (\text{AS} - \text{SU})]^3 + 5.88963 \times 10^{-11} \times e^{(\text{AS} \times \text{SU})} \cdot \\ & 1.38995 \times 10^{-6} [e^{[\text{AS} \times (\text{AS} - \text{SU})]}] + 0.0103637 \times [e^{[\text{SU} \times (\text{AS} - \text{SU})}]] - 16.70984 [\text{AS} \times \text{SU} / e^{(\text{AS} \times \text{SU})}] + 5.67080 \times 10^{-3} [\text{AS} \times (\text{AS} - \text{SU}) / e^{[\text{AS} \times (\text{AS} - \text{SU})]}] \\ & - 1.96079 \times 10^{-5} [\text{AS} \times (\text{AS} - \text{SU}) / e^{[\text{AS} \times (\text{AS} - \text{SU})]}] / 0.104688 \times \sin(\text{AS} \times \text{SU})^3 \times \cos(\text{AS} \times \text{SU})^2 - \\ & 9.53418 \times 10^{-3} \times \sin[\text{AS} \times (\text{AS} - \text{SU})]^3 \times \cos[\text{AS} \times (\text{AS} - \text{SU})]^2 - 0.0848224 \times \sin[\text{AS} \times (\text{AS} - \text{SU})]^3 \times \cos[\text{AS} \times (\text{AS} - \text{SU})]^2 - \\ & 2.08676 \times \sin(\text{AS}) - 2.19124 \times \sin(\text{SU}) + 1.49662 \times \sin(\text{AS} - \text{SU}) - 0.107983 \times e^{\text{AS}} - 0.0504733 \times e^{\text{SU}} + 0.312656^{(\text{AS} - \text{SU})} \cdot \\ & 0.0491919 \times \sin(\text{AS}^3) \times \cos(\text{AS}^2) + 0.0434771 \times \sin(\text{SU}^3) \times \cos(\text{SU}^2) - 1.81229 \times 10^{-4} \times \sin(\text{AS} - \text{SU})^3 \times \cos(\text{AS} - \text{SU})^2 - \\ & 2.01703 \times 10^{-4} \times e^{(\text{AS} + \text{SU})} \times e^{(\text{AS} - \text{SU})} \end{aligned}$$

[28]

where:

AS=Ash

SU=Sulfur

Developed by Matheron and D. Krige in 1962 as the regionalized variables theory, both investigators knew from experience that living beings and physical variables

are distributed neither uniformly nor at random. Somehow, ensemble observations are correlated to sampled ones on the basis of location distance. Thus, the foundation of this technique is the rate at which variance between samples changes over space in a physical continuity context. If conventional statistical methods assume that all data are independent and identically distributed, samples taken close to one another may be expected to have more similar values than samples taken farther apart. “This phenomenon is usually referred to as spatial dependence, spatial continuity or autocorrelation” [Skinner, 1999].

Matheron’s [1965] theory treats spatial properties as the sum of three components: the deterministic structural component, drift or trend¹¹⁸, the stochastic or random term with local variation and a Gaussian independent spatial residual¹¹⁹. Hence, data equals large-scale variation plus small-scale plus error. Data transformation and removal trend is the concern of the first factor, the variogram involves the second while error measurement is closely linked with the nugget-effect. This chapter addresses the second and the third aspects.

Under the regionalized variable assumption, local variation is not generally unstructured but it is spatially dependent at some scale because points within a given distance apart depend on one another in a statistical sense. The study of such correlation is called structural analysis or variogram modeling. The semivariogram, “usually referred to as variogram” [Nalder and Wein, 1998], is a tool that quantifies Tobler’s Law. Two critical returns may be achieved:

- 1) Geometric prediction quality can be quantifiable by Kriging variance;
- 2) It is possible to take into account clustering and possible anisotropies more easily than other inverse distance techniques.

Univariate statistics cannot handle these restrictions because two distributions may differ in the way they are correlated with each other. For instance, according to Table 9, the observation’s geographical arrangement, their nature and the degree of their spatial dependence estimators are key factors, thus creating a four dimensional index although both mean (13.8) and variance (48.6) are the same.

Location	1	2	3	4	5	6	7	8	9
%H ₂ O (Distribution A)	5	10	15	20	25	20	15	10	5
%H ₂ O (Distribution B)	10	25	15	10	20	5	15	5	20

Table 9– The two geographical arrangements for two datasets.

Regarding statistics¹²⁰, any variable is stationary when its distribution is invariant under translation. “The stationary region is the one where scientists want to make estimates” [Ella, 2000]. “It provides a preliminary model decision for uncertainty about the sample value at a location $z(x)$ ” [Journel, 1997]. For any space lag¹²¹ h increment, in a particular direction, the distribution of $z(x_1)$, $z(x_2)$... $z(x_n)$ observations is the same as that of $z(x_1+h)$, $z(x_2+h)$... $z(x_n+h)$ where $z(x)$ is regarded as the sum of three terms: the mean, drift or trend, a random variable plus error. “Here, random does not mean unpredictable¹²² but rather that it is governed by probability rules that include dependence on neighboring values called autocorrelation” [ESRI, 2001]. “Similarly, the estimation of weighted linear models and estimation error are random variables” [Isaaks, Srivastava, 1989]. Thus, the final surface is the sum of fixed (homogeneous spatial variation over larger areas) and random surfaces (local variation existence).

In its strictest sense, the stationary assumption (spatial homogeneous region) requires the first two moments¹²³ to be constant. The stationary assumption makes random functions homogeneous and self-repeating in space. The variance of the random spatial variable equals $Cov(0)=\sigma^2$ while the covariance between $z(x)$ and $z(x+h)$ is $\sigma_{x,x+h}=\sigma^2 f(h)=Cov(h)$. The stationary covariance has to exist and it only depends on the distance and direction between two samples and not on their locations, the second order stationary assumption (see equation 29 and 30¹²⁴). Oddly, Kriging is based on a random function concept¹²⁵ while unknown values are regarded as a set of spatial dependent variables.

$$\text{If } E(\mu(x))=m \text{ then } z(x)=m+e(x) \text{ and } E[z(x)-z(x+h)]=0 \quad [29]$$

$$E\{[z(x)-m][z(x+h)-m]\}=E[z(x) \times z(x+h)]-m^2=C(h) \quad [30]$$

where:

$z(x)$ is the value at x site;
 $z(x+h)$ is the value at x site plus lag distance h ;
 m represents the overall mean;
 $E()$ symbolizes the expected value.

It often happens that these assumptions are not satisfied due to the presence of a trend or variance change that cannot be assumed to be constant. This is why Matheron developed the intrinsic hypothesis, stationary assumption of the difference or variance of differences, by assuming a weakly stationary situation¹²⁶: the $z(x)-z(x+h)$ increments exist and they are location independent. Quoting Coelho and Pinheiro [1998], “ $z(x)$ variance can be written as a $z(x)-z(x+h)$ vector function”. Although the covariance does not exist, the variance of the difference between two locations is assumed to be present, it does not depend on the location and equals $2\gamma(h)$. This means that once the structural effect has been accounted for, the remaining variation is homogeneous within its variation so differences between sites are merely a displacement function between them. Formally, the intrinsic hypothesis may be stated with Equations 31 and 32.

$$\text{Var}[z(x+h)-z(x)] = E\{[z(x+h)-z(x)]^2\} = 2\gamma(h) \quad [31]$$

$$E[z(x)-z(x+h)] = 0 \quad [32]$$

The $\gamma(h)$ is called the dissimilarity average function and it is the key tool for the structural interpretation of phenomena as well as for estimation and uncertainty. Mathematically, the variogram was defined as:

$$\gamma(h) = 0.5 \times \text{Var}[z(x+h)-z(x)] \quad [33]$$

Since the mean of $z(x+h)-z(x)$ is zero for stationary and intrinsic variables then the variance is just the mean square difference. Consequently, the variogram becomes as follows:

$$\gamma(h) = 0.5 \times E\{[z(x+h)-z(x)]^2\} \quad [34]$$

The variogram indicates how different the values are as distance increases in a specific direction. “On the basis of the method-of-moments” [Cressie, 1993], the variogram estimator becomes:

$$\hat{\gamma}(\mathbf{h}) = \frac{1}{2|N(\mathbf{h})|} \sum_{i=1}^{N(\mathbf{h})} [z(x_i) - z(x_i + \mathbf{h})]^2 \quad [35]$$

where:

$N(\mathbf{h})$ equals the number of point pairs separated by lag $\mathbf{h}$;

$z(x_i)$ is the sample value at the i^{th} location;

$z(x+\mathbf{h})$ is the value at the i^{th} site plus lag $\mathbf{h}$.

Two major parameters describe this tool: the range or lag distance at which all successive values are independent of each other¹²⁷ and the sill, the variogram value corresponding to the range¹²⁸. Thus, the range is a length measure of the heterogeneity correlation. According to Sousa [1998], there is equivalence between the variogram $\gamma(\mathbf{h})$ and the covariance $C(\mathbf{h})$ but no covariance exists for a random function that is non-stationary. Supported by Spring [2000], “the covariogram and the correlogram, $\rho(\mathbf{h})=C(\mathbf{h})/C(0)$, may not exist with intrinsic stationary situations although it can account for possible changing means¹²⁹”. Yet, the variogram always exists and this makes it more generally useful¹³⁰:

- $C(0) = \sigma^2$.
- $C(\mathbf{h}) = C(-\mathbf{h})$, the covariance between two given samples does not depend on their positions.
- $|C(\mathbf{h})| \leq C(0)$, the module of the covariance is always below or equal to the variance.
- $C(\infty) = 0$.
- $\gamma(\mathbf{h}) = \gamma(-\mathbf{h})$, the variogram calculated in both directions is the same.
- $\gamma(\mathbf{h}) = C(0) - C(\mathbf{h})$, the corresponding variogram is obtained by turning down the covariance upside down (see Figure 16). Clearly, this relation is only possible with bounded variograms and only stationary regionalized variables hold this feature¹³¹. “Unbounded ones come from non-stationary situations with infinite dispersion” [Monteiro *et al.*, 2001]. In this case, the trend has to be defined as a vicinity position function in the field. Parabolic behavior is a

trend diagnostic. The proof of the relationship between the variogram and the covariance¹³² starts out from the variogram definition:

$$\begin{aligned}
 2\gamma(\mathbf{h}) &= E[\{z(\mathbf{x}+\mathbf{h}) - z(\mathbf{x})\}^2] = \\
 &E[\{z(\mathbf{x}+\mathbf{h}) - m\}^2 + (z(\mathbf{x}) - m)^2 - 2 \times (z(\mathbf{x}+\mathbf{h}) - m) \times (z(\mathbf{x}) - m)] = \\
 &E\{Z^2(\mathbf{x}+\mathbf{h})\} + E\{Z^2(\mathbf{x})\} - 2E\{z(\mathbf{x}+\mathbf{h}) \times z(\mathbf{x})\} = \\
 &\sigma^2(z(\mathbf{x}+\mathbf{h})) + \sigma^2(z(\mathbf{x})) - 2C(\mathbf{h}) = 2C(0) - 2C(\mathbf{h})
 \end{aligned}$$

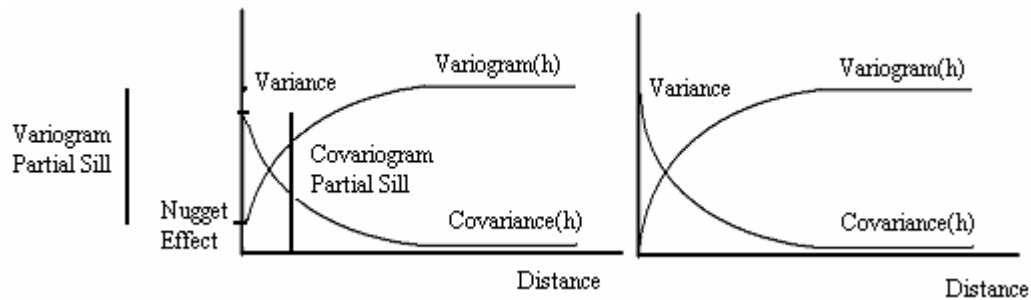


Figure 16 – The variogram and the covariogram function with and without the nugget-effect.

A gradual change in the $C(\mathbf{h})$ function is notorious near the origin indicating strong spatial continuity, while a sudden decrease in the variogram points to scale variability. If space is continuous then $E[z(\mathbf{x}+\mathbf{h})-z(\mathbf{x})]^2$ should be zero when $\mathbf{h}=0$. The only reason for a nugget-effect, $\gamma(\mathbf{h}) \xrightarrow{h \rightarrow 0} C_0$, greater than zero should be measurement error (see Section 7.2.4 for a further discussion). Yet, it is also admitted that the micro-scale variation leads to a highly irregular variogram pattern at a very short range. Because nothing can be said about variograms at lag distances smaller than $SSI = \min\{\|z(\mathbf{x}+\mathbf{h})-z(\mathbf{x})\|\}$ then it is not possible to tell whether micro-scale variation is continuous or not. “Therefore, the nugget-effect equals $C_{\text{Micro-scaleVariation}}$ plus $C_{\text{MeasurementError}}$ ” [Cressie, 1993]. This dramatic short scale variability has a serious impact on the accuracy of future estimates¹³³. At its extreme, this situation concerns the positive constant value of the variogram¹³⁴ which means uncorrelated locations for any distance $\mathbf{h}$.

“Besides the structure component and local variation, the residual random noise with zero mean, constant variance and zero covariance also exists” [Boykova, 1999, Spring, 2000]. In exact terms, this translates into Equation 36.

$$z(x) = \mu(x) + \varepsilon(x) + \xi \text{ or } z(x) = \mu(x) + \gamma_1(\mathbf{h}) + \dots + \gamma_n(\mathbf{h}) + \xi \quad [36]$$

where:

$\mu(x)$ is the deterministic function that describes the structural component of Z at location x ;

$\varepsilon(x)$ equals the stochastic term whose variance is the nugget-effect;

$\gamma_1(\mathbf{h}) + \dots + \gamma_n(\mathbf{h})$ represents different variogram structures;

ξ defines the spatially independent Gaussian noise term.

As expected, if a positive spatial autocorrelation is presented, similar values will be given on the variogram plot. Hence, the question is to decide whether a particular variable can be considered stationary or not. “No right or wrong judgment should be made, only appropriate or inappropriate” [Isaaks, Srivastava, 1989, Ella, 2000]. In practical situations, the variogram is only used up to a certain distance. Therefore, the problem is to decide which series of moving neighborhoods can be found with a constant expected value and enough data to give meaningful estimates. The quasi-stationary assumption is a homogeneity scale and sampling density compromise.

In Figure 17, there is a clearly increasing trend along the distance, a non-stationary situation. However, looking at both blow-ups of the central section, fluctuation appears to be stationary in the short run (constant mean with the local search neighborhood). The assumption of its being quasi-stationary does not apply to the entire dataset but only to the search neighborhood under which the estimation model is fitted. Thus, “these sub-areas meet the quasi-stationary assumption for the average and the variogram” [Soares, 2000]. One common way to test a stationary pattern, for instance, is the variogram cloud: selecting pairs of points with large variance at small distances reflects mean and variance discontinuity. In addition, the mean-variance computation in a moving window context can help to find spatial heterogeneity.

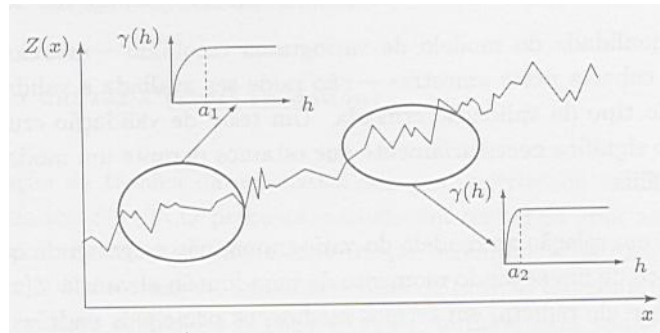


Figure 17 – The long trend versus short range stationary [Soares, 2000].

If there is a spatial trend then an OK variogram is no longer the appropriate tool to model patterns that result from the expected random fluctuation in a stationary stochastic local process. Some alternatives are possible: Universal Kriging (UK)¹³⁵, stratified regionalization, Simple Kriging (SK) with varying local mean (based on secondary information) and external drift (based on the residuals of the secondary linear variable function). Another strategy is to model the global trend component using a multi-linear coordinate regression to remove the drift component before OK is carried out on autocorrelated residuals (see Section 3.1.2 for a further discussion). However, “if data residuals are not normally distributed then interpolation will be biased” [Beckers, 2000]. Because this detrended process requires parameter estimation, to detrend with lower-order global polynomials is recommended. Yet, Goovaerts [1997] considers this approach unreliable because the estimated residuals strongly depend on the algorithm used to estimate the trend component. For this author, OK with a local search neighborhood achieves the best results. Theoretically, the main reason to detrend within Kriging is to satisfy stationary assumptions. Stationary assumption is a random statistical inference condition although, quite often, a stationary decision for the entire dataset is taken because the user cannot afford to split data into smaller subsets.

3.3. EXPLORATORY (A)SPATIAL DATA ANALYSIS

Anisotropies, outliers, discontinuities and variogram parameters (directions, lags distance, azimuth angle and bandwidth) can be described with spatial descriptors and spatial autocorrelation measures. To reduce input errors is also essential for

propagation errors. The second step with Kriging is assessment of univariate and bivariate descriptive tools and spatial description such as nearest neighborhood analysis. “Familiarity with the dataset is an asset in any statistical study” [Isaaks, Srivastava, 1989, Hock, Jensen, 1999]. The next three tables try to summarize some of the essential exploratory aspatial analyses. “Although these analyses are time-consuming, they are nevertheless vital” [ESRI, 2001].

Distribution graphs measures: General and normal QQ plot, PP plot, histogram and cumulative frequency table. For the number of classes, researchers suggest between 5 and 20. For Clark and Hosking [1986], the approximate number of classes equals $1+3.3 \times \log_{10}(\text{number\_of\_scores})$. Ferreira and Simões [1994] suggest $5 \times \log_{10}(\text{number\_of\_scores})$. Avoiding empty classes is equally important.
Location measures: Arithmetic, harmonic, weighted mean, geometric mean, median, mode, minimum, maximum, range, 1 st , 2 nd and 3 rd quartiles, deciles and percentiles.
Spread measures: Variance, standard deviation, inter-quartile range, quartile deviation, mean deviation from the mean and the median, standard mean error, standard deviation, variation and total variability.
Shape measures: Skewness coefficient or symmetry, coefficient of variation or tail length and kurtosis.

Table 10– Univariate statistics.

The scattergram shows anomalous data and relationships between variables. Cross-tabulation, biplots and contingencies tables can also be useful for comparing different spatial clusters.
The determination coefficient indicates the proportion of the total variation explained by the linear regression. Its range is [0...1]. The linear Pearson correlation varies between -1 and +1 ¹³⁶ . The nonparametric Spearman is only valid for rank data.
Linear regression: $y=b_1x+b_0$
$b_1 = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2}$
$b_0 = \bar{y} - b_1 \bar{x} \quad \text{where } \bar{y} \text{ is the dependent variable mean and}$
$\bar{x} \text{ is the independent variable mean.}$

Table 11– Bivariate statistics¹³⁷.

Spatial description maps:
1) Data posting is plotted with its corresponding data value.
2) Indicator maps reveal what data are above or below a certain threshold.
3) Flux movement maps and raster and vector contour maps are other possibilities.
4) Graduated point symbol and cartograms where the circle area is proportional to the grade.
5) H-scatterplot for a particular direction (similar distribution when plot pairs are close to the x=y line).
6) Cloropleth mapping. A valuable extension is the stepped statistical surface, a 3D column map whose height is proportional to the score.

7) Moving rectangular windows (Matos [2001] calls it the focal average) allow assessment of summary descriptive statistics and check on data heteroskedasticity (proportional effect stands for situations where the mean and the standard deviation vary in the same sense).

Nearest neighborhood analysis:

- 1) The arithmetic mean center (AMC) identifies the point with minimum squared distance deviation and is the most effective location for any company that provides a dispersed population service. If AMC is weighted by the magnitude of each variable, it represents the population centroid.
- 2) The observed mean distance of the nearest neighbor.
- 3) The expected mean for a random arrangement.
- 4) The standard error of the observed distances.
- 5) The $R_{nearest} = \text{observed/expected}$ index (a cluster zone is indicated by a 0 value while a dispersed one equals 2.149).

Table 12– Univariate spatial descriptors¹³⁸.

3.3.1. Distribution Standardization

Another concern regarding random spatial processes is to confirm Gaussian spatial distribution. Kriging as a predictor does not require that data follows a Normal distribution because Kriging is considered the best linear unbiased predictor. However, “if data pursues a Normal one then Kriging becomes the best unbiased predictor of all” [ESRI, 2001]. Moreover, normality is necessary to obtain probability maps for OK, SK and UK. “If sample values come from a lognormal distribution¹³⁹, the variance assessment from raw untransformed data is synonymous with outright stupidity” [Clark and Harper, 2000]. The lognormal distribution variance is a skewness measure, not variability, and if the goal is to measure changeability some transformation should be produced for a better-behaved distribution, particularly for a very small sample size.

“Data transformation often makes variance to be more constant and allows the correction lacking of homoskedasticity” [Goovaerts, 1998]. In addition, “the multiple linear regression model states clearly that the relationship between each independent and dependent variable should be linear and it can be achieved with data standardization” [Clark and Hosking, 1986]. Hence, the farther the random distribution is far from the Gaussian one, the worse the optimal linear prediction

performs. “The Gaussian model is unique in statistics for its extreme analytical simplicity and for being the distribution limit of many analytical theorems, globally known as central limit theorem” [Gslib, 2002].

The easiest way to check if samples respect the ‘bell’ curve is given by the cumulative probability plot or fractal diagram: the closer the data values are to the 45° line, the better. An additional method is to verify the overall similarity shape between the Gaussian distribution and the sample histogram. Also, a classical goodness-of-fit test is the χ^2 statistic based on the frequency distribution of (nc-3) classes¹⁴⁰. If the resulting statistic is less than $\chi^2[p=95\%,df=nc-3]$, for instance, then the rejection of the null hypothesis becomes true and the observed distribution does not differ from the Normal one:

$$\chi^2 = \sum_{j=1}^{nc} \frac{(O_j - E_j)^2}{E_j} \quad [37]$$

where:

O_j - Observed value of class j ;

E_j - Expected value of class j ;

nc – Number of classes.

“If the histogram is highly positively skewed then the trick is to persuade samples to act normally” [Murteira, 1993]. But positively skewed distributions do not imply a lognormal transformation. If data is only skewed, the probability logarithm transformation plot will be curved, kinked or just weird revealing the presence of two or more populations¹⁴¹. However, the second order stationary assumption forces the data values to be viewed as an outcome of a random function model that has the same mean and same covariance function between pair of values that are distance h away from each other. When more than one population appears to be present in the dataset, it is common practice to look for reasons that could have created the multiple populations. Experience in earth science indicates that mixtures of observations coming from different rock types or human activity normally result in multiple populations. “Some other approaches are required such as division of the datasets by soil domains” [Goovaerts, 1997].

The straight line of the cumulative probability plot on a logarithmic scale requires a logarithmic transformation adjusted by an additive constant, if necessary. “Yet, the issue is that users do not really care what the logarithm mean is. Users are only concerned about knowing what the average reef thickness is” [Clark and Harper, 2000]. However, “the lognormal distribution estimation is quite a complex procedure, based on Sichel tables” [Rossiter, 1999]. For instance, the Sichel optimal estimator for the arithmetic mean of the lognormal population is:

$$\text{mean\_back\_unbiased} = e^{\log(\text{mean})} \gamma_n(v) \quad [38]$$

where:

e represents the neperian constant;

$\gamma_n(v)$ is the bias correction factor based on natural logarithm variance, v , and n observations (Appendix D).

It is imperative to stress that the direct exponential back-transform is biased since the logarithm mean is the logarithm of the median and not the mean. “When anti-logged, large values will become exponentially larger while small values will become exponentially smaller and cannot go below zero” [Clark and Harper, 2000]. The over and under estimations will no longer cancel out when back-transformed. Therefore, it is not possible to write the regression relationship of Equation 39 and expect unbiased answers.

$$e^{(y)} = e^{(b_0 + b_1 x)} \quad [39]$$

As confirmed by both authors, the best back-transformed estimate would be:

$$e^{(y)} = e^{(b_0 + b_1 x + 0.5\sigma_y^2 + 0.5\sigma_x^2)} \quad [40]$$

As confirmed by ESRI, [2001], Geostatistical Analyst[®] back transformation may give biased results. For Cressie [1993] and Kravchenko and Bullock [1999], back transformation can be accomplished with the following formula, if LK is used:

$$z^*(x) = e^{y \times (x) + 1/2\sigma^2(x) - \Psi + 1/2\gamma(0)} \quad [41]$$

where:

$z^*(x)$ is the original back-transformed estimate at site x ;

$y^*(x)$ signifies the LK estimate;
 $\gamma(0)$ represents the nugget-effect;
 $\sigma^2(x)$ is LK variance;
 Ψ is the LaGrange multiplier.

Using GSLib[®] with P and K soil tests from 30 agriculture fields in Illinois, Indiana and Iowa, USA, Kravchenko and Bullock [1999] showed that for lognormal data, LK produces better results than OK, a conclusion based on the mean absolute error and on the G goodness of prediction index¹⁴² (the higher, the better):

$$G_I = \left\{ 1 - \frac{\sum_{i=1}^n (x_i^* - x_i)^2}{\sum_{i=1}^n (\bar{x} - x_i)^2} \right\} \quad [42]$$

where:

$\bar{x}$ is the sample mean;
 n is the number of observations;
 x_i^* represents the estimated i^{th} value;
 x_i equals the sample value.

“However, the log transformation advantage disappears with an increase in dataset size, particularly with 200 data points more” [Beckers, 2000]. In addition, “the LK plot suggests biased estimates for datasets with a low coefficient of variation” [Beckers, 2000]. For logarithmic data, Cressie [1993] supports the spatial median polish procedure for detrend and relative variograms while Olivier [1996] recommends Hermite polynomials for DK or the Normal Score Transformation (NST)¹⁴³. Controversially, citing Passarella *et al.* [1999], “we have accepted the bias introduced by the skewed data because log transformation before the final Kriging exponential estimate would result in another bias and we see no reason to prefer this latter bias”.

3.3.2. Sampling Considerations

Samples should be independent, representative of a very large population and truthful. Regarding probabilities properties with two independent events, $P(e_1 \cap e_2) = P(e_1) \times P(e_2)$ and $P(e_1 \cup e_2) = P(e_1) + P(e_2)$, geographical sampling is rarely concerned with events that are not mutually exclusive. If sampling with replacement constitutes a random sample, it is common practice not to evaluate the same element twice due to time and money reasons and to evaluate as many elements as possible. In addition, samples like people or animals are not stationary in space but, when surveyed, they are considered to be for a particular time.

It is possible to improve estimation by applying CK with previous temporal subsets in a co-regionalization framework. “The measurements collected at different time's t_i can be interpreted as a particular realization of temporally random functions k but spatially correlated” [Passarella *et al.*, 1999]. Hence, data are assumed to come from the same distribution with no significantly different statistical properties. When dealing with the nitrate concentration measures of May (47 wells), July (28 wells) and November (27 wells) for the Luca Plain, Italy, Passarella *et al.* [1999] use co-regionalization CK¹⁴⁴ and reach an improved spatial distribution estimation variance (16% on average) in relation to OK. Furthermore, Leenaers *et al.* [1997] use the cheap 463 elevation points of the Netherlands topographic survey for the prediction of zinc pollutant in the River Geul. The lognormal linear correlation coefficient between elevation and zinc was -0.48 :

$$\log_{10}Zn = 3.48 - 0.17 \times \text{Elevation} \quad [43]$$

With an exponential model for elevation ($C_0=9\%$) and zinc ($C_0=25\%$), the cross-variogram presented a lower absolute error with a 70% estimation variance improvement in the southern region due to data scarcity. However, if spatial variability and process scales do not match then it is complicated to combine various datasets for spatial analysis since the measurement of the sampling interval must capture the same distribution nature for the spatial phenomena. “In practical

terms, the variogram should be the same” [Soares, 2000]. According to Burrough and McDonnell [1998], the survey of fifty points taken in the Netherlands to measure soil surface temperature was carried out using two different supports: a Daedalus airborne multispectral scanner and Heimann ground radiometer. It also became clear that the two variogram structures show little correlation to each other: as expected, spatial variation varies inversely with support size.

Samples hold greater variances than blocks while their distribution is less symmetric, making the support effect another conditional bias for incorrect threshold values although as the data becomes more continuous, the support effect decreases. Regarding variogram regularization, Armstrong [1998] demonstrates that the variogram of a new regularized variable is the corresponding variogram of other available supports¹⁴⁵ plus the Journel and Huibregts (1978) table factor. “The affine correction or indirect lognormal corrections are other possibilities for an upscaling operation” [Isaaks, Srivastava, 1989].

Concerning sampling size, no rule could be provided. Only some principles can be given. “For demographic and urban studies, 3% is acceptable but no fixed minimum can be stated” [Clark, Hosking, 1986]. “A sample size of at least 30 is employed in most situations” [Griffith, 1996] although the central limit theorem guarantees normality for a large number of observations. Gaudard *et al.* [1999] argued that Bayesian methodology reduces the number of sample points required for the estimation of a reliable variogram but paid off by the greater computational complexity. Nevertheless, time and money dictate sample size (Appendix F). As a common sense rule, the greater the available samples the more reliable are the estimations although their location can be critical for the analysis that follows. Clustered samples will not improve reliability as dispersed ones do. In fact, OK variance (see Section 3.6 for a further review) takes this factor into consideration. Looking more closely at the OK variance equation, there is a system balance between getting the samples close to the estimated one and keeping them as far from one another as possible. While the second term calculates how closely the samples are related to one another, the last one assesses the relationship between

samples and estimations. Hence, sample density can turn out to be only a rough guide to quality interpolation.

Also, uniform or systematic sampling imposes a limit on the minimum distances among points, the distance interval in which capturing spatial dependency is most important. Hence, random location provides better estimates on the variogram for distances close to zero but yielding a less representative depiction of a surface¹⁴⁶. According to Griffith and Layne [1999], “if a choice is to be made between these two sampling types then the systematic one is to be preferred as a more accurate surface representation and reflected on the Kriging prediction errors”.

Isaaks and Srivastava [1989] demonstrated that additional sampling is most necessary in anomalous areas since it improves estimates where the proportional effect makes them least accurate. Even so, variograms with a proportional effect associated with clustered samples might not work well either, due to the apparent hole-effect, a dip variance at distances greater than the range¹⁴⁷. Certainly, a major variogram limitation is that only one single dominant correlation scale is assumed. In addition, heterogeneity on a larger scale is not apparent while heterogeneity on small scales is reflected in the nugget-effect. The variogram is a middle-aged man with difficulty seeing close objects and far away ones.

Groenigen, Siderius and Stein [1999] also report simulated annealing optimization sampling with Ping, Thailand river sand. By taking into account soft information and variogram shape, the maximum Kriging variance decreased from 146% to 69% with more observations along the boundaries and smaller spacing in the shortest-range direction.

3.4. THE VARIOGRAM

The variogram quantifies Tobler’s Law on all scales by summarizing the degree of similarity between data values for all possible data pairs as a distance function. Most variograms show a behavior like the one presented in Figure 18. The upper

limit is the sill which implies no spatial dependence between data points because all variances are invariant with the sample separation distance. “If the variogram has a sill then this behavior tends to support the assumption that samples come from a stationary distribution with a fixed mean and standard deviation” [Clark and Harper, 2000]. Sill fluctuations also correspond to periodic structures. In addition, “sill rescaling by 10 or 100 would not change prediction values” [Kriging Description, 2000]. Maps would still look the same although changes would be reflected in prediction variances: the higher the sill, the higher the prediction variances.

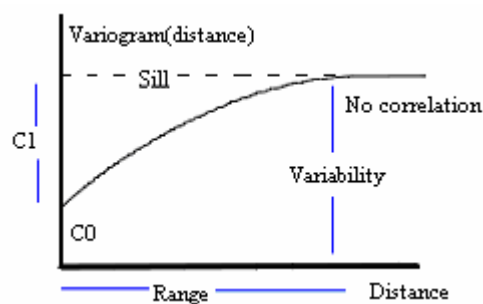


Figure 18 – The variogram¹⁴⁸.

The separation distance at which samples are spatially auto-correlated is the range that gives the answer to the short scale stationary question posed in Figure 17¹⁴⁹. “If the distance between the estimation and the sample is greater than the range then those samples make no useful contribution to final interpolation: it is too far” [GASA, 2001]. In theory, its weights are zero. Yet, quoting Stein and Zaadnoordijk [1999], “samples deletion beyond the range is inappropriate even when those values suffer from the screen-effect”¹⁵⁰. As expected, stronger spatial autocorrelation increases the range of the variogram and provides a fairly smooth map since Kriging weights become less powerful with respect to closer samples. On the contrary, as the range becomes shorter, more variation is expected. As Olivier [1996] argues, “bounded variograms imply transition features, patches or zones”¹⁵¹ whereas unbounded ones suggest continuous change over a region”.

“The nugget-effect represents the measurement error variance and the spatial variation at distances much shorter than sample spacing, which cannot be

resolved” [Spring, 2000, Liebhold, 1995]. Nugget-effect measures the unreliability of the sample value and the lack of a micro-scale model. GASA [2001] confirms this belief, stating that the nugget-effect is part of the variation that we cannot predict because it derives from spatial measurement error¹⁵² and purely random variation. Theoretically, the variogram is strictly zero for zero distance since there can be no variability between it and the sample point. Quoting Clark and Harper [2000], “if we do not accept zero at zero then what we are basically saying is that we do not believe the data values”. Controversially, once a uranium sample has been drawn, it is not there anymore for another replica. Also, distinguishing an outlier cluster and a non-stationary pocket or a sampling variation from a true trend can lead to a problematical situation. If UK handles true trend, BK applies to noisy circumstances because of lower variance values.

According to Figure 19, there are four types of behavior regarding C_0 :

- Quadratic (i): The regionalized variable is highly continuous and differentiable with the presence of a drift.
- Linear (ii): The regionalized variable is continuous but not differentiable and, thus, less regular.
- Discontinuous(iii): The variable is highly irregular at short distances¹⁵³.
- Flat (iv): Pure randomness or white noise. The regionalized variables $z(x)$ and $z(x+h)$ are uncorrelated for all values no matter how close they are. This is the extreme case of a total lack of structure where the best estimate for an unknown value would be the population average.

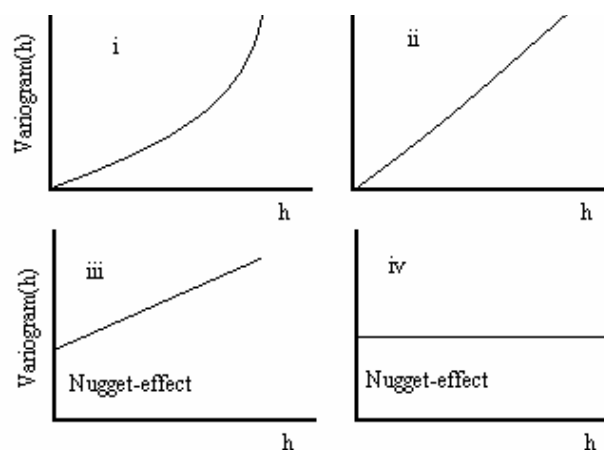


Figure 19 - The four type of behavior near the origin.

Although insufficient data may create a deadlock situation for the variogram estimation, Aunón and Hernandez [2000] confirm the creation of local variograms for specific areas according to their local mean. By using local models whenever possible, they demonstrate how stratification reduces interpolation errors based on prior information¹⁵⁴. But should anyone risk an inappropriate but well defined global variogram in preference to several poor local ones? Besides, the recognition of different populations might not also contribute to the study goal, either. For instance, if the fluid flow in a petroleum reservoir within several rock type environments is the main research aim then this last distinction will not contribute to the flow properties analysis. Goovaerts [1998] reports that experimental values can be combined into a single pooled stratum variogram. However, sudden changes at the strata boundaries can be found because available data are based on two different sets.

If the stationary assumption is at risk then the relative variograms for each separate population with a constant local mean can be found by Equation 44.

$$\gamma_{LR}(\mathbf{h}) = \frac{\sum_{i=1}^n N_i(\mathbf{h}) \frac{\gamma_i(\mathbf{h})}{m_i^2}}{\sum_{i=1}^n N_i(\mathbf{h})} \quad [44]$$

where:

$\gamma_1(\mathbf{h}), \dots, \gamma_n(\mathbf{h})$ are the local variograms from n regions;

$m_1, \dots, m_n$ are the local region mean;

$N_1(\mathbf{h}), \dots, N_n(\mathbf{h})$ are the number of sample pairs from each region.

Since local variograms are based on small populations, another alternative is to create a general relative variogram (see Equation 45), adjusted by the overall mean. “This solution is particularly useful for variograms that are built on different sample sets with dissimilar proportional effects” [Clark and Harper, 2000]. Quite often, Kriging stratification estimations and variances are less smooth than non-stratified ones.

$$\gamma_{GR}(\mathbf{h}) = \frac{\frac{1}{2|N(\mathbf{h})|} \sum_{i=1}^{N(\mathbf{h})} [z(x_i) - z(x_i + \mathbf{h})]^2}{\left[\frac{1}{2|N(\mathbf{h})|} \sum_{i=1}^{N(\mathbf{h})} [z(x_i) + z(x_i + \mathbf{h})] \right]^2} \quad [45]$$

Another option presented by Isaaks and Srivastava [1989] is the pairwise relative where the denominator serves to reduce the influence of very large differences (see Equation 46). Note that if $z(x_i)$ and $z(x_i + \mathbf{h})$ are zero then dropping that calculation for the next pair is the solution. Geostatisticians may also use the covariance function¹⁵⁵, the madogram and the correlogram model¹⁵⁶.

$$\gamma_{PR}(\mathbf{h}) = \frac{1}{2|N(\mathbf{h})|} \sum_{i=1}^{N(\mathbf{h})} \frac{[z(x_i) - z(x_i + \mathbf{h})]^2}{\left[\frac{z(x_i) + z(x_i + \mathbf{h})}{2} \right]^2} \quad [46]$$

Disregarding the variogram choice, the removal of the most erratic values can be a plus even if this might be a questionable solution. “In mining excavation, an unusually large value may be a cause for celebration” [Wang, Zhang, 1999]. If outlier removal is crucial in generating negative spatial autocorrelation¹⁵⁷ then small high-grade deposits can equally make the difference between opening the mine and keeping it closed. Moreover, with high-skew distributions such as with gold grades, it is not possible to cut them in an arbitrary way. Cressie [1993] proposes an OK resistant algorithm by using neighboring values to determine how much of the outlier should be down weighted (see subsection 2.4.1 for further discussion). Data standardization is an option although this may create data artifacts. For instance, if a variable holds a zero value and a logarithm is applied then a problem may emerge. A plausible solution may be to set 0 to 0.00000001, with log -8, or 0.01, with log -2. Even so, either decision may or may not change the variogram shape, depending on the variable range context.

If high-grade samples are close together in a specific sub-area, eliminating those samples may not be a good idea either. “The model variogram is the spatial averaging result of obscuring regional variations and differences caused by local

anomalies” [Isaaks, Srivastava, 1989]. Therefore, it is essential to examine spatial contiguity and short-range differences before modeling variograms with first-class spatial exploratory software including spatial autocorrelation and association measures.

Another critical issue is anisotropy, different spatial autocorrelation behavior of the variogram in several directions, versus trend influences due to a physical process, e.g., wind, runoff or a geological formation¹⁵⁸. If the global trend affects all measurements due to a known physical process in a deterministic global polynomial framework, anisotropy differs from drifts since the directional influence process is not usually known. Thus it is quantified as a random error process with directional autocorrelation variograms. “Anisotropy is a random process feature that shows higher autocorrelation in one direction than in another” [ESRI, 2001]. In addition, this factor affects circle shape regarding the local search neighborhood window. Major literature recommends computation in four directions to divide the omni-directional search into four subsets, though this causes an increase in noise by a factor of four: 0° (North-South), 45° (NorthEast-SouthWest), 90° (East-West) and 135° (NorthWest-SouthEast). Hence, defining the type, sill, nugget-effect and range with directional variograms can create greater obscurity. Nevertheless, the ranges of each of them can confirm or refute the assumption of the anisotropy. When plotting a rose diagram, for instance, circular contour lines are indicative of an isotropic surface whereas an elliptic shape points to an anisotropy occurrence. Another option is to draw the contour map of the variogram surface. As expected, more weight will be given to samples that lie in the direction of maximum continuity and considerably less where minimum continuity is involved.

Concerning the elliptic or geometric anisotropy presented in Figure 20 (left), the variogram holds the same sill for major directions although their ranges are different. Then the overall variogram becomes:

$$\gamma(\mathbf{h})=\gamma(\mathbf{h}_x,\mathbf{h}_y)=\gamma_1(\sqrt{h_x^2+r_{ani}^2h_y^2}) \quad [47]$$

where:

$\mathbf{h}_x$ and $\mathbf{h}_y$ are the cosine and sine of direction $\mathbf{h}$;

a_1 and a_2 equal the length of the major and minor axes of the resulting ellipse;

$r_{ani} = \frac{a_2}{a_1}$ represents the amount of survey space that has to be stretched in direction $\mathbf{h}$ so the

variogram in direction $\mathbf{h}_2$ is the same as that in direction $\mathbf{h}_1$.

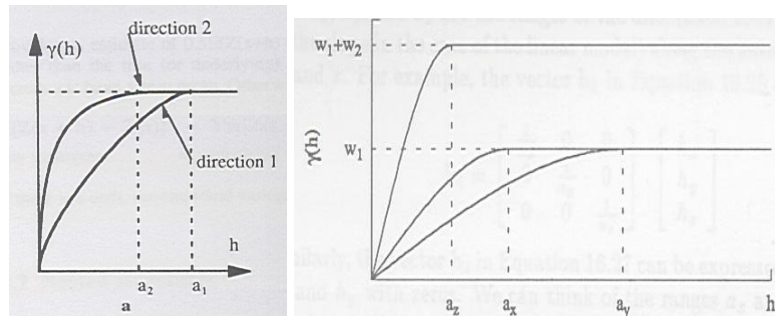


Figure 20 – Geometric (left) and geometric with zonal (right) anisotropy.

“Sometimes, the variogram exhibits different ranges and sills, thus creating an overall variogram division into two linear components¹⁵⁹” [Isaaks and Srivastava, 1989]:

$$\gamma(\mathbf{h}) = w_1\gamma_1(\mathbf{h}_1) + w_2\gamma_1(\mathbf{h}_2) \quad [48]$$

where:

$\gamma_1(\mathbf{h}_1) = \gamma_1\left(\sqrt{\left(\frac{\mathbf{h}_x}{a_x}\right)^2 + \left(\frac{\mathbf{h}_y}{a_y}\right)^2}\right)$ must be an isotropic model whose sill equals w_1 and a range of 1.

$\gamma_1(\mathbf{h}_2) = \gamma_1\left(\frac{\mathbf{h}_z}{a_z}\right)$ is the equivalent model whose sill equals w_2 and only exists in the z direction.

In the case of the proportional effect of Figure 21 (note that the proportion of averages, m_p and m_s , and sills equals three to one), the overall model can be found dividing each local variogram by the local mean square average. Quite often, this effect is a result of sampling clusters leading to bias statistical indices such as the average or the histogram shape.

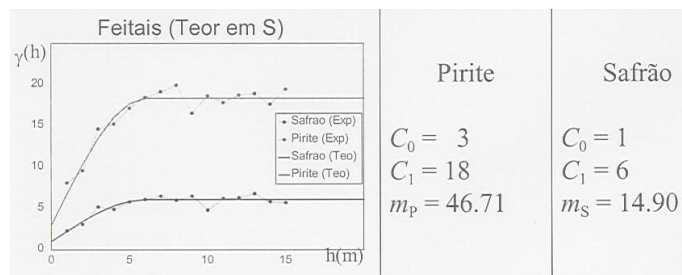


Figure 21 – An example of the proportional effect [Sousa, 1998].

“Theoretically, this last anisotropy should not occur if the dataset is the same for three main reasons” [Clark and Harper, 2001]:

- The influence of the range in one of the directions has not been fully exploited because the distance axis of the variogram graph is not long enough.
- Skew data is present and the variogram is interpreting the proportional effect due to different sample sets, according to different directions, and not the spatial pattern.
- Violation of the stationary assumption like the change of the population characteristics in a particular direction.

3.4.1. SAKWeb[®] Isotropic Variogram Models

A mathematical model has to be fitted to a quantitative regionalized variation description so that autocorrelation can be characterized. The ultimate purpose is to obtain the covariance values to setup the **A** and **C** matrices within equation $\mathbf{A} \times \mathbf{B} = \mathbf{C}$ (see Section 3.5 for a further discussion). Therefore, the model variogram equation is critical to the interpolation quality. “It is imperative that this model fits experimental values well and avoids negative variances, the conditional negative definite property” [Armstrong, 1998]. The next equations present the three admissible transitive variogram models considered for SAKWeb[®]. Other research papers mention other models but are not reviewed here, e.g., the cardinal sine, wave-hole (the only variogram that reflects negative spatial autocorrelation), intrinsic power function, nugget-effect, prismato-magnetic, Cauchy, Bessel, cubic and prismato-gravimetric model. Non-transition variogram models used to model

intrinsic random functions are not enclosed either. The variograms covered by SAKWeb[®] are:

Spherical: $\gamma(h) = (C_0 + C_1) \left(\frac{3|h|}{2a} - \frac{1}{2} \left(\frac{|h|^3}{a^3} \right) \right), 0 \leq |h| \leq a$ [49]

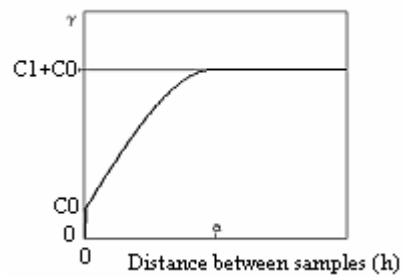


Figure 22 – The spherical model¹⁶⁰.

Exponential: $\gamma(h) = (C_0 + C_1)(1 - e^{-(h/a)})$, $|h| > 0$ ($a' = 3a \rightarrow \gamma(a') = 0.95(C_0 + C_1)$) [50]

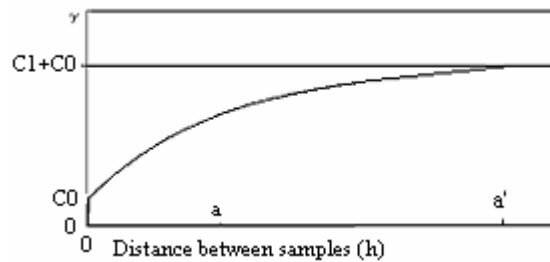


Figure 23 – The exponential model¹⁶¹.

Gaussian: $\gamma(h) = (C_0 + C_1) \left(1 - e^{-\left(\frac{|h|}{a}\right)^2} \right)$, $|h| > 0$ ($a' = \sqrt{3} a \rightarrow \gamma(a') = 0.95(C_0 + C_1)$) [51]

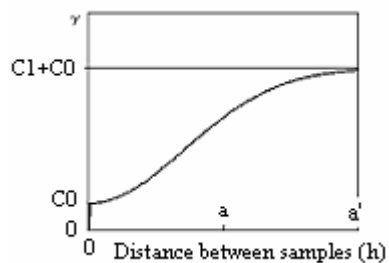


Figure 24 – The gaussian model¹⁶².

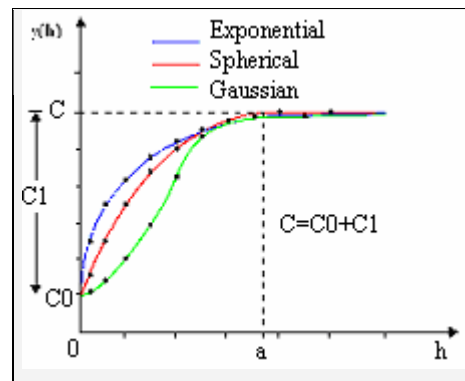


Table 13 – A global behavior comparison of previous variogram models [Geostatistics FAQ, 1997].

3.4.2. Computational Variogram Issues

Any experimental variogram must be aware of observations that layout in an irregular network. This is the reason for the associated tolerance angle. If the two major directions are perpendicular then the angular tolerance should be less than 45° to avoid points overlapping. It is also crucial that samples with mistaken measurements should not be substituted by zero because it would distort the true variability.

Webster and Olivier [1992] suggest between 150 and 200 observations to produce a reliable variogram while Englund and Sparks [1991] and Geostatistics FAQ [1997] present some other practical tips to improve a variogram:

- 30 to 50 pairs at each lag distance¹⁶³;
- Half of the maximum lag distance;
- Similar lag space should be applied for different anisotropy axes;
- Data transformation for skewed distributions;
- An omni-directional variogram (angular tolerance of 90° degrees on either side) before reducing into directional ones. In fact, an omni-directional variogram contains more sample pairs than any directional one and is, therefore, more likely to show a clearly interpretable long-range structure;

- Try several different lag intervals for plotting the experimental variograms to obtain the maximum detail at small distances without being misled by structural artifacts due to the particular interval used.

For Isaaks and Srivastava [1989], the search radius for irregular grid samples should be slightly larger than the average spacing between samples although additional sampling will always improve the estimation regardless of the distance from the point being estimated¹⁶⁴:

$$\text{search\_radius} = \sqrt{\frac{\text{region\_area}}{n}} \quad [52]$$

According to Griffith and Layne [1999], the tolerance angle should be 23 degrees (see Figure 25). Moreover, large lag tolerance leads to nugget-effect overestimation and masked short-range spatial autocorrelation. On the other hand, small lag tolerances may generate many empty bins.

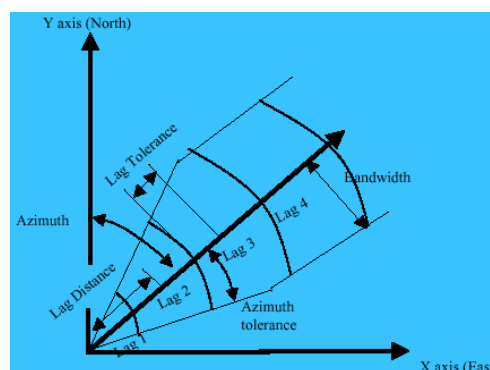


Figure 25 – The setup factors for the variogram construction with regard to the sector binning method.

If samples are located within a regular grid, the grid spacing is the lag itself. Otherwise, according to Englund and Sparks [1991], “the maximum pair should be divided by two and then subdivided by ten equal distance classes”. “The lag tolerance should be half of the lag spacing” [Geostatistics FAQ, 1997]. For Armstrong [1998], “this tolerance should not be too large, in order to avoid the anisotropy blurring and not too small, to avoid a diminutive number of pairs”.

“Modeling anisotropies is done by trial and error using a graphic terminal” [Soares, 2000]. For Armstrong [1998], it is not a good idea to apply least squares or other automatic methods to fit variogram models. First, the model must be a positive definite function to guarantee having one and only one stable solution and the Kriging matrix has an inverse. Polynomials obtained by least squares regression will rarely satisfy these conditions. Second, least squares assume that sample points are independent observations, which is clearly not true. Third, critical variogram behavior close to the origin is not known and least squares do not take this into account. Outlier adjustment can also be complicated. In addition, it does not take into account additional information that may justify the choice of one model instead of another. SAKWeb[®] pursues this philosophy but presents one variogram fitness measure based on the squared difference in the theoretical model and the experimental variogram lags.

For automatic fitting, Olivier [1997] supports WLS using the Ross Maximum Likelihood Program with weights that are directly proportional to the number of paired comparisons in the estimates. Goovaerts [2000] uses the minimally weighted sum of the squares of the differences between the experimental variogram and the theoretical one. Clark and Harper [2000] suggest Cressie’s [1993] modification of the automatically fitting modeling by replacing the scaling division of the total number of pairs¹⁶⁵ (see equation 53). Basically, this formula quantifies the squared difference between the theoretical and real variograms but scaled by the model itself: the lower this index becomes, the better the model turns out to be. In addition, each point is weighted according to the number of pairs available: the more samples available, the smaller the error that is tolerated.

$$\text{Automatic\_fitness\_index} = \frac{1}{\sum |N(\mathbf{h})|} \sum N(\mathbf{h}) \left[\frac{\gamma^*(\mathbf{h}) - \gamma(\mathbf{h})}{\gamma(\mathbf{h})} \right]^2 \quad [53]$$

where:

$N(\mathbf{h})$ equals the number of point pairs separated by lag $\mathbf{h}$;

$\gamma^*(\mathbf{h})$ represents the experimental variogram value for lag $\mathbf{h}$;

$\gamma(\mathbf{h})$ is the theoretical variogram value for lag $\mathbf{h}$.

3.5. ORDINARY KRIGING

The difference between IDW and Kriging derives from the mathematical process that minimizes the minimum squared error variance¹⁶⁶ of the estimation. The equation system can be obtained from the following two algebraic expressions [Soares, 2000]:

$$\sum_j w_j \gamma(x_i, x_j) - \Psi = \gamma(x_i, x_0), i=1, \dots, n \quad [54]$$

$$\sum_i w_i = 1 \quad [55]$$

where:

w_i is the weights to be assigned to the i^{th} observation;

Ψ represents the LaGrange multiplier or slack value required for forcing total weights to 1;

$\gamma(x_i, x_j)$ is the variogram value between sample i and j ;

$\gamma(x_i, x_0)$ equals the variogram value between sample i and the estimation x_0 .

In terms of covariance, the Kriging system becomes¹⁶⁷:

$$\sum_j w_j C(x_i, x_j) + \Psi = C(x_i, x_0), i=1, \dots, n \quad [56]$$

$$\sum_i w_i = 1 \quad [57]$$

where:

$C(x_i, x_j)$ is the covariogram value between sample i and j ;

$C(x_i, x_0)$ equals the covariogram value between sample i and the estimation x_0 .

Like Thiessen and B-Splines (see Section 3.1 for a further review), Kriging is an exact interpolator in the sense that the sample and estimation are equal. This happens because the variogram value between the sample k and the estimation x_0 , $\gamma(x_k, x_0)$, equals zero since the estimated site is the sample itself. Therefore, its Kriging variance equals zero. By resolving the Kriging system (see Figure 26)¹⁶⁸, the weight w_k is one while all others weights and the LaGrangean dummy parameter are zero.

$$A = \begin{bmatrix} \text{Cov}(x_1, y_1) & \dots & \text{Cov}(x_1, y_n) & 1 \\ \vdots & \ddots & \vdots & \vdots \\ \text{Cov}(x_n, y_1) & \dots & \text{Cov}(x_n, y_n) & 1 \\ 1 & \dots & 1 & 0 \end{bmatrix} \quad B = \begin{bmatrix} \text{Cov}(x_1, x_0) \\ \vdots \\ \text{Cov}(x_n, x_0) \\ 1 \end{bmatrix} \quad W = \begin{bmatrix} w_1 \\ \vdots \\ w_n \\ \psi \end{bmatrix}$$

Figure 26 – The Kriging system¹⁶⁹.

Once again, the goal is to find the weighted average estimator that provides an unbiased estimator with the smallest estimation error variance using the LaGrange multiplier and the constraint of the weights sum to one. Assuming an unknown constant mean and intrinsic variables¹⁷⁰, the general arrangement to setup the Matheron equations are as follows:

$$\begin{aligned} w_1\gamma(x_1, x_1) + \dots + w_n\gamma(x_1, x_n) - \psi &= \gamma(x_1, x_0) \\ w_1\gamma(x_2, x_1) + \dots + w_n\gamma(x_2, x_n) - \psi &= \gamma(x_2, x_0) \\ &\dots \\ w_1\gamma(x_n, x_1) + \dots + w_n\gamma(x_n, x_n) - \psi &= \gamma(x_n, x_0) \\ w_1 + \dots + w_n &= 1 \end{aligned} \quad [58]$$

In an elegant matrix layout, each interpolated value is calculated as the sum of weighted known points whose weights are calculated from the (n+1) simultaneous linear equations set: $A \times W = B$ or $W = A^{-1} \times B$. The statistical distance¹⁷¹ between sample points and distances from each sample to the grid point are used to compute the model variance reproduced on matrices **A** (between samples) and **B** (between each sample and the estimated location). While A^{-1} underlies the declustering factor, **B** represents the structural distance between the estimation and all samples. In addition, the product of A^{-1} by **B** adjusts the raw inverse statistical distance weights in matrix **B** to account for possible redundancies between samples. As expected, if no spatial autocorrelation is found among the available samples then the Kriging estimator equals the sample average.

With zero nugget-effect, the diagonal of matrix **A** equals zero while Kriging weights decrease as the dataset location gets farther from the estimated one. Hence, major software computation uses the covariogram model to setup matrices **A** and **C** in order to avoid the diagonal zeros of **A**. After weights are calculated, the estimation of the regular grid surface can be achieved with a great deal of calculation. In fact, “the computation intensity to solve OK weights is proportional

to the cube of the number of observations although a reduction can be achieved by applying the sequential Kriging approach” [Yeh, Guzmán, 1999].

3.5.1. The Cross-Validation Procedure

Before the final surface is produced, consistency and the absence of biased estimations can be tested by temporarily dropping any sample from the dataset while its value is re-estimated using the remaining samples. Repeating this cross-validation procedure for all observations, it is possible to obtain the experimental mean error¹⁷², the root mean square prediction error¹⁷³ and the average Kriging standard error¹⁷⁴. If the average standard error is close to the root-mean-square prediction error then the user is assessing the prediction variability¹⁷⁵ correctly. However, “if the average standard error is greater or less than the root mean square prediction error then the user is overestimating or underestimating estimation variability” [ESRI, 2001].

“It is hard to judge the influence of variogram parameters on cross-validation errors” [Soares, 2000]. In fact, Isaaks and Srivastava [1989] proved that if a relative nugget-effect of 4% lead to a lower cross-validation residuals for the dataset of 470 samples, the original model performed better than the one whose nugget-effect was automatically adjusted on the basis of cross-validate residuals. Also, if the original dataset is spatially clustered then the cross validation pursues the same trend leading to a situation where residuals become more representative of certain regions. Hence, a good statistical model does not signify that a good spatial genuine model is good.

The QQ plot comparison layout between the true and the estimation with the linear correlation index or the residual map in order to check location bias are the common graphic alternatives. The SAKWeb[®] approach is to outline a validation mapping with an extra dataset test based on three nugget-effect approaches plus the conventional cross-validation.

3.6. ASSESSING UNCERTAINTY

Although spatial data lives with uncertainty, science needs safe foundations. Uncertainty is a dimensionless parameter for which high values are bad and lower ones are optimal. Thus, uncertainty must be space geometry dependent because areas away from sample locations hold higher uncertainty. It must also take into account the variability of sample values. Furthermore, since different interpolation procedures may give dissimilar results and ground truth can never be known, it may be useful to know what the predicted chance of exceeding a given upper limit is, so decisions about expensive cleanup operations can be well founded, for instance.

With agricultural applications, administrators might be interested to know how much of the whole population would give a higher return than the value of a certain crop while, within environmental issues, supervisors might be looking at toxicity levels. The same question arises when the fire department calculates tree density in tropical forests, when the fishery service computes water salinity and the density of shellfish or when engineers evaluate slope stability conditions to decide on the best route for a new road. Irrespective of the circumstances, the question is to determine how much of the population is likely to lie above or below a cutoff value.

Two issues emerge from this perspective. First, the choice of the probability threshold can be subjective. A given contamination level may be unacceptable for residential areas but tolerable for industrial yards. Second, the estimation error may be ignored. “A contaminated location can be declared safe on the basis of an estimate of pollutant concentration which is incorrect but slightly less than the regulatory threshold” [Goovaerts, 1998]. If the true and the estimated values belong to quadrant II on an Estimated versus True grade plot, a good opportunity to invest can be missed¹⁷⁶. If it falls within quadrant IV, expensive consequences can be expected¹⁷⁷.

Uncertainty about cell size and error has a close relationship with cleanup costs too. Small blocks are desirable because cleanup can cost less money than in larger ones. Nevertheless, if errors are too large then there is a realistic possibility that blocks below threshold level are treated and blocks that are really above that level are missed. According to Isaaks and Srivastava [1989] with their Walker Lake study, 7% of the area for a threshold value of a thousand was considered while only 2% would have been suitable if the support had increased to $20 \times 20 \text{m}^2$.

How can we quantify these uncertainties? The misclassification risk associated with a particular physical cutoff definitely increases at threshold location boundaries¹⁷⁸. “If the goal of a manager’s decision is to minimize unnecessary cleansing and ill health costs (in conjunction with a pre-setup deterministic cost) then it is possible to layout the total spatial health and remediation costs based on the resulting expected false negative error and false positive error models” [Goovaerts, 1997].

With regard to economic land evaluation, linear programming including sensibility analysis is a possibility. Burrough [1991] presents a gradual deterministic response of PH crop illustrating soil acidity impact on crop growth:

- No crops, if $\text{PH} \geq 7$;
- Normal growth, if $\text{PH} \leq 5$;
- $(7 - \text{PH})/2$ of crops, if $5 < \text{PH} < 7$.

Seven years latter, Burrough and McDonnell presented the Kenya annual soil erosion simulation using the Universal Soil Loss Equation for rainfall-rainoff impact, yielding an annual soil loss rate of $9 \pm 6 \text{cm}$ in 40 years:

$$R_e(297 \pm 72) \times S_e(0.1 \pm 0.05) \times S_l(2.13 \pm 0.045) \times S(1.169 \pm 0.122) \times C(0.63 \pm 0.15) \times P(0.5 \pm 0.1) \quad [59]$$

where:

- R_e is rainfall erosion;
- S_e is soil erodibility;
- S_l is slope length;
- S is slope;
- C is cultivation;

P is practice.

By assuming an unknown constant mean¹⁷⁹, non-linear PK¹⁸⁰ and IK¹⁸¹ present an alternative uncertainty approach whose final estimates show the probability mapping for exceeding a given cutoff¹⁸². Because of the smoothing effect of Kriging (lower estimate variance when compared with the real sample variance), the local distribution of Kriging estimates is conditionally biased leading to a false and biased probability distribution, especially when the cutoff value becomes very high or very low. Juang and Lee [2000] point out that PK accuracy is much higher than that of IK in their probability estimation of heavy-metal concentration in Taiwan. According to both authors, “it yields more space variability and it behaves better with a screen-effect situation by reducing the risk of getting inconsistent probabilities”. In addition, with both approaches, the local uncertainty cumulative distribution is taken into consideration which avoids the strong assumption about the spatial distribution.

Geostatistical simulation tries to reflect the mean, the histogram, the covariance structure and the spatial data variance characteristics of the original dataset while, at the same time, it makes the simulation value close to the real one. “By generating multiple and unique interpretations that respect the spatial dataset, simulation generates multiple configurations of possible realities, a realism issue¹⁸³, based on the search window Gaussian distribution but not on the optimum estimation” [Wang, Zhang, 1999]. For Soares [2000], “the greatest potential of geostatistical simulation lies in the production of uncertainty estimations for a given cutoff value and, therefore, the assessment of impact costs”.

On the basis of the assumption of null dependency among points, classical statistics presents another option:

- 1) Estimation of the best unbiased estimator for the true population standard deviation.

$$\sigma^* = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} \quad [60]$$

where:

x_i is the i^{th} sample value;

$\bar{x}$ equals the sample mean;

n is the sample number.

- 2) Computation of the best standard deviation confidence interval based on the true unbiased population assumption:

$$\sqrt{\frac{(n-1)\sigma^{*2}}{\chi^2_{\text{lower}}}} < \sigma < \sqrt{\frac{(n-1)\sigma^{*2}}{\chi^2_{\text{upper}}}} \quad [61]$$

where:

σ^2 is the estimated sample variance;

χ^2_{lower} equals the lowest χ^2 for a certain confidence level (df=n-1);

χ^2_{upper} equals the upper χ^2 for a certain confidence level (df=n-1).

- 3) Assessment of the mean confidence interval using the t-Student distribution.

$$\bar{x} + t_{\text{df,Lower}} \frac{\sigma^*}{\sqrt{n}} < \mu < \bar{x} + t_{\text{df,Upper}} \frac{\sigma^*}{\sqrt{n}} \quad [62]$$

where:

$t_{\text{df,Lower}}$ equals the lowest t value for a particular confidence level (df=n-1);

$t_{\text{df,Upper}}$ equals the upper t value for a particular confidence level (df=n-1);

σ^* is standard deviation dataset.

- 4) Estimation of the probability that is likely to lie above or below a threshold value (based on the Normal distribution).

$$x = \frac{\text{threshold} - \bar{x}}{\sigma^*} \quad [63]$$

where:

$\bar{x}$ is the estimated sample mean;

Finally, the uncertainty layout of the conventional Kriging software is closely related to OK variance¹⁸⁴:

$$\sigma_e^2 = C_{00} + \sum_{i=1}^n \sum_{j=1}^n w_i w_j C_{ij} - 2 \sum_{i=1}^n w_i C_{i0} = -Y_{00} - \sum_{i=1}^n \sum_{j=1}^n w_i w_j Y_{ij} + 2 \sum_{i=1}^n w_i Y_{i0} \quad [64]$$

where:

C_{00} is the variance of the estimated point value¹⁸⁵;

C_{ij} is the covariance between the i^{th} and j^{th} sample¹⁸⁶;

w_i and w_j are the OK weights;

C_{i0} represents the covariance between the i^{th} sample and the unknown value being estimated¹⁸⁷.

According to Soares [2000], “if the sum of the weights is one¹⁸⁸ and the average estimation error equals zero then the Kriging error variance becomes” (no measurement error is included):

$$\sigma_{OK}^2 = \sum_{i=1}^n w_i \gamma(x_i, x_0) + \Psi \quad [65]$$

where:

Ψ equals the LaGrange multiplier of the OK system;

w_i is the OK weights;

$\gamma(x_i, x_0)$ is the variance between the i^{th} and the estimated point.

Hence, if errors respect the ‘bell’ curve then real values will fall within the $\text{Kriging\_predictor} \pm 2\sigma_{OK}^2$ interval for a 95% confidence level¹⁸⁹. However, uncertainty is not included with variogram estimation and prediction variance is underestimated. But even more critically, OK variance is not sensitive to local error for two major reasons: 1) it is based on the same global variogram and 2) distances among locations are the only relevant factor. OK variance is mainly a geometry-dependent measure heading the assumption that an OK true error map is a better substitute. OK variance is too much of a spatial operation.

3.6.1. The Classical Local Spatial Variance

To overcome this latter uncertainty issue and to avoid the heavy computational issue of stochastic simulation, the classical local spatial variance could be defined as:

$$\sum_{i=1}^4 \frac{C(h)}{\sum C(h)} (x_i - x_0^*)^2 * \frac{\text{local\_area}}{\text{total\_area}} \quad [66]$$

where:

$\frac{C(h)}{\sum C(h)}$ represents the covariogram ratio of the nearest four samples on a quadrant search between

the true sample and the estimate location¹⁹⁰;

h corresponds to the spatial distance between x_i and x^* ;

x_i equals one of the four nearest true samples;

x_0^* is the estimation value at that particular location;

local_area is the region defined by the four nearest samples to the estimation point (the local_area/total_area ratio represents the region spread factor).

This direct conventional and geostatistic adaptation contains six central properties:

- The quadrant search reflects the non screen influence effect.
- It ensures unbiased results because the $\frac{C(h)}{\sum C(h)}$ sum weight factor equals one.
- $\frac{C(h)}{\sum C(h)}$ is always positive.
- It uses the rescaled variogram structural distance whose influential data location receives a directly proportional weight.
- It includes sample and estimation values since neighborhood variability are crucial to local uncertainty assessment.
- This local variance index equals zero when x_0^* equals x_i .
- The variogram range with the highest Moran I achieves the best results¹⁹¹.

To confirm this reliable measure against OK error, Appendix O presents the GS+[®] Pb default dataset to verify this index, although the following twenty-one random observations are used for validation purposes: 7, 12, 20, 23, 36, 44, 51, 56, 58, 64, 65, 70, 85, 94, 105, 112, 113, 116, 117, 119 and 121. The mean, variance, skewness and kurtosis are 0.3756, 0.0368, 0.97 and 1.54, respectively. The R index equals 1.1825, its centroid is located at coordinates (39, 38) while its dataset range varies between 0.060 and 1.14. The local absolute true error¹⁹² presents a null Pearson correlation with OK variance, a local non-uncertainty measure. Yet, the right-hand scattergram in Figure 27 shows a reasonable linear relationship with the

classical local spatial variance whose Pearson correlation reaches 42% for a covariogram range of fifteen units. If the covariogram range equals seventy-three, the Pearson falls to 22.5%.

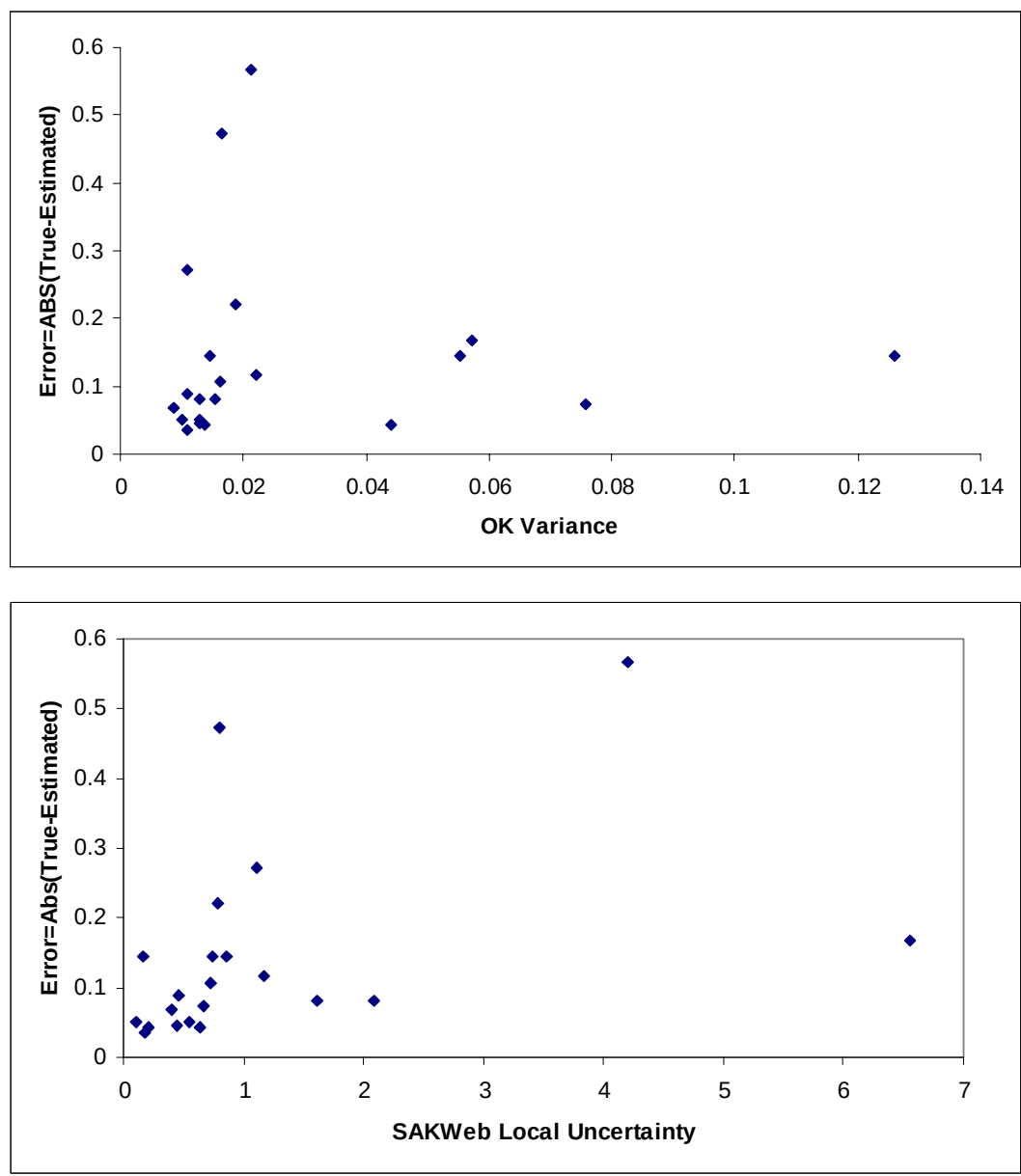


Figure 27 – The absolute true error against OK variance (left) and local variance (right).

With the lead substance of the GeoEAS[®] dataset¹⁹³, the next nine samples were used for validation purposes only: 1, 7, 15, 21, 23, 29, 54, 56 and 58. Based on the exponential variogram and confirmed by Pearson, both results suggest that this classical local spatial variance is a more sensitive uncertainty model than OK variance because it is data dependent. The Pearson correlation of the classical local

spatial variance equals 65% against 38% of OK variance (see Figure 28). It is hoped that future research may improve this promising uncertainty index. For instance, changing the constant neighborhood samples must be considered. It is known that this number must vary according to the location estimation. If this point is located in an erratic area then the vicinity observations considered must increase while, with stable regions, this number should decrease. To consider only the samples that are under the range with the highest Moran I is an option too. Because OK variance is a good geometric uncertainty measure, another possibility could be the rescale of the original variogram to a sill of 1, calculate the OK variance for each point and substitute this factor by the $\text{local\_area}/\text{total\_area}$.

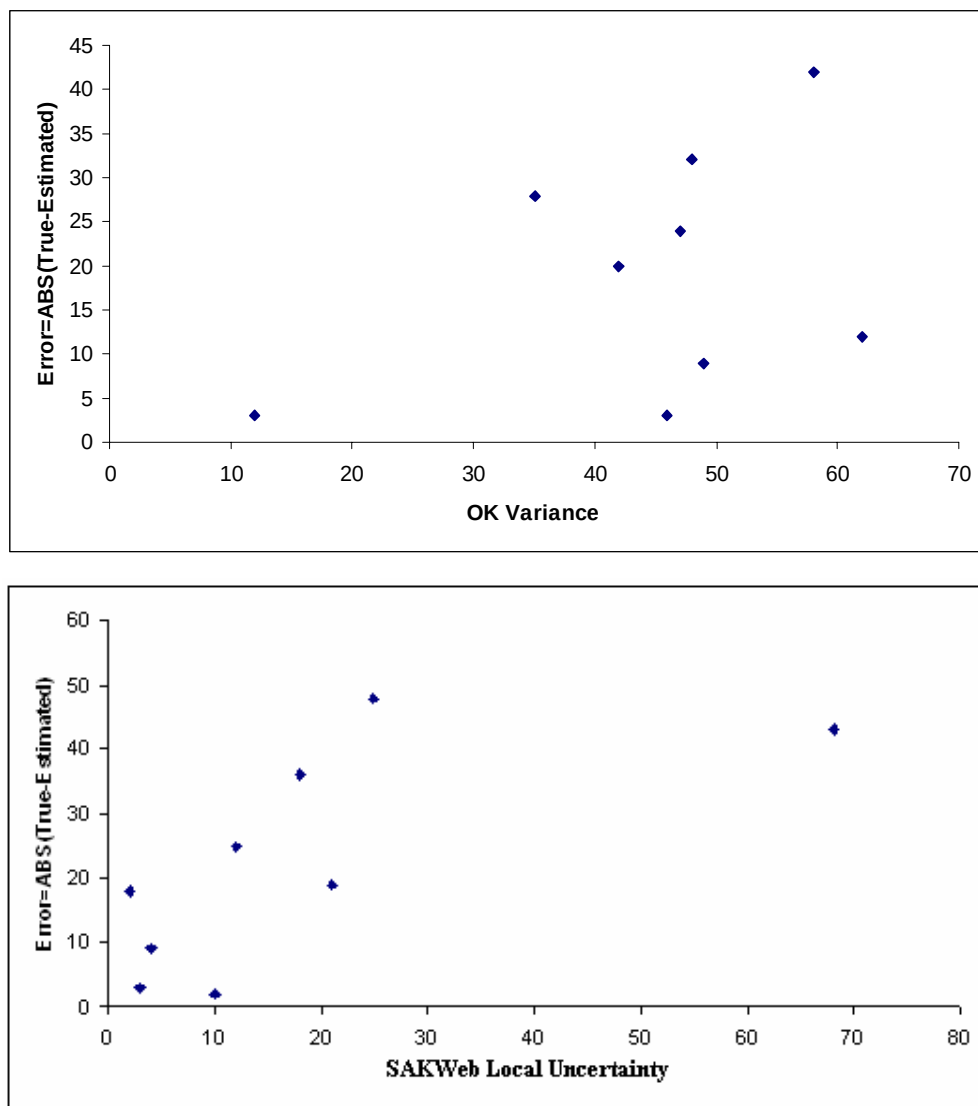


Figure 28 – The absolute true error against classical local spatial variance (left) and OK variance (right).

3.7. CONCLUSIONS

If we forget the future and consider the issues that surround spatial analysis today, the gulf of ignorance between the known and the unknown is a difficulty not yet resolved. Of the approaches presented above, Kriging seems to achieve better interpolation results (see Table 14). As confirmed by Burrough [1991, 1998], “B-Splines result in visually attractive maps with limited computing while Kriging computation, despite the costs, leads to better results”. Certainly, the variogram is a major breakthrough for finding the variation structure that underlies space.

Kriging	Deterministic Interpolators
Samples weights are based on the spatial autocorrelation function.	Samples weights are based on distance.
The search neighborhood is setup by variogram range.	The search neighborhood is arbitrary.
Anisotropy is detected.	Anisotropy is ignored even quadrant and sectors approach may help.
It takes accounts of sample clustering.	Redundancy is ignored.
Geometric uncertainty can be assessed.	Uncertainty cannot be computed.

Table 14 – The major differences between Kriging and other interpolation methods.

If samples come from a non-Gaussian distribution, the variance assessment from a raw untransformed dataset is a skewness measure, not variability. Hence, some change in the transformation should be carried out. Data transformation provides geographic data to produce statistically appealing estimates, particularly important to t-Student and F tests. On the other hand, this transformation viewpoint focuses solely on attribute information and completely ignores location. This omission of the location component may also create biased results. For instance, according to Griffith and Layne [1999], “Puerto Rican agriculture study found significant differences among the administrative regions in the mean production based on raw data¹⁹⁴ but not a significant one with transformed data”.

A review is provided of the initial univariate and bivariate analysis stage¹⁹⁵, sampling design, deterministic interpolation techniques, the nature of theoretical regionalized data, uncertainty assessment and variogram-related computational

issues. Finally, a new uncertainty measure, the classical local spatial variance, is introduced with promising results, although further improvement and confirmation is needed. Parallel to the concluding section of Chapter 3, the new developments regarding Kriging within SAKWeb[®] are presented in Chapter 7.

4. Spatial Analysis: Implementation Review

Chapter 4 is divided into six major sections. The first three sections focus on statistical packages for spatial data, GIS modules and independent spatial analysis software while Section 4.4 stresses the close link potential between spatial analysis and GIS topology information. E-learning and GUI major features are presented in Section 4.5. Last section draws the conclusions for this chapter.

4.1. STATISTICAL PACKAGES

One software implementation solution is to relate GIS to existing statistical packages such as SAS-GIS^{®196}, SPSS-X[®], Glim[®], Systat[®] or Minitab[®] (see Appendix A for a further evaluation). In his hydraulic oxidized conductivity study¹⁹⁷, where deeper layers exhibited shorter scale variation than the uppermost oxidized one, Ella [2000] used GeoPack[®] statistical software¹⁹⁸ for variogram nonlinear least squares minimization while GEO-EAS[®] performed cross-validation and the Kriging estimation. Disregarding GIS technology exclusively, Zimmerman *et al.* [1999] applied S+SpatialStats[®] procedures¹⁹⁹ in their experimental comparison interpolation research according to four factors: surface type (plane, sombrero, Morrison), sampling pattern (hexagonal, inhibited, random, clustered), error variance (0.0036 and 0.0576) and small-scale spatial correlation strength among residuals (4 and 10 units). All these illustrations involve ASCII files exported from the GIS to the statistical software. New trends in software development point include data exchange where a set of independent portable spatial analysis macro and routine libraries co-exist with the target GIS.

Following this philosophy, Griffith and Layne [1999] present the SAS[®] code for Moran I, Geary C and spatial autoregression modeling where the spatial

autocorrelation connectivity matrix is introduced by a CON or NEI file²⁰⁰. Waugh [1986] refers Geolink[®] which allows linkage through ASCII files to any GIS while Walford [1995] calculates the mean center and the variance-mean indices with Minitab[®]. According to Levine [1997], Ripley and Vanables [1995] have developed a collection of S-plus[®] spatial functions²⁰¹ creating an interactive interface between point coverage and the concentration estimation of the K-function²⁰². Using the United Kingdom campylobacter and salmonella dataset in the Lancaster District for the K-function point clustering, Rowlingson and Gatrell [1996] report the Mikhail S-plus[®] module that reads ArcInfo[®] point coverage and plots the output by the ArcGraph[®] routine. Finally, Flowerdew and Green [1996] mention the complexity of linking ArcInfo[®] with Glim[®] with regard to the polygon overlay algorithm.

The difficulty involved in this question is that statistical packages do not provide the GIS functionality needed because they were not developed to handle spatial data or, particularly, the structure of spatial data switching back and forth. Furthermore, the large volume of spatial data²⁰³, the complexity of the topological structure and the lack of interface transparency are a reality (see Table 15). To display GIS maps within statistical packages is not sensible either. In his spatial analysis of net migration in Rwanda, for instance, Brown [1996] complains about the effort required to export and import data from ArcInfo[®] to Spacestat[®]. The same consideration can be assumed with Skinner's prediction [1999] for Colorado grasshopper research where spatial analysis functions were performed with S-Plus[®] 4.0 while ArcInfo 3.5[®] did the mapping. It is certainly not possible to ask GIS users to buy S-Plus[®] while it is unlikely to have interfaces for all GIS and statistical products. The requirement to learn more than one package can also be problematic.

Many objects and points	Errors need to be random
Many variables	Not samples in usual sense
Large data volumes	Surrogate variables abound
Spatial autocorrelated values	Non-linearity is the norm
Spatial data error sources	High-degree of complexity
Non-conformity with standard statistical distributions	Modifiable areal unit, scale and aggregation effects
Data precision can hold spatial structures	Mixtures of measurements

Table 15 – Spatial data characteristics by Openshaw [1994].

SAAP[®] and Spacestat[®] implementation rectify this situation by replicating all data structures, although software duplication, manuals and topological structures are a questionable and costly effort. Fortran[®] routines integrated into Glim[®], allowing access to ArcInfo[®] data structures, is a fascinating project from Lancaster University to overcome this issue. Another example, presented by Collins [1998], is NO₂ urban air pollution modeling at Huddersfield, England, through the Gaussian dispersion model, CALINE3[®], and Kriging. This linkage with ArcInfo[®] was made through AML[®]/Fortran[®] programs and ASCII files. At the MAGI project, SPSS[®] produces all the descriptive statistics, correlations and multiple regression analysis. Anselin [1998] reports S+Gislink[®] close coupling between S-Plus[®] and Unix[®] ArcInfo[®] where commands are called from another system in a seamless inter-process communication. S+ArcView[®] is another example of close coupling between S-Plus[®] (server) and ArcView[®] (client) where S-plus[®] commands are entered within ArcView[®] while spatial data are passed by an intermediate ASCII file in order to calculate the Moran scatterplot and the variogram cloud. However, quoting Anselin [1998], “this linkage is not fully reliable since a built-in mechanism for a consistent identifier regarding each observation in both systems is lacking thus resulting in limited dynamic linkage among graphs”. In a September 2001 survey of Portuguese GIS users (Appendices B and C), 30% of the responses on the use of topological information within statistical software²⁰⁴ were affirmative.

4.2. GIS MODULES

A second implementation strategy is to add statistical functionality into GIS by modules integration²⁰⁵. A major advantage of full integration of spatial analysis into commercial GIS is good software documentation and ready availability of spatial analysis functionality for all GIS users. This solution is more secure for users and simultaneously saves time. Moreover, users prefer to buy complete systems²⁰⁶. According to a January 2003 survey (Appendix H), 80% of users feel more comfortable working with a known commercial product (statistical and GIS modules) than unknown independent software. It may be considered that the greatest benefit is the software company trademark attached to the GIS module. The common user trusts it blindly because of the brand marketing that lies behind the software, as confirmed in the same survey: 95% of common users hold the opinion that the software of commercial corporations is trustworthy and reliable²⁰⁷. In addition, direct GIS spatial queries can be enhanced. Concerning spatial autocorrelation and autoregressive models, they could also benefit from direct access to topological information. Finally, macro GIS languages can analyse spatial data in a much easier way with fewer code instructions.

Macmillan and Pierce [1996] reported that the most powerful option of vector TransCAD^{®208} is the allowance of C[®] and Fortran[®] procedures for the electoral redistricting module of simulated annealing, Anneal. Once the procedure is selected from the procs.mnu procedure, the proc.cmd prompts the user for any required input parameters. The DOS[®] proc.exe executes the algorithm and generates a .sol file while the proc.bck sends this output to the appropriate TransCAD[®] database within the chart window. Thus, it is simple to add new modules or reuse older ones²⁰⁹. Because a binary connectivity matrix is an Anneal requirement, TransCAD[®] is also able to generate it in ASCII format.

The major GIS module deficiencies are their high costs in terms of software, hardware and maintenance and their slow scripting language execution²¹⁰, particularly in the presence of large datasets. “Avenue[®]”, for instance, can only use

List objects to store matrices in one dimension” [ESRI, 1997, 2001]. Another major issue is the lack of GIS developers for spatial analysis issues because profitable GIS companies do not seem involved in a spatial analysis research effort. For instance, the NCGIA initiative [1992] had 35 participants but only two software corporations took part: ESRI[®] and Caliper[®], responsible for TransCAD[®]. Participants were mainly universities and research institutes. According to the analysis of the geostats site in April 2003 (<http://ai-geostats.jrc.it/software/>), over 70% of spatial correlation analysis, geostatistical simulation, multivariate geostatistics and interpolation are created by universities and research institutes. Quoting Dubois [2002], “the development of many add-on modules during these last five years seems to have helped the GIS industry to realize that there is a need for more integrated tools”. The recent release of the beta version of Arc/Info[®] 8 with a new module for geostatistics is a great step in this direction. At last, the availability of new tools will depend on software developers, primarily, and not on the research community²¹¹.

It is known that several theoretical research tools are already available in laboratories but the practical implementation involved if they are to reach the GIS community is still complex. This situation is changing with the new millennium, as evidenced by ESRI[®] [2001] with ArcGIS Geostatistics Analyst[®], for instance, and it is expected that other international corporations will follow this posture. Thus, the accomplishment of SAKWeb[®] also relies on the implementation philosophy of theoretical research papers. Quite often, after years of studies leading to beautiful conclusions, their practical application for the common GIS user is almost nil. These investigations reports and research papers end up on a library shelf. It is imperative to invert this tendency.

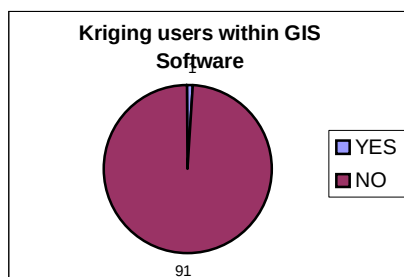


Figure 29 – The number of users that use Kriging within GIS modules revealed by the first survey.

4.3. INDEPENDENT GEOSOFTWARE

Bailey [1996] corroborates the vision of statistical deficiency stating that ArcInfo[®], Spans[®] or GenMap[®] only offer support for basic statistical summarization, data transformation or simple graphics like data posting and frequency distribution. Kollias *et al.* [1999] in their clay research at the Pinios river basin, Greece, used ArcInfo[®] extension for soil data storage, while Variowin[®] carried out the Kriging and IDW interpolation. Meanwhile, both fuzzy maps had to be programmed with new AML[®] commands at ArcInfo[®] level. In contrast, the Regard[®] independent software (see Section 5.5 for a further review) allowed several data layers to be associated into a single map. It also permits calculations to be carried out between themes and regions, hence facilitating spatial analysis. Certainly, the major problem is the effort required to write a complete package from scratch, reflecting a well-known dilemma: customization versus the time and money savings of existing software modules. Regard[®] is considered by experts as belonging among the best-specialized Kriging software although it is not well known by Portuguese GIS users. According to Figure 30, GeoMS[®] is only known at IST, while VarioWin[®] is preferred by FCUL students.

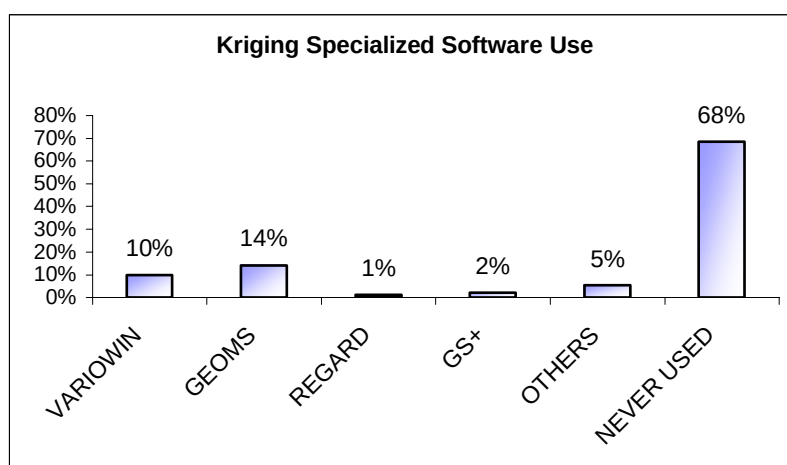


Figure 30 – The Kriging software awareness revealed by the September 2001 survey.

The same procedure was applied by Goovaerts [2000] with Geostatistical Software Library, GSLib[®], in his multivariate geostatistical prediction by incorporating raster DEM into rainfall for the Algarve, Portugal. Since correlation between rainfall and elevation ranges from 33% to 85%, larger prediction errors were obtained by algorithms that ignore the elevation factor, like Thiessen or OK than CK. Similarly, Nalder and Wein [1998] used the GStat[®] Kriging package with Visual Basic[®] to generate the required data and command files for GStat[®] while Chainey and Stuart [1997] report the use of Turbo Pascal[®] to develop the Voronoi interpolator. Stein and Zaadnoordijk [1999] run a groundwater flow simulation package, TRIWaco[®], to model the dual aquifer of Goeree in the Netherlands. Boykova [1999] used GeoEAS[®] to estimate the depth discontinuity at the Moho Balkan Peninsula²¹², a seismically unstable area. After a strong spherical geometric anisotropy had been setup with a low nugget-effect, Boykova [1999] confirmed a different map from those published including new spatial structures, particularly in southeast Greece. “Geolith[®] is also referenced by the Chevron Petroleum Corporation” [Frank *et al.*, 1999]. Contrary to common users, it seems that researchers do not use proprietary GIS technology and choose to write their own programs or prefer specialized software in order to carry out their own research.

An extension of this approach is the close coupling of GIS products and specialized ones such as ArcInfo[®] and Sage[®] of Sheffield University. “Due to performance limitations, ArcInfo[®] acts as a server while spatial analysis acts as a client, allowing client-server processes to be distributed around the computer network using RPC²¹³ technology” [Haining *et al.*, 1997]. Still, one drawback with Sage[®] remains: the need to own a SunSparc[®] workstation with Solaris[®] 2.4, Motif[®] 1.2 shared library, NAG Fortran[®] library and Arctools[®] 7.0.2. Anselin [1998] reports another RPC mechanism replica in ArcView[®] 3.0 from Iowa State University and AT&T Labs, the regression exploratory XploRe[®] and XGobi[®] (see Figure 31). XGobi[®] is a dynamic graphic program that allows multivariate data to be explored within Unix[®] ArcView^{®214}. STAC[®], a spatial temporal crime analysis tool²¹⁵ for hot spot search developed by the Illinois Criminal Justice Information Authority, produces ASCII result files that can be displayed into ArcInfo[®], MapInfo[®] and Atlas[®] GIS. Fragstats[®] for ArcInfo[®], an analysis of landscape

patterns, Gstat[®] for IDRISI[®] and PC Raster[®] or Vertical Mapper[®] for MapInfo[®] are other GIS software extensions with capabilities in the fields of statistics, distance, overlay, database queries, geostatistics, import/export, image and hyper-spectral processing, Fourier, TIN and simulation.



Figure 31 – The XGobi[®] menu within ArcView[®].

4.4. SPATIAL ANALYSIS VERSUS TOPOLOGY LINKAGE: A NEED

If the main part of truly spatial analysis is performed by independent software, the effort to link spatial analysis to GIS can be compensated for by the direct access to topological and spatial objects with a comfortable WIMP interface. “When housing sellers need to compare socio-economic areas that contain private swimming pools, this relationship is beyond the traditional SQL database operation” [Korth, Silberschatz, 1994]. For instance, “AML[®] and Fortran[®] ZDES[®] multilevel modeling software is able to define housing market areas in Hampshire, England, on the basis of property attributes and spatial location” [Bullen, 1998]. Thus, the key question is how to combine GIS spatial structures with spatial analysis tools.

For simple descriptive analysis, the potential of spatial analysis linkage to GIS is of little interest. Regarding nearest neighbor methods, connectivity measures directly derived from the topology matrix can be extremely important. Road network distances may be more relevant than straight line distances in the case of traffic accidents. Because spatial correlation techniques take account of the spatial covariance structure to find spatial patterns, outliers and unstable regions to satisfy the stationary assumption, the greatest linkage potential regarding spatial autocorrelation and autoregressive methodology is the **W** specification (see

Section 2.7 for a further discussion). Since the topological relationship is already stored within GIS, this goal can be achieved with little effort. Otherwise, this debate is quite irrelevant. Splitting a map into different regions graphically and comparing classical statistics results among them can be accomplished by any either approach separately. Computing local variograms with common boundaries including a specific spatial object inside can only be achieved with topological information.

SpaceStat[®] has the capability of creating ArcInfo[®] boundary length weights within its GIS interface. Taking into account road distances or distance-time²¹⁶, this methodology can lead to other improved possibilities for the **W** specification. According to Figure 32, if the **W** time-distance matrix is based on the train access time then its specification will be quite different when compared with that of the common map of the Portuguese provinces, particularly Trás-Os-Montes.

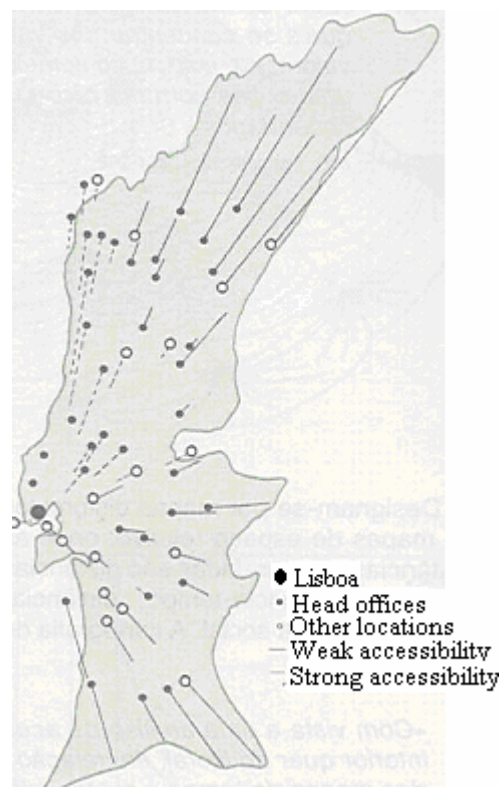


Figure 32 – The train access time from Lisboa [Figueira, 2000].

4.5. E-LEARNING AND INTERFACES

Quite often, common users request spatial analysis knowledge in a self-learning view because of global cost reduction, both time and money. In addition, “implementing the technological structure that supports e-learning platform is a scalable solution” [Painho *et al.*, 2002]. Spatial analysis wizards, multimedia tools (including animation and hyperlinks), on-line help, software courses, videos, e-Learning²¹⁷ and M-Learning with WML technology are ingredients of this demand. CRM of the Universidade Autónoma de Lisboa and GIS&Sc Master of ISEGI, certified by UNIGIS, are some examples of the Learning Space strategy (<http://www.lotus.com/home.nsf/welcome/learnspace>) [Semana Informática, 2002]. The International Center for Distance Learning (<http://www-icdl.open.ac.uk>), the AT&T Learning Network Virtual Academy (<http://www.att.com/learningnetwork/virtualacademy>), La Escuela de Negocios a Distancia de la Universidade Politecnica de Madrid (<http://www.cepade.es>), Le Centre National d’Enseignement à Distance (<http://www.cned.fr>) and the Universidade Aberta (<http://www.univ-ab.pt>) are others. As confirmed by Santos [2000], five million US students were taking distance degrees in 1994 and two million in the EC in 1997.

In Europe, this learning trend has been successfully tapped into by EC programs such as Delta (1988-1990), Telematics Applications (1994-1998) and Multimedia Task-Force (1999). Strong future development is therefore expected, supported by current brain-ware know-how:

- The UNAVE project from the Universidade de Aveiro (<http://www.unave.pt/fd>) reveals greatly encouraging results with its multimedia Web interfaces program²¹⁸: 70% of the 350 students passed, 20% were advised to repeat and 10% failed.
- With PROF2000 LAN/WAN management classes²¹⁹, the final grades of nineteen high-school professors ranged between 77% and 90% [Santos, 2000].

- During 1999, the estimate for Portugal Telecom's cost savings with e-Learning (<http://formare.ptinovacao.pt>) reached €270,000 for ISDN, LAN/WAN and telecommunications courses.
- Santos [2000] presents a p-Fisher statistical study comparing conventional (276 participants) and distance learning (42) methodologies on the basis of age (26-35, 36-45 and 46-55) and school qualifications (grades 6, 7-9, 10-11 and 12), revealing a positive homogeneity of both approaches with a 95% confidence level. This conclusion was based on two representative populations of the same number, 30. A positive Pearson correlation was also found between final grades and school qualifications (+0.313) and a negative one with age (-0.413).

Nevertheless, users should understand that in spite of any help that any teaching technology may give, the theoretical background of any problem will be best solved in an ongoing practical context. "Geostatistics owes much to practice" [Isaaks, Srivastava, 1989, ESRI, 2001]. This observation does not suggest that GUIs should not be improved or e-Learning should not be adopted, but non-specialists should be aware that insufficient practical experience cannot lead to high-quality answers. It is imperative not to forget the traditional, live classroom teacher and the work-group discussions required for the preparation of a paper. From previous research, it is possible to infer²²⁰ that the main negative aspect of e-Learning is the lack of class attendance to satisfy the demands of the human tutor/student factor.

From the MS-DOS[®] to the X-Windows, Motif[®] or Open Look[®], these interfaces offers integration, opening architecture, imagination, look and feel features since the nineties. For Negreiros [2000], "the best human computer interaction is the kind that allows users to concentrate on the job in the fastest way, the easy-to-use menu-driven interface". Restaurants with touch screen applications are a good example where the keyboard is already history. It is crucial that unfamiliar users can use intuition and human common sense to infer the purpose of a particular graphic object, because if the human aspect is neglected then people will waste time and energy understanding the computer interface. "A good error message

should give clues to overcoming the problem while the resolution should rely on a question-answer approach” [Negreiros, 2000].

In addition, operating manuals documentation should be considered as learning as educational books based on a non-theoretical methodology. Certainly, GUIs should be evaluated by end-users since software designers only trust their own intuition. The simultaneous use of mouse and keyboard for accounting software is a good example of a time-consuming operation in a field changing operation within forms. According to FDTI [2001], “poor control panel design was officially reported to be a major factor in the nuclear power station disaster of Three Mile Island”. “A number of studies have found that close to 50% of Web shoppers give up purchasing items on-line because the site was too difficult to navigate and they could not locate products on the Web site” [Laudon, Laudon, 2002]. Model Builder[®] from ArcView[®] Spatial Analyst 2 tries to overcome these problems. “With an interactive flow chart modeling tool, its wizard helps spatial data choice, spatial analysis selection and data output definition” [ArcUser, 2000]. It is also always better to represent continuous variation on the red-green-blue/red-orange-yellow spectrum or using one color only but varying from the darkest to the lightest shade. According to Longley *et al.* [2001], “most users are unable to differentiate between more than seven ordinal categories”.

Comprehensive help facilities using Web hypertext and Macromedia Flash[®] (see Section 7.1.8 for a further review) offer new opportunities within geocyberspace although VRML and virtual reality hold a great potential for geographical applications in the near future. Quite often, W3 animation with interactivity is perceived with special attention. SAKWeb[®] help was built over this platform. Professionally speaking, this is how Web designers call consumers’ attention to their products (see Table 16 for some relevant Web economics indices from <http://www.internetindicators.com>). “Bandwidth is increasing by at least 300% per year at the moment and it can be expected to accelerate to the point where physical access to all data is essentially local” [ArcNews, 2000]. The Internet2 infrastructure is already on the way based on a series of interconnected gigaPoPs which are regional high-speed points-of-presence that serve as aggregation points

for traffic from participating institutions. Also, today's internet transmissions are best effort packets of data because they arrive without any regard to the priority of their contents. Under Internet2, different types of packets could be assigned different levels of priority as they travel over the network. Email messages, for instance, could be delivered when capacity is available, a situation not pursued by videoconferencing packets.

Concerning e-Learning, it is important to guarantee a professional site with this capability but without forgetting the virtual environment, e.g., the feeling of entering a virtual classroom and the possibility of knowing the list of participants, accessing hypermedia contents including downloads in a sequential learning framework, retrieving notes and curriculum timetables, visiting a virtual chat bar for students to exchange ideas, editing historical messages and clarifying questions with a tutor in an email-IRC forum. The community of SAKWeb[®] news and chat community pursues this belief.

Five million Web servers with 800 millions Web pages.
Ten million networked homes in the US by 2003.
Sixty-four million U.S. adults were regular Internet users in 2000.
The U.S. Internet Service Provider market generated \$15.1 billion in 1999, a 45% increase over 1997.
In Europe, the ISP market generated \$4.3 billion in 1998 while Internet advertising makes \$1.92 billion.
Seventeen million US households shopped online in 1999.
Total E-commerce and E-business reached US\$1 trillion in 2000.
U.S. high-tech industry employed 4.8 million workers in 1998.

Table 16– Internet economic indicators: The evidence²²¹.

4.6. CONCLUSIONS

In a speculative rather than definitive way, the future development of spatial analysis depends on the implementation of spatial techniques born in university

research laboratories. It is well known that independent geosoftware and statistical packages solutions are preferred by most spatial analysis experts, while GIS modules continue to manage large volumes of spatial data rather than carry out sophisticated statistical analysis. For this reason, it is hard to compare independent geosoftware and statistical packages with GIS modules. In terms of the chronology time of this dissertation, for instance, this evaluation could only be performed with the third survey (see Section 7.3 for a further discussion) because the ESRI[®] Geostatistical Analyst module was introduced to this dissertation to the author in the second quarter of 2002.

The reality is that experts and traditional users have different software preferences for spatial analysis tools. Although no survey of experts like Goovaerts, Sousa or Anselin was carried out, major articles and books such as Geocomputation, A Primer, only refer to non-commercial software for the computation of Kriging, risk cost maps and spatial autocorrelation research. This is understandable because there are almost no commercial GIS modules with these capabilities. In addition, experts are not particularly concerned about interfaces. It can also be hard to include particular parameters within GIS modules because of the black-box strategy of commercial companies. Finally, their constant use of Unix[®] and Macs[®] platforms and free GSLib[®] software are major evidence of this small, well-defined population. In contrast, according to the third survey, these common users know Windows[®] pretty well while 20% are aware of Mac[®]. The Linux[®], Sun[®] Unix and IBM[®] operating systems are completely unknown platforms.

To invert this solution and put both communities on the same computation level, it is urgent to upgrade the subjects in the geography curriculum at universities in order to create pressure for new and revised GIS modules. Ultimately, as Section 4.4 stresses, access to topology may be a basic requirement for spatial analysis tools. Following the spatial analysis classification, Section 4.1 (statistical packages), 4.2 (independent geosoftware) and 4.3 (GIS modules) describes each one while Appendix A²²² introduces a table summarizing geostatistical software available nowadays. Finally, Section 4.5 reaffirms the use of e-Learning and Web help facilities to disseminate spatial analysis issues around the globe.

5. Spatial Analysis: Software Evaluation

Most spatial analysis computer programs are available on the ‘World Wide Wait’ leading this platform to a new GIS one. Ultimately, this world expects fast answers because if it works then it is obsolete. Surface[®] III, Regard[®], Ecosse[®], GeoEAS[®], GeoMS[®], GS+[®], Kriging for Berkley Unix[®], Surfer[®], GMS[®], Uncert[®], VarioWin[®], GSlib[®], SpaceStat[®] and GeoPack[®] are analyzed in the next sections. The ESRI Geostatistical Analyst[®] module and Kriging ArcScripts[®] written in Avenue[®] are also reviewed.

5.1. SURFACE III[®]

Developed by the Kansas Geological Survey at the University of Kansas, this software is available for Macintosh[®] III+ version. Its input is a text file containing the coordinates and data values. If a particular data location has multiple values then those values can be averaged or deleted. Contour lines and data postings can also be overlaid on shade maps but with the possibility of adjusting labels so that overprinting is minimized (see Figure 33). In terms of the view of the graphics, the Azimuth command controls the direction²²³, Distance identifies the space between the viewpoint and the block center while Elevation defines the perspective angle above the horizon.

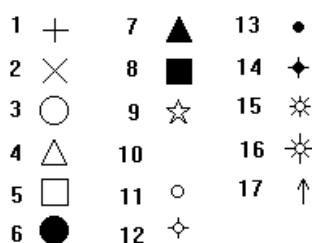


Figure 33 – The Surface III[®] data postings.

Surface III[®] produces spatial statistics including:

- A map of the distance to the nearest, farthest and average neighbors.
- The standard deviation for the distance and the number of the vicinity observations for a particular point.
- The number of octants and quadrants and the relative contributions of the nearest, farthest and average neighbors to the final estimate.

Regarding gridding algorithms, numerical arrays from uneven space data and square patterns are possible. Naturally, if a large grid cell size is chosen then the resulting map presents a low resolution and a coarse appearance but it can be computed quickly. In contrast, if grid cells are small then contour maps have a finer appearance with time consuming expenses. Transects and moving averaging gridding are further possibilities. Figure 34 introduces quadrant and octant searches, two methods that provide a search procedure for finding data points within a particular radius of the estimation point²²⁴.

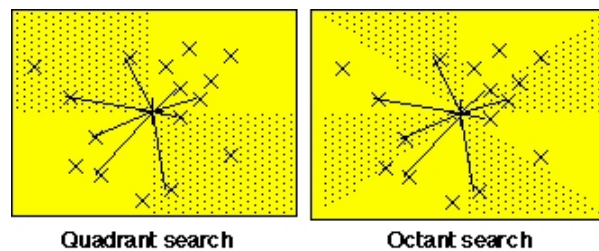


Figure 34 – The Surface III[®] search neighborhood.

5.2. VARIOWIN[®]

Created in 1995 by Pannatier at the University of Lausanne, Switzerland, this geostatistical shareware can be found at <http://www.springer-ny.com> (986k). Written in Borland C++[®], this application is divided into four modules: Prevar2D, Vario2D, Model and Grid Display. The ability to handle datasets with UTM coordinates²²⁵ and the capability to read data files with empty lines are two attractive features of Prevar2D whose major goal is to create the .PCF file from the .DAT. Another extended feature is to set limits for the active subset in relation to

the x-axis, y-axis and other variables. The default estimator is the traditional variogram although the madogram²²⁶, the correlogram²²⁷ and the covariance can be plotted too. The Model unit includes interactive geometric and zonal modeling with a particular characteristic: the Indicative Goodness-of-fit checks different nested models without forgetting the best fit model²²⁸. The chosen model is also saved in an ASCII file but it includes the capability to produce a variogram model that can be copied into Gslib[®].

5.3. GS+[®]

Geostatistics for the Environmental Sciences is a VBasic[®] program for Windows[®] platforms designed by Contact Gamma Design[®]. Some major features are offered, e.g.:

- The Integrated Data Worksheet allows the user to edit and filter data temporarily or permanently from the original dataset. Missing values are ignored during analysis. Field assignments are made on the worksheet by clicking on the top of the row.
- It is possible to specify a log-normal or a square-root transformation for skew distributions, prior to Kriging analysis. Also, the GS+[®] Kriged report can be either the transformed or the back transformed²²⁹ data.
- A number of different file types can be imported/exported into/from GS+[®]. These include spreadsheets (Excel[®] and Lotus[®]), relational databases (Access[®], DBase[®], Paradox[®], FoxPro[®]), GeoEAS[®], Surfer[®], ArcView[®], HTML and text files.
- The 3D anisotropic semivariance surface provides a visual image in every compass direction.
- 3D map images can be rotated to be viewed at any angle.
- Five isotropic models (spherical, exponential, linear, linear to sill or Gaussian) can be selected, including a major or minor range definition for anisotropic regions. Analogous to SAKWeb[®], GS+[®] provides two extra statistics to help the interpretation of the model output: 1) The Spatial

Structure Proportion, $C1/(C0+C1)$, provides a proportion measure of the sample variance ($C0+C1$) that is explained by the spatial structured variance $C1$; 2) The Reduced Sums of Squares gives an exact measure of how well the model fits the variogram data: the lower the reduced sums of squares, the better the model becomes.

- The Non-Uniform Interpolation Grid specifies the locations to be Kriged.
- The latest version (5.1) includes CK and IDW.
- The Mapping window specifies the break points and colors for mapping contour intervals.
- By placing the mouse on the top of individual points within the Variance Cloud window, it is possible to determine which pairs of points are suspect.
- The Polygon Outline Map window contains an outline polygon map used to include (blue) and exclude (red) areas from default Kriging interpolation.
- The Graph Print Dialog window allows users to specify printing parameters such as the device or file, the place on the page and the size of the image.
- The Moran I Analysis window allows visualization of this spatial autocorrelation index according to several lag distances.
- With cross-validation analysis, each point measured is individually re-estimated and compared with the real one.
- GS+[®] interpolates punctual and block Kriging using the nearest neighbor values weighted by distance. The default is 16 points with no restrictions for radius.

5.4. GEOMS[®]

Developed in 1998 by the Mines Department of IST, Lisboa Geostatistical Modelling Software[®] runs on Windows[®] and is divided into thirteen modules:

- geoDATA[®] for the analysis of univariate and bivariate statistics;
- geoVAR[®] for the calculation of experimental variograms;
- geoMOD[®] for variogram model creation with user interactive fitness;
- geoKRIG[®] for OK and SK with global mean;

- geoKED[®] for Kriging with External Drift and SK with local mean;
- geoCoKrig[®] for CoKriging and Collocated CK;
- geoSGS[®] for Sequential Gaussian Simulation;
- geoSIS[®] for Sequential Indicator Simulation;
- geoCLASS[®] for Multiphase Classification;
- geoSA[®] for Simulated Annealing;
- geoPFS[®] for Probability Field Simulation;
- geoVIEW[®] for displaying color-filled contour or hypsometric maps;
- geoTRDATA[®] for data transformations.

Location and spread measures²³⁰ are the univariate statistics presented by GeoMS[®]. The box plot and cumulative histogram distribution is drawn too. After the azimuth angle, directions, tolerance, lag and cut distances have been setup within geoVAR[®], the variogram structure is setup by the user within geoMOD[®]. It also provides both isotropic and anisotropic variograms where users have complete control over interval lags.

5.5. REGARD[®]

Developed in 1990 in Pascal[®] at Trinity College Dublin for Apple Macintosh[®], Regard[®] seems to be distinguished Kriging software that allows simultaneous hot-linked interaction between data layers and views, including satellite imagery maps. This dynamic graphic capability is extremely useful for examining local variability and anomalies. Quoting Haslett *et al.* [1992], “we conclude that the most important single data view is the Map View: all others must be cross-referred to this one and future applications must encourage this concept”. It includes overlay plots on the background and brushing with many update links maps. ArcGIS[®] Geostatistical Analyst²³¹ pursues the same philosophy.

The Map View and the Histogram View help the user to visualize potentially significant spatial patterns. “The recognition of spatial discontinuities and other

anomalous behavior makes one reluctant to rely overly on any interpolation procedures at an initial stage for most data” [Haslett *et al.*, 1992]. According to Gunnink and Burrough [1996] and regarding the 100 zinc concentration measurements at the river Meuse, Netherlands, the highest histogram selection lies along the river borders (see Figure 35).

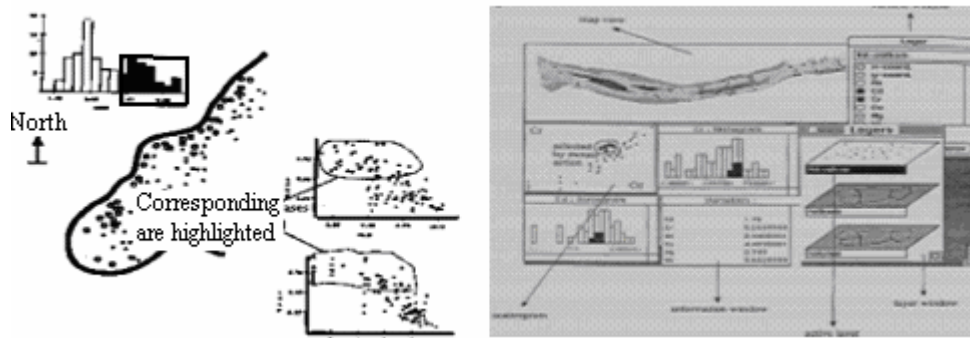


Figure 35 – The Regard[®] brushing and linking features.

The Map View also provides a mechanism for the user to define a window and to drag it, under mouse control, over an appropriate region. As it passes over data points, statistics are produced, like the constant updated mean or variance in the Trace View window. Hence, the direct control of spatial analysis and encouragement to experiment is stressed. SAKWeb[®] pursues this trend too.

It is possible to select data points on the variogram cloud and highlight them on the Map View. Also, the Scatterplot Matrix View among variables can suggest the possible existence of two populations. For instance, according to the Pb, Cu and Zn logarithm pairwise scatterplots of Figure 36, the top-right one reveals two different populations. Undoubtedly, it would be more difficult to apply CK to Pb and Zn than Zn and Cu.

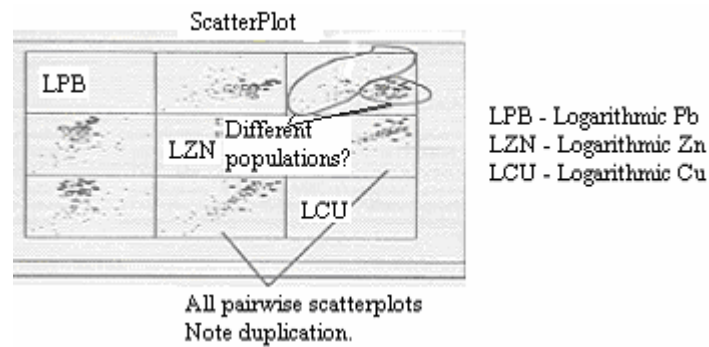


Figure 36 – The Regard[®] brushing and linking features within the scatterplot view.

The Rescaled Variogram Cloud View uses the robust estimator proposed by Cressie and Hawkins in 1980 (see Equation 2). The pairs selection of the Map View rescaled cloud that are closed in space but different in values are another attractive feature because these are the ones that most contribute to the nugget-effect. If these pairs are well scattered in space then outliers can be found. If these samples are concentrated in a particular region then local pockets of non-stationary emerge.

5.6. GMS[®]

GMS[®] is a comprehensive ten-module package (43.9MB²³²) that provides tools for groundwater simulation including site characterization, model development, post-processing, calibration and visualization. It also supports TINs, 2D and 3D geostatistics and other groundwater simulation techniques. Some other attributes are offered:

- Any dataset simulation can be converted to an animation film loop saved in AVI format.
- Clough-Tocher interpolation, the moving weighted average and mapping elevations are incorporated.
- Data Calculator allows the comparison of two interpolation dataset fields.
- IDW, OK, UK, IK and Indicator simulations are covered by GMS[®].
- Imports/exports includes raster and vector GIS data from/to ArcInfo[®], ArcView[®], GRASS[®] and DXF[®] files for AutoCAD[®].

- The PC version runs on Windows[®] 2000 while the Unix[®] version is available for Compaq[®] Alpha, HP[®] Unix[®], SGI[®] Irix and Solaris[®] platforms.
- Images can be copied to the clipboard and pasted into other applications for report generation.
- The Mt3dms module is particularly suited for migration simulation of contaminant plumes over time.

5.7. GSLIB[®] AND GSTL[®]

The Geostatistical Software Library[®] is a complete collection of geostatistical programs developed by the Stanford faculty over the last 15 years, with two major goals: 1) to provide free geostatistical and simulation software to the public domain including Fortran[®] 77 source code²³³; 2) to supply WinGSLib[®], a commercial version for Windows (US \$200). Several aspects can be enhanced:

- Utilities programs (cell declustering, Normal Score Transformation, bivariate normal distribution for DK).
- Relative, general and pairwise variograms, madogram and Indicator variogram.
- SK, OK, UK with nine monomials, IK, CK and SK with external drift.
- Stochastic simulation (Gaussian, Sequential, Indicator, Boolean, probability field, ellipsoid and simulated annealing).
- PostScript plotting utilities (normal QQ and PP plots, gray/color scale maps, Normal and lognormal probability plots).

GsTL[®] is a C++ library of generic geostatistical programming for researchers and not a software collection for end users. Its aim is to provide tools for building new geostatistics algorithms, eliminating for the need to reinvent the wheel each time a Kriging routine is needed. Hence, this library is flexible and extendable enough to develop new algorithms, including source code and a detailed description of the requirements. Remy [2001] illustrates examples of this new approach. One of these is presented next²³⁴. The goal is to estimate CK weights of the point $z(0,0)$ from

$z(2,3)=0.21$, $z(4,-7)=0.09$ and one secondary piece of data, $s(1,3)=2.1$. After the `Kriging_weight` function has been called, the vector weights result in yield four elements: the first three are the weights corresponding to $z(2,3)$, $z(4,-7)$ and $s(1,3)$ while the last one is the LaGrange parameter.

```
/* Module inclusion */
typedef std::vector<geo_value2d> neighborhood;
/* Main program */
int main() {
/* Coordinates of input and neighborhood definition */
location2d u1(2,3); location2d u2(4,-7);
geo_value2d z1(u1,0.21); geo_value2d z2(u2,0.09);
Neighborhood neighbors1, neighbor2;
neighbors1.push_back(z1); neighbors1.push_back(z2);
/* Coordinates of secondary point and neighborhood redefinition */
location2d u3(1,2);
geo_value2d s1(u3, 42.1);
neighbors2.push_back(s1);
neighborhood*neigh_array[2]={&neighbors1, &neighbors2};
location2d u(0,0);
/* Definition of new vector and weights assessment */
std::vector<double> weights;
double kvariance=Kriging_weights(weights, u, neigh_array, 2, LMC_covariance, OKconstraint); }
```

5.8. KRIGING FOR BERKLEY UNIX[®]

The goal of this project, created by Lang [1997] at the Computer Science Department of Cornell University, was to implement an OK module using spherical and exponential variograms for Unix[®] IBM RS/6000 and HP-700[®], a system V of the Berkley[®] Software Distribution version [Negreiros, Amador, Castro, 1998]. According to Figure 37, the left image illustrates the estimated layout while the right one uses a color map to display error variance. Since the location of the white dots represents input samples, it becomes clear that the error variance is smaller in the presence of abundant observed data.

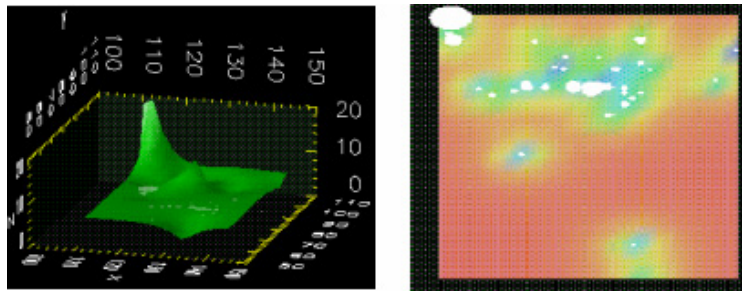


Figure 37 – The Kriging for Berkley Unix[®] OK interpolation and variance.

5.9. UNCERT[®]

According to Wingle *et al.* [1999], Uncert[®] is 2D and 3D geostatistical software applied to groundwater flow and contaminant transport modeling for Unix[®] workstations that run X-Windows and Motif^{®235}. Besides the traditional GUI, it also works from the command line or in batch mode, which is particularly useful for large datasets or to process the Monte Carlo simulation. This software is organized into several modules:

- Plotgraph is an X-Y graphic tool that performs least-squares regression for up to tenth order polynomials.
- Histo displays univariate statistical data and, with up to 20 populations, it can be compared simultaneously.
- Distcomp allows different data populations to be displayed together using histograms and general QQ plots²³⁶.
- Vario calculates experimental variograms and other spatial relations such as covariance, correlograms, log variograms, madograms, indicator variograms and cross-variograms.
- Variofit fits the experimental variogram using the least squares regression techniques and the jackknife techniques. Spherical, exponential and Gaussian models with up to four nested structures plus nugget-effect are also included.

- Contour provides a 2D contouring utility for regular grids including posted labels, direction arrows and line work representing roads, buildings or formation contacts, for instance.
- Analogous to SAKWeb[®], Surface treats grid values as elevations to create an undulating 3D surface.
- Block is a utility for viewing 3D scattered and gridded data.
- Modmain and mt3dmain offers a complete GUI interface for MODFLOW[®] and MT3D[®] simulation models with animation.
- Array is a tool for mathematical calculations on gridded data like variance estimation or cell frequency for a specific range. Constants can be added, subtracted, multiplied or divided for every cell. In addition, cells can be reclassified on the basis of their value.
- Grid module is responsible for interpolating data using three methods: IDW (residuals are evaluated at each sample location), trend surface analysis (up to fifth order) and Kriging (SK, OK and UK) including the cross-validation procedure. Kriged residuals from trend surface analysis are also possible to compute.

5.10. GEO-EAS[®]

Geostatistical Environmental Assessment is DOS[®] interactive software developed by the U.S. EPA Environmental Monitoring Systems Laboratory in Las Vegas, Nevada, in 1990 [Englund and Sparks, 1991]. The length of the GeoEAS.zip file equals 1248 KB making it, thus, the smallest independent geostatistical software available at present. The next items make up the executable file list needed to be called by Geoeas.exe:

1. Dataprep: Data preparation utilities;
2. Trans: GeoEAS[®] data file variable transformations;
3. Stat1: Univariate statistics;
4. Scatter: Data posting;

5. Prevar: Computation of pair comparison and used prior to the Vario unit;
6. Vario: Variogram analysis and modeling (see Figure 38);
7. Xvalid: Kriging cross-validation;
8. Krige: Ordinary and simple Kriging;
9. Postplot: Produces metacode graph files;
10. XYGraph: Line and scatterplots;
11. Conrec : Contours for grided data;
12. View: Utility to display metacode graph files;
13. HPSetup: A batch file to redirect printer output to COM1;
14. HC: A utility to change modes on Hercules video cards;

O P T I O N S					
Variable :Cadmium		Pair File : Example.pcf			
Minimum : .000		Min. Distance : 6.00			
Maximum : 16.7		Max. Distance : 302.			
Direction		Lag	Distance	Lag	Distance
		1	15.194	13	197.523
Direction : .000		2	30.300	14	212.717
Tolerance : 90.000		3	45.582	15	227.911
Max Bandwidth: MAX		4	60.776	16	243.105
		5	75.970	17	258.299
		6	91.164	18	273.493
		7	106.358	19	288.687
Lag Spacing		8	121.553	20	303.881
		9	136.747	21	
Minimum : .000		10	151.941	22	
Maximum : 302.362		11	167.135	23	
Increment : 15.194		12	182.329	24	
Direction New Lags Change Lags Post Plot Execute Quit					
Specify pair orientation (selection) criteria					

Figure 38 – The GEO-EAS[®] Vario options.

5.11. SURFER[®]

Surfer[®] 8 is a program for 3D contouring surface mapping that runs on Windows²³⁷ with the capability of converting spatial data into outstanding contour maps and surface plots. It also supports surface vector maps by showing the

direction and the data magnitude at data points, overlay maps and USGS DEMs.

Additional utilities include:

- Map exports into DXF, SHP, BNA, BLN, MIF, GSI, GSB, EMF, WMF, CLP, CGM, TIF, BMP, JPG, TGA, PNG, PCX, DCX, WPG and PCT file extensions;
- Windows Clipboard to support copying maps to other applications;
- Volume and surface area computation;
- Reshaping tools to edit areas and curves;
- File importing from a diverse range of formats such as DAT, TXT, SLK, XLS, WKx, WRx, CSV, BNA and BLN.
- Variogram models can be setup from the exponential, Gaussian, linear, logarithmic, nugget-effect, power, quadratic, rational quadratic, spherical, wave, pentaspherical and cubic models.

5.12. ECOSSE[®]

Developed by Alloa[®] Business Center, this remarkable geostatistical software was written in Visual Fortran[®] 90 for Windows^{®238}. If sample distribution follows the ‘bell’ curve, EcoSse[®] presents the ‘grade/tonnage’ curve whose cutoff probability plot is introduced and based on the estimated mean, standard deviation and probability population above the cutoff limit (see subsection 3.6 for further details). Based on Sichel’s correction factor, the lognormal transformation is also covered, including the plotting of the cumulative density function and other descriptive statistics. In addition, the linear regression relationship based on logarithms and ranked transform in accordance with nonparametric statistical analysis is possible.

The linear, quadratic and cubic polynomial trends can be used within EcoSse[®] as a global estimator or as a detrend Kriging process. Particularly, it is the random residuals variogram that confirms the detrend polynomial decision. If the remaining residuals show rises, peaks and falls, then an over-fitting data surface

was made and, hence, a lower polynomial model is needed. “The variance in residuals and the corresponding variogram sill should also be identical suggesting that the variogram will be a good variation indicator” [Clark and Harper, 2000]. Another two single features highlighted by EcoSse[®] are:

- The nearest neighbor option gives the the average and the closest distance among samples.
- The geographical area definition specifies the area through a file that contains the boundary information about the field for sampling limitation purposes.

5.13. GEOPACK[®]

GeoPack[®] is a US\$200 menu-driven package for conducting analyses of spatial variability with up to ten variables. The basic statistics includes mean, median, variance, standard deviation, skewness, kurtosis, maximum/minimum and coefficient of variation. It also includes linear and polynomial regression, the Kolomogorov-Smirnov test for Normal distribution and percentiles calculation. According to Ella [2000], “the variogram models can be fitted using the fitting procedure of nonlinear least-squares developed by Marquardt or the traditional manual interactive approach”. GeoPack[®] embraces OK and CK along with their associated estimation variance and conditional probability for a specific cutoff level. IK, OK and BK with geometric anisotropy are also integrated. Nonlinear estimators such as the parametric DK and disjunctive CK can also be assessed with the estimation variance. Up to ten cutoff levels are allowed.

5.14. ARCGIS GEOSTATISTICAL ANALYST[®] MODULE

This new ESRI[®] extension involves variogram modeling which includes graphical and automatic anisotropy assessment²³⁹, covariance surface layout²⁴⁰ and neighborhood search size²⁴¹ to a limited number of points used in prediction. Its

wizard includes four major steps, thus allowing new users to produce an easy surface with the default parameters:

- 1) Input and methodology;
- 2) Variance/covariance modeling;
- 3) Search neighborhood;
- 4) Cross-validation and validation with a new test dataset.

As with other geosoftware, it offers histograms and a normal QQ plot for a Gaussian distribution check, OK, UK and DK²⁴², trend analysis in a 3D environment for global trend detection and Voronoi as a spatial variability assessment. This includes the nugget-effect division into microstructure and measurement error.

Concerning the cross-validation procedure, four statistics are presented: the mean prediction errors, the root mean square, the average standard error and the root mean square standardization. Cross cross-validation based on an alternative new sampling dataset is an option. The prediction standard error surface for each location can be plotted. In conjunction with other boundaries layers such as hill shade, grid, transparency and filled contours, IK and PK for threshold probability mapping are also feasible. Finally, probability, quantile and prediction standard error maps can be computed from SK, OK, DK, UK and CK.

Regarding deterministic approaches, five Spline functions, IDW, and ten global (fitting a polynomial to the entire surface) and local (fitting many polynomials, each within specified overlapping neighborhoods) polynomial orders are covered. According to ESRI [2001], “IDW optimal power is determined by the minimized root mean square prediction error in a cross-validation framework”. In other words, Geostatistical Analyst[®] tries different powers and, according to the resulting root mean square prediction error, the power that provides the smallest one is the optimal exponential.

Another single feature of ESRI[®] GIS extension is the Normal Score Transformation dialog box, a ranked-based method including direct (using the

observed cumulative distribution), linear (fitting lines between each step of the cumulative distribution) and Gaussian (based on a smoother probability density function) approximation resolution. In addition, the conventional declustering methodologies of cell and Voronoi polygons can be used. Some other major features are offered:

- Similar to Regard[®], the brushing operation is applied intensively between the ArcMap[®] and the corresponding selected points in the histogram, normal QQ plot, variance cloud, Voronoi map and two dataset cross-covariance cloud views.
- The capability of computing measurement error, if the user possesses multiple measurements per location, makes this module unique too.
- Data classification is viable through the creation of range classes.
- The search neighborhood includes directional influences, the maximum and minimum number of locations to use and the sectors division number.
- The next variogram functions are provided: circular, spherical, tetrasherical, pentasherical, exponential, Gaussian, rational quadratic, wave-hole, K-Bessel, J-Bessel and stable.

This extension is revolutionary because it fills the gap between geostatistics and GIS since geostatistical software has never been completely integrated within GIS by the same corporation. In combination with ESRI[®] ArcCatalog[®] for metadata purposes, layers operations can also be accomplished such as zooming, panning, layer name changing, copying, removing, and exporting geostatistical layers to raster or vector formats.

5.15. KRIGING ARCSRIPTS[®]

Another possibility for accomplishing Kriging within ESRI[®] GIS concerns the Kriging Data Interpolation Extension implementation for ArcView[®] Spatial Analyst (<http://gis.esri.com/arcscripts>). For extension implementation within Windows[®] NT, three steps must be followed:

- 1) Kgdi.avx must be placed in the ArcView[®] Ext32 folder;
- 2) Avdlog.dll should be copied to the ArcView[®] Bin directory;
- 3) Avdlog.dat must exist within the ArcView[®] Lib folder.

According to its author, Marco Boeringa, the Kriging Interpolation²⁴³ option is available from the Analysis menu (see Figure 39). Some of the other features implemented are:

- OK comprises spherical, circular, exponential, Gaussian and linear variograms.
- UK includes linear and quadratic drift.
- The facility to set either a fixed or variable radius.
- The ability to export the variogram data to DBase[®] and text files during Kriging.

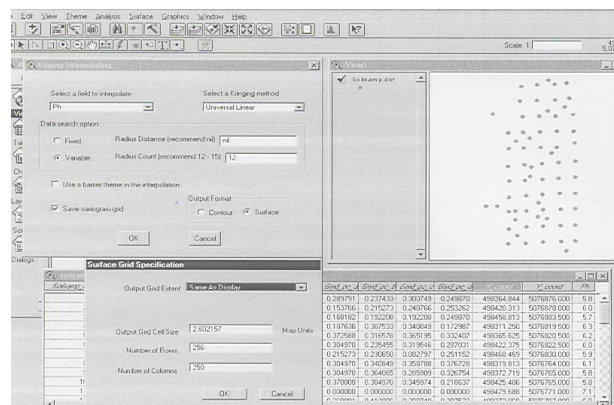


Figure 39 – The Kriging ArcScript[®] window.

5.16. SPACESTAT[®]

SpaceStat[®], a self-contained Windows[®] 2000 and Millennium[®] software package written in GAUSS[®] for lattice data, is organized into four modules. The data module allows merging files, to recoded variables²⁴⁴ and to create dummy categorical indicator variables for a given range. Note that the input files (contiguity, spatial weight and auxiliary) must be converted into GAUSS[®] binary

format for a post-processing operation. This includes principal component analysis.

The spatial tools module for spatial weight manipulation consists of:

- Creation of the **W** matrix²⁴⁵ based on the Euclidean and great circle distance²⁴⁶ with standardization and the capability of dissolving areas units;
- A characteristic report for the **W** matrix²⁴⁷;
- Information on the connectivity structure²⁴⁸;
- Construction of the contiguity weights using the rook/bishop/queen criterion (the Raster Wts option);
- The **W** eigenvalues²⁴⁹ and the trace index calculation²⁵⁰ [Anselin, 1992].

If the contiguity matrix only contains {0,1} information then SpaceStat[®] follows the GAL format (FMT files). Thus, the corresponding ASCII input file contains the number of observations and, for each observation, it lists its identifier, the number of contiguous observations and their respective identifiers. Regarding the weights of the sparse format, this input file contains the number of observations and, for each non-zero weight element, it lists the identifier for the row and column followed by the weighted value. Remark that the GAUSS[®] matrix format is unreadable by standard editors or word processors. The symmetry lacking due to the typographical errors can be checked. The contiguity structure between observations can also be recorded visually by inspecting a map or by running an AML[®] macro within ArcInfo[®]. Another possibility is to define the lower and upper bounds for a distance band definition.

The Explore unit for spatial association measures presents a diversity of spatial autocorrelation measures such as the Moran scatterplot, Moran I, the spatial correlogram, Gary C, global G(d) and local G_i by showing the ten highest and lowest values. The join-count statistics for spatial autocorrelation as an ESDA tool is also undertaken with the Data-Var Create-Create Dummy Variables option.

The generic OLS and robust OLS regression for non-homoskedasticity situations are computed within the spatial regress module. As expected, the R² and adjusted

R^2 indices, the t-Student for regression coefficients and the global F test are also reported. Although both OLS techniques lead to the same coefficients and prediction values, their variances are quite different if spatial autocorrelation among errors or heteroskedasticity exists. An alternative method to overcome heteroskedasticity is based on the Regress Heterosked Error²⁵¹ option, a GLS procedure. The following figure presents the principal techniques and spatial analysis offered by SpaceStat[®].

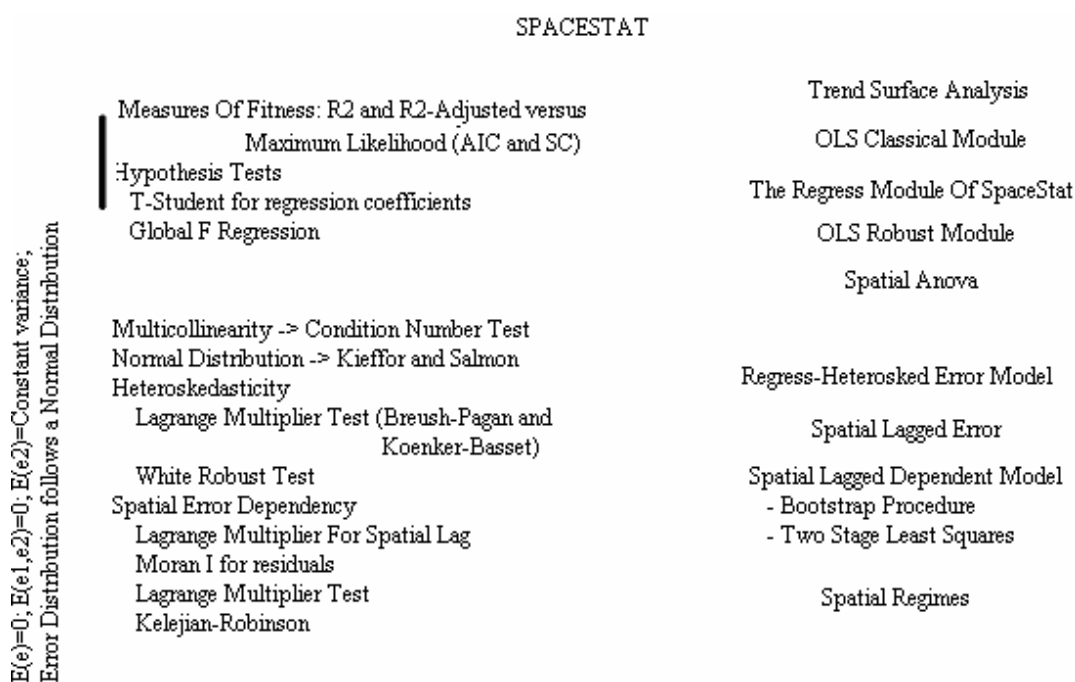


Figure 40 – The SpaceStat[®] Regress module framework.

Another version is the SpaceStat Extension[®] for ArcView[®] (see Figure 41). Some of the important developed features are the capability to compute centroid coordinates and to add them to the ArcView[®] table as X_Coord and Y_Coord variables, to export polygon boundary coordinates from an ArcView[®] shape file to an ASCII one²⁵², to present a quartile map for a selected variable including highlighted outliers, to display a bar chart map showing the observed/predicted values and to map the standard deviational residuals of spatial regression.

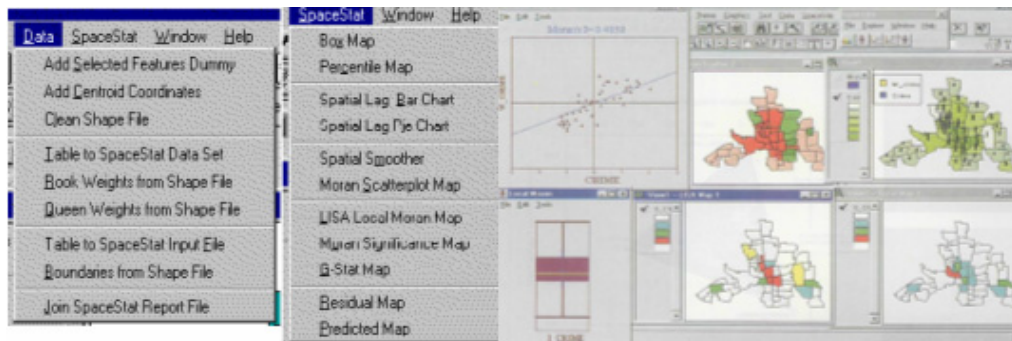


Figure 41 – The SpaceStat[®] Extension for ArcView[®].

Appendix A shows other software packages reviewed though not discussed in this chapter, including GridStat[®], GeoStat[®], GViz[®], EasyKrig[®], GeoR[®], MATLAB 1.1[®], SPlus-X[®], SADA[®], Isomap[®], S+Spatial[®], ISATIS[®], DACE[®], XGobi[®], CrimeStat[®], S-MATLab[®], SGS[®], SAS/STAT[®], Drill Guide[®] and Rooks[®]. The following table presents a feature summary of all the packages analyzed.

Software	Major Features to be Analyzed																
	Product		ESDA Features				Stochastic Interpolation							Operating Platform			
	Commercial Company	Researchers and Universities	Brushing and Linking	Nearest Neighbors Analysis	Conventional Moran scatterplot	Spatial Autoregressive (AR, SAR)	Variography and Kriging	Sill Rescaling	Measurement Error	Nugget-effect comparison on the fly	CK and Geostatistical Simulation	Deterministic Interpolation	e-Learning framework	Type Of Software ²⁵³	Windows [®]	Unix [®] or Macs [®]	Web Platform
Surface [®]		X					X					X		3		X	
VarioWin [®]		X					X							3	X		
GS+ [®]	X						X				X			3	X		
GeoMS [®]		X					X				X			3	X		
Regard [®]		X	X				X				X			3		X	
GMS [®]		X		X			X				X	X		3	X	X	
GSLib [®]		X					X				X			3	X	X	
Berkley Unix [®]		X					X							3		X	
Uncert [®]		X					X					X		3		X	
Geo-EAS [®]		X					X							3	X		
Surfer [®]	X			X			X					X		3	X		
EcoSSE [®]		X		X			X				X	X		3	X		

GeoPack [®]	X						X				X	X		3	X		
ArcGIS [®]	X		X	X			X		X		X	X		2	X		
ESRI Scripts [®]		X					X							2	X		
SpaceStat [®]		X		X	X	X								1	X		
GridStat [®]	X						X				X			3	X	X	
GeoStat [®]	X						X				X			3	X		
GViz [®]	X			X			X				X	X		3	X		
EasyKrig [®]		X					X					X		1	X	X	
GeoR [®]		X					X				X			1	X	X	
MATLab 1.1 [®]		X					X				X			1	X		
SPlus-X [®]		X		X			X				X			1	X	X	
SADA [®]		X		X			X					X		3	X		
Isomap [®]	X						X					X		3	X		
S+Spatial [®]	X			X		X	X				X	X		1, 2	X	X	
ISATIS [®]		X		X			X				X	X		3	X	X	
DACE [®]		X					X					X		1	X		
XGobi [®]		X	X				X							2		X	
CrimeStat [®]	X			X										3	X		
S-MATLab [®]		X				X						X		1	X		
SGS [®]		X					X				X	X		3		X	
SAS/STAT [®]	X						X							1	X	X	
Drill Guide [®]	X						X				X			3	X		
Rooks [®]		X			X						X			1 ²⁵⁴	X		

Table 17 – Analysis of geostatistical software.

5.17. CONCLUSIONS

According to the previous sixteen sections, Appendix A and Table 17, eighteen major inferences can be drawn:

- Commercial costs arise with some of them, like the ESRI[®] Geostatistical module, although available demos are available for the rest of the majority;
- No commercial software offers its computer code (which varies from C++[®] to Pascal[®], Fortran[®], C[®], Excel[®] macros, statistical scripts, Gauss[®] and VBasic[®]), a trend that is not followed by a few non-commercial options;
- No software establishes a close link between Kriging and the Moran scatterplot;

- No software concerns itself with the variability factor that underlies the Moran scatterplot;
- No statistical packages that cover spatial autocorrelation issues such as Rooks[®] or Spacestat[®] handle simple observations with no defined boundaries;
- Only SpaceStat[®] and S+Spatial[®] deal with spatial autoregression and autocorrelation issues;
- No software takes into consideration the variogram rescaling²⁵⁵ and the region confidence interval;
- All Kriging packages compute the OK variance, cross-validation procedure and classical statistics indices while the K-density and variable transformation functions are only available in a few ones such as SPlus-X[®], CrimeStat[®] or GS+[®];
- Measurement error is only considered by the ESRI[®] Geostatistical module;
- No software considers declustering based on the nearest neighborhood analysis;
- No software compares final estimations on the fly according to several nugget-effect solutions;
- No software presents an e-Learning framework for educational purposes;
- Interfaces range from a window and batch mode to command lines. In terms of operating systems, they vary from DOS[®] and Windows[®] to Unix[®] and Macs[®] platforms while only three, Regard[®], ArcGIS[®] and XGobi[®], allow the brush and direct map linkage features;
- Most software pursues the black-box strategy, like Variowin[®] for example, while others demand a complete understanding of geostatistics and computers, like GSlib[®];
- All software has a specific installation procedure, depending on the operating system;
- No software uses the browser interface;
- Most software has a specific input format although interoperability, particularly with ASCII files, can be achieved;
- No software considers the Web as a valuable spatial analysis platform.

6. Spatial Analysis: User Know-how

Chapter 6 is divided into four sections. The first section focuses on the theoretical academic subjects needed for spatial autocorrelation and Kriging issues. The results of the two surveys (Appendix B and G) testify to this situation. Section 6.2 accentuates the close relationship between statistical and geographical knowledge to achieve higher-level solutions, while the last section focuses on the new Web GUI role in implementing spatial analysis. It also establishes the main reason for the SAKWeb[®] choice of this new platform. The conclusion section presents a summary of the current chapter.

6.1. MODEST EXPERTISE IN SOPHISTICATED SPATIAL ANALYSIS

If the universities' core GIS curricula are not updated with new subjects and spatial applications then the danger of its misuse increases. As with SAP[®], Microsoft[®], AutoDesk[®], Adobe[®] and Oracle[®], certification in GIS should be mandatory. For instance, stochastic spatial analysis for geographers is essential as a new subject in both undergraduate and graduate studies. "Error propagation in the photogrammetry process is also a quite complex procedure" [Matos, 2001].

The aim of the Appendix B questionnaire is to confirm the following basic assertion: common Portuguese GIS users do not possess the skills necessary to make use of sophisticated tools such as Kriging, spatial autoregressive models, autocorrelation measures and simulation uncertainty, thus creating a limitation to the understanding of further explanations to their problem. Ninety-two people were surveyed at six different institutions:

- Nineteen employees from the Instituto Geográfico do Exercito (IGE);
- Seven final-year students from ISIG²⁵⁶;

- Five undergraduates from the Faculdade de Ciências da Universidade de Lisboa, FCUL²⁵⁷;
- Five researchers from the Centro Nacional de Informação Geográfica (CNIG);
- Eleven GIS master's degrees students from the Instituto Superior Técnico (IST);
- Fourteen persons from the Laboratório Novas Tecnologias, LabNT-ISEGI;
- Sixteen undergraduate students²⁵⁸ (4-year course) from ISEGI;
- Fifteen undergraduate students²⁵⁹ (2-year course) from ISEGI.

Their main fields of study are statistics, GIS, economics and computing, confirmed by the high percentage ($\geq 80\%$) covering computing (databases, programming languages), statistics (linear relationships, hypothesis test and descriptive analysis), GIS (mainly ESRI[®] and Intergraph[®]) and familiarity with statistical software (primarily SAS[®], SPSS[®] and Minitab[®]) (see Appendix C for further results). Note that the 'all methods' response (see Figure 42) was only given by six people: one from CNIG and LabNT-ISEGI and four from FCUL. Although network and spatial analysis modules are the most widely known, 50% of GIS users do not have any spatial interpolation knowledge:

- Five from IGE (26% of the total IGE interviewees);
- Two from CNIG (40%);
- Eight from LabNT-ISEGI (57%);
- Five from COCITE (71%);
- Twelve from ISEGI-T4 (85%);
- Fourteen from ISEGI-T2 (87%).

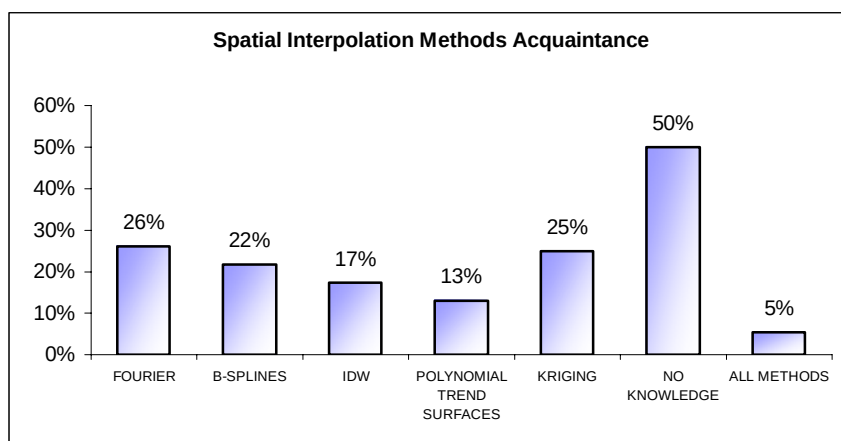


Figure 42 – Portuguese GIS users' acquaintance with spatial interpolation.

The same situation can be assumed regarding spatial autocorrelation measures. The First Law of Geography is only known by 5% of all respondents²⁶⁰, with 20% possessing theoretical knowledge²⁶¹ of some kind of spatial autocorrelation and clustering measures:

- Variogram cloud (18% of the total population);
- Moran I (8%);
- $G_i(d)$ and Geary C (5%).

According to Figure 43 and although the majority of the respondents are aware of network and spatial analysis modules²⁶², this trend does not include spatial autocorrelation and Kriging software. Indeed, implementing spatial analysis requires highly trained end-users to use it with a minimal degree of safety.

Following this perception, a second questionnaire (Appendix G) was applied among high-school teachers (Grades 7-12) after fifty hours of a GIS Prodep III course (March 2002) which was comprised of five ArcView[®] modules. Although these stochastic subjects are not covered by the Portuguese Ministry of Education syllabi, the interest in focusing on this population target arises from the fact that these teachers have an environmental or geography licenciatura degree (4-years first degree) taken in various prestigious universities, e.g., the Faculdade de Ciências da Universidade de Lisboa, Universidade Nova de Lisboa, Universidade do Porto, Universidade de Évora and Universidade do Minho in the latest 80s and

the 90s. It is expected that this homogeneous group can be representative of a larger one with no GIS knowledge. Also, after this educational module, it can be said that those teachers are able to be part of the common GIS user community. For these reasons, their responses in the second and third surveys carry valuable credibility.

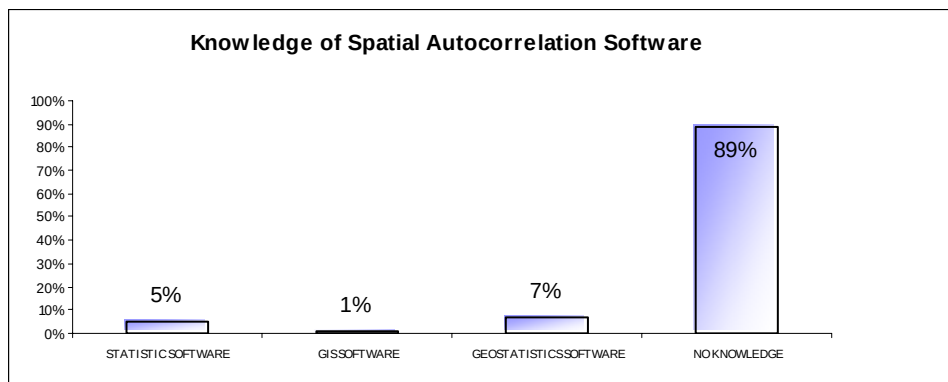


Figure 43 – Portuguese GIS users' knowledge of spatial autocorrelation software.

Among the group (14 females, 6 males), 36% hold some knowledge of relational databases and computer programming, 100% are familiar with descriptive statistics, 65% have already worked with GIS but none of them registers any knowledge of spatial interpolation and Tobler's Law. There are four major explanations²⁶³:

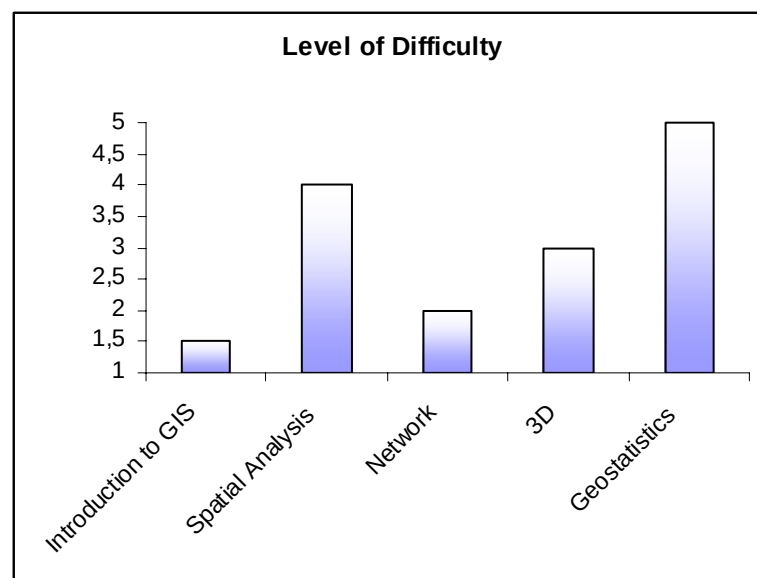
- The present teaching subjects do not involve these recent quantitative geography matters.
- Their licenciatura did not cover these subjects either. Constant training and updating are an absolute GIS requirement.
- There is a total lack of difficulty of access to GIS software in high-schools, colleges and universities. Moreover, the hardware requirements to run these types of packages can be high.
- Geostatistical foundations are statistically complex. According to Figure 44, the Geostatistical Analyst® module was the most difficult one to learn although it represents 44% of the total lesson time.

Predictably, 95% of the teachers consider acquiring Geostatistical Analyst® very difficult since it is quite impossible to copy it or to buy the licenses. All

respondents believe that the implementation of a spatial analysis Web platform is a wonderful strategy for obvious reasons:

- Direct and easy access from any computer worldwide;
- Easy use for training and educational purposes;
- No need to run the installation process;
- No minimum hardware requirements and no marginal software costs;
- No need for software management.

With regard to the major characteristics that Web geostatistical software should possess, the ones selected are user-friendly GUIs, hypermedia help, an e-Learning framework, validation of input variogram parameters including short messages for default values, inclusion of spatial autocorrelation measures as an ESDA tool, capability to read different input formats or a worldwide standard one and attractive layout maps with brushing, rotation and zoom operations. Surprisingly, the four most advanced users of this group request simulation techniques and comparison of Kriging results when sill, range, anisotropy, nugget-effect or variogram models change.



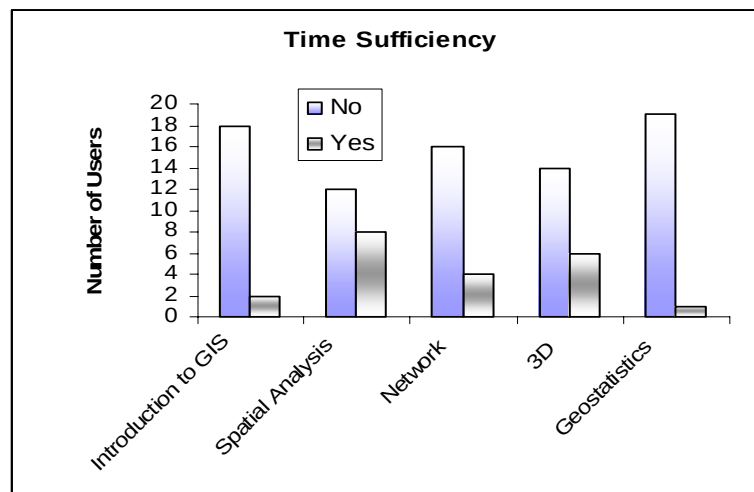


Figure 44 – The high-school teacher’s questionnaire: final answers after fifty hours of five ArcView® modules.

6.2. STATISTICAL AND GEOGRAPHICAL KNOWLEDGE: THE PERFECT MARRIAGE

Self-inferences of data choice, weight assignment, and geographical knowledge, inclusion of soft information, procedure selection and previous user experience are central factors in guiding the process of spatial analysis. For instance, “the long-range nickel variogram structure for the Jura region in Switzerland is closely related to the control asserted by the rock type, while the short-range one for cadmium suggests the local impact of man-made contamination” [Goovaerts, 1998]. Kriging does not consist of throwing spatial data into a black-box response program. Several decisions must be taken during the Kriging procedure like keeping strong outliers or choosing a variogram type. For instance, the organic dataset presented in Appendix N was given to the previous teachers of Survey H and different Kriging decisions emerged, as Table 18 shows.

Major Factor Differences among the 10 Kriging Predictions using Ecosse® Software
Original Data Standardization (Detrend -> 7 groups; No Transformation -> 3)
Kriging Model (Exponential -> 5; Linear -> 5)
Range of Influence (With linear models, the ranges varied between 110 and 150)
Anisotropy (Two of them included this feature)

Table 18– The Kriging strategies for the same Nebraska (USA) dataset.

If outliers are clustered (a minor increase in the local average) or distributed randomly (a local mean and variance increase) then interpolation becomes distinct. “Local regions of greater variability overestimate the nugget-effect that only represents particular areas. If spatial distribution is negatively skewed then local averages decrease with increased variance involving a major interpolation uncertainty” [Soares, 2000]. Also, stratification by region can be based on geographical knowledge with secondary categorical information such as land use, soil or geological maps because the appliance of Kriging should consider the landscape structure and the differences of soil recognition to avoid violation of the stationary assumption. Linear variogram use versus detrend polynomial approach or the threshold setup for IK can also turn out to be a complicated issue.

Hence, the justification of each decision with validation datasets is the best way to test and rank each decision. It is also important do not trust primary and secondary data sources that do not have a reputation for good quality to obtain final uncertainty estimates or to confirm the methodology requirements. Technology should not substitute education and training because decisions should often be based on the scientific problem. A detrend operation has the disadvantage of more parameters to estimate. However, humidity, rainfall and temperature have a major agriculture influence and it makes sense to detrend these factors before anyone wants to compare agricultural production over regions. It is not satisfactory to separate spatial analysis functions from the geographical understanding. “Neither statistical nor geographical knowledge are by themselves sufficient, both are necessary” [Openshaw, 1993, 1998].

“Another interpolation field is provided by geologists who play a key role in seismic research by correlating seismic attributes²⁶⁴ with reservoir properties²⁶⁵ for individual oil wells” [Hart, 1999]. Once the property has been defined for all hydrocarbon wells, the seismic attributes are extracted and correlated against each other in an effort to derive a statistically significant empirical relationship that will predict well properties as a function of seismic features. Quoting Schultz [1997], “the greater the number of characteristics derived, the greater the chances of

finding a statistically significant but spurious correlation”. Kalkomey [1999] strongly recommends, especially when there are few wells, that only those seismic attributes that have a physically justifiable relationship with the reservoir property be considered as candidates for predictors. “There must be a known and suspected physical mechanism underlying that statistical connection” [Hart, 1999]. Therefore, by integrating geological knowledge into research work of seismic attribute, more satisfactory study results can be achieved. Blind reliance on statistics is not a good practice.

Isaaks and Srivastava [1989] raise the same question regarding the definition of a single search strategy for an entire area under study. Both authors report a situation of lead concentration enclosed by a smelter whose major density varies between 59 and 694 ppm. However, the inclusion of a 10400 ppm sample due to leakage from discarded car batteries depends on whether the study goal involves all possible sources or not. Deciding which samples are relevant for the estimation of a particular point can be more important than the choice of an estimation method.

The secret is some kind of mix and match strategy in which both the analysis machine and the analyst concentrate on doing what each is best at. “The effectiveness of spatial analysis requires an intelligent user, not just a powerful computer” [Longley *et al.*, 2001]. Although COTS solutions (see Section 5.15 as an example) are regarded as a full black-box, this should be interpreted from the viewpoint of computer knowledge because geostatisticians should not be concerned about input-output data flows, intermediate calculus, the choice of programming language or other technical matters.

6.3. SPATIAL AUTOCORRELATION AND KRIGING: THE WEB SOLUTION

“We usually tend to overestimate progress in the next 2-3 years but underestimate changes over the next 10 years or so” [Gates, 1995]. It is commonly accepted that GIS is an integrated technology that has resulted from the fusion of various

systems²⁶⁶, although the vision of the future foresees a new computer and statistical conceptions moving from a cartography view to a modeling of virtual reality. At first, spatial analysis moved from a classical exploratory and inferential statistical framework to a spatial analysis methodology embracing non-spatial data only (Phase 0). Overlay, networking, measuring, raster and vectorial buffering, raster spreading, spatial search or contouring operations were, quite naturally, the next step in a common sense spatial approach (analytical cartography analysis). In its second phase, in the new century, advanced spatial analysis is in demand from new quarters, such as MAUP, interpolation, extrapolation, stochastic simulation and spatial patterns search where the use of both topology and non-spatial information is mandatory. Spatial data is special data. Without it, spatial analysis loses its identity.

Thus, should any GIS contain a comprehensive statistical and image processing toolbox or should any GIS simply act as a spatial data organizer, data viewer and file server that interfaces with independent COTS solutions to carry out specialized analysis? According to Unwin [1998] and Anselin [1998], spatial analysis may well be best accomplished outside of a GIS but with relatively seamless data transfer from the statistical to the geographical software, since spatial analysis requires a highly flexible and interactive computing environment. Subroutine libraries, command sequences or menus are the conventional answers to the linkage itself. Olivier [1996] corroborates this design stating that GIS needs to have explicit links to other software such as statistical packages and other spatial analytical programs. A good example is presented by Agterberg, Bonham-Carter and Wright [1992] in their gold mineralization study at Nova Scotia, Canada, using multiple regression analysis and conditional probability modeling. According to these authors, spatial data integration is greatly facilitated using GIS in association with other statistical applications. It seems that geographical analysis can be undertaken without GIS aid.

The viewpoint of Griffith [1993, 1996] is to convert spatial statistical techniques into GIS toolbox functions included in spatial modules, parallel to TIN or network analysis. Yet, his spatial autocorrelation and autoregressive solutions presented

three years latter were all programmed within SAS[®]. According to Figure 45, the dynamic linkage between specialized spatial autocorrelation software coupled within GIS is a non-solution for Portuguese GIS users. For Openshaw [1994, 1998], “standalone packages to investigate the leukemia cluster should not be incorporated into a standard GIS tool kit for three reasons”:

- First, this method has been developed in a quite specific context and may lack the generality that is implied in the idea of becoming part of any module.
- Second, vendor developers have no knowledge of this issue.
- Third, GIS users in both industry and academia do not know how to use them. It seems the real world is driven by market concerns while spatial research tools are limited by market size.

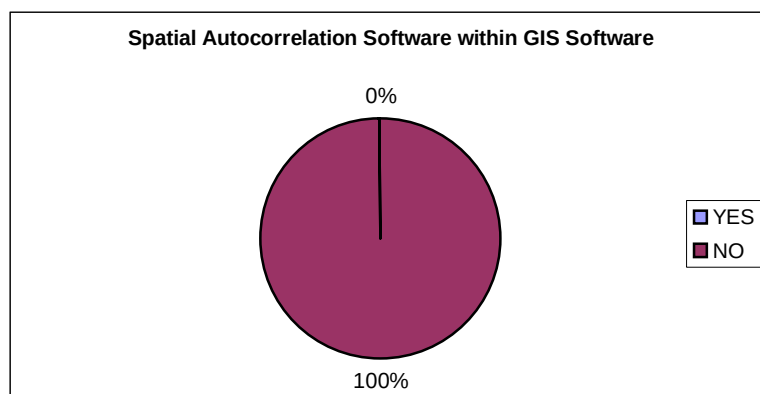


Figure 45 – User agreement (%) concerning the linkage of spatial autocorrelation and GIS software.

A careful review of all pros and cons of the three implementations (see Sections 4.1, 4.2 and 4.3 for further discussion) does not give much of a clear trend for the near future and, in the end, the future is controlled by the past. Taking into consideration the previous reference to the black-box, the solution starts at the end-user position. “Because of the object embedding capability of Windows[®], Linux[®] and Macintosh[®], the potential of computer languages of today is overlapped quite often” [Trezentos, Cardoso, 2000], leading to the possibility, for instance, of assessing regression parameters or visualizing the same plots and maps with different computer codes. “All visual languages have magnificent graphics features and great custodial database access” [Vilar, 2001]. In fact, the recent trend with

GIS and other software is becoming a reality: Seagate Crystal Reports[®] within ArcView[®] and MapInfo[®] or VBasic[®] and Access[®] incorporated into AutoDesk[®] and ArcInfo[®] are good examples. Although certain tools fit best for a specific goal, programmers always use their background and field of expertise in their creation process and, thus, reflect on several types of implementation.

As the new academic methods for spatial analysis emerge, more programming is needed. The ranking of the three kinds of implementation presented in Chapter 4 for new spatial analysis tools is quite unfeasible. Yet, in terms of speed and performance, executable languages²⁶⁷ are certainly the programmers' best choice especially when integrated with the standards of Microsoft COM[®] and its .NET successor and Javasoft's Java Beans[®], SOAP[®] and CORBA[®]. From the user's viewpoint, this fuzzy trend becomes clear, too. The aims for non-expert and expert users are different: the former are looking for simple descriptive statistical tools and data property summarization while the latter requires robust model fitting tools, an inference framework, spatial autoregression and hypothesis testing. "A GIS for utility applications also requires network processing operations while geological ones demand 3D views" [Worboys, 1996]. Thus, the same GUI should offer various possibilities for a wide range of users.

Software should encourage scientific computing visualization where statistical views are interpreted, validated and explored in a dynamic graphic map including zoom, pan, brushing and 3D simulation. Separately, a statistical view has a significant value but this value is enormously enriched when views are linked together emphasizing computation of interaction between human cognition and dynamic statistical graphics. Quoting Anselin [1998], "the state-of-the-art implementation consists of dynamic linked windows in which the user is able to delete points, highlight data subsets, establish links between data points in different graphs and rotate, cut through and project high-dimensional data" (see Section 5.5 for a further review). Still, these graph operations on the Web are still quite difficult to achieve although some of these techniques have been implemented within SAKWeb[®]. Another Web complexity is access to topology information by independent software (see Section 4.4 for further discussion).

The environment should be such that the individual stages retain their distinct identity and are not embraced in a completely automatic procedure. Users must be aware of decisions at each stage of being confronted with an explicit interface. As a result, the stages have to respect a linear sequence where users begin at the beginning and end at the end, progressing from left to right across menu choices. If any stage is inadvertently bypassed, the user is warned to restart and to redefine the operation that has been missed with on-line help. SAKWeb[®] pursues this technique consistently.

Therefore, what the users appreciate and interact with is the GUI²⁶⁸, because they do not care about the technical structure, as long as the results are trustworthy, prompt and compatible with their operating system and hardware. Users want intuitive and easy-to-use software in order to give immediate results without having to read pages of documentation. The standard Web browser, irrespective of the background computer code adopted, fulfills this strategy quite well. It is cost free and already provided in any operating system. In other words, does the user need to see a Word[®] document? It can be viewed in a browser. Does the user need to work on an Excel[®] spreadsheet? It can be opened in the browser. Does the user need to find local, network and Internet files? The browser can search for them. Ultimately, the user will work with all available software in the same way, regardless of the location of the data or its purpose:

- In the beginning, the trend was to find the data with the browser;
- Today, download the software and installed it;
- In the near future, just run the programs from the Internet.

This represents a significant advance in the user interface because users will no longer have to worry about the software location and the technical knowledge necessary to connect to the data. Even today, the keyboards available on the market hold special keys for the browser such as home, search, back, mail and refresh. The standard Web browser is the future interface and the one adopted by SAKWeb[®].

The solution relies on an enhanced standard front-end acting as a protected wall against the computer code²⁶⁹ in an embeddable, extendable and reusable development. The capability to undercover technical implementation to the final user via WWW is essential. After struggling for years for digital information, spatial analysis needs to concentrate on what the information means, sharing it through the W3 and distributed architectures. The shift from thick servers to IIS[®] and Apache[®] intranets and Internet wireless transmission via PDAs, laptop or smart phones, for instance, with authorization, updates, backups, error logs, authentication features and spatial analysis capabilities is already the way.

Software independency ('write once, read anywhere'), XML applications and WAP technology are important. In effect, the wireless Web is developing furiously nowadays. Although Web-enabled wireless devices will not replace the PC, it will enable millions of people to access the Internet while on the go, such as:

- Wireless Web can provide access to employees to access information wherever they are, and make decisions instantly without being tethered to a desk or computer.
- Drivers could use this capability to obtain local weather data and local traffic information along with alternate route suggestion and descriptions of nearby restaurants.

Therefore, improving data exploration by applying the available tools of spatial analysis and back-office technology integration with the Internet (mobile or not) is essential. Technically, this challenge relies on factors such as standardized data models, mobile computers²⁷⁰, full acceptance of open networks by an HTTP memoryless protocol²⁷¹, full independence of computer languages from platforms, nanotechnology and, particularly, an increasing in bandwidth²⁷² for W3 services²⁷³. Better quality service transmission by improving protection, priority, link recovery from failures, information overload, errors, return delays and impedances should be network's first concern.

The second problem regards the use of a single and unified wireless Web standard because, at present, there are two main ones:

- Wireless Application Protocol (WAP) that uses Wireless Markup Language (WML). When the micro-browser sends a Web request, the WAP gateway retrieves the information from an Internet server in either standard HTML format or WML²⁷⁴.
- I-mode, a rival standard developed by Japan's NTT DoCoMo mobile phone network, uses compact HTML to deliver content.

“The third issue is the low data transfer speed on existing wireless networks, ranging from 9.6 to 14.4 Kbps” [Laudon, Laudon, 2002]. Hence, the present Web content for wireless devices is in the form of text with very little graphics and a main difficulty for GIS, in particular.

The respectability of the academic research with regard to implementation should be a reality where the results are displayed in appealing pictures while the arduous work takes place beyond the stage in a W3 wireless environment. Whether this belief prevails or not, it is better at all costs to avoid a spatial tool with all theoretical spatial routines because of the unbounded nature of spatial analysis²⁷⁵. It is better to have a few methods that work reliably. Most definitely, this emerging digital world characterized by the uncoordinated proliferation of datasets and software across the ‘World Wide West’ does not help a peaceful and ordered scheme.

At the opening of the 19th Annual Conference of ESRI[®] in 1999 in San Diego, CA, the president Jack Dangermond refers to the geography channel like the weather channel where all monitoring data and all analyses will be reported daily [ArcNews, 2000]. Only time will tell. After all, there are few regions that have not been touched by spatial analysis. “Spatial prediction is just as important as temporal prediction because people living in cities and rural districts have the same right to know how little or how much their water or their air is polluted” [Cressie, 1993]. From a positive perspective, the recent technological trend is moving from centralization towards distributed systems including interoperability as a requirement. “The AlintaGas project in Australia is a full example where the SAP[®] Plant Maintenance module interfaces with GNIS” [ArcUser, 2000].

Spatial analysis within the Web environment²⁷⁶ is the step now necessary for the geographical channel proposed by Dangermond. Other topics introduced in the course of this dissertation substantiate this Web trend:

- The present catastrophic situation of advanced spatial analysis techniques for GIS users, confirmed by the first questionnaire (Appendix B), can be modified with W3 access in an e-Learning context, to increase the knowledge of true spatial analysis. As confirmed in Sections 4.5 and 6.1, researchers suggest that this methodology is a promising one for the near future.
- The capability of reading several data formats and focusing on the process in a straight forward and reliable way is a major key to success. The greater popularity of Microsoft[®] products like MS-Excel[®] relies on the ease of using them in spite of the faster Linux[®] performance²⁷⁷ [Negreiros, 2000].
- Forgetting older markup languages²⁷⁸ and protocols²⁷⁹ and allowing the educational community to contribute to the implementation of the research is significant. Materializing the findings of this dissertation (see Section 7.2 for further discussion) on certain issues introduced in Section 1.1 to the real world (in order to create an impact on how users do spatial analysis) was not forgotten either.
- As Table 17 shows, the Web solution was never undertaken. The close linkage between the Moran scatterplot and Kriging was never considered either (see Section 5.18 for further discussion).
- In the end, the common user will inherit all WWW intrinsic features such as installation procedures they can forget about, the absence of site management obligations, personal customization and lower marginal costs as the second survey (Appendix G) reveals. This questionnaire also quite regularly brings to light the traditional users' difficulty in accessing and setting up spatial analysis software, a situation easily overcome by the Web technology.

6.4. CONCLUSIONS

This chapter works as a bridge between the previous chapters and the following one: SAKWeb[®] developments and design. After the review of spatial autocorrelation and Kriging theory in Chapters 2 and 3, Chapter 4 dealt with software implementation. Chapter 5 evaluates some independent and statistical packages available at present. This information is complemented by Appendix A and Table 17.

Section 6.1 emphasizes the need to introduce the components of the missing data issue to GIS users and the need to update geography and GIS curricula on stochastic issues. The section that follows it states that statistical and geographical knowledge are both necessary to obtain high-level solutions. It is also crucial to say that any spatial problem may result in different Kriging results while the best solution relies on the best justification model. The last section validates the new Web solution as plausible (the first of the key specifications for SAKWeb[®]), highlighting, for instance, the browser interface, the brushing feature, the hypermedia help, the interoperability of data formats and the process view instead of a simple function.

7.SAKWeb[©]

“The death of specialized software is often cited as one of the main reasons for the lack of acceptance of spatial data analysis by empirical analysts” [Anselin, 1992]. Although this situation has much improved in the last ten years, SAKWeb[©] is the first Web software in operation that provides access to a wider audience because of the lacking of geostatistical knowledge revealed on the first and second surveys (see Section 6.1 for a further review).

SAKWeb[©] version 2.0 is not a comprehensive statistical package in the traditional of solving everyone’s problems. Written for an IIS[®] environment, it was developed with the philosophy that spatial autocorrelation and Kriging software is needed as a learning tool by individuals with limited geostatistical knowledge. Yet, it can also satisfy the needs of individuals with more training. SAKWeb[©] deals with Kriging interpolation in conjunction with spatial association measures in a Web continuum process instead of a loose local spatial function. From this view point, an element of its originality and innovation can, thus, be appreciated.

To make this project come to life, several W3 technologies were used. ASP[®], PHP[®] and Dreamweaver[®] are the main SAKWeb[©] development set in a WWW application context. Data Access Components[®], WebChart[®] from Component One[®], FrontPage[®] Server Extensions Service Release, Flash[®], VBScript[®] and JavaScript[®] are some of the other components required. At present, there are two Internet techniques to build dynamic applications: client-side and server-side. The aim of any Web server is to publish HTML contents in order to reply to any request through port 80 (443, if SSL is used). This type of solution, where the HTML code is interpreted by the browser, is the client-side strategy (see Figure 46) that includes JavaScript[®], Java Applets[®], VBScript[®] and ActiveX[®].

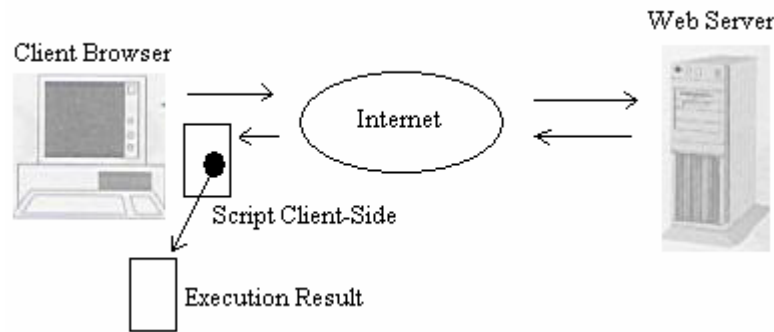


Figure 46 – The client-side technology.

If the contents are dynamically generated by the Web server through the execution of local programs and scripts then the server-side is chosen (see Figure 47). In this domain, the use of the Common Gateway Interface (mainly written on Perl®, script shell or C® for Unix® platforms) was the first standard communication between the Web server and the programs running on the server-side. For instance, the most widely known CGI program was the linkage between the submission of an HTML form (using the Get and Post methods) and a file or database server to save the results. Note that CGI is not a language. Nor is it a program. CGI is a process, an interface which provides well-defined rules for creating partnerships. The benefit is that if everyone respects the rules of the interface then everyone can talk to each other. “One alternative, especially for Netscape® servers (iPlanet®), is Livewire®, a development environment that allows the use of JavaScript® on the server-side” [Coelho, 2002].

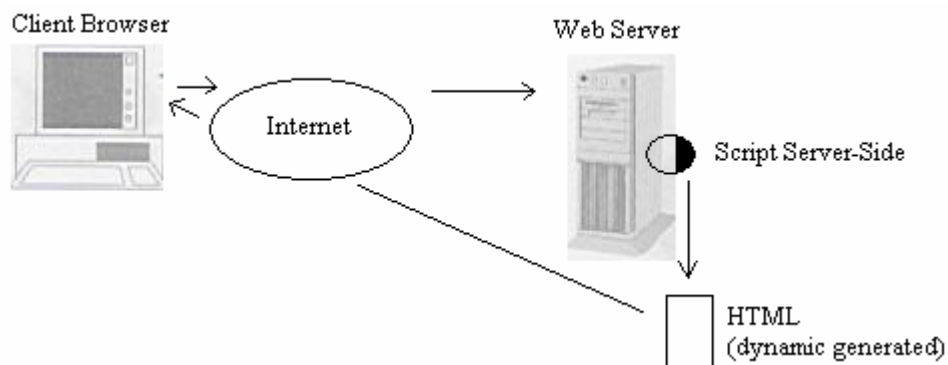


Figure 47 – The server-side technology.

In this new phase, the present development is the server-side script languages that allows the generation of dynamic contents such as PHP®, Java Server Pages

(JSP[®]), Java Servlets^{280®} and ASPs[®]. As expected, the use of this server-side scripting has its pros and cons, e.g.:

- Since the client has no processing to do, the loading of any page requires less time for the browser user. Yet, the responsibility of the Web server to generate the HTML must be taken into account and, hence, robust hardware for the server is mandatory.
- Since the client only receives HTML, there are no special requirements and plugins for the browser.
- The server script code cannot be viewed by the user, a situation not considered by client-side technologies.
- More complex solutions can be built on account of the access to ActiveX[®] components and other internal objects.

Nevertheless, for the majority of Web applications, the server-side versus client-side issue does not make much sense because both methods have specific and defined goals. The client-side components are used to validate and process local pages while the server-side ones are used to produce dynamic contents. Recapping Figure 6 of subsection 1.5.4, SAKWeb[®] version 2.0 follows this trend because both technologies are applied extensively with these two purposes.

Accordingly, Section 7.1 briefly reviews these components, without forgetting their use within SAKWeb[®]. Section 7.2 focuses on major innovations regarding spatial autocorrelation and OK (see again Figure 7 of subsection 1.5.4):

- The Moran location scatterplot;
- The Moran variance scatterplot;
- The sill rescaling of the OK variance and interpolation confidence interval, both global and local;
- The four nugget-effect strategies;

Section 7.3 contains a description of SAKWeb[®] functionalities and interfaces while Section 7.4 centers on the third questionnaire whose goal was to evaluate SAKWeb[®] version 2.0 with the co-operation of the twenty respondents of the second survey. The concluding section summarizes the main inferences in this chapter.

7.1. SAKWEB[®] TECHNOLOGIES

7.1.1. Internet Information Server[®]

During the last decade, there have been great developments with the Internet. To shop for groceries, to transfer virtual money or to exchange bonds are some of the possibilities for each a client-server framework has been mandatory. “This model is divided into three layers²⁸¹: interface, application and data access” [Tereso, 2000]. Compared with stand-alone and two-tier structures, the Web application emerges as a new paradigm by integrating complex and boundless technologies such as CGI, ISAPI²⁸², Coldfusion^{®283}, PHP[®], JSP[®] and ASP[®]s.

“At the heart of any Web application, a Web server is a requirement and this includes FTP²⁸⁴, IRC²⁸⁵, mail, news, telnet and proxy servers” [Baptista, 2002]. IIS[®] is SAKWeb[®] choice. First and foremost, IIS[®] is a protocol server that implements the most common Internet protocols including HTTP²⁸⁶, FTP, NNTP²⁸⁷ and SMTP²⁸⁸. It also offers standard APIs for extending and customizing the server's capabilities. “Its management console can be found in the Administration Tools” [Loureiro, 2000]. Regarding its architecture, it is based on WAM²⁸⁹ technology, a COM²⁹⁰ with ISAPI functionality. When the IIS[®] receives an HTTP request, its job is to return the request resource such as a static or dynamic page. Depending on the resource extension name, it loads the appropriate ISAPI extension, forwards the request, receives a return call and returns that data stream to the requesting user's browser. “Specifically, when an ASP[®] page request is received, IIS[®] forwards the request to the ASP.DLL²⁹¹, processes the page and sends the return data to the inetinfo.exe file which simply returns the data to the user” [Tereso, 2000].

With regard to the SAKWeb[®] site setup, the next four items depicts its major steps:

- The home directory is c:\inetpub\wwwroot.
- An IP and port address is required with a DNS²⁹² name²⁹³.
- Login.asp is the default home page.
- The ./executables folder holds execute permission while the ./data folder retains write authorization. All other folders hold the default right permissions for read and run scripts.

Implementing security and privacy at IIS[®] can never be overstated. Basically, this Web server provides four types of authentication:

- Basic – The logon name and password specified by the user are used for authentication purposes before processing a request.
- Digest – This sends the user's credentials for validation in an encrypted form.
- Integrated windows – They use the account credentials of the Windows XP[®] domain to authenticate the user.
- Anonymous – This type uses a built-in user account to request resources from the Web server. Note that this option must be disabled if anyone of the previous three methods are used (the SAKWeb[®] choice).

To secure a virtual directory, the Web manager needs to invoke the Internet Services Manager and, after the directory in question is selected, the Security tab allows permissions to specific computers to be granted or restricted. Another option is the Directory tab that allows selection of the appropriate permissions for the Web application such as Read, Write, Script source access, Directory browsing and Log visits. As to this last issue, IIS[®] provides a number of formats for the Web server log files:

- Microsoft[®] IIS[®] log – This records basic information about Web requests, including IP address, date, time and the number of bytes exchanged while processing the request (the SAKWeb[®] choice);
- NCSA common log – This option records fewer details of requests than the previous one;

- W3C extend log – This format records the most detailed description available of a Web request.

In addition to IIS®, SAKWeb® implements its own authentication mechanism based on the default authentication form (see Figure 48). If the user is not authorized then the Web application directs the user to a pre-defined logon refused page. Otherwise, he or she is redirected to the main Web page.

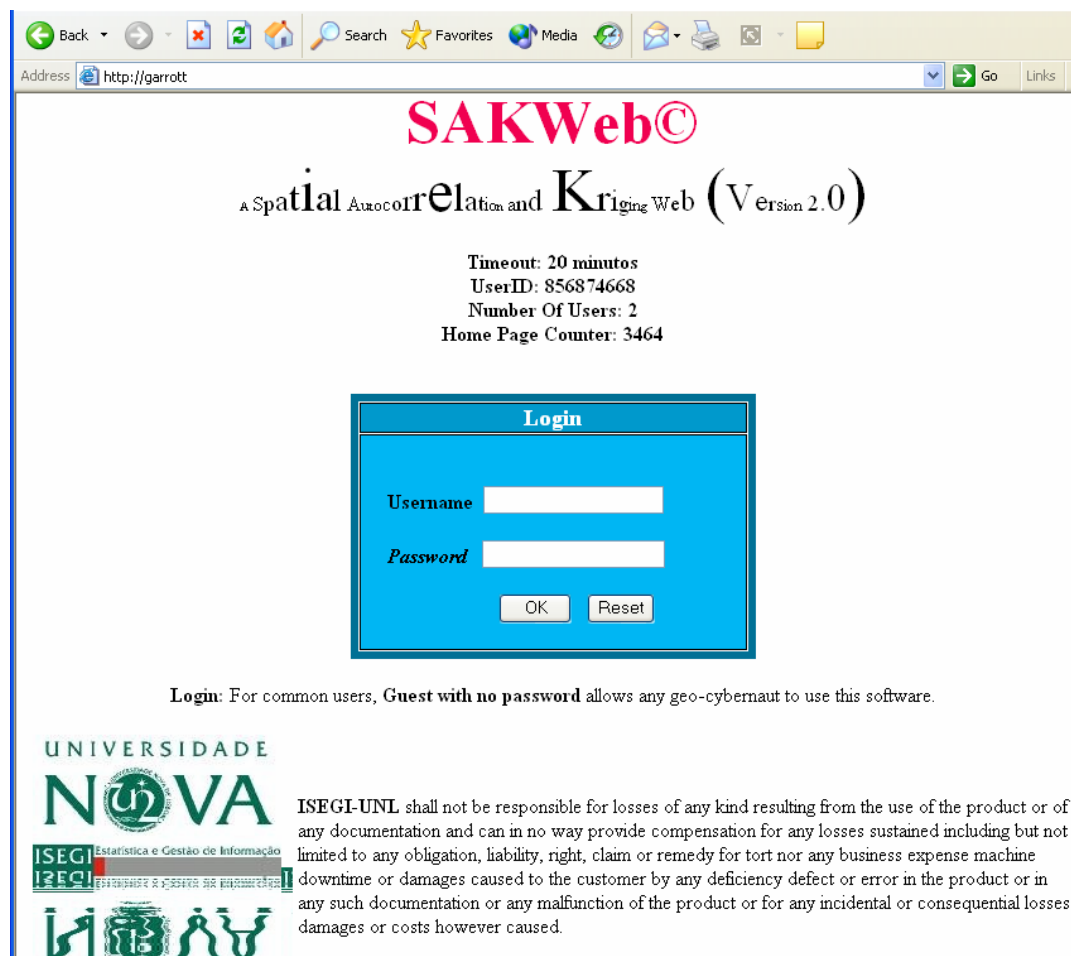


Figure 48 – The home and authentication page of SAKWeb® that includes the keywords meta tag for search engines robots.

7.1.2. From CGI to ASP[®] and PHP[®]

In the dim past, Web developers²⁹⁴ turned to CGIs and the Perl[®] language to introduce some sort of interaction. “Since CGI programs are executable applications, it is the equivalent of letting the world run a program on the system, which is not the safest thing to do” [Negreiros, Ferreira, 1999]. Accordingly, certain security precautions are needed, for example that CGIs need to reside in a special directory²⁹⁵. “This directory is usually under direct control of the Webmaster, prohibiting the average user from creating them” [McBride, 1998].

With CGI applications, the client request is first sent to the server over the Internet via HTTP. The server receives it, determines which program needs to be run and writes the information to an input file. Then, the server launches the CGI program. This program reads the input file, writes the output to another one and terminates itself. “The Web server process, which has been waiting, reads the output file and finally sends it back to the client via HTTP” [Loureiro, 2000].

“CGI applications are easy to write but scale very poorly within Windows[®]“ [Stephens, 2001]. Because a separate process is spawned for each client request, hundreds of clients create hundreds of instances of the CGI, each requiring its own memory space and system resources. “This is not such a bad thing on Unix[®], which is designed to handle multiple processes with very little overhead” [Morgan, 1996]. However, Windows^{®296} expends more system resources when creating and destroying application instances. One way to get around this problem is to embed this processing into the Web server itself. Rather than rely on another layer, most Web servers provide several ways of extending the Web server itself by adding logical and processing power. The earliest technology to take advantage of this idea was SSI²⁹⁷. The concept of SSI is simple (see Figure 49). Special tags that are inserted into the HTML document are understood by the Web server and can be translated on the fly by the Web server as the HTML document passes through on its way to the browser.

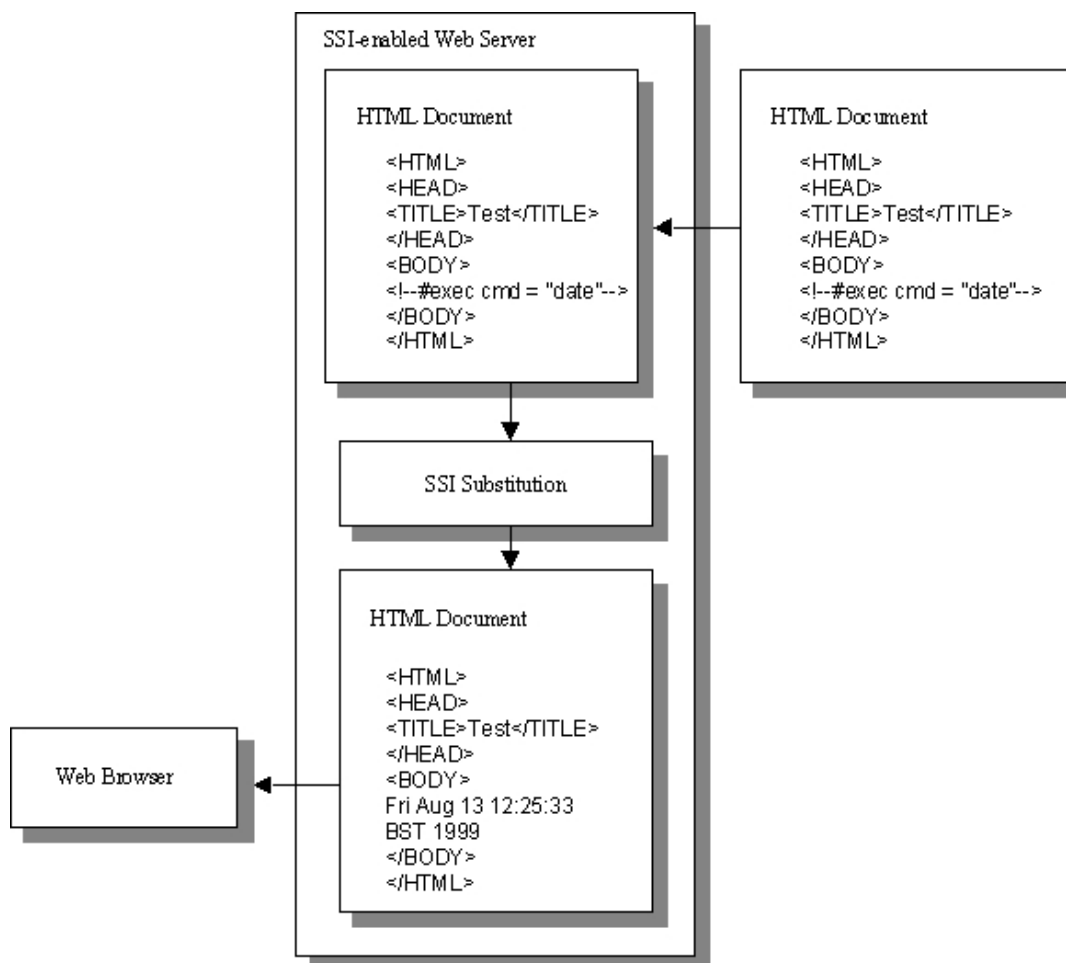


Figure 49 – An example of an SSI that returns the date and time of the Web server.

“ISAPI²⁹⁸ came and went primarily because it required more knowledge to create a dynamic filter than Web programmers were prepared to learn” [Pandey, 2002]. Finally, Web scripting languages emerged including ASP[®]s, a server-side scripting technology for building Web pages that are both dynamic and interactive.

“This new technology shifts the focus from desktop to distributed computing where a number of applications are integrated to provide a mechanism solution” [Marques, Mendes, Henriques, Sena, 2001]. Thus, resources at remote locations can be integrated with servers such as Exchange Server[®], Internet Security[®] and SQL Server[®]. In addition, PHP[®] and ASP[®] offers an open server-application environment where the user combines HTML, server-side scripts and reusable COM server components to create a dynamic and powerful Web-based business

solution. Hence, after the server-side script runs, the results are returned to the client browser in the form of a standard HTML document.

This server-client framework has some remarkable features:

- The intrinsic Response and Request²⁹⁹, Server³⁰⁰, Scripting³⁰¹, Application and Session³⁰² objects allow dynamic information access to the ASP[®] Object Model hierarchy.
- The capability of including standard scripting languages like VBScript[®], JScript[®] and JavaScript[®] and the potential to incorporate standard COM objects (.dll or .exe files) from third party companies (Presentation³⁰³, Business³⁰⁴, Data³⁰⁵ and Utility³⁰⁶ Components) are available.
- The potential to access internal objects such as the Dictionary³⁰⁷ also exists.

“A Web application is, therefore, a collection of ASP[®] pages, server components and a Website where the distinction among the applications is made by the root directory within the site” [Boavida, Monteiro, 2001]. All content within this directory structure is considered part of the scope of the same application. Furthermore, each application has its own set of variables and attributes that define its current state and these are maintained throughout the application lifetime³⁰⁸. “Another primary concern with Web based applications is that HTTP has no memory and retains no information from one client request to the next” [Coelho, 2000]. ASP[®] gets around this issue by using Application and Session objects to store information during a user’s session. Also, both the application and session are initialized and destroyed by the global.asa file found in the root directory of any Web application. This global.asa is an optional file that contains declarations of objects, variables, events, library declarations and the #include directive.

The Application object holds information that will be used by many pages as a database connection. This means that any user can access that information from any page. It also means that any user can change that information. Thus, the user must protect it with the application.lock method before any change occur. When an application is locked by one user, no other users can change the application

variables. As expected, the user can and must unlock a variable with the application.unlock method.

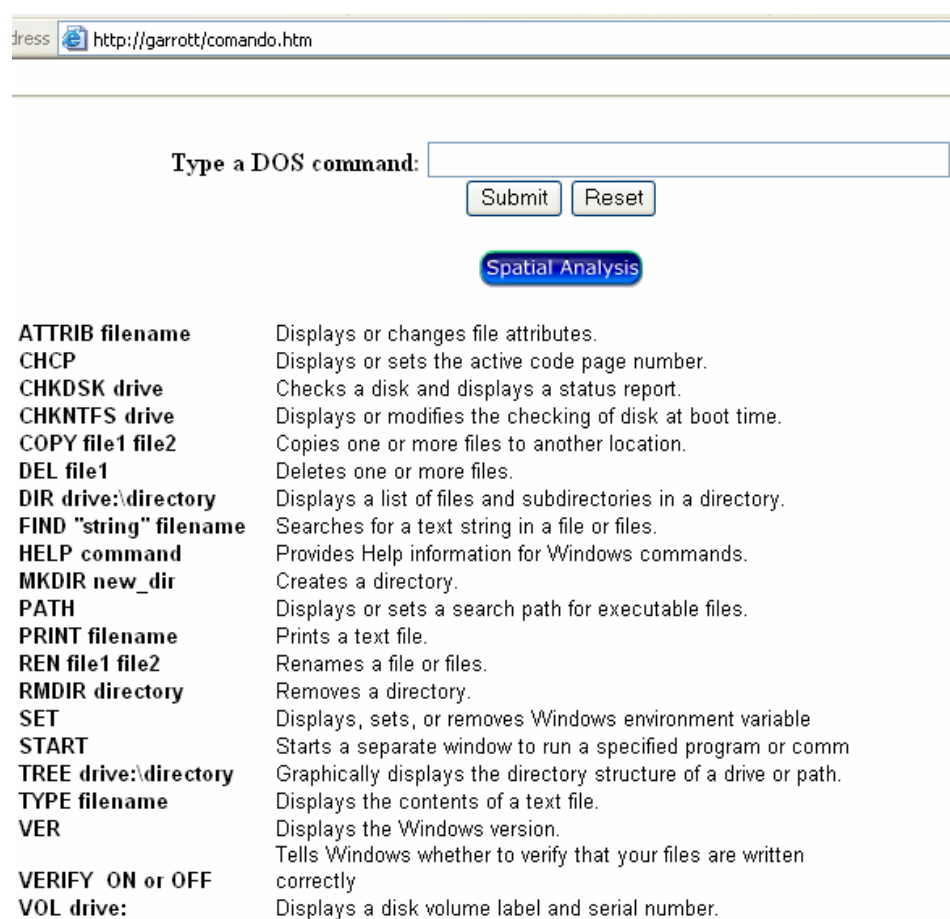
The main SAKWeb[®] code was achieved with this technology in accordance with to the following folder structure:

- /Chat – This directory contains the files³⁰⁹ that allow users to meet in a virtual room. It uses cookies and twelve session's variables although its home is an HTML file.
- /Common – This includes three important files: 1) Lib_data_to_array.asp reads Excel[®] data to an internal dynamic array; 2) Menudynamic.asp redirects the user to the home page or to the main menu; 3) Menudynamic.htm includes VBScript[®] code for the call of a third party component³¹⁰.
- /Content – This folder consists of several types of files whose aims are Excel[®] management, variogram fitness procedure³¹¹, Flash[®] executables³¹², univariate, indicator mapping for samples, bivariate data posting, image mapping, local interactive statistics and nearest neighborhood analysis.
- /SK – This folder contains the files that are responsible for Simple Kriging.
- /Data – This one holds all background text files that are created during the SAKWeb[®] interpolation process.
- /Executable – This contains gamv.for, varmap.for, kt3d.for and kb2d.for³¹³ files and their GSLib[®] executable versions³¹⁴.
- /Kriging – The 25 files existing in it are responsible for the control of OK nugget-effect approaches, 3D-2D profile and cross and error validation.
- /Moran – This folder includes the seven ASP[®] files responsible for dealing with the Moran I correlogram and SAKWeb[®] Moran scatterplot.
- /Olectra - The files 3dconst.inc, 3dgentag.inc, 2dconst.inc, binstream.asp, error.inc, color.inc and 2dgentag.inc are needed for the Web charts presented in this software. This folder also includes the 2D and 3D Olectra .cab, .inf and .lpk files³¹⁵.
- /HelpSAKWeb – This directory is responsible for handling all images, Flash[®] and Java[®] Applets files need it for SAKWeb[®] Help, SAKWeb[®] News, SAKWeb[®] email and other e-Learning features.

An alternative Web development, SAKWeb[®] includes two tools that uses PHP[®] server-side technology:

- The DOS Commands allows the user entering any DOS[®] instruction while the results are shown in the browser itself.

```
<?php
$line1 = exec(escapeshellcmd($_POST['comando']), $output, $error);
while (list($line) = each($output)) {
    echo $line, "<br>"; }
if ($error) {
    echo "Error Code: $error<br>\n";
    exit;}
?>
```



ATTRIB filename	Displays or changes file attributes.
CHCP	Displays or sets the active code page number.
CHKDSK drive	Checks a disk and displays a status report.
CHKNTFS drive	Displays or modifies the checking of disk at boot time.
COPY file1 file2	Copies one or more files to another location.
DEL file1	Deletes one or more files.
DIR drive:\directory	Displays a list of files and subdirectories in a directory.
FIND "string" filename	Searches for a text string in a file or files.
HELP command	Provides Help information for Windows commands.
MKDIR new_dir	Creates a directory.
PATH	Displays or sets a search path for executable files.
PRINT filename	Prints a text file.
REN file1 file2	Renames a file or files.
RMDIR directory	Removes a directory.
SET	Displays, sets, or removes Windows environment variable
START	Starts a separate window to run a specified program or comm
TREE drive:\directory	Graphically displays the directory structure of a drive or path.
TYPE filename	Displays the contents of a text file.
VER	Displays the Windows version.
VERIFY ON or OFF	Tells Windows whether to verify that your files are written correctly
VOL drive:	Displays a disk volume label and serial number.

Figure 50 – The DOS Commands selection of SAKWeb[®] including the PHP[®] code.

- The PHP Configuration option displays the parameters of SAKWeb[®] regarding PHP[®] with the phpinfo() instruction (see Figure 51).

PHP Configuration of SAKWeb.

PHP Version 4.3.3 	
System	Windows NT GARROTT 5.1 build 2600
Build Date	Aug 24 2003 22:01:16
Server API	CGI/FastCGI
Virtual Directory Support	enabled
Configuration File (php.ini) Path	C:\WINNT\php.ini
PHP API	20020918
PHP Extension	20020429
Zend Extension	20021010
Debug Build	no
Thread Safety	enabled
Registered PHP Streams	php, http, ftp, compress.zlib

Figure 51 – Partial view of PHP[®] configuration.

7.1.3. JavaScript[®]

“JavaScript[®] is executed when a document is loaded by the client” [Pandey, 2002]. Often used to create dynamic HTML documents, JavaScript[®] increases the aesthetics and friendliness of Websites by adding events to static pages, referenced by the HTML tag `<script language="JavaScript" src="corefunctions.js"></script>`. In this case, there is an external file whose filename ends with .js. Checking input forms or computing client-side mathematical calculus are two important capabilities of this technique. According to Coelho [2002], this script language includes:

- Arithmetic operators (+, -, *, / and %);
- Logic operators (==, !=, <, >, <=, >=, &&, ! and ||);
- Bit operators (&, |, ^, <<, >> and ~);
- Comments;

- Functions with arguments;
- Event handlers (onClick, onChange, onFocus, onBlur, onMouseOver, onMouseOut, onSelect, onSubmit, onResize, onLoad, onUnload);
- High-language statements such as if(condition){...}else{...}, while(condition){...} or for(begin; end;increment).

Another important capability of JavaScript[®] is the possibility to control the browser itself through the browser objects and their methods:

- The history object allows working with the local accesses made by the user. For instance, history.go(-1); is equivalent to pressing the Back button of the browser while history.toString(); displays an HTML table with the history of the browser links.
- The document object allows access to the properties of the Web page when it is loaded. For instance, document.bgColor="FF0000"; sets up the background color to red while document.clear(); clears the content of the active Web page.
- The location object provides information regarding the current URL. For instance, location.reload(); refreshes the current document while location.hostname; displays the name of the remote host.
- The window object includes the document and other objects where the user has the capability to manage the browser. For instance, window.open("http://www.nasa.gov") loads the home page of NASA into a new window while window.alert("Your input value is out of range!"); generates a pop-up alert message whose content is its input parameter.

JavaScript[®] 1.2 also introduces two new functions that are very useful for the programmer, concerning the debugging and error corrections within programs. The watch() method is applied to a certain variable and, if something has changed, a particular procedure can be triggered. For instance, if the instruction watch('x',change_var); is setup then an alert message will be displayed every time the x variable changes its content and the function change_var() includes the window.alert(); method. As expected, the unwatch() method turns off this debugging effect.

It is imperative to say that JavaScript[®] cannot create files or establish network connections with other remote hosts. In addition, JavaScript[®] is case sensitive. Finally, Java[®] and JavaScript[®] are two distinctive technologies in spite of the resemblance of their names. While the former is a complete language like C++[®], the latter can only be applied to a W3 environment. Second, Java Applets[®] suffer from have a pre-compilation state that generates the binary bytecode while JavaScript[®] is interpreted on the fly by the user browser. Third, Java[®] can be understood as a new set of object oriented classes that will match requirements for a specific programming situation while JavaScript[®] only holds the most important objects and classes that concern the WWW client-side environment. “Nevertheless, it is possible to create a degree of interaction between both technologies” [Coelho, 2002]. As an authentic example of the use of this technology within SAKWeb[®], the following code illustrates how JavaScript[®] modifies the browser status bar (see Figure 52).

```
<HTML><head><title>Status Clock</title></head>
<body>
<script>
    function relogio(){
        var hoje = new Date();
        window.status = hoje.toString();
        setTimeout("relogio()",1000); }
    relogio();
</script>
</body></HTML>
```

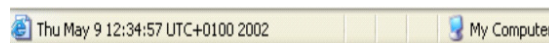


Figure 52 – The JavaScript[®] clock output.

7.1.4. ActiveX[®] and Component One WebChart[®]

As with JavaScript[®], VBScript[®] allows developers to link a wide variety of objects in Web pages including ActiveX[®] controls and Java[®] Applets that are then loaded

and registered into the user's system. Yet, compatibility with existing browsers is a problem for VBScript[®]. “Although IE[®] supports both VBScript[®] and JavaScript[®], VBScript[®] is not currently supported by Netscape[®]” [Pandey, 2002]. Therefore, SAKWeb[®] requires IE[®].

VBScript[®] is fully dissected with the menudynamic.htm, for instance. This Web page includes a Component One[®] ActiveX^{®316} interface that communicates with the local user by exposing the properties and methods to fire events. “This also includes an assigned CLSID for an entry into the Windows[®] registry, allowing the client browser to obtain, register and load the control” [Vieira, 2000]. Technically, the <object> tag is inserted into the body of the HTML page and not into the page header like VBScript[®] and JavaScript[®] declarations.

```
<HTML><head><title>Main Menu</title></head>
<OBJECT classid="CLSID:5220CB21-C88D-11CF-B347-00AA00A28331" VIEWASTEXT><PARAM
    NAME="LPKPath" VALUE="ActiveBar.lpk"></OBJECT>
<OBJECT classid="CLSID:E4F874A0-56ED-11D0-9C43-00A0C90F29FC" codeBase=actbar.cab#Version=
    1,0,6,5 id=ActiveBar1 style="LEFT: 0px; TOP: 0px" width=32 height=32 VIEWASTEXT> <param
    name="_ExtentX" value="847"><param name="_ExtentY" value="847"></OBJECT>

<body>
<script language="VBScript">
<!--
Sub window_onLoad()
    ActiveBar1.Attach
    Create_Tools
    Create_Bands
end sub
Sub Create_Tools()
    iCat = 100
    Set Tool = ActiveBar1.Tools.Add(iCat + 1, "mnu1")
    Tool.Caption = "&Data View and Tools"
    Tool.SubBand = "mnu1"
    Tool.Category = "Menus"
    ...
end sub
Sub Create_Bands()
    Set b = ActiveBar1.Bands.Add("mnu1")
    b.Type = 2 ' ddBTPopup
    b.Tools.Insert b.Tools.Count, ActiveBar1.Tools("descriptivemeasures")
    b.Tools.Insert b.Tools.Count, ActiveBar1.Tools("dataposting")
```

```

        b.Tools.Insert b.Tools.Count, ActiveBar1.Tools("dataposting1")
        b.Tools.Insert b.Tools.Count, ActiveBar1.Tools("inputreview")
        b.Tools.Insert b.Tools.Count, ActiveBar1.Tools("tools1")
        ...
        ActiveBar1.Bands("mnuMain").GrabHandleStyle = 0
        ActiveBar1.BackColor = 33792
        ActiveBar1.ForeColor = 16777215
        ActiveBar1.MenuFontStyle = 1 ' ddMSCustom
        ActiveBar1.Font.Name = "Arial"
        ActiveBar1.Font.Bold = True
        ActiveBar1.Font.Size = 10
        ActiveBar1.RecalcLayout
        ActiveBar1.Refresh
    End Sub
    Sub ActiveBar1_DataReady()
        ActiveBar1.RecalcLayout
    end sub
    Sub ActiveBar1_BandOpen(Band)
        If Band.Name = "SysCustomize" then Band.Tools.Remove (Band.Tools.Count - 1) end if
    end sub
    Sub ActiveBar1_Click(Tool)
        if Tool.Name="descriptivemeasures" then window.location=http://garrott/content/univariate.ASP" end if
        if Tool.Name="dataposting" then window.location="http://garrott/images/3dscatter.ASP" end if
        ...
    end sub
-->
</script>
</body></HTML>

```

Glancing at previous code, the client-side ActiveBar[®] interface was included to layout the output presented in Figure 53 which has a unique identifier, the 128-bit CLSID. It is the use of this CLSID that avoids the possibility of name collisions among classes because CLSIDs are in no way connected to the names used in the underlying implementation. “So, two different vendors can write classes with the same name but each one will have a unique CLSID” [Côtés, 2000]. With the SAKWeb[®] host system, two ActiveX[®] names were registered with the respective CLSIDs³¹⁷: ActiveBar[®] and WebChart[®].

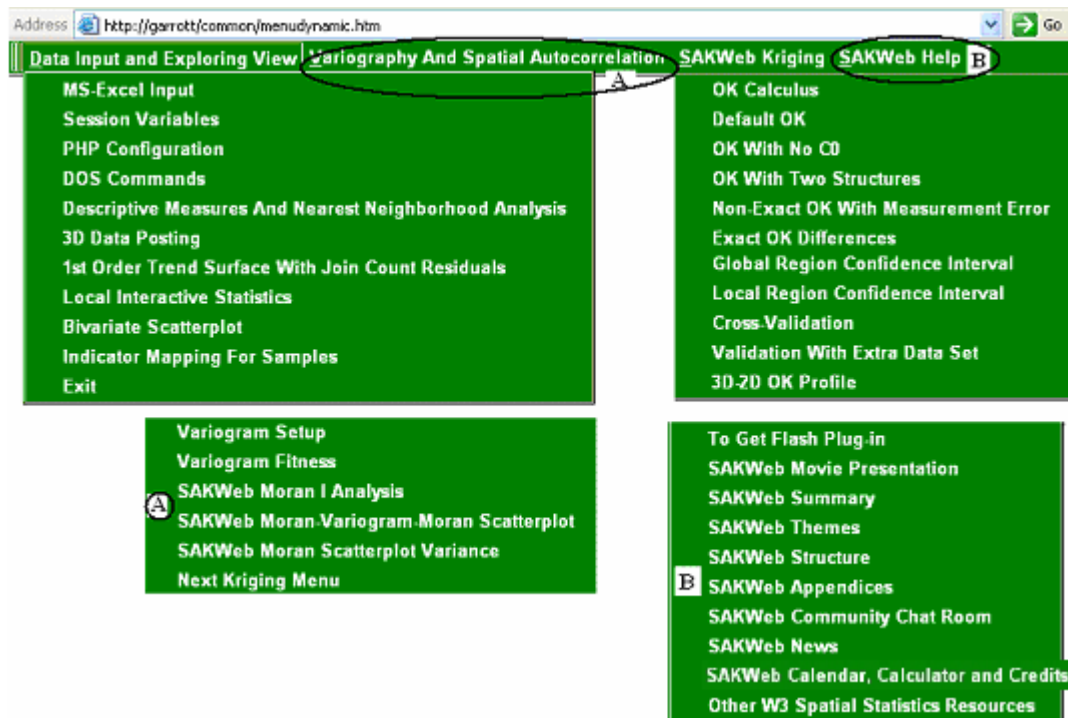


Figure 53 – The SAKWeb[®] main menu.

The 2D and 3D WebChart[®] gives power to any ASP[®] environment by generating different graphic charts. Internally, the server-side component shares a common API with the client-side that generates an .OC2 or .OC3 control to be passed to the client with the appropriate HTML tags. If the client does not have olch2x8.cab and olch3x8.cab, the server cab file generates a runtime-only copy of WebChart[®] on the client's machine (these CAB files are compressed versions of Component One[®] Chart). “As expected, a runtime license for the client’s computer is created if and only if there is a registered version on the developer's computer” [Côtés, 2000]. The generation of the HTML tags to support this Web functionality is contained in two routines of the 2dgentag.inc and 3dgentag.inc files: OlectraChart2D_GenerateTag_Control and OlectraChart3D_GenerateTag_Control. It is these procedures that allow ASP[®] developers to customize the way that the tags are generated.

As a SAKWeb[®] example, the following ASP[®] code introduces the generation of the linear trend surface (see Figure 54) with the following features:

- Data values are stored in d(30,30) array;
- The chart size was setup to 400×300 pixels;

- The chart is a 3D bar type;
- The background color is white while the chart axis is blue;
- The shading surface and zone colors are included;
- The IsBatched=true property does not allow changes to the chart while other calculations are being executed;
- No captions are shown;
- The Call OlectraChart3D_GenerateTag_JPEG(Chart3D,"Chart3D") instruction spawns responsible for spawning the chart image to the client.

'Setup Chart

```
Set Chart3D = Server.CreateObject("C1Chart3D8.ASPComponent")
```

```
With Chart3D.ChartGroups(1)
```

```
With .ElevationData
```

```
.ColumnCount = 30
```

```
.RowCount = 30
```

```
For i = 1 To 30
```

```
For j = 1 To 30
```

```
.Value(i, j) = d(i,j)
```

```
Next
```

```
Next
```

```
End With
```

```
End With
```

```
With Chart3D
```

```
.Width = 400
```

```
.Height = 300
```

```
.IsBatched = True
```

```
.ChartGroups(1).ChartType = oc3dTypeBar
```

```
.Interior.BackgroundColor = ocColorWhite
```

```
.Interior.ForegroundColor = ocColorBlue
```

```
With .ChartGroups(1)
```

```
.Styles(1).Symbol.Size = 2
```

```
With .Contour
```

```
.Levels.NumLevels = 20
```

```
.IsZoned = True
```

```
End With
```

```
.Elevation.IsShaded = True
```

```
End With
```

```
.Legend.IsShowing = False
```

```
.IsBatched = False
```

```
End With
```

```
With Chart3D
```

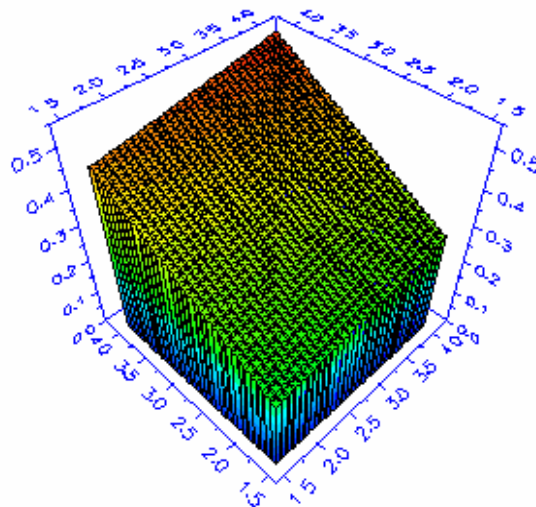
```

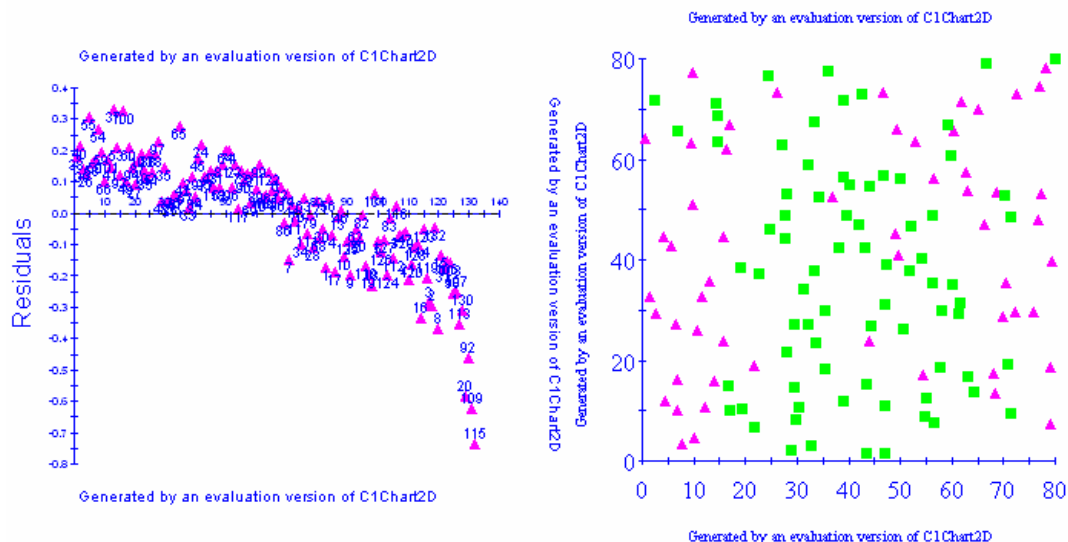
Response.Write("<p align=center>")
dispChart = Session("DispChart")
If dispChart = "Jpeg" Then
    Call OletraChart3D_GenerateTag_JPEG(Chart3D,"Chart3D")
ElseIf dispChart = "Png" Then
    Call OletraChart3D_GenerateTag_PNG(Chart3D,"Chart3D")
    ElseIf dispChart = "Png-BinaryWrite" Then
    Call OletraChart3D_GenerateTag_PNG_BinaryWrite(Chart3D,"Chart3D")
    ElseIf dispChart = "Jpeg-BinaryWrite" Then
    Call OletraChart3D_GenerateTag_JPEG_BinaryWrite(Chart3D,"Chart3D")
Else
    Call OletraChart3D_GenerateTag(Chart3D,"Chart3D")
End If
Response.Write("</p>")
End With
Set Chart3D = Nothing
    
```

SAKWeb General Linear Trend Model

$b_0 = 0.2006$
 $b_1 = 0.0011$
 $b_2 = 0.0034$
 Total variance=5.4354 (df=131)
 Regression variance=0.8952 (df=1)
 Residual variance=4.5401(df=130)
 Adjusted $R^2=0.1519$
 $R^2=0.1647$
 $F=12.9156$ (df=2,131)

Generated by an evaluation version of C1Chart3D





Legend: Green (negative residual) and Pink (positive residual)

Join Count Statistics of Residuals

Number of joins: PP 10.0000 NN 23.0000 PN 15.0000

Expected number of joins for an irregular lattice: PP 16.3333 NN 8.3333 PN 23.3333

Expected variances: PP 293.1351 NN 110.1490 PN 12.1933

[Spatial Analysis](#) [Data Review](#) [Image Map](#)

Figure 54 – The SAKWeb® trend surface analysis of the GS® + Pb dataset.

7.1.5. Java Applets®

Developed in 1994 by Sun Microsystems®, Java® is an object-oriented programming language derived from C++® with an extended library. “Programs written in Java® and interpreted by browsers are called Applets®” [Negreiros, 2001]. Hot Java® was the first browser to support this technology but, nowadays, Netscape Navigator® and IE® also include this capability. “Usually, an Applet® has several input settings that allow any user to customize it” [FDTI, 2000]. For instance, if an Applet® is inserted as a menu, it is possible to specify which options should be in the menu and which pages should be loaded upon any click. Figure 48 (above) is a SAKWeb® live example where the `<applet code=“Lake.class” width=“370” height=“200”><param name=“image” value=“ISEGI1.gif”>`

</applet> tag allows the insertion of the *ISEGI Applet*[®] with the following attributes:

- The name of the executable Java[®] is Lake.class;
- Its layout requires an image size of 370×200 pixels;
- The name of the image file is ISEGI1.gif.

Java Applets[®] have an independent platform feature because of the Java Virtual Machine that runs the interpreted program (see Figure 55). This is crucial because if the language can not run on any machine, it cannot be used on the Web, which must service every machine and imaginable environment. Thus, “the programmer only needs to be concerned with the generic programming and compile it into byte code to be assured that this version will be executed correctly whether the clients are Macs[®], Windows[®], Unix[®] or anything else” [Boavida, Monteiro, 2001].

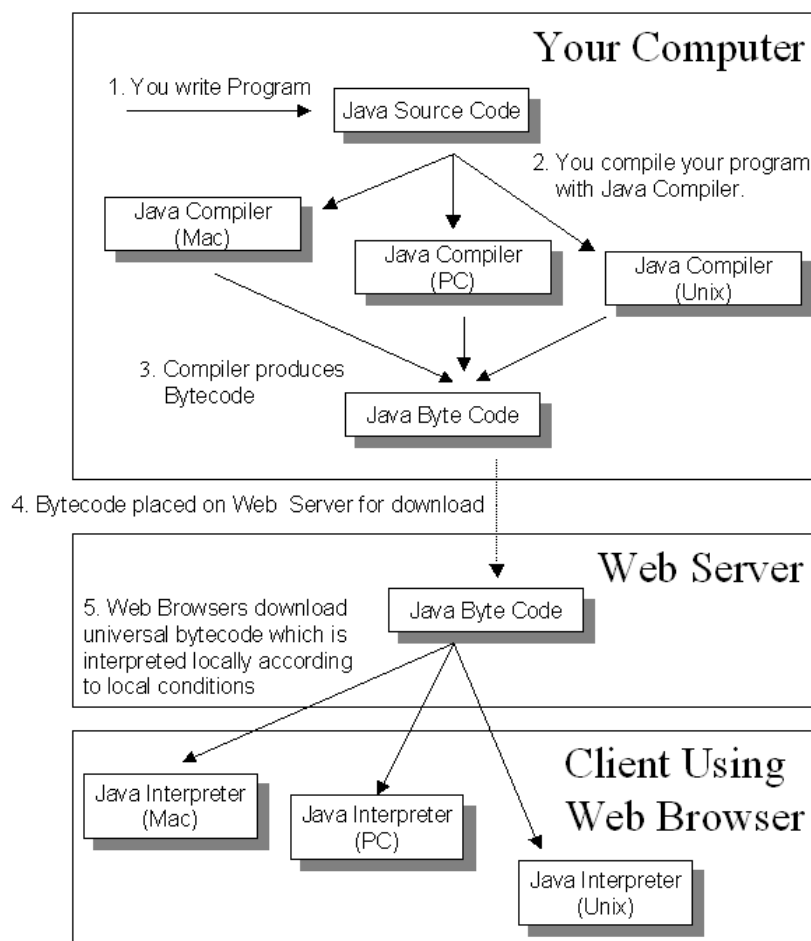


Figure 55 – The independent structure of Java[®] language.

Java[®] is case sensitive, too. Another benefit is the multithreading characteristic, to interact with real-time behavior. In fact, an Applet[®] begins its life when the Web browser loads its classes and calls the `init()` method. Once the initialization is complete, the continuum process as follows:

- The Web browser requests the `start()` method. At this point, the user can begin to interact with it;
- When another Web page is called, the Web browser calls the `stop()` method;
- If the user returns to the initial Applet[®], the Web browser asks for the `start()` method again and the user backs into the program;
- Finally, if the user decides to quit then the `destroy()` method frees up the system resources by killing the Applet[®] before it closes.

Another crucial feature of Java Applets[®] is the TCP/IP mechanism to establish connections between two hosts. Based on the `java.net` package, it is possible to create a client-server applications using the `URL`, `URLConnection`, `Socket` and `ServerSocket`, `DatagramPacket` and `MulticastSocket` classes. Briefly, the creation of the I/O socket object that allows the writing and reading of any type of data is made with the following two syntaxes: `Socket c = new Socket(computer_name, port);` and `DataInputStream i = new DataInputStream(c.getInputStream());`. Thus, the I/O communication socket can now be used as an input file to the client application while the `ssock = new ServerSocket(port);` instruction creates it at the moment the server program is launched. Yet, the communication is only recognized when `csock = ssock.accept();` becomes true. At this moment, the reading and the writing on the socket can be made up by both sides with the two directives that follow: `PrintWriter out=new PrintWriter(csock.getOutputStream(), true);` and `BufferedReader in=new BufferedReader(new InputStreamReader(csock.getInputStream()))`.

However, Java[®] still has serious restrictions. Of particular concern are the security restrictions built into Java[®] such as the fact that Java Applets[®] cannot write files to the local hard drive or get data from servers other than the ones they came from. While this may make the public more confident about the language, it makes Java[®]

programs fairly useless for the average developer who absolutely needs such capabilities to create full featured applications. Furthermore, Java[®] programs with a lot of logic take longer to download. Similarly, “because the programs run on the clients’ machines, they do not have access to resources on the server” [Coelho, 2002]. Thus, without a proxy, a Java[®] program cannot even query simple flat files or databases located remotely. Finally, there are also security issues. For example, the JDBC³¹⁸ standard will not necessarily help a corporate environment in which the proxy protects the database server. Not only will Java[®] be blocked by a firewall but it won’t be able to use standard encryption standards to provide secure and encrypted transactions either.

7.1.6. DHTML

“Dynamic HTML is typically used to describe the combination of HTML, style sheets and scripts that allow documents to be animated” [Ferreira, 2000]. “It is used quite often within SAKWeb[®], allowing a piece of text to change from one size and color to another” [Ferreira, Azevedo, 2001]. Thus, DHTML gives authors the creativeness to manipulate any page element and change styles, positioning and content at any time. The three major components function are as follows:

- The style sheets let the user specify the stylistic attributes of the typographic elements of the Web page.
- The content positioning allows a Web developer to animate any element such as moving pictures and text. Frontpage[®], for instance, presents this concept through the Format/DHTML effects option.
- The dynamic content and data binding (IE[®] technology) or downloadable fonts (Netscape[®] technology) change words, pictures or multimedia on a page without going back to the Web server. However, Microsoft[®] and Netscape[®] use two different techniques for DHTML. Netscape[®] bases much of its approach on the new Layer tag while Microsoft[®] bases its approach on CSS (Cascading Style Sheets) elements and ActiveX[®] controls. In general, a layer refers to elements that can be positioned at exact coordinates on the

Web page. These elements can be defined with the Div, Span, Layer or Ilayer tags. Note that layers created with the Div and Span tags are referred to as CSS layers because their properties are specified by the W3Consortium. This specification also defines style properties (font, color, padding, margin and word spacing) and the positioning properties (top, left, z-index and visibility). On the other hand, Microsoft IE[®] allows the user to change style properties on the fly with VBScript[®] language.

Another perspective of this concept is provided by Dreamweaver[®] software. “Timelines use DHTML to change the properties of layers and images over a given time” [Dreamweaver MX Help]. Timelines also allow the user to change the position, size, visibility and stacking order of a layer or other actions that occurs after a page is loaded. For example, timelines can change the source file of an image tag so different images appear in the Web page over a given time.

7.2. SAKWEB[®] DEVELOPMENTS

7.2.1. The SAKWeb[®] Variogram - Moran I - Moran Scatterplot Optimal ESDA

Binary contiguity fits the **W** matrix relatively well for delimited regions. However, much of traditional sampling does not have boundaries. In addition, recapping subsection 2.4.6, Getis and Ord [1992] mention the difficulty of finding the appropriate optimal neighborhood distance of the local $G_i(d)$ with the North Carolina dataset in order to detect some spatial autocorrelation. Ding and Fotheringham [1992] also report the sensitivity of the G_i statistic to the proximity definition in their population growth dataset for China, with regard to SAM software integration within Unix[®] ArcInfo[®].

Under SAKWeb[®], the global Moran I is setup by the covariogram contiguity where each weight is based on the best variogram fit model. Its range indicates vicinity boundaries. If the distance between the location and its neighbor is larger than the variogram range then that particular neighbor is not included for the weighted vicinity average (see Figure 56). By taking into account spatial anisotropy, the behavior of spatial autocorrelation and the influence of spatial range on the **W** matrix specification, the suggestion is also made that the initial spatial autocorrelation division into neighborhood and distance views is not a clear cut-issue (see Section 2.4 for a further review)³¹⁹. The restriction of the total number of neighbors considered by distance decay functions does not affect this view, either (see Section 2.7 for a further discussion)³²⁰.

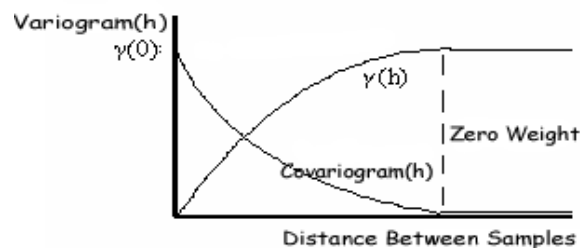


Figure 56 – The general covariogram contiguity plot.

7.2.1.1. Case Study: The San Diego Dataset

With the San Diego dataset, the best variogram fitted model for the uniform interval of three units and an active lag distance of thirty-one was an exponential one:

$$\gamma(\mathbf{h}) = 6497 - 6497 \times (1 - \exp(-|\mathbf{h}|/5.91)) \quad [67]$$

where:

$\gamma(\mathbf{h})$ represents the variogram value;

$\mathbf{h}$ is the distance lag.

The Moran I reaches +0.2631 (see Appendix K), a close value to the binary **W** matrix approach. In terms of linear regression, the same equivalent situation can also be found:

$$\text{Binary\_Vicinity\_Costs} = 0.336 \times \text{Housing\_Costs} + 124.384 \quad (R^2 = 0.238) \quad [68]$$

$$\text{Covariogram\_Vicinity\_Costs} = 0.267 \times \text{Housing\_Costs} + 131.941 \quad (R^2 = 0.239). \quad [69]$$

The covariogram contiguity produces a similar neighborhood range to the binary one: [283, 126] versus [299.6, 115.5]. The sum of the difference between the districts and their vicinity confirms the same suggestion: 468 against 427. Oddly, the districts Jamul (34) and El Cajon East (23) are not considered neighbors according to the covariogram contiguity because their centroids distances are greater than the range. However, Coronado (43) has four natural neighbors, Hill Crest (29), Downtown (40), Ocean Beach (41) and Point Loma (43), as against the zero of the binary contiguity. As expected, the Moran I decreases as distance increases (see Figure 57).

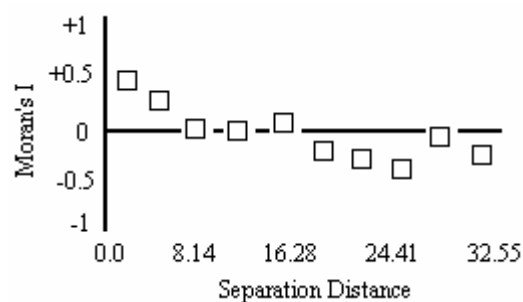


Figure 57 – The Moran I correlogram for the San Diego dataset.

7.2.1.2. Case Study: The Africa Dataset

The following situation concerns the GDP per capita and the population density in African countries. Some evidence of spatial association is expected. “For individual regional members, the actions of their neighbors matter a lot because they are closely associated by such interaction as trade, formal alliances, migration, disputes over territories or the arms races” [Anselin, O’Loughlin, 1991]. The question is to know where those countries are located. Africa is the victim of the most conflict³²¹ and the poorest region in the world, as measured by the GNP per capita. Since population density can be closely related with these considerations, the interest of this small but complex dataset makes Africa an

attractive study region to explain the differences between US and Soviet Union client-coterie in the post-independence process. For instance, Anselin and O’Loughlin [1991] report that post-colonial periods and small states have a strong temporal and spatial relationship as regards war conflicts situations.

Regarding population density, the mean, range, skewness and kurtosis equal 36.23, [2, 228], 2.79 and 8.60, respectively (see Appendix L). The best variogram fitted model was a spherical one³²². The Moran I reveals no spatial autocorrelation for the previous range distance and, as expected, a flat linear regression was found³²³. In addition, four inferences can be drawn:

- 68% percent of the countries are below the mean which lead to the assumption that some other countries have a high population density: Sierra Leone (11), Ghana (12), Togo (13) and Nigeria (15) in West Africa and Burundi (25), Uganda (22), Malawi (31) and Rwanda (26) in Central Africa. Note that those countries are small in terms of total area.
- In contrast, their neighbors, like Angola (21), Kenya (23), Tanzania (24), Congo (20) and Gabon (16) are located in quadrant II.
- Three quadrant IV outliers can be found: Gambia (1), Lesotho (34) and Tunisia (39).
- Note that Egypt (42) does not have any neighbors according to this framework in spite of its medium-high population density: 48 persons per km².

7.2.1.3. Log Transformation Impact on Moran I

The Moran I computed by the binary **W** matrix (+0.0643) reaches a different value than the covariogram contiguity (+0.2327). This supports the non-convergence of the **W** matrix for the Moran I and Moran scatterplot when using both approaches, a situation not pursued in the San Diego case. As the vicinity gets wider, the smoothing effect gets stronger and the Moran I tendency becomes zero. “If the Moran I decreases with increasing distances then the spatial pattern is one of regularly distributed patches” [Upton, Fingleton, 1994]. If the variogram range

increased for seven, eight, nine, ten, eleven, fifteen or nineteen units, the Moran I would decrease to +0.5982, +0.2607, +0.2409, +0.2005, +0.1839, +0.1601 and +0.0079. Since this spatial dataset does not respect a Normal distribution, a lognormal transformation was applied leading to a new distribution with mean=3.15, range=[1.38, 5.43], skewness=0.01 and kurtosis=-0.62. The spherical variogram changes to ($R^2=0.57$)³²⁴:

$$\gamma(\mathbf{h})=0.001+1.085\times\left(\frac{3}{2}\frac{|\mathbf{h}|}{31.2}-\frac{1}{2}\frac{|\mathbf{h}|^3}{31.2^3}\right) \quad [70]$$

where:

$\gamma(\mathbf{h})$ represents the variogram value;

$\mathbf{h}$ is the distance lag.

The Moran I reveals a negative spatial autocorrelation (-0.1263) for the default distance range while the binary approach reveals a strong positive one of +0.3891. Note that different sills do not affect this measure. Regarding the GDP per capita, the mean, range, skewness and kurtosis equal 1803, [478...7380], 1.81 and 2.15, respectively. The best variogram model was a spherical one ($R^2=0.88$):

$$\gamma(\mathbf{h})=230000+4151000\times\left(\frac{3}{2}\frac{|\mathbf{h}|}{87}-\frac{1}{2}\frac{|\mathbf{h}|^3}{87^3}\right) \quad [71]$$

The lognormal transformation presents a range of 95.6 for its spherical variogram, with the lowest Moran I being -0.3149, against the highest of +0.5098, for an eighteen range. It seems that the non-Gaussian distribution produces a narrow range for the global Moran I.

7.2.1.4. Moran I as a Scale Pattern Factor for the Moran Scatterplot

The neighborhood concept is diffuse: Euclidean distance, travel time, boundary length, the rush hour factor, migration flows, telephone calls, arc distance or vehicle speeds are some aspects that may change the **W** specification and, therefore, dissimilar results can be expected. “Deriving geographical connectivity

matrices is somewhat subjective, an art rather than a science” [Griffith and Layne, 1999]. Yet, if the spatial distance between samples is the only factor to take into account then the variogram becomes the best tool available at present to understand this constraint, thus, making this methodology suitable for the **W** matrix. Because traditional sampling does not have defined boundaries like districts, wards or countries, the use of this tool to create **W** can, thus, be refreshing.

With the San Diego dataset, the binary and the covariogram contiguities match the same neighborhood distance and, therefore, similar results are reached. Nevertheless, the Africa dataset does not respect this pattern. The covariogram contiguity presents a wider neighborhood radius and, thus, the Moran I results become weaker and negative. There are no setup rules to compute the best radius neighborhood. Therefore, its range must be determined in the context of an adjusting mechanism because spatial autocorrelation is scale dependent. Table 19 confirms the need to search for the best neighborhood radius in an interactive process.

Range	Moran I	Range	Moran I	Range	Moran I
10	+0.019	50	+0.095	90	-0.139
20	+0.6092	60	+0.0054	100	-0.161
30	+0.2680	70	-0.0593	110	-0.17
40	+0.2081	80	-0.0981	120	--0.147

Table 19 – Moran I according to various covariograms ranges for the Africa GDP dataset.

SpaceStat[®] has the capability to compute global indicators of spatial autocorrelation and association using the first three orders of contiguity to overcome this problem. For instance, with the 1990 housing dataset of West Virginia, USA, Moran I and Geary C indicates a highly significant positive spatial autocorrelation³²⁵ while the third order contiguity no longer yields a significant spatial association³²⁶. With the Africa GDP per capita dataset, the strongest Moran I equals +0.6092 for a 20 units range. Compared with the binary Moran I (+0.3891), a higher Moran I is expected with an interactive methodology and, hence, spatial patterns and outliers can be better enhanced.

In terms of computation, after the covariogram discontinuity has been setup, SAKWeb[®] maps the Moran scatterplot for the highest Moran I found according to various distances where the maximum active lag distance equals the variogram range. The lag distance is setup by the software itself³²⁷. The purpose is to find the interval distance which maximizes Moran I but does not exceed the variogram range because the distance at which the curve levels off, indicates the scale of the spatial pattern³²⁸. Also, since nature pursues quite often a positive spatial autocorrelation pattern, the prevalence of smaller radii is expected. Nevertheless, short ranges danger of omitting a good percentage of the original samples within the Moran scatterplot because of the lack of neighbors. Thus, the first SAKWeb[®] range is one and half times the mean nearest distance.

According to Section 2.7.1, observations can only exchange positions between quadrants I and IV or quadrants II and III in a vertical movement: a negative Moran I highlights repulsive spatial relationships (quadrants II and IV) while positive ones stress quadrants I and III (high-high and low-low patterns). A zero Moran I produces a random and dispersed outcome throughout all quadrants, the least favorable circumstance for anyone whose objective is spatial interpolation. So the highest and the lowest Moran I will layout the extreme positions that locations may hold on the Moran scatterplot.

If quadrant I highlights two separate regions, with quadrant III locations in the middle, then a second order polynomial can be applied. If samples are plotted around the center of the Moran scatterplot then the intrinsic hypothesis is valid for all regions. If quadrant I is followed by quadrant III then UK may be appropriate. If no spatial autocorrelation is present then it is meaningless to create an interpolated surface. In conjunction with the appropriate visual mapping, the variogram - Moran I - Moran scatterplot combination can provide a powerful confirmatory ESDA tool.

Another feature of this view is the capability to plot the Moran I correlogram in order to give hints for the nugget-effect, sill and range factors of the variogram

itself. As the modeling of variograms is regarded as a trial and error exercise in SAKWeb[®], the following suggestions may provide important inferences to take into consideration in any variography re-estimation modeling:

- If the Moran I correlogram confirms no spatial autocorrelation for all distances then the variogram model is a pure nugget-effect.
- If the Moran I correlogram presents a short range of strong positive spatial autocorrelation then the variogram should reflect this characteristic with a low nugget-effect and a similar range.
- If the Moran I correlogram shows a long distance of low positive spatial autocorrelation then the variogram range must reflect this aspect with a higher nugget-effect, too.
- Stronger positive spatial autocorrelation leads to a lower variability between samples and, thus, a lower sill.

7.2.1.5. The Issue of True and False Outliers

Appendix M presents the GS+[®] Pb default dataset³²⁹ where conventional boundaries do not exist. The highest Moran I equals +0.3812³³⁰ for a seven units range although smaller values can be found for shorter distances:

- +0.14 (for a three units distance);
- +0.16 (four);
- +0.28 (five);
- +0.34 (six).

The fifteen unit range achieves a Moran I of +0.37 and since this last range does not exclude any observation³³¹, Figure 58 layouts the Moran location scatterplot. As an ESDA tool, it is clear that two different patterns emerge from quadrants I and III while quadrants II and IV present a random design. Some of these samples are true outliers in need of inspection while others are false. Because major interpolation processes are particularly sensitive to true outliers, it is important to locate them, as with the samples of group's s1, s2 (positives) and s4 (negative):

- Group s1 (records 92 and 93) shows 0.49 and 0.4 values while their four direct neighbors average $(0.26+0.30+0.15+0.19)/4=0.225$.

- Group s2 (record 56) indicates 0.41 while their four natural neighbors average $(0.23+0.25+0.14+0.15)/4=0.1925$.
- Group s4 (records 67 and 82) registers 0.17 and 0.32 while their four natural neighbors average $(1.14+0.72+0.71+0.94)/4=0.877$.

Visually, it is clear that those samples are situated within the lower and higher spatial clusters, respectively. False outliers³³² are located at transition regions reflecting the nature of spatial Mother continuity as with the s3 (positive-negative) and s5 (negative-positive) groups:

- Group s3 (records 40 and 48) shows 0.42 and 0.44, respectively. Their five direct neighbors present two different averages according to their location, $(0.29+0.13)/2=0.21$ in the north and $(0.61+0.52+0.54)/3=0.556$ in the south.
- Group s5 (records 54 and 59) exhibits 0.30 and 0.29, respectively. Their seven natural neighbors have two different averages, $(0.94+0.52+0.49)/3=0.65$ in the north, $(0.54+0.42)/2=0.48$ in the south and $(0.13+0.39)/2=0.26$ in the east.

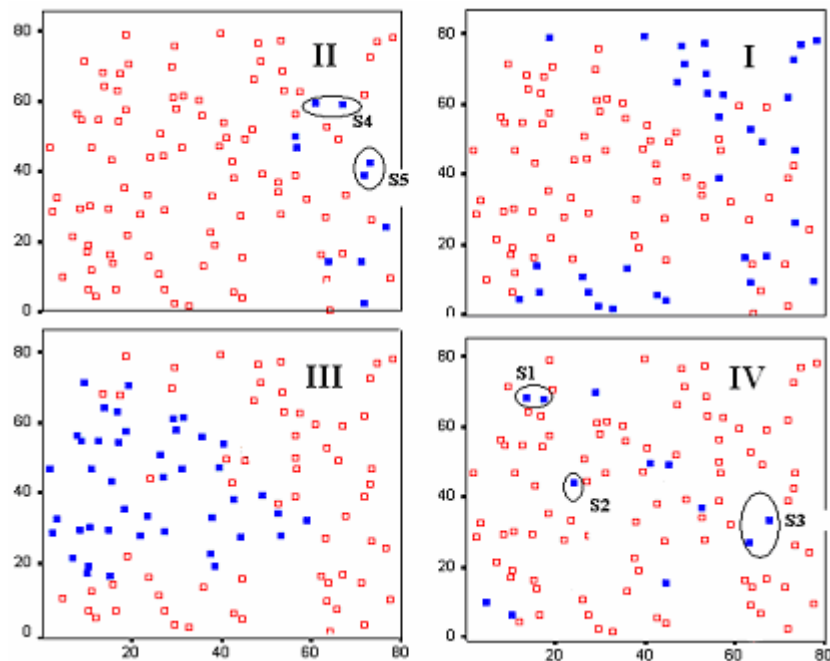


Figure 58 – The Pb dataset results with the Variogram - Moran I - Moran scatterplot optimal technique.

7.2.1.6. The Undercover Uncertainty Region of the Conventional Moran Scatterplot

The samples of quadrants I and III increased from 66 to 78 with a range of 15 against the range of 40. Samples of quadrants II and IV also decreased from 41 to 29. However, the major reward of choosing the neighborhood radius with the highest Moran I is a reduction in the uncertainty samples of the fuzzy region as a result of the neighborhood mean cutoff value³³³ and **W** specification uncertainty³³⁴. If the first factor results from the 'cold' clear threshold then the second difficulty is caused by the ambiguity in **W** specification. According to Figure 59, for instance, points 1, 2 and 3 are located in quadrant II and, therefore, the traditional Moran scatterplot identifies them as negative outliers. Yet, looking carefully at those points, they characterize three different situations:

- (1) represents an average sample surrounded by a high neighborhood values. This means that this sample is a false negative outlier or a shift region between a positive and a negative pattern.
- (2) stands for a high and true negative outlier.
- (3) represents a very low sample value whose neighborhood is slightly above the overall mean. This indicates that this sample is a false negative outlier or a shift region between a negative and a positive pattern.

The same can be said of points 4, 5 and 6 located in quadrant IV:

- (4) represents an average sample surrounded by very low neighborhood values. This implies that this sample is a false positive outlier or a shift region between a negative and a positive pattern.
- (5) stands for a high and true positive outlier.
- (6) reveals a very high sample value whose neighborhood is slightly below the overall mean. That is, this sample represents a false positive outlier or a shift region between a positive and a negative pattern.

In fact, this is a major flaw in the traditional Moran scatterplot because it cannot recognize these situations in any of the quadrants. As with the GIS boundaries, this

clear cut issue must be converted to a fuzzy area for those points close to both axes (the v1-v2 and h1-h2 grey areas in Figure 59), hence reflecting some vagueness in the Moran scatterplot. Stochastically, an option is to use the standard error deviation of the mean for the fuzzy range limits. SAKWeb[®] setups 7.5% of the samples mean for each side to limit this shift area³³⁵.

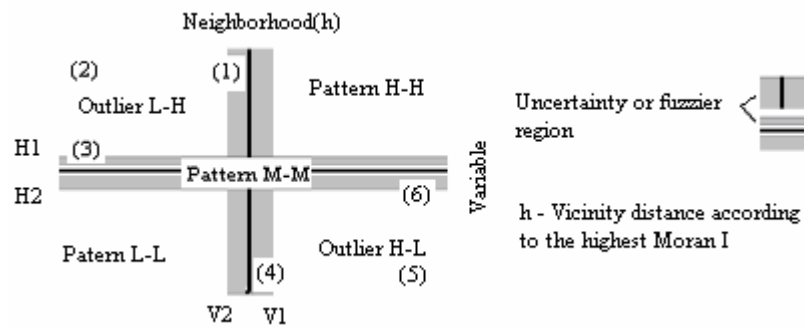


Figure 59 – The location samples of the transition regime.

To locate the samples in space is the next step. Recapping on the Pb dataset, if the [0.362, 0.402] domain³³⁶ is setup to define the fuzzy boundaries of the Moran scatterplot, for instance, the samples 5, 6, 10, 11, 12, 13, 19, 20, 27, 28, 30, 38, 39, 46, 50, 51, 52, 53, 58, 63, 65, 66, 67, 74, 75, 76, 81, 82, 83, 89, 90, 94, 95, 96, 98, 99 are the thirty-six observations that lie within the uncertainty region (see Figure 60, right). As expected, fewer samples (1, 5, 10, 12, 13, 15, 16, 24, 27, 30, 53, 58, 74, 88 and 102) situated in the uncertainty region are expected for the optimal range (see Figure 60, left).

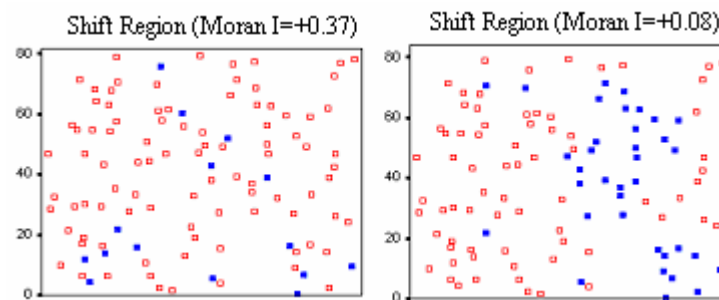


Figure 60 – The Pb location samples of the transition regime.

Therefore, if the samples of the conventional Moran scatterplot are subtracted from the ones of the fuzzier region then the resulting map will reveal outliers, trends and

patterns with a certain confidence level. The ones that are classified in a particular quadrant while bordering on another are stressed too. It is also to be noted that the number of shift observations increases with a lower Moran I. Hence, the total number of outliers decreases and shrink patterns becomes real. Since SAKWeb[®] chooses the Moran I with the highest value, the Variogram - Moran I - Moran scatterplot combination can reach a deeper level of knowledge in the detection of local spatial pockets, trends, true outliers and transition samples by selecting the best spatial **W** definition based upon the highest Moran I.

7.2.1.7. The SAKWeb[®] Moran Location Scatterplot

The remaining problem with the conventional Moran scatterplot is related to its final layout. This spatial autocorrelation measure plots the observation number on a four quadrant mapping but it forgets the true location of each sample. The user quite often has difficulty visualizing them in space unless an additional location map with brushing and linking capabilities is installed³³⁷. Thus, if each quadrant is plotted in the traditional way but on the basis of the coordinates for each sample and the latter are highlight according to their classification then this new design may lead to a more explicit ESDA tool than the conventional one.

Although the information is the same, the aim of the Moran location scatterplot is to plot the samples site in the same four-quadrant layout while emphasizing each sample according to the rules discussed in this section. According to the third survey (see Section 7.3 for further details), for instance, the new Moran location scatterplot is regarded as a better approach than the conventional one by all participants. In summary, the SAKWeb[®] algorithm for the Moran location scatterplot is as follows:

1. A variogram model is setup based on the available samples;
2. The global Moran I is computed for ten lag distances. Their search range is between 3/2 of the average distance among samples (a nearest neighborhood analysis parameter) and the previous variogram range. The goal is to identify the best scale of autocorrelation;

3. If the highest Moran I found corresponds to a very small distance that omits more than 10% of the samples within the Moran scatterplot (due to the absence of neighbours) then the second highest Moran I takes place. The corresponding distance is defined, therefore, as the neighborhood boundary of each sample;
4. Respecting Tobler's Law, the **W** matrix is based on the best covariogram contiguity of step 1 while the range is defined by the distance indicated in step 3;
5. Since variogram modelling is not an exact science, an option is to recheck the variogram nugget-effect, model, sill and range factors (setup in the first step) where the user can modify them to reflect the information revealed by the Moran I correlogram (see subsection 7.2.1.4 and Section 8.4 for a further discussion).
6. The four uncertainty limits (the h1-h2 and v1-v2 grey areas in Figure 58, above) of the Moran scatterplot are computed as:
 - a. h1 equals the samples mean multiplied by 92.5%;
 - b. h2 equals the samples mean multiplied by 107.5%;
 - c. v1 equals the neighborhood mean multiplied by 107.5%;
 - d. v2 equals the neighborhood mean multiplied by 92.5%;
7. A conventional Moran scatterplot can now be plotted.
8. The final step is to draw the four scatterplots on the basis of the Cartesian coordinates of the available samples:
 - a. A red star always symbolizes a medium pattern (the neighborhood and the sample itself are similar to the overall mean).
 - b. The first scatterplot on the top-right hand side corresponds to quadrant I, where the dark blue plots represent positive patterns, the green squares are uncertainty regions due to the cutoff mean (the higher these points are located on the vertical axis and the closer to the y-axis, the greater the shift patterns becomes slope) and the red diamond indicates uncertainty points due to **W** misspecification (points located on the horizontal axis and closer to the h1 limit);
 - c. The second scatterplot on the top-left hand side matches quadrant II, where the dark blue circles represent negative outliers, the green squares

suggest false negative outliers (as in the corresponding situation in quadrant I, these points represents shift patterns) and the red diamonds identify uncertainty points due to **W** misspecification;

- d. With the third scatterplot, the dark blue circles represent a negative pattern, the green squares are uncertainty regions (the lower these points are located on the vertical axis closer to the neighborhood mean, the greater the shift patterns slope) and the red diamond shows points of ambiguity;
- e. With quadrant IV, the dark blue symbols represent true negative outliers, the green squares suggest false positive outliers (these points represents shift patterns) and the red diamond reveals transition samples.

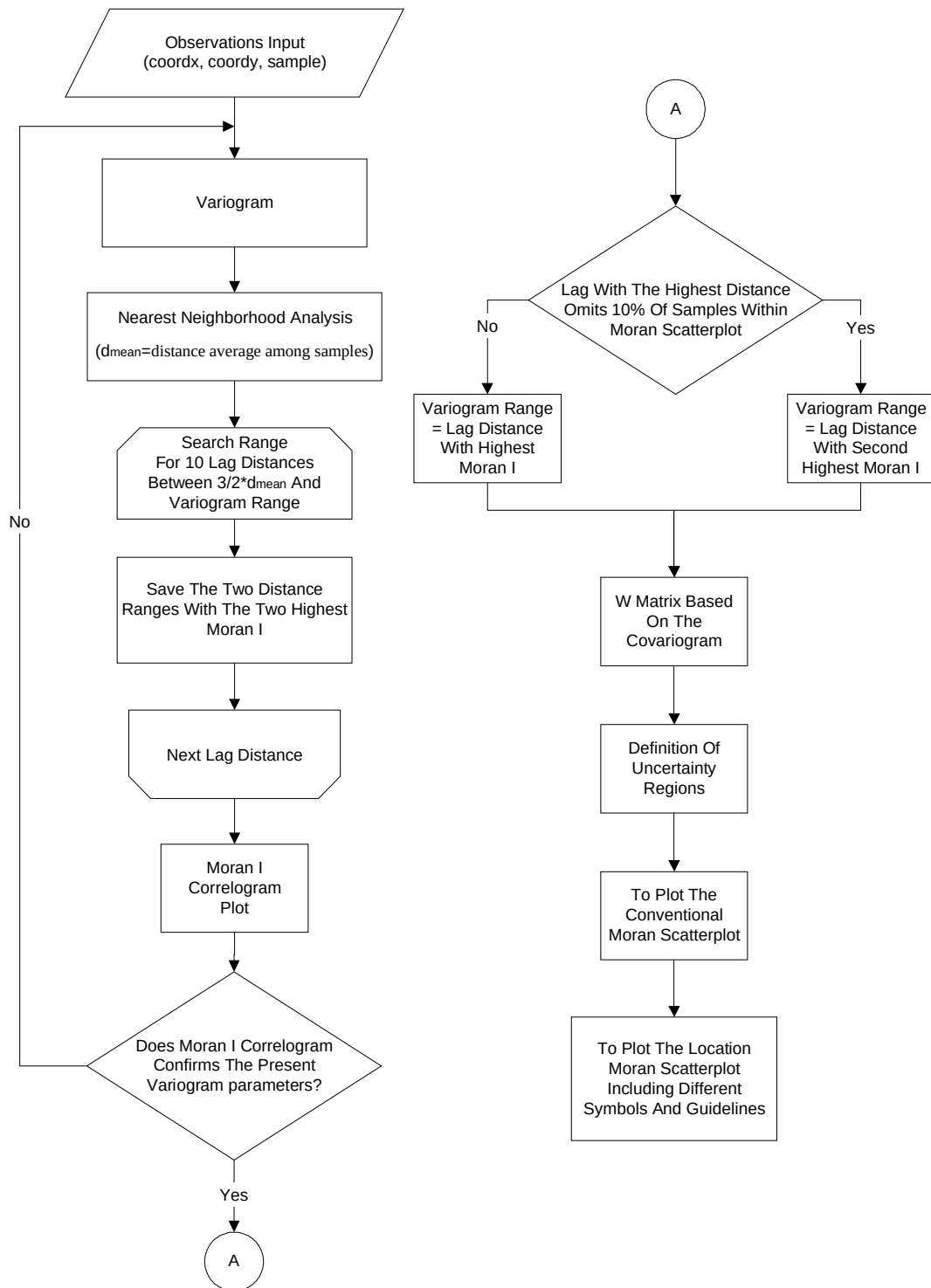


Figure 61 – The optimal Moran scatterplot flowing chart of SAKWeb[®].

The optimal SAKWeb[®] Variogram - Moran I - Moran scatterplot approach (see Figure 62 and 63) was applied to the San Diego case confirming old conclusions and suggesting new ones:

- High cost housing is located along the sea coast (Encinitas-1, Cardiff-2, Solana Beach-3, Del Mar-4, Lake Hodges-5, La Jolla-11, Ocean Beach-41 and Point Loma-42) or in the north (Poway-8 and Scripps Ranch-9).
- Five changeover regions (Rancho Bernardo-7, Clairemont-13, Beaches-14, Del Cerro-19 and Chula Vista West-45) are revealed by the green squares of quadrants I and II³³⁸.
- The green squares of quadrant III (Mission Valley East-28, College-26 and Mission Valley West-17) indicate an average housing with low vicinity costs (uncertainty regions due to **W** misspecification). It should be noted that quadrant IV does not show any counties.
- Two positive outliers were found: Hillcrest-29 and Bonita-35.
- Three negative outliers are revealed by the layout of quadrant II: Bay Park-15, Downtown-40 and Otay Mesa-47.
- A pattern of low housing costs can be found in the center and south-east of San Diego.
- Counties 18 (Tierrasanta), 27 (Kensington) and 45 (Chula Vista West) of quadrant III and IV represent uncertainty regions due to **W** misspecification³³⁹.

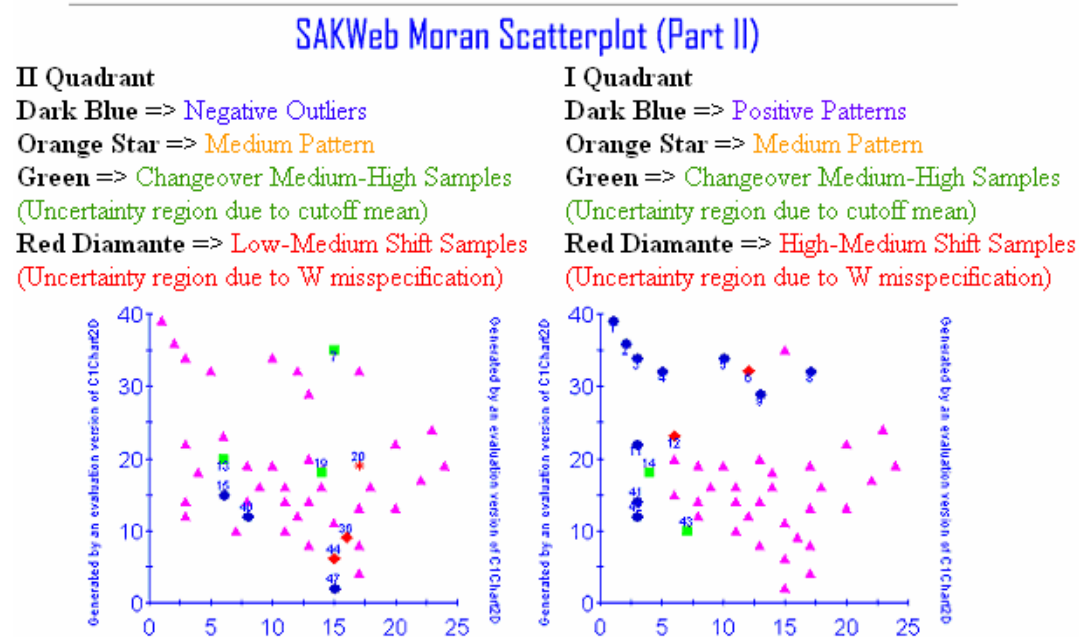


Figure 62 – The SAKWeb[®] view concerning the quadrant I and II of the San Diego dataset.

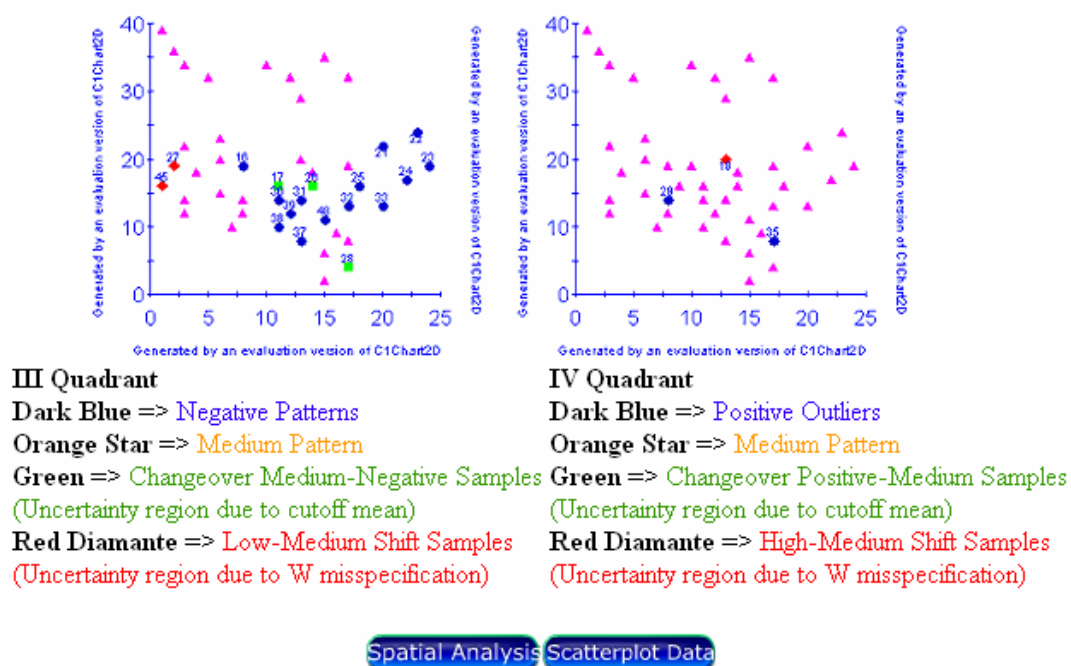


Figure 63 – The SAKWeb® view concerning the quadrant III and IV of the San Diego dataset.

7.2.2. The SAKWeb® Moran Variance Scatterplot ESDA

The conventional Moran scatterplot and the spatial lag pies rely on the neighborhood average. This obscurity reduces neighbor locations to the single mean where variability among them and between the central site and its vicinity are totally lost. With the five situations presented in Figure 64, the neighbour average is thirty in spite of the homogeneity lacking among them³⁴⁰, a heteroskedasticity issue. Although the present Moran scatterplot does not hold a solution to this problem, this issue must be taken into account by spatial autocorrelation measures as an ESDA tool. For instance, “OLS and MLE models are biased with heteroskedastical regions³⁴¹ due to the systematic regional differences, spatial drifts and irregular spatial units” [Anselin, 1998]. With OK, locating unpredictable and unstable regions is very important because of the stationary assumptions validation (see Section 3.2 for further discussion).

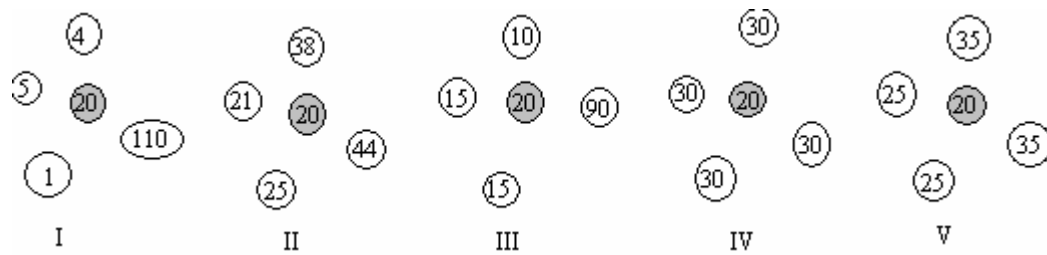


Figure 64 – The heteroskedasticity question, a problem not resolved by the Moran scatterplot.

SAKWeb[®] presents the Moran variance scatterplot to resolve this topic, where the x-axis represents the variable itself and y-axis reflects the neighborhood homogeneity/heterogeneity between the sample and its neighbor's n :

$$x\_axis = \frac{\sum_j w_{ij}(x_i - x_j)^2}{n} \quad [72]$$

where:

w_{ij} represents the binary cutoff distance or the covariogram neighborhood contiguity around the location;

n is the total number of observations;

x_i is the site itself;

x_k and x_j corresponds to the neighbors of a particular site.

The z-axis represents the neighborhood homogeneity/heterogeneity among neighbors:

$$y\_axis = \frac{\sum_{j,k,k < j} w_{ij}(x_k - x_j)^2}{n*(n-1)} \quad [73]$$

A low y-axis value indicates a stable pattern³⁴² while a high one, in conjunction with a high z-axis value, reveals an erratic region. Yet, if a low z-axis occurs then an outlier emerges. Briefly, the Moran variance scatterplot emphasizes outliers and stable/unstable regions although positive and negative patterns can also be identified in spite of the scatter overlapping. This happens because both patterns are constant regions and, thus, they layout on the same position. In order to highlight them, a cutoff value or a point-to-point selection must be setup. The smaller the threshold becomes, the greater the number of outliers (yellow), high

variability regions (black) and medium central-neighbors (green) that can be found. Plausibly, the number of observations classified as a stable region (blues dots) will be less. Also, if the outliers highlighted by this approach do not completely match those stressed by the Moran location scatterplot then the cutoff value must be lower or higher³⁴³.

According to Table 20, five comments can be about the five situations above:

- Case I represents a highly unpredictable region.
- Case II indicates some stability over the area. The smaller the y-axis, the more constant the region is.
- Case III characterizes a medium-high change zone.
- Case IV corresponds to a negative outlier situation³⁴⁴ in a stable region where all neighbors are almost equal.
- Case V is slightly more erratic than the previous situation.

Samples	Case I	Case II	Case III	Case IV	Case V
Overall Mean	28	28	28	28	28
X-axis	20	20	20	20	20
Y-axis	2210	181	1262	100	125
Z-axis	6631	190	2952	0	66

Table 20 – The Moran variance scatterplot results for the five cases above.

To avoid the 3D view, the 2D can be displayed where the x-axis ceases to exist, leading to the conventional orthogonal axes. Hence, the renewed x-axis represents the mean variance between the central site and their neighbors while the y-axis corresponds to the mean variance among all neighbors. In addition, the Moran variance scatterplot can enclose the sample number (it helps to locate it in the map) and the original value, giving its magnitude with regard to the total mean. Figure 65 displays the Moran variance scatterplot layout for the five samples above although the axes are not on a true scale.

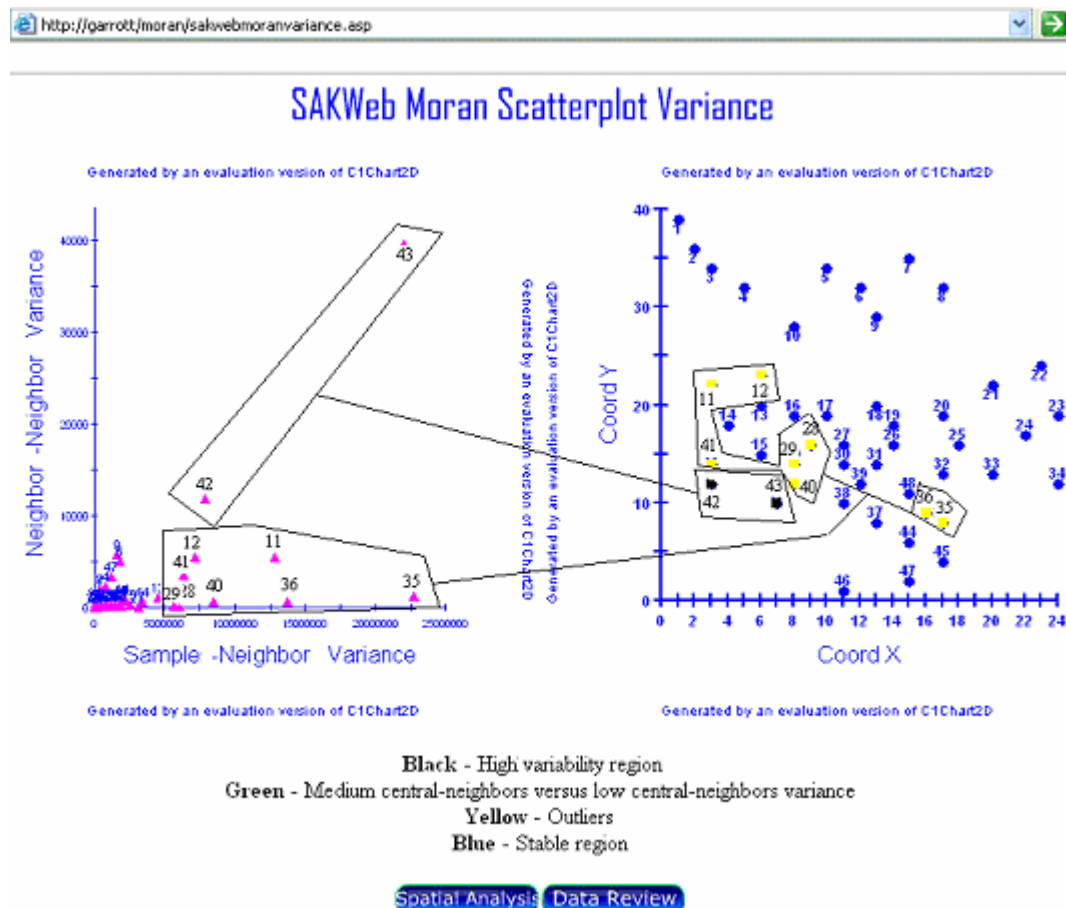


Figure 66 – The Moran variance scatterplot for the San Diego dataset³⁴⁵.

7.2.2.2. Case Study: The GeoEAS® Lead Dataset

With the GeoEAS® lead dataset³⁴⁶, there are clearly five variability groups (see Figure 67):

- Negative outliers 8 and 9 are shown with the symbol .
- The sign  underlines a stable pattern among neighbors with some variability for the central sample.
- Two areas of medium-high unpredictability (samples 1, 2, 3, 5 and 10 versus 32 and 36) are highlighted with large black dots (it is clear a medium location bias).
- The low-medium variability patterns (samples 32, 36, 6, 14 and 4) are indicated with the symbol .

- Samples 7, 22, 38, 49, 50 and 51 does not layout within the Moran variance scatterplot because the forty-five units range does not enclose, at least, two neighbours.
- The remaining samples can be considered a stable region.

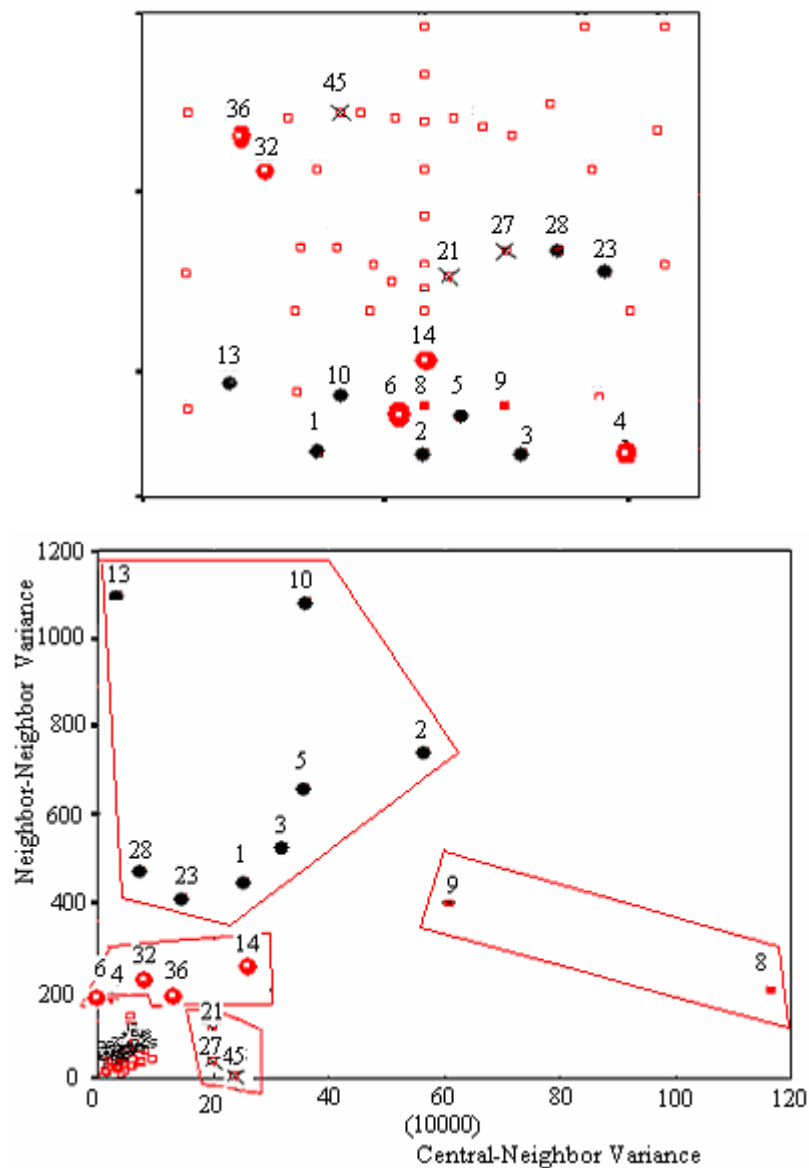


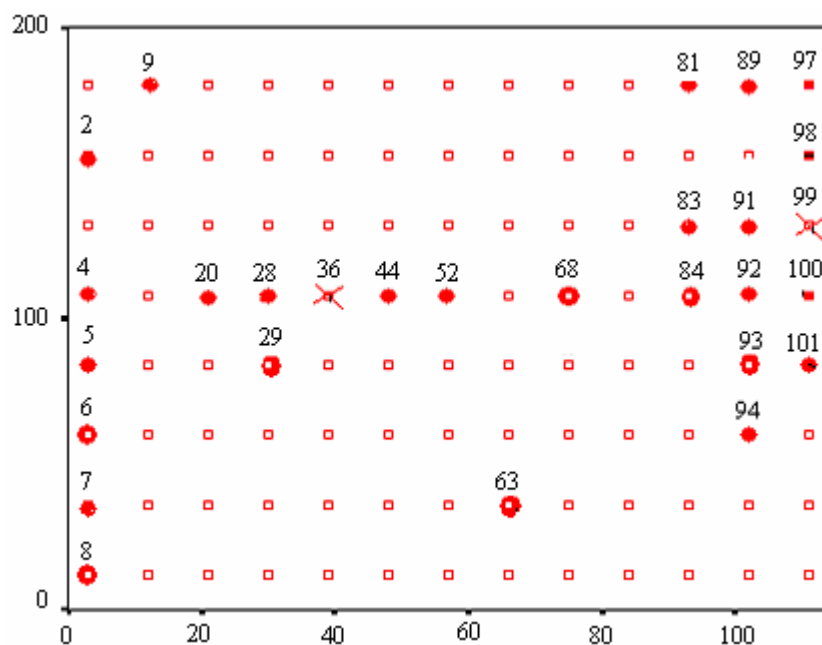
Figure 67 – The Moran variance scatterplot for the GeoEAS[®] lead dataset.

7.2.2.3. Case Study: The Organic Dataset for the Soil in Nebraska

The Moran variance scatterplot can also be applied to residuals in a detrend operation to expose errors homoskedasticity in space, a major classical statistical

assumption (see Section 2.8 to review other assumptions). Using the soil organic dataset for the Nebraska, USA, a linear detrend was carried out, for which the analysis of variance table is depicted in Appendix N. The Moran variance scatterplot was, then, applied to locate heteroskedasticity among the residuals in order to validate the stationary assumptions of Kriging theory and Anova assumptions. According to Figure 68 and similarly to the previous case, five groups emerge:

- Both negative outliers, 99 (-6.05) and 36 (-6.3), are shown with the symbols . It is better to recheck these samples.
- Three medium-high erratic residuals are identified: 100 (-0.9), 98 (1.57) and 97 (0.67). The stationary assumption may be at risk.
- Seven residuals (6, 8, 29, 63, 68, 84 and 93) are pointed out as a stable pattern among neighbors with some changeability against the central error ().
- Sixteen residuals (2, 4, 5, 7, 9, 20, 28, 44, 52, 81, 83, 89, 91, 92, 94 and 101) reveal a low-medium variability pattern ().
- All other residuals have a constant variability.



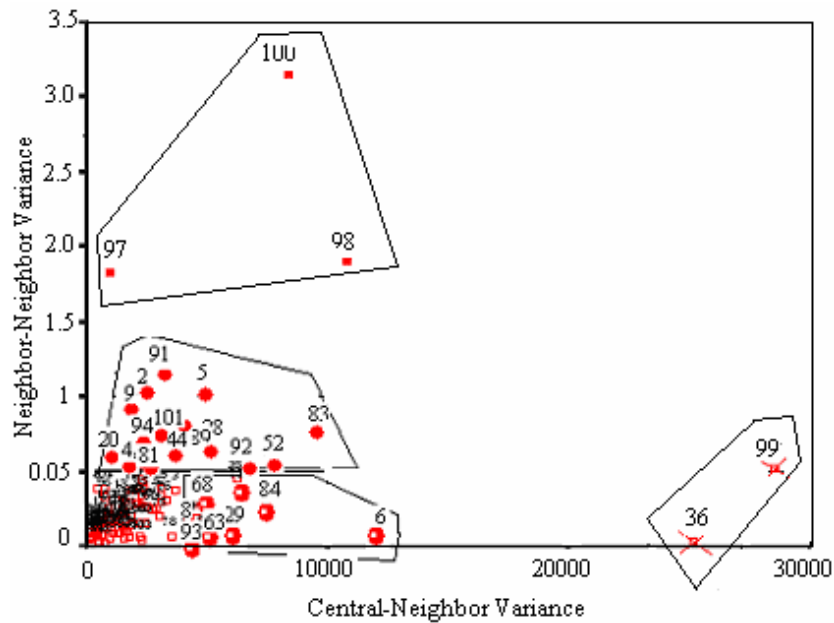


Figure 68 – The Moran variance scatterplot of the first polynomial residuals for the Nebraska soil organic dataset.

7.2.3. SAKWeb[®] Variogram Sill Rescaling

“The variogram sill is not a good global variance estimator because it is often higher than the global variance” [Cressie, 1993]. If preferential sampling in rich areas makes the classical arithmetic mean a poor global mean estimate then preferential sampling also affects variability measures. “Regrettably, it is almost inevitable in mining that a geologist will schedule more samples in rich regions than in poor ones” [Gslib, 2002]. Hence, inaccuracy in the global mean and the global variance will inflate leading to a higher variogram sill unless different weights are assigned to samples. BK should also include this discrimination issue since the global estimate is computed as an even linear combination of block estimates. This type of Kriging must take into account the area of the samples considered as a proportion of the total block area. Likewise, if SK is considered then the computation of the known mean must include some type of declustering to avoid preferential sampling. In addition, for the normal score transformation, it is important that the samples histogram reflects the histogram of the whole population. Therefore, the solution for preferential sampling is to weighting them where representative regions of dense observations receive less weight.

The Kriging_predictor $\pm \sigma_{ok}^2$ confidence interval (see Section 3.6 for a further review) should also embrace this feature. According to the third survey for this research, 85% of the responses confirmed that OK variance is an important aspect of any Kriging software, although it is not a good uncertainty measure. If this is the case, some type of transformation must be carried out to take into account the discrepancy between the sample variance and the variogram sill.

Conventional cell declustering assigns the weights according to the cell size but because each sample receives an inversely proportional weight to the number of points that fall within the same cell then several outcomes may emerge (see Figure 69-A). “With Geostatistical Analyst[®], the cell size is determined from the maximum value of Morisita’s index” [ESRI, 2001]. With the polygonal method, it may also generate several layouts and, therefore, different results can be given (see Figure 69, B and C). As confirmed by Goovaerts [1997], “when sampling size does not suggest a natural cell dimension, several cell sizes and origins must be tried”. In addition, both methods have a difficulty in defining the limits across the research area edges, quite often leading too excessive weighting. With Geostatistical Analyst[®], for example, the outer rectangle boundary is formed by taking the largest (x,y) coordinates plus:

$$\frac{\sqrt{\text{area}/2}}{2} \quad [74]$$

where:

area is the area of the region;

n equals the total number of observations.

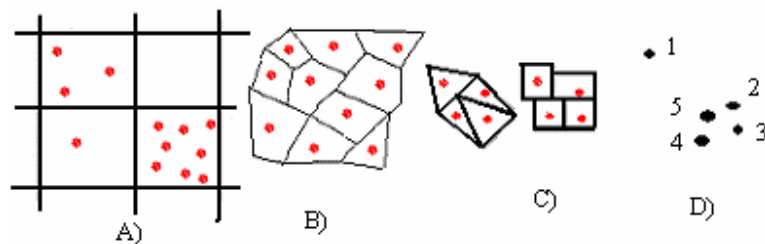


Figure 69 – The weight assignment methodologies.

For the sampling global variance, SAKWeb[®] is based on the nearest neighborhood analysis where each sample weight relies on the nearest neighbor distance among all samples and the estimated one. According to design D) in Figure 69 (above), sample 1 should hold the highest influence, quantified by the distance between samples 1 and 5. As expected, the nearest neighbor distances among samples 2, 3, 4 and 5 are smaller and, thus, their weights for representativeness are also smaller. Political boundaries or coastlines delimitation areas do not affect this methodology, either, because nearest neighbor distance only depends on the samples coordinates.

Nevertheless, the main cause of this new declustering approach is closely related with its assessment because it is difficult to implement the polygonal approach over the Web. Also, cell declustering depends heavily on the size of the cell grid, thus, creating a serious decision for the user. To avoid both issues, the nearest neighborhood analysis was then adopted.

A last topic is related with this approach ranks in terms of accuracy. Considering the D) layout of figure 69, for instance, the samples weights³⁴⁷ fluctuate quite a lot: 0.44, 0.14, 0.14, 0.14 and 0.14 for the nearest analysis versus 0.23, 0.3, 0.135, 0.2 and 0.135 for the polygonal approach versus 0.25, 0.125, 0.125, 0.25 and 0.25 for cell declustering. Considering the grasshopper case from Colorado, presented in Chapter 8, for instance, it is clear that the samples layout is a U shaped. With the polygonal approach, the samples that are located on the inner border are given too much weight because they represent a vast area without any observations. Cell declustering and nearest neighborhood analysis do not reflect this issue. However, it is imperative to stress that none of these approaches is considered the best in terms of final results unless the exhaustive dataset is available. Still, “it can be guaranteed that any these declustering methods for weighting will always lead to a tremendous improvement in local and global distribution estimates, for instance, when compared with the traditional statistical approach where all observations have the same weight” [Isaaks, Srivastava, 1989].

7.2.3.1. Case Study: The San Diego Dataset

The dataset for San Diego housing costs follows the non-convergence trend regarding the overestimation of the sill variogram for the classical global variance: 6497 (sill variogram) against 5523 (standard variance). When the nearest neighborhood distance is applied to the samples weights, the estimated global mean, EGM, and the estimated global variance, EGV, are:

$$EGM = \sum_{i=1}^n w_i x_i / \sum_{i=1}^n w_i = 199 \quad [75]$$

$$EGV = \sum_{i=1}^n w_i (x_i - EGM)^2 = 5379 \quad [76]$$

Although the sill rescaling operation does not affect estimation accuracy, the ratio computation between the EGV and the variogram sill can lead to a reliable improvement in the OK variance. The initial variogram is as follows:

$$\gamma(\mathbf{h}) = 6497 \times (1 - \exp(-h/5.31)) \quad [77]$$

Therefore, if the ratio index between the sill and the EGV equals 82% then the new variogram becomes (with OK variance decrease of 7.86%, 5272, on average):

$$\gamma_1(\mathbf{h}) = \gamma(\mathbf{h}) \times 0.82 = 6497 \times (1 - \exp(-h/5.31)) \times 0.82 = 5379 \times (1 - \exp(-h/5.31)) \quad [78]$$

7.2.3.2. Case Study: The Africa Dataset

With the Africa GDP per capita and population density datasets, both variograms confirm a higher sill against the classical variance: 4151000 and 2170 against 3230837 and 1949, respectively. Taking into consideration the first dataset, the EGM and EGV indices are 1495 and 3599804, leading to a ratio scaling of 86%³⁴⁸. Thus, the rescaled variogram becomes:

$$\gamma_1(\mathbf{h})=4.151\text{e}06\times\left(\frac{3}{2}\frac{|\mathbf{h}|}{87}-\frac{1}{2}\frac{|\mathbf{h}|^3}{87^3}\right)\times 86\%=3599804\times\left(\frac{3}{2}\frac{|\mathbf{h}|}{87}-\frac{1}{2}\frac{|\mathbf{h}|^3}{87^3}\right) \quad [79]$$

Its OK variance decreases, on average, 6% (75437)³⁴⁹. With the Africa density dataset, its EGM and EGV are 35 and 1707 and, thus, the EGV/sill ratio equals 78%. The new spherical variogram then becomes:

$$\gamma_1(\mathbf{h})=2178\times\left(\frac{3}{2}\frac{|\mathbf{h}|}{26.5}-\frac{1}{2}\frac{|\mathbf{h}|^3}{26.5^3}\right)\times 78\%=1707\times\left(\frac{3}{2}\frac{|\mathbf{h}|}{26.5}-\frac{1}{2}\frac{|\mathbf{h}|^3}{26.5^3}\right) \quad [80]$$

Its OK variance decreases, on average, 14%. Noticed that with these three examples, there was no nugget-effect. If there was, this factor had also to be multiplied by the same EGV/sill ratio. In a theoretical framework, an EGV based on the nearest neighbor distance and the traditional variance gives similar results if the sampling adopted is a perfect grid layout. The organic soil dataset of Nebraska, USA, confirms this pattern. However, high and low outliers increase the nearest neighbor EGV index if samples are located on the edge of the map. For instance, the GeoEAS[®] lead dataset pursues this trend³⁵⁰:

- EGV=4219 against classical variance=3637;
- EGM=54 against classical mean=53.

7.2.3.3. SAKWeb[®] Global Confidence Interval

Using the GS+[®] Pb dataset, it is possible to generate, for instance, the Kriging_predictor⁺₋1.645×σ_{OK} confidence interval for a 90% confidence level or the Kriging_predictor⁺₋1.282×σ_{OK} for an 80% level. On the basis of the Normal error distribution, rescale of the sill variogram and OK variance, SAKWeb[®] simulates the minimum and maximum plume for three confidence intervals (80%, 90% and 95%) and three OK models (OK with two structures and OK with and without C0). Hence, nine configurations can be achieved. For the same confidence level, SAKWeb[®] also computes the largest and the smallest area above a specific threshold (indicated by the user). According to Figure 70, for instance, the left map

depicts a wider region than the right one for a cutoff limit of 0.5 and a confidence level of 80% (GS+® Pb dataset).

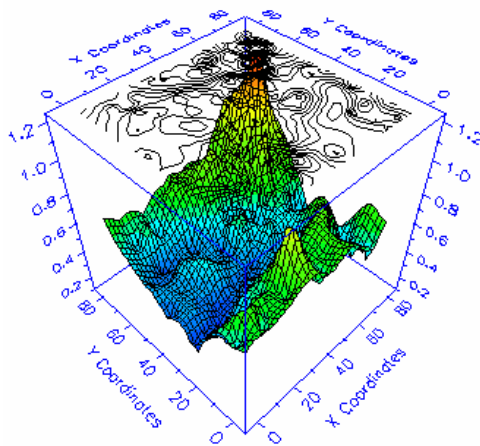
SAKWeb Global Region For A Confidence Level

Confidence Level For The Minimal And Maximal Plume: 80% OK model: OK with C0

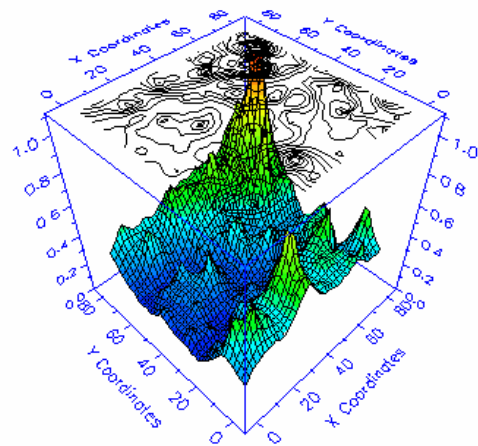
Reset Submit

Spatial Analysis

Highest Plume



Lowest Plume

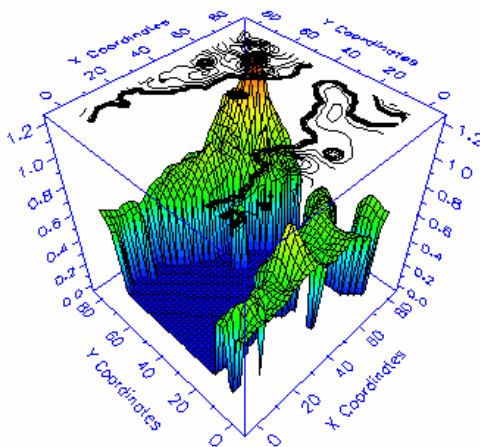


Minimal Value: 0.06 Maximal Value: 1.25

Threshold Value: 0.5 Reset Submit

Spatial Analysis

Highest Plume



Lowest Plume

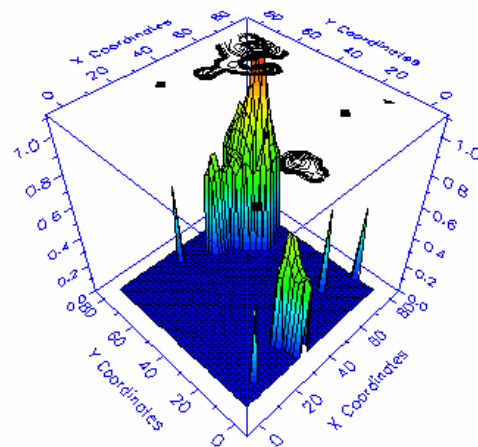


Figure 70 – The Global Region Confidence Interval option of SAKWeb®, a three step process.

7.2.3.4. SAKWeb[®] Local Confidence Interval

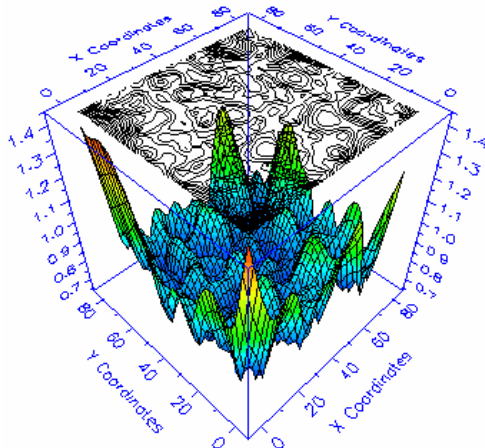
Another possibility to enhance the previous confidence interval is based on the local error variance as a true representative of the local pattern of spatial continuity. Since the shape of the variogram is the same for the whole study region, a hard assumption, to rescale the local variogram should be computed to reflect local spatial variability. According to Isaaks and Srivastava [1989], “the error variance of a relative variogram with a sill of one multiplied by the local variance achieves this aim in a proper way”. In practical terms, to produce a variogram sill of one, each coefficient of the original model should be divided by its sill. With regard to the local variance, for both authors, “moving neighborhood statistics can help this estimation, especially with the relationship between local mean and local variance”.

Under SAKWeb[®], this local variability will be computed based on the local variance whose local range equals the greatest value between the lag with the highest and second highest found in the Moran I correlogram³⁵¹. This happens because the proportional effect is not always true for all datasets and it is central for the local error variance that this factor becomes the most accurate as possible. Once the local error variance has been assessed, thus, the same Normal 80%, 90% or 95% confidence intervals for the three OK models can be assigned but reflecting local conditions. The user can also setup a threshold limit to assess the highest and lowest confidence plume as Figure 71 shows (compared with the previous one, it becomes obvious a less smooth effect and a major variability of this latter approach).

SAKWeb Local Region for a Confidence Level

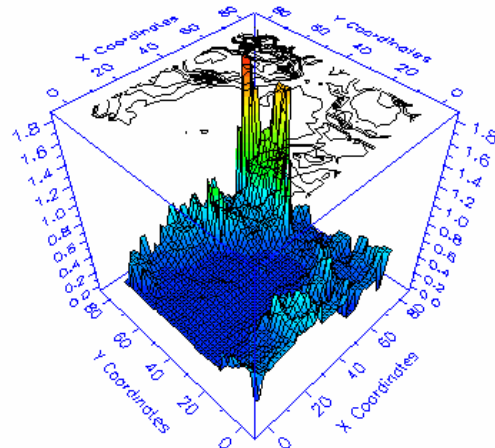
Generated by an evaluation version of C1Chart3D

The Error Variance of the Relative Variogram (1)

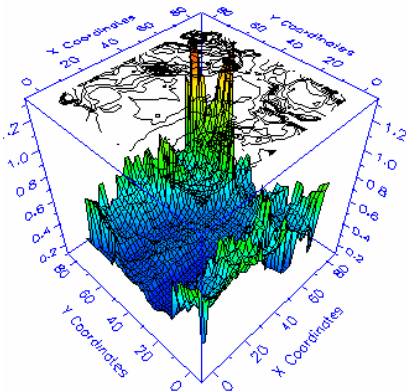


Generated by an evaluation version of C1Chart3D

The Conventional Local Error Variance (2)



The Local Standard Deviation (3)



(1) - A relative variogram with a sill of one.

(2) - The local variance equals $\frac{(\hat{\alpha}_1 - \bar{x})^2}{(n-1)}$ whose range is 8.3503.

(3) - The uncertainty measure of Local Standard Deviation equals the relative error variance multiplied by the estimate of the local variance.

Confidence Level For The Minimal And Maximal Plume: 80% ▾

OK Model: OK with C0 ▾

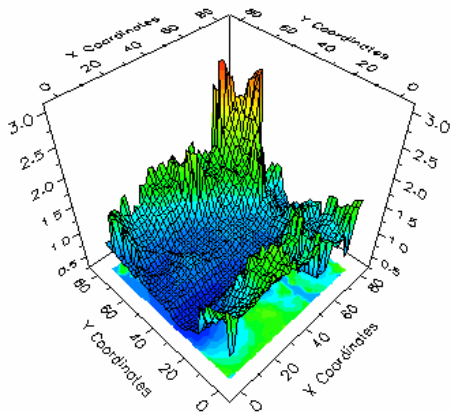
Reset

Submit

80%
90%
95%

Spatial Analysis

Local Confidence Highest Plume

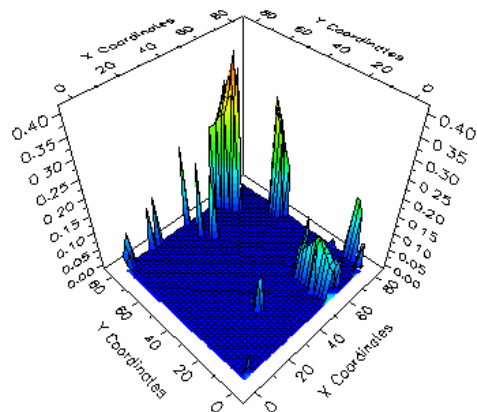


Reset

Rotation

Zoom

Local Confidence Lowest Plume



Reset

Rotation

Zoom

Threshold Limit: 0.5

Reset

Submit

Spatial Analysis

SAKWeb Local Region For A Confidence Level

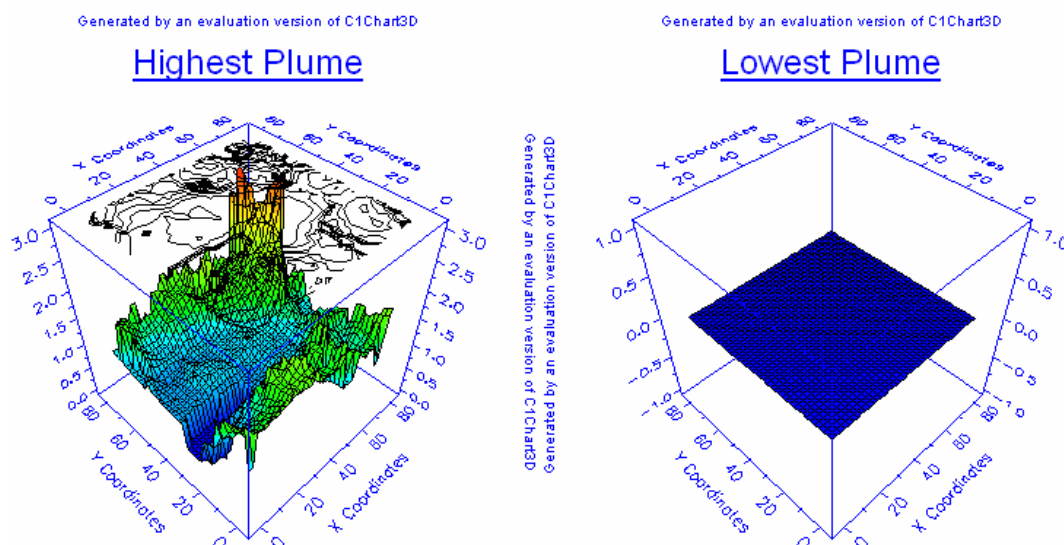


Figure 71 – The three step process of the Local Region Confidence Interval option.

7.2.4. SAKWeb[®] Nugget-effect

When no nugget-effect is introduced or when the phenomenon is well structure, e.g., a quadratic model near the origin, negative weights are more likely [Armstrong, 1998]. “The screen-effect and clustered points are also likely to produce negative weights” [Cressie, 1993] and, thus, the possibility to estimate beyond the available sample extremes becomes real and negative concentrations can be achieved³⁵². This includes an increase in the mean squared prediction error. “This usually happens if you have one or more data points at the same location or the data points are located close to each other” [Geostatistic FAQ, 1997]. Statistically, according to Sousa [1998], the following positive defining function might not be guaranteed (in practical terms, if a new variable results from any linear combination of other random ones then the variance of the new variable is always positive):

$$\sum_i \sum_j w_i w_j C(h) \geq 0 \text{ and } \forall w_i, w_j, h \quad [81]$$

The discontinuity of the variogram origin should not occur in most spatial processes because space is continuous most of the time, although particular earth processes such as gold mineralization do not respect this pattern. “Nevertheless, the variogram should be continuous at the origin” [Yeh, Guzman, 1999]. “Remember that one of our basic assumptions is physical continuity of the phenomena being measured” [Clark and Harper, 2000]. “Furthermore, a free version of the measurement error should be predicted because when predicting at a particular location, the investigator wants to know what actually exists at that location and not the value distorted by, for example, laboratory measurement error” [Cressie, 1993].

Another constraint is the constant nugget-effect regardless of the variogram direction. Existing software also considers $\gamma(0)=0$ although the variogram does not tend towards zero as the lag distance approaches zero, therefore, leading to a sharp rise for very small distances. Yet, Kriging becomes an exact estimator reflected in a strange looking map due to a sudden jump in the predicted collected samples³⁵³. “If the nugget-effect is treated as an intercept with the variogram axis so that $\gamma(0)=C_0$, two samples taken at the same location could have a variance of $2C_0$ ” [Clark and Harper, 2000]. Thus, another possibility to handle this factor is to consider $\gamma(0)=C_0$ while $2C_0$ must be added to the final Kriging variance. However, “Kriging will become a non-exact interpolator, a condition sometimes not required by mining companies” [Hock, Jensen, 1999].

Other interpretations of the nugget-effect can be found in specialized literature like Statistics for Spatial Data [Cressie, 1993]. For Atkinson [1997], “if the source of measurement error is external, e.g., the mechanical noise, removing it is to subtract from the initial variogram some value from all lags. If the source of measurement error is internal, e.g., site position error, removing it is to subtract some function type that decreases from the variogram value at lag of zero to zero at some positive lag” (see Figure 72). In extreme cases, if the variogram is a pure nugget-effect or the variogram range is less than the distance between samples, the best estimation would be the sample arithmetic average because adjacent points are uncorrelated. With regard to major circumstances, the shift of massive weights towards the

closest sample becomes a reality, including the tendency of the LaGrange multiplier factor towards negative values (see Section 3.5 for a further review). Certainly, the nugget-effect is of critical importance in spatial estimation.

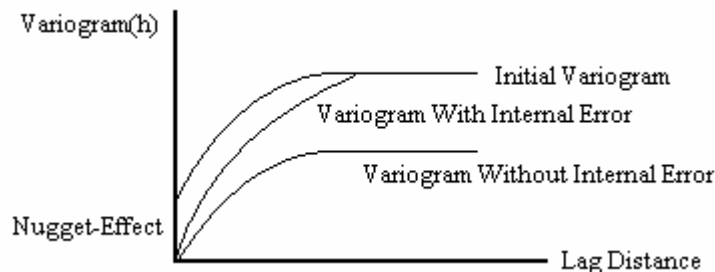


Figure 72 – The Atkinson [1997] internal and external error view.

As reported by Hock and Jensen [1999] in their valley research in Sweden, the Kriging interpolation of glacier mass reveals a considerable difference with respect to the spatial distribution up to 0.66m. It is clearly a wide estimate range without the nugget-effect (see Figure 73). For Isaaks and Srivastava [1989], “particular differences in the nugget-effect at different thresholds will cause IK with different variogram models at different cutoffs to produce results that are quite different from those produced by median IK”.

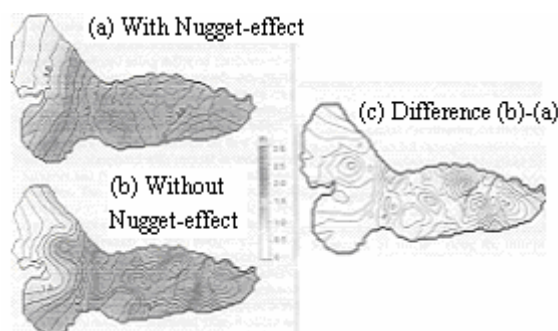


Figure 73 – The Kriging interpolation of a Swedish glacier mass with and without the nugget-effect³⁵⁴.

Unless some additional close samples are available, geostatisticians must guess the variogram shape near the origin, especially when the distance from the point being estimated to a particular sample is small. Once again, “the nugget-effect and the origin slope are the most important features for fitting the variogram” [Gslib, 2002]. In mathematical terms, the larger the nugget-effect the greater the number of equal weights that will be given among all samples³⁵⁵, making the OK

procedure similar to the simple average. Therefore, less Kriging variance and fewer weights for the central points will be assigned. With well structured variograms, most weights are concentrated in locations around the estimated one³⁵⁶. With poor situations, Kriging improves as the range increases or more samples are added, leading to a smoother variogram. According to GeoEAS[®], for instance, a rough rule of thumb is that eight samples are included when the nugget is near zero, increasing to twenty when it is more than 50% of the sill.

“Because transition models depends on the variogram behavior near the origin” [Goovaerts, 1997], SAKWeb[®] presents four strategies for the nugget-effect³⁵⁷. The default is to consider that observations are precise and accurate, although sudden jumps at the variogram origin may emerge. This means forcing the nugget-effect to zero at zero distance:

- $\gamma(0) = 0$;
- $\gamma(\mathbf{h}) = \text{Spherical, Exponential and Gaussian model, if } 0 < h \leq \text{range}$;
- $\gamma(\mathbf{h}) = \text{sill} = C_0 \text{ (nugget-effect)} + C_1 \text{ (partial sill), if } h > \text{range}$.

Another possibility is to consider $\gamma(0)=0$ in a continuous mode, without the internal error of the Atkinson [1997] variogram model. Two extra approaches are considered: the first regards the micro-scale component and the second includes the measurement error:

- $\gamma(0)$ division into two variograms: $\gamma_1(0) = 0$; $\gamma_1(0 < h \leq \text{shortest sampling interval}) = \text{Spherical, Exponential and Gaussian first model}$; $\gamma_2(\text{shortest sampling interval}^{358} < h \leq \text{range}) = \text{Spherical, Exponential and Gaussian second model}$. The first micro-scale range equals the shortest sampling interval (SSI) lag, its nugget-effect is zero while its total sill matches the extrapolated value of the second variogram structure³⁵⁹, that is, the given value of the second variogram structure at SSI is the first variogram sill parameter. This vision reflects different processes of variability on different scales³⁶⁰: $Z(s) = \mu(s) + W(s) + \eta(s) + \varepsilon(s)$ where $\mu(s)$ is the large-scale deterministic variation, $W(s)$ is the smooth small-scale variation (the intrinsically stationary process whose variogram range exists and is larger than $\min\{\|s_i - s_j\|\}$), $\eta(s)$ is the microscale variation whose variogram range

exists and is smaller than $\min\{\|s_i - s_j\|\}$) and $\varepsilon(s)$ is a zero-mean white-noise process, independent of $\mathbf{W}$ and η [Cressie, 1993]. Quoting Soares [2000], “linear combinations or sums of basic correlation structures are often used to model multiple ranges in nested models”. “Moreover, the sum of known positive definite models is also a positive definite one, leading to great flexibility in modeling variograms” [Gslib, 2002]. Notice also that the total sill equals the sum of both sills including the nugget-effect of the long-range variogram³⁶¹.

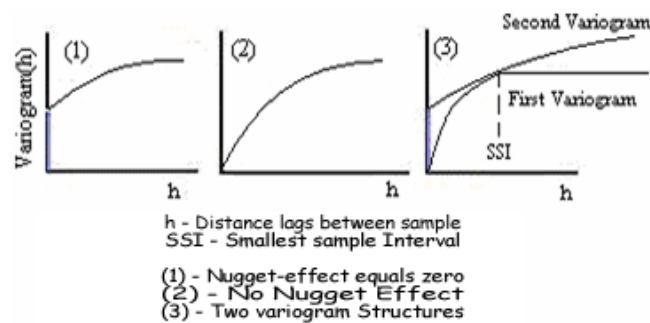


Figure 74 – The three exact nugget-effect approaches of SAKWeb[®].

- If measurement error³⁶² is given, the C_{me} factor is included within the $\gamma_B(x_0, x_i)$ matrix³⁶³ (see Section 3.5 for a further review). In this case, the Kriging $\mathbf{B}$ matrix is the same unless it coincides with the samples locations, a non-exact version of the interpolator. Therefore, the measurement error must be setup instead of zero of the exact version³⁶⁴, and a smoother prediction of the Kriging estimations thus becomes real for the available samples³⁶⁵. Otherwise, the OK estimations are the same whether the measurement error is considered or not (see Figure 75, left) and this is the main cause of the SAKWeb[®] choice. Instead of the final OK map being laid out, only the available samples are estimated.

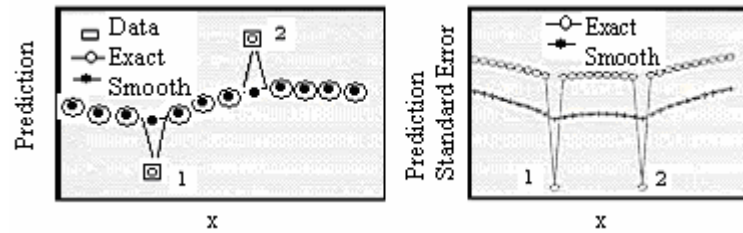


Figure 75 – OK comparisons between the exact version and the one with all measurement error (1 and 2 represents data locations).

This factor must also be subtracted from the variance prediction error, leading to a lower value and a smooth standard error surface. Moreover, the prediction standard error of the samples is zero when no measurement is considered, a situation not contemplated when the measurement error is included (see Figure 75, above, right, or Figure 76), for instance, due to last night's rainfall. As confirmed by Cressie [1993], “it is important to include measurement error (which is often the case) within the Kriging system”.

$$\sigma_{OK}^2 = \sum w_i \gamma(x_0, x_i) + \Psi - C_{me} = \sigma^2 - \sum w_i C(x_0, x_i) - \Psi - C_{me} \quad [82]$$

where:

w_i is OK weight;

$\gamma(x_0, x_i)$ equals the variogram value between the estimated point and the sample;

C_{me} represents the measurement error;

σ^2 signifies sample variance;

$C(x_0, x_i)$ stands for the covariance between the estimated point and the sample;

Ψ is the LaGrange parameter.

Suppose that the user has five samples: (1,5,100), (3,4,105), (1,3,105), (4,5,100) and (5,1,115). His goal is to estimate the (1,4) unsampled location with OK, where the linear variogram equals $\gamma(\mathbf{h})=2+13.5 \times \text{distance}$. The $\mathbf{A}^{-1}$, $\mathbf{B}$ and $\mathbf{W}$ matrices are as follows:

$A^{-1} =$	0	32.19	29	42.5	78.37	1	=	-0.02392	0.006659	0.013706	0.006126	-0.00257	0.329226
	32.19	0	32.19	21.19	50.67	1		0.006659	-0.04069	0.009486	0.019637	0.004903	-0.18427
	29	32.19	0	50.67	62.37	1		0.013706	0.009486	-0.02445	-0.00358	0.004838	0.176614
	42.5	21.19	50.67	0	57.66	1		0.006126	0.019637	-0.00358	-0.02633	0.004141	0.265995
	78.37	50.67	62.37	57.66	0	1		-0.00257	0.004903	0.004838	0.004141	-0.01131	0.412432
	1	1	1	1	1	0		0.329226	-0.18427	0.176614	0.265995	0.412432	-42.8173
$B =$	15.5												
	29												
	15.5												
	44.69												
	69.5												
	1												
$W =$	0.459169877												
	0.104453911												
	0.461557935												
	-0.01380446												
	-0.01137726												
	0.230774028												
OK Estimation=	102.8902												
sample1	100												
sample2	105												
sample3	105												
sample4	100												
sample5	115												

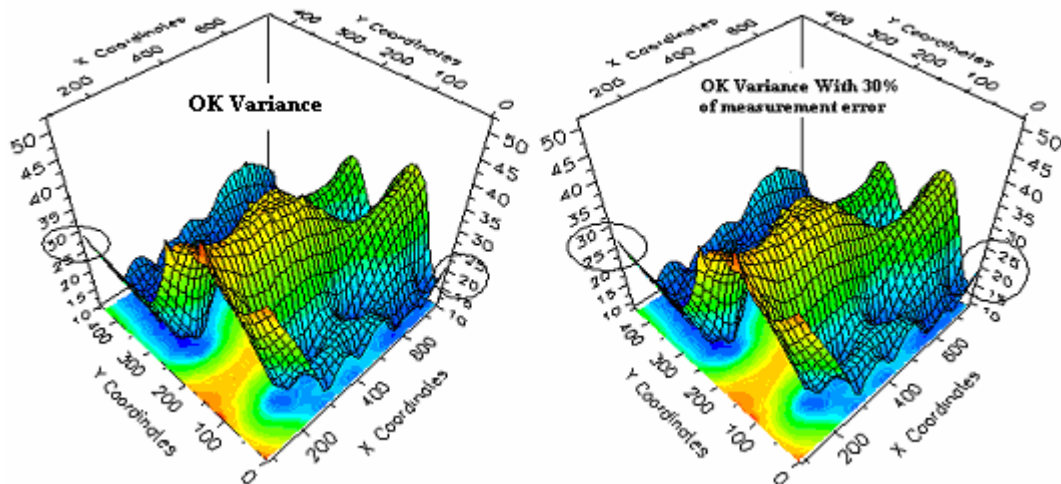
Figure 76 – The OK system of the working example.

The final prediction equals 102.8202 with a LaGrange factor of 0.2307. Since no samples have a zero distance from the interpolation site, the OK with measurement error does not have any impact on the **A** and **B** matrices. However, if the estimated site becomes the first sample, then both OK versions yield different results. Considering a 30% measurement error ($0.3 \times 2 = 0.6$), for instance, the **B** matrix and the final estimate becomes different, as Figure 77 shows.

$A^{-1} =$	0	32.19	29	42.5	78.37	1	=	-0.02392	0.006659	0.013706	0.006126	-0.00257	0.329226
	32.19	0	32.19	21.19	50.67	1		0.006659	-0.04069	0.009486	0.019637	0.004903	-0.18427
	29	32.19	0	50.67	62.37	1		0.013706	0.009486	-0.02445	-0.00358	0.004838	0.176614
	42.5	21.19	50.67	0	57.66	1		0.006126	0.019637	-0.00358	-0.02633	0.004141	0.265995
	78.37	50.67	62.37	57.66	0	1		-0.00257	0.004903	0.004838	0.004141	-0.01131	0.412432
	1	1	1	1	1	0		0.329226	-0.18427	0.176614	0.265995	0.412432	-42.8173
$B =$	0												
	32.187												
	29												
	42.5												
	78.368												
	1												
$W =$	0.999985816												
	0.000113306												
	-4.1178E-05												
	-7.0745E-05												
	1.28018E-05												
	-0.00044976												
OK Estimation=	100.000												
sample1	100												
sample2	105												
sample3	105												
sample4	100												
sample5	115												
$A^{-1} =$	0	32.19	29	42.5	78.37	1	=	-0.02392	0.006659	0.013706	0.006126	-0.00257	0.329226
	32.19	0	32.19	21.19	50.67	1		0.006659	-0.04069	0.009486	0.019637	0.004903	-0.18427
	29	32.19	0	50.67	62.37	1		0.013706	0.009486	-0.02445	-0.00358	0.004838	0.176614
	42.5	21.19	50.67	0	57.66	1		0.006126	0.019637	-0.00358	-0.02633	0.004141	0.265995
	78.37	50.67	62.37	57.66	0	1		-0.00257	0.004903	0.004838	0.004141	-0.01131	0.412432
	1	1	1	1	1	0		0.329226	-0.18427	0.176614	0.265995	0.412432	-42.8173
$B =$	0.6												
	32.187												
	29												
	42.5												
	78.368												
	1												
$W =$	0.985632768												
	0.004107719												
	0.008185204												
	0.003608176												
	-0.00153387												
	0.197263267												
OK Estimation=	100.236												
sample1	100												
sample2	105												
sample3	105												
sample4	100												
sample5	115												

Figure 77 – The OK system of the working example, considering with and without measurement error.

Note that the final interpolation of the exact version computes the sample itself (the OK variance equals zero), a situation not followed in the non-exact version (the OK variance becomes 0.5913). Also, the OK variance of the case above is reduced from 16.1235 to 15.5235, a generic situation confirmed by SAKWeb[®] with the 1993 grasshopper dataset (see Figure 78).



Sample ID	Sample Value	Estimation with 30% of measurement error	OK Variance
14	0	0.1548	1.3708
17	0.2	0.2933	1.7737
2	0.3	0.8854	2.2987
10	0.3	0.7692	1.5446
80	0.3	0.5360	2.0167
87	0.3	0.4703	1.9213
93	0.3	0.7093	1.9562
77	0.5	0.7159	2.0351
9	0.5	0.8193	1.6721
83	0.5	0.9083	1.9969
6	0.5	0.6498	1.5716
63	0.5	1.8643	2.0277
95	0.5	0.6204	1.8804
96	0.5	0.6083	1.8459
99	0.5	0.6658	1.9528
100	0.5	0.6462	1.9161
53	0.7	1.0509	2.3142
1	0.7	0.8127	1.9075
7	0.7	0.7577	1.5566
27	0.7	1.7538	2.0191
37	0.7	1.5269	1.7835
86	0.7	0.8140	1.9406
102	0.7	0.9878	2.5226
71	0.9	1.2541	2.1252
97	0.9	0.9129	1.8606
89	0.9	0.9714	1.9695
76	0.9	0.9630	2.1057
70	0.9	1.0814	2.1675
58	1.1	1.6589	1.8221
101	1.1	1.1509	2.2650
18	1.1	1.0403	1.6628
85	1.3	1.2741	1.9256
8	1.3	1.9904	2.5465
41	1.3	2.3852	1.9347
88	1.3	1.2823	2.0299
44	1.3	1.7501	1.7492
91	1.3	1.4339	2.0332
29	1.3	2.5035	1.9944
73	1.5	1.7967	2.0212
84	1.5	1.5350	1.9589
15	1.5	1.3621	2.0289
81	1.5	1.4469	2.0249
19	1.7	2.2074	1.9166
16	1.7	1.5718	2.2195
90	1.9	1.8293	1.9948
75	1.9	1.7331	2.1478
36	2.1	3.4031	1.9216
98	2.1	1.9310	1.9994
69	2.1	2.1507	2.1629
42	2.3	3.5415	1.9397
39	2.3	2.9580	2.1245
49	2.5	2.7784	1.9600

72	2.5	2.6033	2.0787	13	5.4	5.6418	3.8831
40	2.5	2.6490	4.6205	32	5.8	5.8612	2.0823
79	2.5	2.4627	2.0386	48	5.8	5.2959	1.9305
47	2.7	2.9586	1.8171	82	6.5	5.6052	2.0354
94	2.7	2.4471	1.9401	5	7.2	6.8489	1.0454
51	3.2	4.3494	2.0567	55	7.6	7.4993	2.1103
59	3.2	3.0968	2.7502	31	8.1	7.7117	2.0299
78	3.2	2.9085	2.0570	26	9	8.1973	2.0169
74	3.2	3.1818	1.9040	21	9	8.6746	2.1828
24	3.2	4.0792	2.1339	56	9.5	9.4817	2.0531
46	3.5	3.4861	1.7767	28	10	10.3082	2.0612
61	3.7	4.4208	2.0967	22	10	9.2501	2.0239
20	4.3	4.2566	1.9180	33	10.6	9.8648	2.1110
64	4.3	4.9173	2.0599	11	10.6	9.8145	3.1774
66	4.3	4.3688	2.0633	38	11.2	9.8095	1.7910
68	4.3	4.1451	2.1364	65	11.2	9.8808	1.9576
67	4.5	4.3662	2.1157	3	11.9	12.0857	4.2089
35	4.5	5.1598	2.1363	62	12.6	11.5010	2.0474
12	4.8	4.6785	1.2670	52	12.6	11.6842	2.0663
92	4.8	4.2002	2.0075	60	15.3	13.7783	2.0451
50	4.8	5.3441	2.1694	25	17.7	16.1566	2.1876
30	4.8	5.2357	1.9933	34	21	18.4134	2.0362
103	4.8	4.2924	4.0123	43	21	18.1261	2.0180
23	5.1	5.4839	2.0446	4	23.4	21.9415	2.2908

Figure 78 – Comparison between the exact version of SAKWeb[®] OK and the one with 30% measurement error³⁶⁶.

It is clear that the 30% measurement error considered has an impact on the OK estimation in the original observations: 58% of the estimates are higher than the original samples and, therefore, 42% are lower. According to Figure 79, no spatial patterns can be found for the measurement error effect but lower samples leads to higher estimations and vice-versa³⁶⁷:

- The sites with grasshopper infestation of less than 1.1 (thirty sites) gives higher OK estimates;
- The twenty-one samples that register densities higher than 5.8 produce lower estimations;
- Of the fifty-one observations whose values vary between 1.1 and 5.8, 55% have lower densities than the OK estimates.

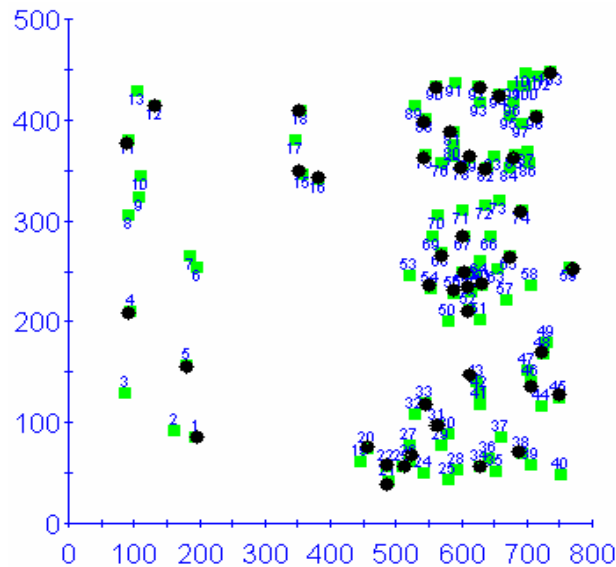


Figure 79 – The spatial location of the grasshopper densities in 1993 where the dark dots represent the sites whose OK estimates with a 30% measurement error are lower than the original samples.

7.2.4.1. Case Study: The GS+® Pb Dataset

The spherical model adopted was $C_0=0.0148$, $C_1=0.0319$, $\text{range}=73.9$ ($R^2=94.5\%$) but five variants were assessed³⁶⁸:

- No nugget-effect with higher sill (SAKWeb_1) $\Rightarrow C(\mathbf{h})=C(0)-\gamma(\mathbf{h})=0.0467-0-0.0467(1.5h/73.9-0.5h^3/73.9^3)=0.0467-0.0467(1.5h/73.9-0.5h^3/73.9^3)$ for $h \geq 0$.
- Nugget-effect with sudden jump (SAKWeb_2) $\Rightarrow C(\mathbf{h})=C(0)-\gamma(\mathbf{h})=0.0467-0.0148-0.0319(1.5h/73.9-0.5h^3/73.9^3)=0.0319-0.0319(1.5h/73.9-0.5h^3/73.9^3)$ for $h > 0$ and $C(\mathbf{h})=0.0467$ for $h=0$.
- Measurement error equals $25\%C_0=0.0037$ (SAKWeb_3) $\Rightarrow C(\mathbf{h})=C(0)-\gamma(\mathbf{h})=0.0319-0.0319(1.5h/73.9-0.5h^3/73.9^3)$ for $h > 0$ and $C(\mathbf{h})=0.0467-0.0037=0.043$ for $h=0$.
- No nugget-effect with lower sill (SAKWeb_4) $\Rightarrow C(\mathbf{h})=C(0)-\gamma(\mathbf{h})=0.0319-0.0319(1.5h/73.9-0.5h^3/73.9^3)$ for $h \geq 0$.
- Nested micro and macro structures (SAKWeb_5) $\Rightarrow C(\mathbf{h})=C(0)_1+C(0)_2-\gamma_1(\mathbf{h})-\gamma_2(\mathbf{h})=\gamma_1(\text{ssi})+\gamma_2(73.9)-\gamma_1(\mathbf{h})-\gamma_2(\mathbf{h})$ for $h > 0$ and $C(\mathbf{h})=\gamma_1(\text{ssi})+\gamma_2(73.9)$ for $h=0$. The chosen microstructure $\gamma_1(\mathbf{h})$ is an exponential one, $0.0161(1-$

$\exp(-h/2.08))$, whose sill equals $\gamma_2(ssi=2.08 \text{ units})=0.0161$. The chosen macrostructure is the spherical one of SAKWeb_2.

According to the OK results (see Table 21) and the log file created by the C program, several conclusions could be drawn:

- SAKWeb_1 and SAKWeb_4 show the same OK results but different OK variances (lower values are obtained with SAKWeb_4 because of the existence of a lower sill).
- SAKWeb_2 and SAKWeb_3 exhibit the same descriptive OK measurement. Interpolation values are only different if any original observations are included in the validation dataset.
- The lowest OK estimates presents more similar values among the five approaches than the higher estimates.
- No neighborhood pattern indications can be found with regard to which nugget-effect fits best (see Appendix O for spatial locations of the validation dataset).
- The correlation between the true error and the OK variance are negative for the estimates whose Pearson varies between -0.001 and -0.192.
- As expected, SAKWeb_2 and SAKWeb_3 reveal that most of their Kriging estimates are closer to the dataset mean (0.3756) than those of SAKWeb_1 and SAKWeb_4.
- The lowest error sum is achieved by SAKWeb_2 and SAKWeb_3 (3.34) followed by SAKWeb_1 and SAKWeb_4 (3.36). The worst result comes from the SAKWeb_5 (3.44) approach.

ID	X	Y	Pb	SAKWeb_1	SAKWeb_2	SAKWeb_3	SAKWeb_4	SAKWeb_5
1	7.8	3.5	0.37	0.57	0.48	0.48	0.57	0.48
2	9.7	51.1	0.47	0.48	0.42	0.42	0.48	0.42
3	11.6	32.7	0.01	0.51	0.43	0.43	0.51	0.44
4	14.6	68.8	0.3	0.34	0.4	0.4	0.34	0.39
5	24.6	46	0.34	0.28	0.28	0.28	0.28	0.3
6	27.7	48.9	0.37	0.24	0.27	0.27	0.24	0.28
7	31.1	34.3	0	0.19	0.22	0.22	0.19	0.23
8	39.1	11.9	0.23	0.19	0.18	0.18	0.19	0.19
9	35.4	20.9	0.15	0.21	0.22	0.22	0.21	0.22

10	36	77.8	0.23	0.34	0.39	0.39	0.34	0.41
11	43.5	1.8	0.15	0.22	0.2	0.2	0.22	0.21
12	41.8	47	0.27	0.36	0.32	0.32	0.36	0.33
13	51.8	38	0.2	0.32	0.34	0.34	0.32	0.34
14	56.1	48.8	0.2	0.48	0.47	0.47	0.48	0.48
15	60.4	65.5	0.65	0.28	0.57	0.57	0.28	0.49
16	65	70	1.25	0.83	0.69	0.69	0.83	0.69
17	86.6	79	0.52	0.83	0.68	0.68	0.83	0.65
18	72.3	29.8	0.59	0.52	0.45	0.45	0.52	0.45
19	70.5	35.3	0.58	0.52	0.46	0.46	0.52	0.48
20	70.3	52.7	0.33	0.53	0.57	0.57	0.53	0.56
21	70.1	7.3	0.51	0.38	0.38	0.38	0.38	0.39

Table 21 – The SAKWeb[®] OK estimations for the Pb dataset.

7.2.4.2. Case Study: The San Diego Dataset

Table 22 and Appendix O presents some additional results regarding the San Diego housing costs³⁶⁹. The exponential model implemented was $C_0=1220$, $C_1=5271$, range=5.91 ($R^2=80.8\%$) and, parallel to the case above, five modifications were computed:

- No nugget-effect with higher sill (SAKWeb_1) $\Rightarrow C(\mathbf{h})=C(0)-\gamma(\mathbf{h})=6497-6497(1-\exp(-h/5.91))$ for $h \geq 0$.
- Nugget-effect with sudden jump (SAKWeb_2) $\Rightarrow C(\mathbf{h})=C(0)-\gamma(\mathbf{h})=5277-5277(1-\exp(-h/5.91))$ for $h > 0$ and $C(\mathbf{h})=6497$ for $h=0$.
- Measurement error equals $25\%C_0=305$ (SAKWeb_3) $\Rightarrow C(\mathbf{h})=C(0)-\gamma(\mathbf{h})=5277-5277(1-\exp(-h/5.91))$ for $h > 0$ and $C(\mathbf{h})=6497-305=6192$ for $h=0$.
- No nugget-effect with lower sill (SAKWeb_4) $\Rightarrow C(\mathbf{h})=C(0)-\gamma(\mathbf{h})=5277-5277(1-\exp(-h/5.91))$ for $h \geq 0$.
- Nested micro and macro structures (SAKWeb_5) $\Rightarrow C(\mathbf{h})=C(0)_1+C(0)_2-\gamma_1(\mathbf{h})-\gamma_2(\mathbf{h})=\gamma_1(ssi)+\gamma_2(5.91)-\gamma_1(\mathbf{h})-\gamma_2(\mathbf{h})$ for $h > 0$ and $C(\mathbf{h})=\gamma_1(ssi)+\gamma_2(5.91)$ for $h=0$. The selected microstructure $\gamma_1(\mathbf{h})$ is also an exponential one, $2341.74(1-\exp(-h/1.41))$. The chosen macrostructure is the exponential one of SAKWeb_2.

The major inferences are:

- SAKWeb_1 and SAKWeb_4 present equal OK interpolation values. In addition, SAKWeb_1 and SAKWeb_4 exhibit the highest and the lowest interpolation range (260.35 and 101.2, respectively).
- Again, neighborhood patterns cannot be found regarding which approach fits best (see Appendix O).
- The Pearson indices between the true and the SAKWeb[®] estimates are: SAKWeb_1 and SAKWeb_4 (70.1%), SAKWeb_2 and SAKWeb_3 (72.4%) and SAKWeb_5 (74%).
- The lowest error sum is achieved by SAKWeb_5 (261.34), followed by SAKWeb_2 and SAKWeb_3 (264.46). The highest are SAKWeb_1 and SAKWeb_4 (297.67). In contrast to the case above, the sum of the two variogram structures fits better to this particular dataset.
- SAKWeb_2 and SAKWeb_3 also reveal Kriging estimates closer to the mean of this dataset (192.81) than SAKWeb_1 and SAKWeb_4.

Site	X	Y	Housing	SAKWeb_1	SAKWeb_2	SAKWeb_3	SAKWeb_4	SAKWeb_5
2	2	36	260	260.35	260.07	260.07	260.35	255.63
10	8	28	162	256.32	252.42	252.42	256.32	240.47
16	8	19	138	143.95	164.26	164.26	143.95	169.8
22	23	24	147	169.08	176.1	176.1	169.08	182.39
30	11	14	152	146.26	142.71	142.71	146.26	146.78
33	20	13	150	201.74	195.69	195.69	201.74	192.06
40	8	8	175	257.24	233.86	233.86	257.24	232.24
45	17	4	184	194.45	187.48	187.48	194.45	186.1
48	15	11	126	101.2	124.71	124.71	101.2	130.69

Table 22 – The SAKWeb[®] OK estimations for the San Diego housing costs.

7.3. SAKWEB[®] FUNCTIONALITIES

SAKWeb[®] is not a full geostatistical and simulation software package. Its functionality is depicted in Figure 80 (see also Figure 53 in subsection 7.1.4) on the basis of four critical procedures:

- Data Input and Exploring View, which focuses on MS-Excel[®] input, control management of the user session and exploratory data analysis.
- ESDA and Variography, which concerns variogram setup, the Moran I correlogram, the Moran location scatterplot and the Moran variance scatterplot. This includes an option to jump directly to a new menu with SK, UK, IK and CK.
- SAKWeb[®] Ordinary Kriging, which concentrates on OK calculus and surface mapping in accordance with four nugget-effect strategies. Validation with an extra dataset, 3D-2D surface profiles, cross-validation and region plumes based on threshold values and confidence levels are also included.
- SAKWeb[®] Help, which presents ten options regarding help with the software and additional e-Learning tools.

Reiterating what was said in subsection 7.1.1, it is significant to stress that SAKWeb[®] implements its own authentication mechanism based on the default form of authentication. If the user does not login with the Guest username and a password then the logon process is refused. Although SAKWeb[®] has no future strategy regarding the user access, technically, this control can easily be implemented by using the current internal database control. Also, the default home page has the capability to check the browser version of the user and check it according to the following two rules: 1) if the Web access is made with IE[®] (version 5.0 or greater) then the user may use the site; 2) Otherwise, an error message is displayed.

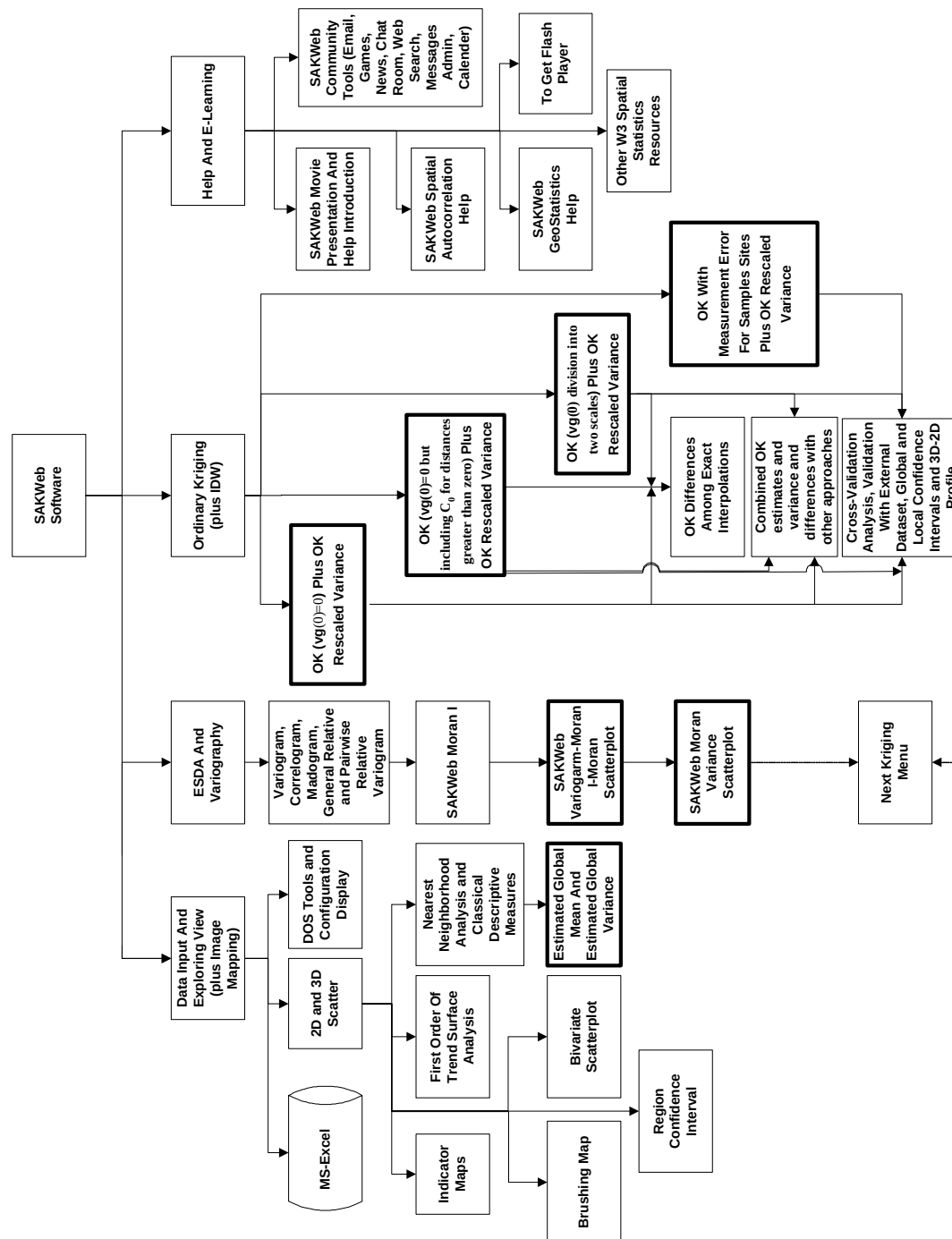


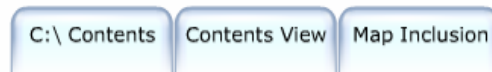
Figure 80 – The SAKWeb[®] overall structure.

7.3.1. Data Input and Exploring View

The first task regards the direct capability to retrieve data from MS-Excel[®], for which the dataset must be defined by the Insert/Name/Define option (see Figure 81). The records to be processed must follow the next pattern: Sample ID,

Coordinate x, Coordinate y and Sample value where each cell of each row separates each field. Notice that SAKWeb[®] only processes datasets with less than 500 samples.

SAKWeb Excel Data Input



Complete path of MS-Excel data file:

Definition region cells name:

Warning: The MS-Excel datasheet must hold a definition region setup by the Insert/Name/Define option (see Figure below).

Warning: The MS-Excel definition must hold the following data name fields: sample_id (autonumber), coord_x (number), coord_y (number), sample_value (number)

Warning: The Bivariate Scatterplot and the Non-Exact OK With Measurement Error options are not available for datasets greater than 250 samples.

Warning: No option is available for datasets with more 500 observations.

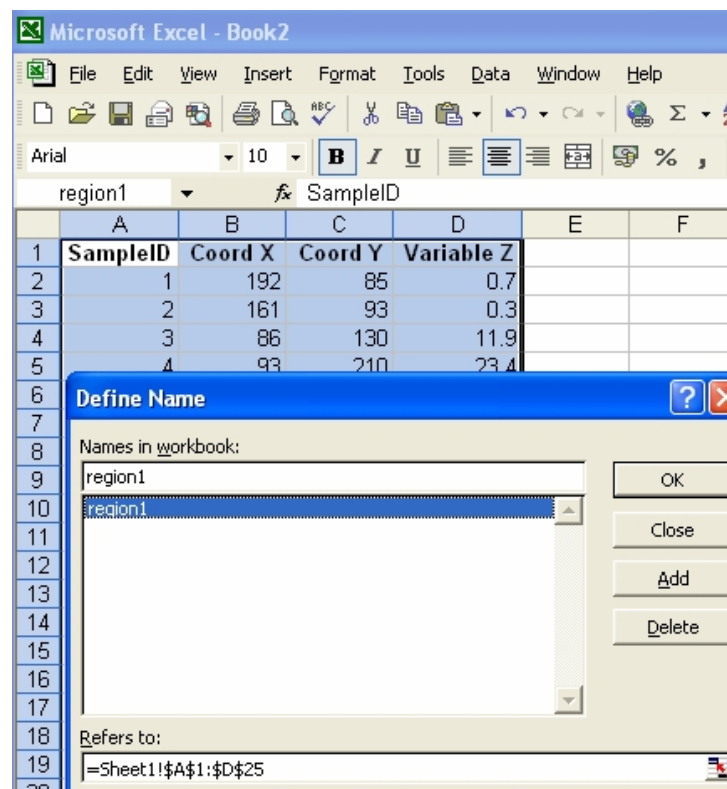


Figure 81 – The MS-Excel[®] data input selection. In addition, a movie was added for beginner's users.

There are three other related features to data input, based on three Flash[®] buttons:

- C:\Contents - The capability to browse the contents of the local hard disk in a new window.
- Contents View – The facility to examine the actual contents of the records.
- Map Inclusion – The potential to include an image (.jpg, .wmf, .bmp or .gif) of the samples site to enhance the realism of the final OK interpolation (see Figure 99 of Section 8.3).

Control management is achieved by means of three options:

- DOS Commands - The capability to browse the contents of DOS[®] commands (see Figure 50 of subsection 7.1.2).
- PHP Configuration - The ability to display the PHP[®] configuration of SAKWeb[®] (see Figure 51 of subsection 7.1.2).
- Session Variables – It gives the CGI parameters and the contents of the session variables created in the current user session, a useful tool for the programmer to control the application flowchart. It also plots a horizontal 3D graphic with the distribution of the visitors to this site throughout the day by dividing it into six groups: 0h-4h, 4h-8h, 8h-12h, 12h-16h, 16h-20h and 20h-24h. Its aim is to give clues to Web users what is the best time to connect SAKWeb[®] (see figure 82).

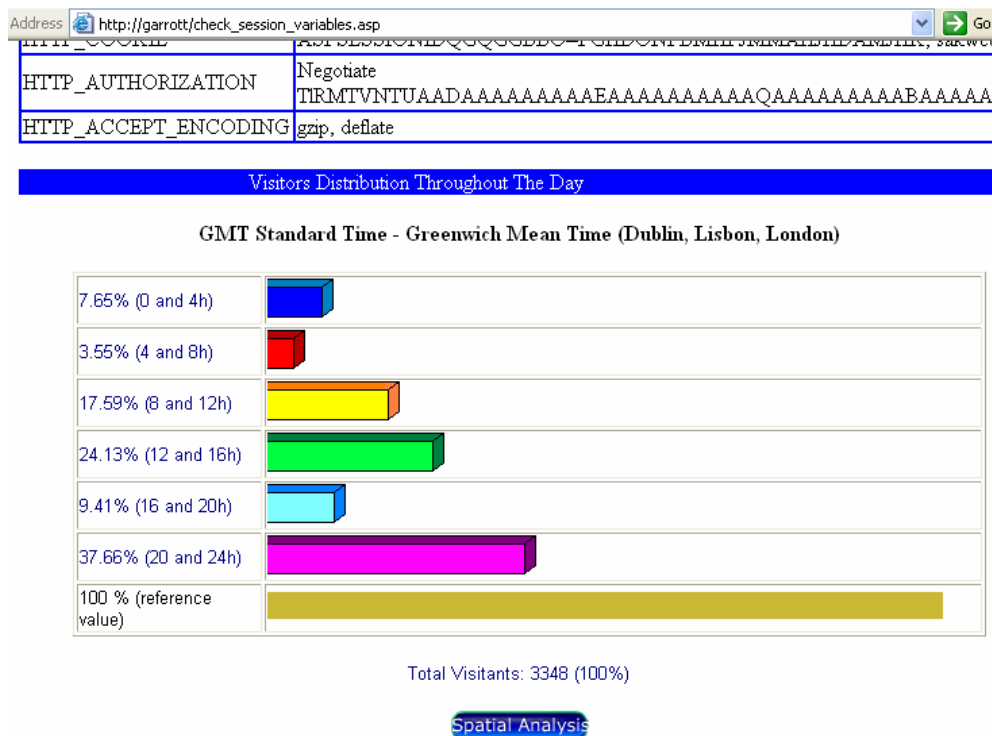


Figure 82 – The visitor distribution graphic of SAKWeb[®].

After the correct input is made, the spatial analysis Flash[®] button leads the user to the main menu. The Data Input and Exploring View offers seven options:

- 1st Order Trend Surface with Join Count Residuals – The general linear trend model is shown in a 3D perspective embracing nine statistics: the b0, b1 and b2 parameters of the linear polynomial regression, total variance, regression variance, residual variance, adjusted R^2 , R^2 and F test (see Figure 54 of subsection 7.1.4). Moreover, the residuals scatterplot with the join-count statistics and discrimination between the positive and negative residuals are depicted in a 2D scatterplot. Note that this option and the following ones have a Flash[®] button that creates an additional window with the site image mapping, setup by the MS-Excel Input selection³⁷⁰.
- Local Interactive Statistics – With the control mouse, as Figure 82 shows, SAKWeb[®] holds the capability to select a continuous subset of points to obtain the local average, the local standard deviation, the local coefficient of variation and the local kurtosis, a significant aptitude to check local heteroskedasticity within the dataset. Note that if the selected dataset is less than three samples then a warning message is displayed.

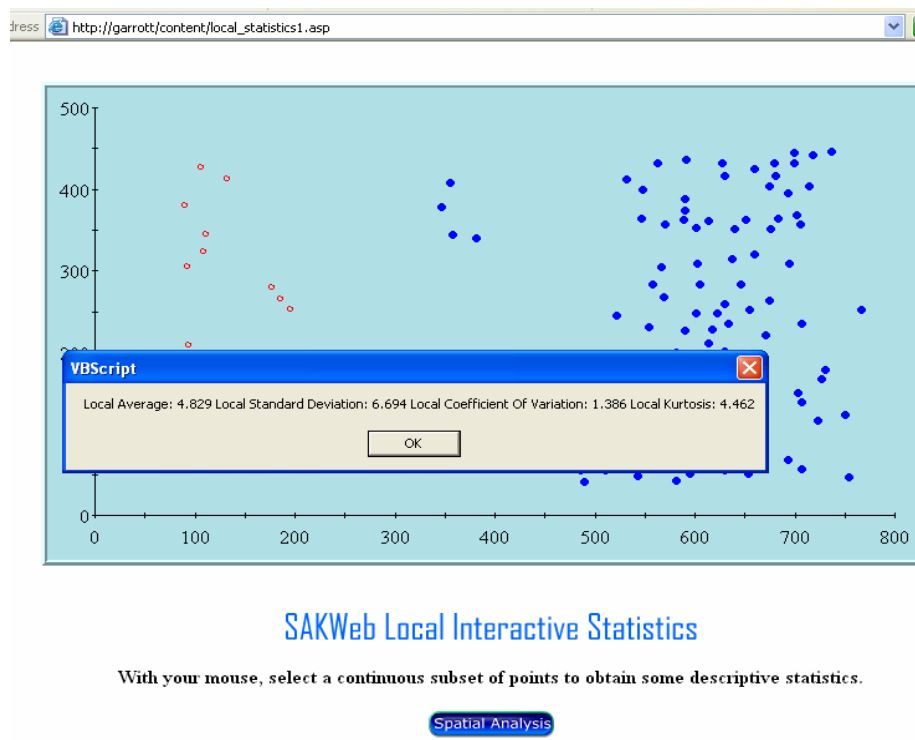


Figure 83 – The local interactive statistics option of SAKWeb[®].

This major technical innovation and capability to select a continuous subset of points to obtain local statistics in a WWW environment was only possible with the merging of server-side and client-side implementations. That is, the ASP[®] server-side produces the dynamic code of VBScript[®], a client-side technology, to be interpreted by the browser. This happens because the original data must be exported to the client in an active way. According to Table 23, two features are enhanced: 1) the WebChart[®] CLSID with its Licence Pack file allows a 2D map interface to be created in the client's browser; 2) a single string called sScript is generated by the ASP[®] code, while the last instruction, response.write sScript, is responsible for the browser script.

```
<html><head><title>Local Statistics</title></head>
<OBJECT classid="CLSID:5220CB21-C88D-11CF-B347-00AA00A28331" VIEWASTEXT>
<PARAM NAME="LPKPath" VALUE="OLCHDemo.lpk"></OBJECT>
<OBJECT CLASSID="CLSID:92D71E93-25A8-11CF-A640-9986B64D9618" CODEBASE=
OLCHDemo.cab ID=Chart2D1 WIDTH=473 HEIGHT=352 VIEWASTEXT></OBJECT>
<body>
<% sScript="<SCRIPT LANGUAGE='VBScript'>" & vbCrLf & _
```

```

"<!--" & vbcrLf & _
"Dim dPlotLeft, dPlotTop, dPlotRight, dPlotBottom" & vbcrLf & _
"Dim lSymbolStartShape, lSymbolStartSize, lSymbolStartColor" & vbcrLf & _
"Dim lSymbolNewShape, lSymbolNewSize, lSymbolNewColor" & vbcrLf & _
"Dim MouseDX, MouseDY, MouseUX, MouseUY, resultado_x, resultado_y, samplesvalues1" &
vbcrLf & _
"Dim aX0, aX1, aY0, aY1, linha, Numpoints, resultado" & vbcrLf & _
"Dim lNumSelPoints, x" & vbcrLf & _
"Sub Window_OnLoad()" & vbcrLf & _
"  Dim l, C, i" & vbcrLf & _
"  numpoints = " & session("numrows") & vbcrLf & _
...
"  With Chart2D1" & vbcrLf & _
"    .AllowUserChanges = True" & vbcrLf & _ ...
"    .Axes(2).Max = aY1"& vbcrLf & _
"  End With"& vbcrLf & _
"End Sub"& vbcrLf & _
"-->"& vbcrLf & _
"</SCRIPT>"
response.write sScript %>
</body></html>

```

Table 23 – The ASP[®] (server-side) code that produces the VBScript[®] (client-side) one.

- **Descriptive Measures and Nearest Neighborhood Analysis** – This displays the histogram and the cumulative one, the univariate descriptive measures (number of points, average, error of the average, variance, standard deviation, maximum, median, medium, mode, coefficient of variation, absolute average of deviation, first, second and third quartile, skewness, Pearson and Kurtosis), nearest neighborhood analysis (EGM and EGV by the nearest analysis, coordinate x and y of mean center and weighted geographical centroid, standard deviation distance, region area, region density, standard deviation ellipse length and its angle, mean nearest distance, standard error of the mean nearest neighbour distance, expected mean nearest distance for a random arrangement, expected nearest distance for a perfect uniform point pattern situation and nearest neighborhood index), the box-and-whisker plot and the 2D posting (see Appendix E). It also comprises a Flash[®] button to review the original samples in a new

window and an HTML form button to assess the anisotropy calculus via a PHP[®] file.

- 3D Data Posting – This 3D view includes drop lines and X-Y-Z axis grid lines. According to its value, data observations are plotted with three different symbols: below 25%, between 25% and 75% and above 75% (see Figure 84).

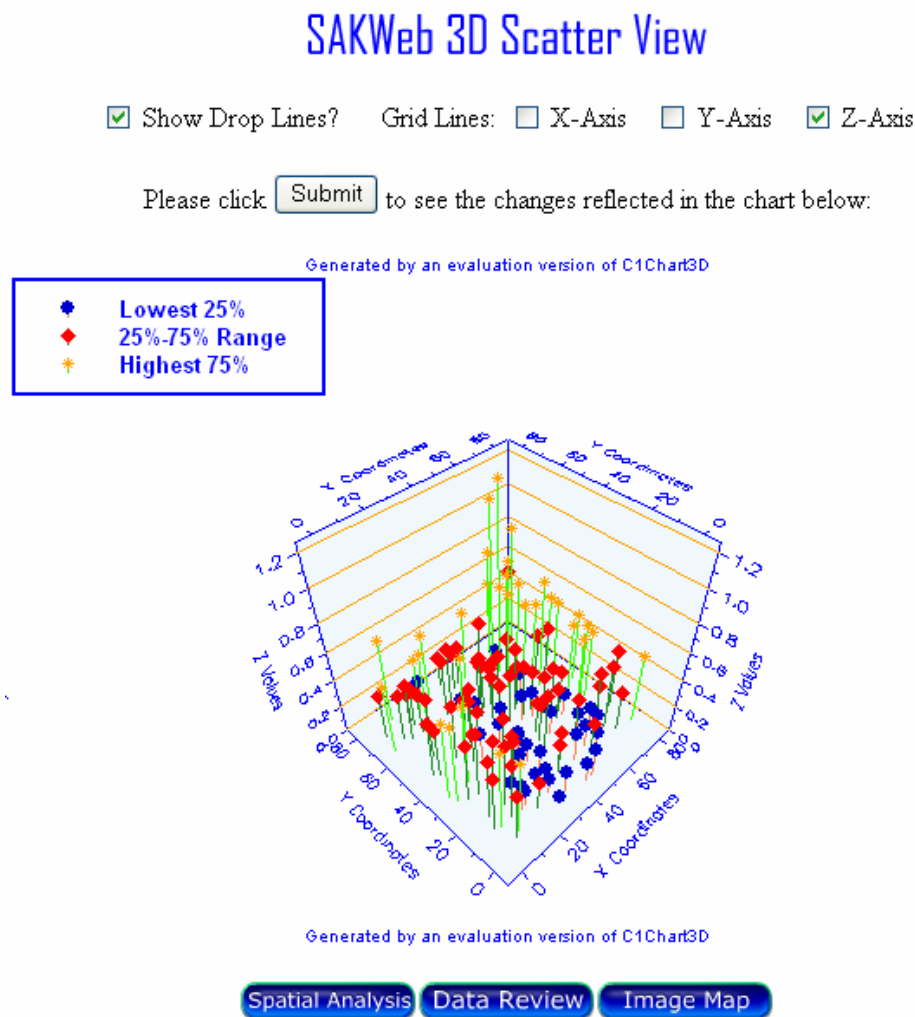


Figure 84 – The 3D scatter view of SAKWeb[®].

- Bivariate Scatterplot – This displays the samples values along the x and y coordinates in order to verify where high, medium and low values are located (see Appendix E).
- Exit – Logs off the current user by calling the SAKWeb[®] home page.

- **Indicator Mapping for Samples** – Centered on the nearest weighted and conventional cumulative histograms. The user can specify a threshold value to reveal the observations that are above and below the given cutoff (see Figure 85).

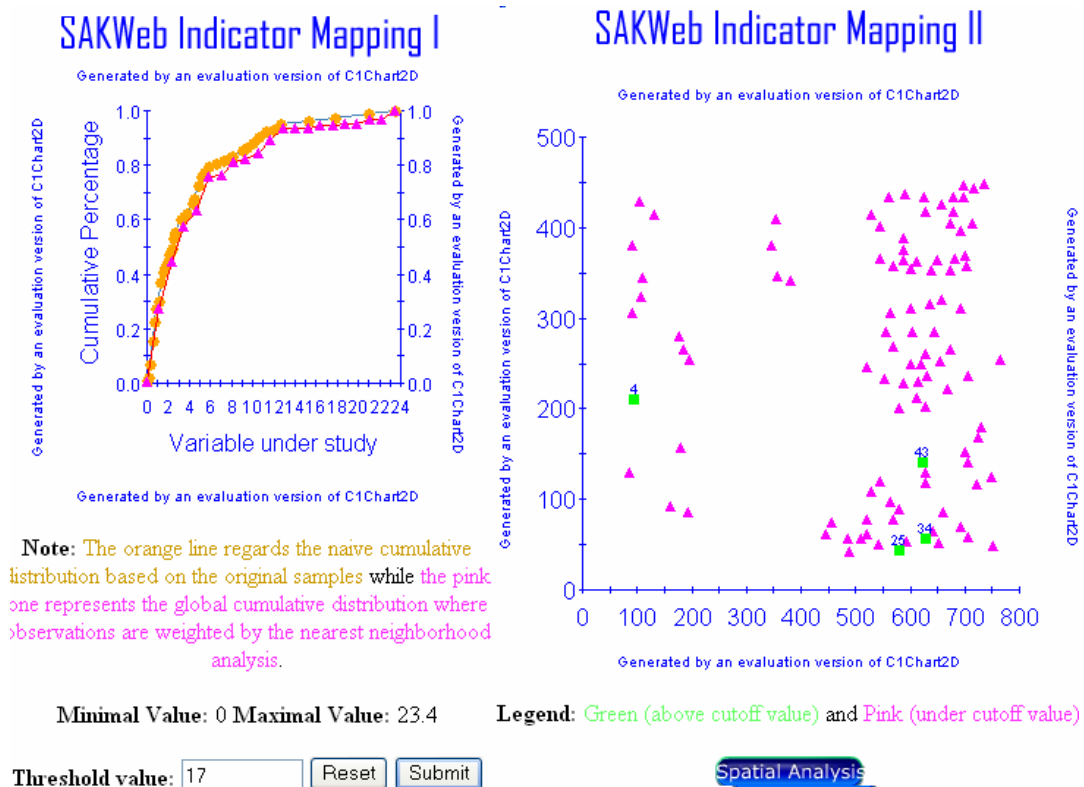


Figure 85 – The SAKWeb[®] indicator mapping in a two step process.

7.3.2. Variography and Spatial Autocorrelation

The second menu focuses on the Variography and Spatial Autocorrelation menu. It comprises:

- The variogram Setup – This option requires six parameters (lag distance, tolerance lag, number of lags, horizontal bandwidth, principal axis angle and tolerance angle) to produce the experimental variograms by means of a PHP[®] file whose fragment is shown next:

<?php

```
$comando = "c:\\inetpub\\wwwroot\\executable\\gamv.exe"
```

```
c:\inetpub\wwwroot\data\".$ COOKIE["sakwebsessionId"]."0.par";
```



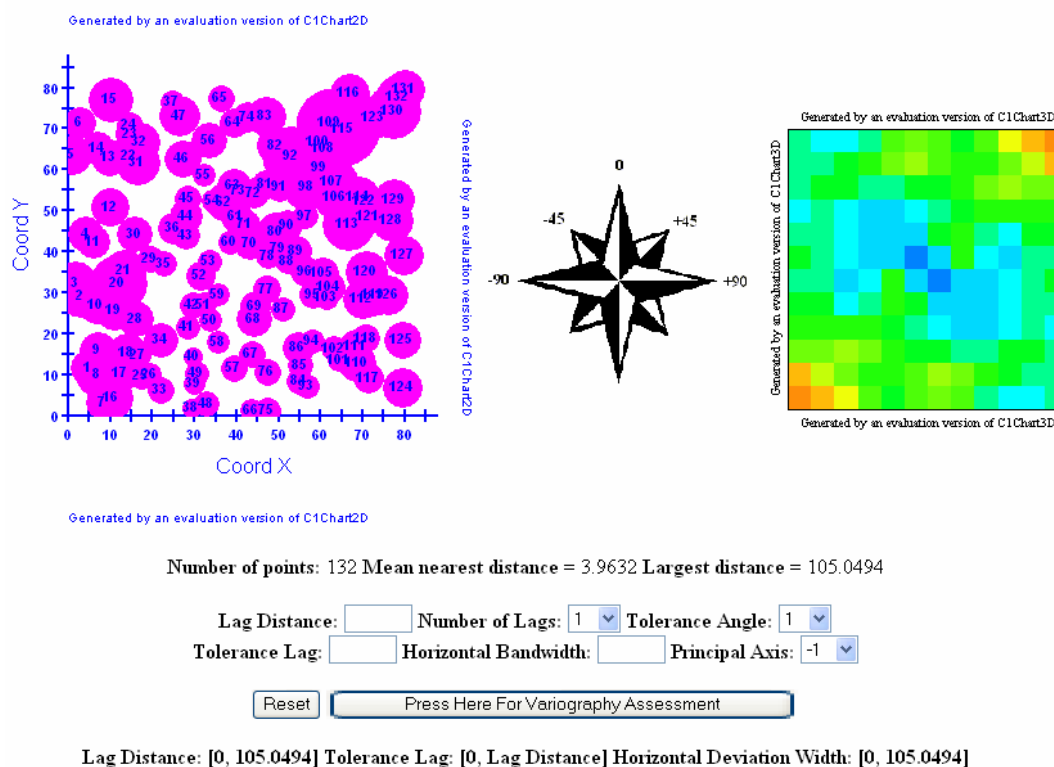
```

echo $comando;
$line1 = exec(escapeshellcmd($comando), $output, $error);
while (list($line) = each($output)) {
    echo $line, "<br>"; }
if ($error) {
    echo "<p>Error Code: $error<br>\n";
    echo "<p>Please, exit your browser and restart again SAKWeb.<br>Remember that this application
    uses Cookies and you can not run two sessions at the same time.";
    exit; }
?>

```

In order to help the common user to setup these factors, different types of advice and assistance are given relating, for example, to 2D data posting, the compass card, the anisotropy map, the mean nearest distance and the largest distance between points (see Figure 86). It is essential to emphasize that beginner' notes and input validation are also offered. Thus, the lag distance and the horizontal deviation width can not exceed the largest distance between samples or fall below zero. Nor can, the tolerance lag exceed the lag distance setup by the user.

SAKWeb Variogram Parameters Setup



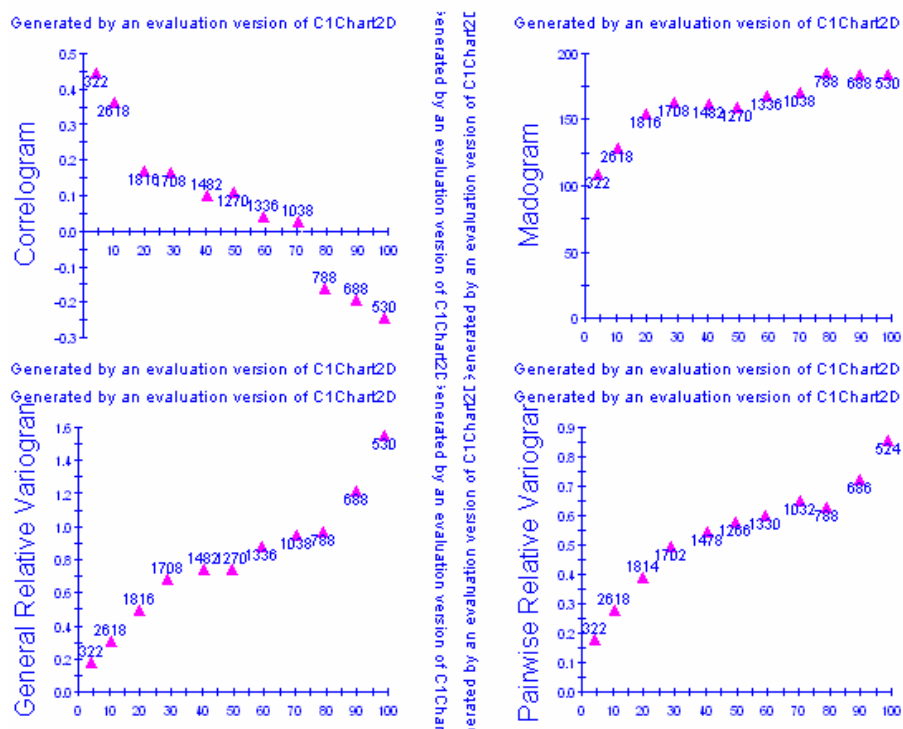
Beginners Notes:

- I. After pressing the submit button, an open window will pop-up in your browser because a batch file must be run. **Press Open option.**
- II. The lag distance represents the variogram class within which each data pair is separated by the same separation distance. Decreasing its size, it increases the number of points on the variogram and thereby its resolution. At the same time, narrowing the interval also decreases the number of data pairs within each class which decreases the likelihood that the average semivariance for that class is accurately estimated.
- III. As a default value, the number of lags might be between **eight and ten**.
- IV. The active lag distance specifies the range over which semivariance will be calculated. The minimum distance for this field is the minimum distance between adjacent points in the data set, while the maximum distance is the maximum distance between points. Specifying a value too large or too small will assign respectively the largest or smallest possible values to this distance. A good reference start could be **80% of the maximum distance among points** (84.0395). Hence, the number of lags multiplied by the lag distance should not exceed it.
- V. An acceptable default value regarding the **horizontal bandwidth** or maximum acceptable horizontal deviation from the direction vector is the **lag distance itself**.
- VI. **The lag tolerance parameter of one half of lag distance** is an excellent initial value.
- VII. The azimuth tolerance determines how closely the alignment between any two points needs to be for those points to be included in the analysis for a given offset angle. Two points will be included in the analysis for a given offset angle if the angle between them is within the offset tolerance of the offset angle. **The angle tolerance can vary between 1° and 90° (omnidirectional one) although 23°** is an acceptable starting point to avoid overlap points among different variogram directions.
- VIII. The last factor defines the appropriate **principal axis (longest continuity)** for defining the anisotropic variogram model. Its range varies between 90 degrees (East-West) to -89 degrees. Naturally, 0 degrees means North-South.

Figure 86 – The SAKWeb[®] experimental variogram setup.

- **Variogram Fitness** – The 0°, 45°, 90° and 135° experimental variograms are depicted with this selection in Part I (see Appendix E). The pairwise relative variogram, the general relative variogram, the correlogram and the madogram are also displayed in Part II (see Figure 87). As expected, five parameters are asked for in order to fit the omni-directional variogram: major and minor range, nugget-effect, sill and variogram type (spherical, Gaussian and exponential). After the submit button is pressed, the red line changes to a new curve according to the chosen parameters in an interactive process (see Table 32 in Section 8.7). Furthermore, a variogram fitness indicator is presented. It is the Save button that confirms the end of this process and saves the final variogram parameters internally.

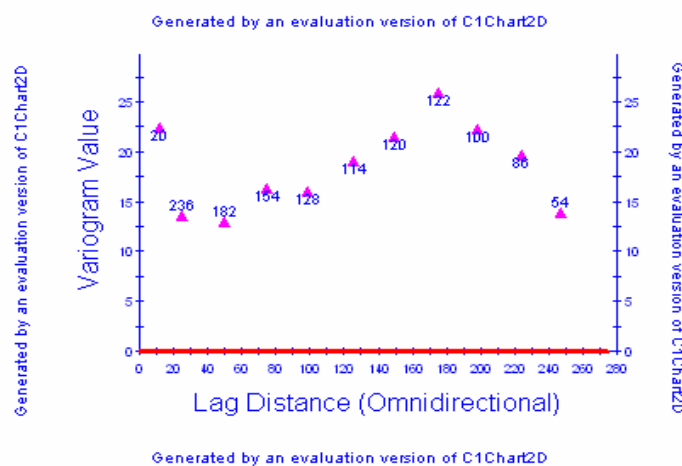
SAKWeb Variogram Fitness Setup (Part II)



Variogram Major Range: Variogram Minor Range:
 Variogram Nugget: Variogram Sill:
☐ Exponential ☐ Gaussian ☐ Spherical

Please click to see the changes reflected in the chart below.

Variogram Range: [0, 275] Variogram Sill: [0, 25.86394] Variogram Nugget: [0, Sill Input]



Variogram Fitness Indicator: 3,955.395050

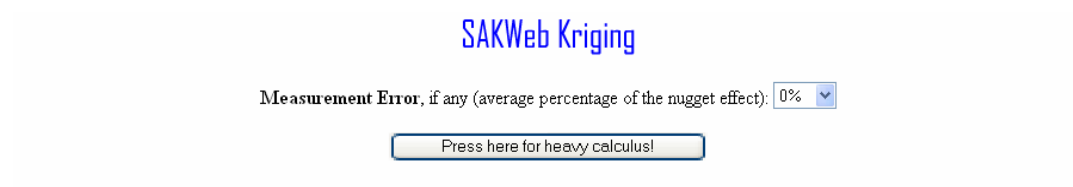
Figure 87 – A partial view of the SAKWeb® variogram fitness process.

- SAKWeb[®] Moran I Analysis – This depicts the Moran I correlogram for several distances (see Table 29 in Section 8.5) and determines the distance lags with the two highest Moran I results. These values are saved internally as session variables since both are needed for other operations.
- SAKWeb[®] Moran–Variogram-Moran scatterplot – This option displays the conventional Moran scatterplot and the Moran location scatterplot based on the algorithm in subsection 7.2.1.7. The legends relating to the different symbols in each quadrant is shown (see Figures 62 and 63 of subsection 7.2.1.7). A Flash[®] button is also included to create a new window that organizes samples according to their classification group.
- SAKWeb[®] Moran Scatterplot Variance – This item plots the samples-neighborhood variance versus neighbors-neighbors variance on the basis of the theory presented in subsection 7.2.2. A 2D scatter view is also shown where the colours black (high variability region), green (medium central-neighbors versus low central-neighbors variance), yellow (outliers) and blue (stable region) highlights different conditions for the samples (see Figure 66 in subsection 7.2.2.1).
- Next Kriging Menu – This leads the user to a new set of choices that includes Simple Kriging (SK), Universal Kriging (UK), Indicator Kriging (IK) and CoKriging (CK)³⁷¹.

7.3.3. SAKWeb[®] (Ordinary) Kriging

The Ordinary Kriging menu follows next. The first option generates the four OK approaches by emphasizing the nugget-effect factor. The IDW global interpolation is shown just to give an idea of how the phenomena behave in space (see Figure 88). The Default OK, OK with no C0 and the OK with Two Structures selections generate the mapping for interpolation and associated variance (see Tables 35, 36, 37 and 38 in Section 8.4). The capability to rotate, to zoom, to reset and to join the 2D and 3D surfaces together in a W3 environment is a remarkable step for future spatial data analysis software. Another implementation issue based on subsection

7.2.4 involves the Non-Exact OK with Measurement Error option where the OK variance surface is shown. Since the original samples interpolate different results against the exact OK version, another table is laid out with the following information: sample ID, sample value and estimation, including the defined measurement error and the respective OK variance.



SAKWeb[®] Kriging approach for nugget-effect is significant. Hence, four default interpolation strategies are presented:

1. **Variogram(0)=0 and variogram(h->0)=C0** (discontinuity pattern)
2. **Variogram(0)=0** (continuity pattern)
3. **Nugget-effect (variogram(0)=0 and variogram(h->0)=C0) with two structures** (micro and long range):
 1. Variogram($0 \leq h_1 \leq \text{shortest sampling interval}$)=Spherical, Exponential or Gaussian model whose sill equals second structure model at ssi lag distance;
 2. Variogram($0 \leq h_2 \leq \text{range}$)=Spherical, Exponential or Gaussian model
4. **Measurement Error** (instrument cause, for instance). Thus, Variogram(0)=Cme=%*C0 (a non-exact Kriging version).

The Global Deterministic Interpolation of Inverse Distance Weighting

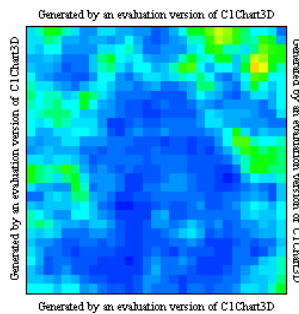


Figure 88 – SAKWeb[®] OK Calculus option³⁷².

The Exact OK Differences option displays the interpolation differences among the three methods in 3D view with the following four statistics (see Appendix E): average difference, maximum and minimum disparity and average OK variance.

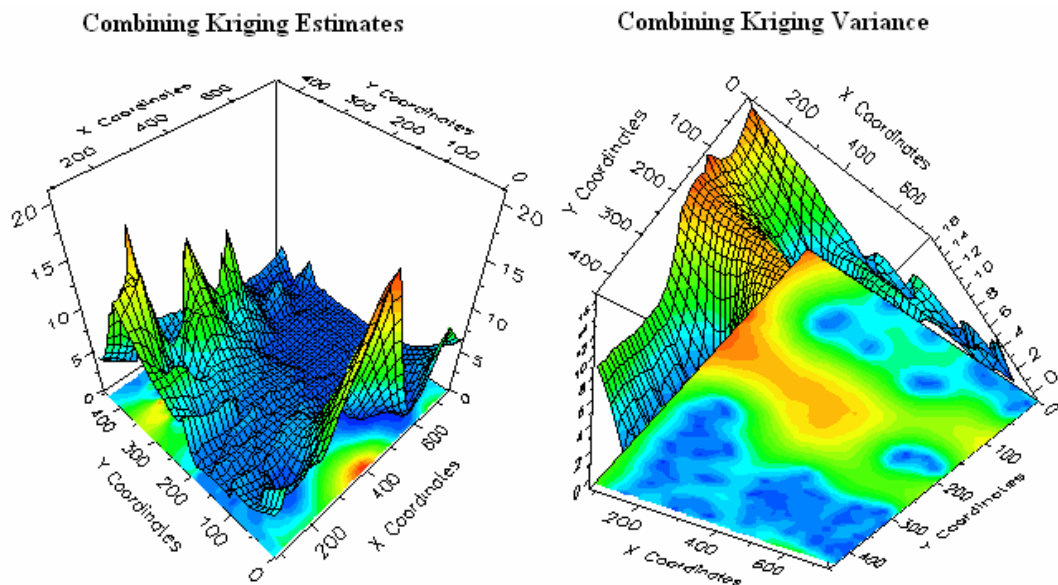
The combined Kriged estimates option is a weighted averaging of the above three exact OK models. It is based on the variance of the Kriged estimates, giving more weight to those values with smaller estimation variance. “Variances will also be combined in order to have a final Kriging estimate value a final variance for each block” [Bryan, 1994]. Equation 83 shows the process of weighting the Kriging

values for the three datasets for certain block $b(x_1, y_1)$ with three Kriging estimation variance values: K_{m_1} , K_{m_2} , K_{m_3} and three variances V_{m_1} , V_{m_2} , V_{m_3} .

$$b(x_1, y_1) = K_{m_1} \left(\frac{\frac{1}{V_{m_1}}}{\frac{1}{V_{m_1}} + \frac{1}{V_{m_2}} + \frac{1}{V_{m_3}}} \right) + K_{m_2} \left(\frac{\frac{1}{V_{m_2}}}{\frac{1}{V_{m_1}} + \frac{1}{V_{m_2}} + \frac{1}{V_{m_3}}} \right) + K_{m_3} \left(\frac{\frac{1}{V_{m_3}}}{\frac{1}{V_{m_1}} + \frac{1}{V_{m_2}} + \frac{1}{V_{m_3}}} \right) \quad [83]$$

The same method is used to calculate the variances (see Equation 84). Notice, in Figure 89, the comparison Flash[®] button that shows the differences between combined and exact OK methods in a two step process. The possibility to cross-validate the combined Kriged estimates is not possible because the Kriging variance equals zero.

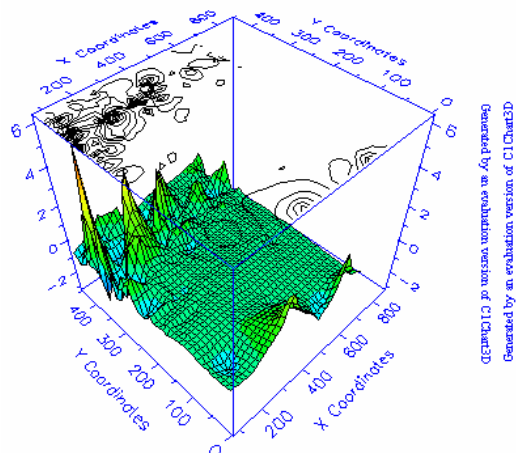
$$V(x_1, y_1) = V_{m_1} \left(\frac{\frac{1}{V_{m_1}}}{\frac{1}{V_{m_1}} + \frac{1}{V_{m_2}} + \frac{1}{V_{m_3}}} \right)^2 + V_{m_2} \left(\frac{\frac{1}{V_{m_2}}}{\frac{1}{V_{m_1}} + \frac{1}{V_{m_2}} + \frac{1}{V_{m_3}}} \right)^2 + V_{m_3} \left(\frac{\frac{1}{V_{m_3}}}{\frac{1}{V_{m_1}} + \frac{1}{V_{m_2}} + \frac{1}{V_{m_3}}} \right)^2 \quad [84]$$



SAKWeb OK Differences Between Combined and Exact Versions

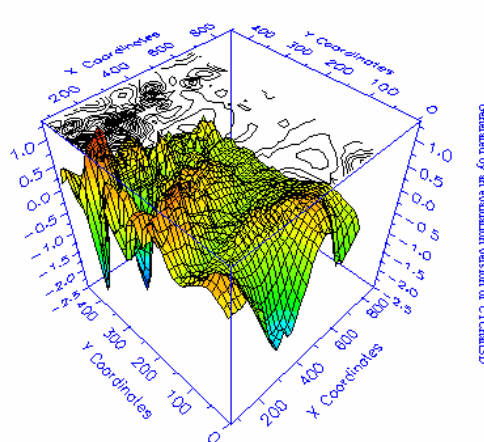
OK difference between Combined and First approach

Generated by an evaluation version of C1Chart3D



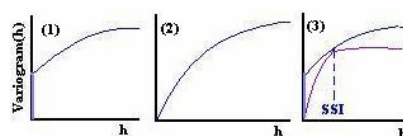
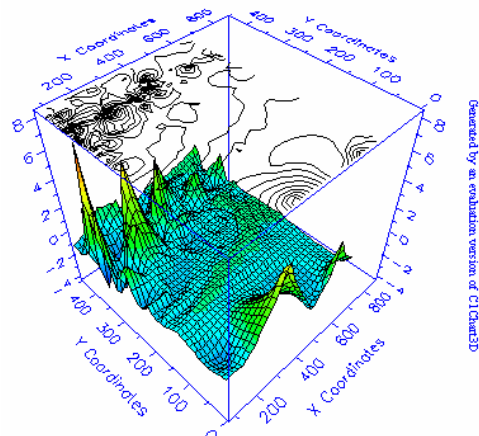
OK difference between Combined and Second approach

Generated by an evaluation version of C1Chart3D



OK difference between Combined and Third approach

Generated by an evaluation version of C1Chart3D



Combined OK Variance Average: 6.35978

Kriging Combined Estimation

$$K_{m,1} \left(\frac{\frac{1}{V_{m,1}}}{\frac{1}{V_{m,1}} + \frac{1}{V_{m,2}} + \frac{1}{V_{m,3}}} \right) + K_{m,2} \left(\frac{\frac{1}{V_{m,2}}}{\frac{1}{V_{m,1}} + \frac{1}{V_{m,2}} + \frac{1}{V_{m,3}}} \right) + K_{m,3} \left(\frac{\frac{1}{V_{m,3}}}{\frac{1}{V_{m,1}} + \frac{1}{V_{m,2}} + \frac{1}{V_{m,3}}} \right)$$

Kriging Estimation Variance

$$V_{m,1} \left(\frac{\frac{1}{V_{m,1}}}{\frac{1}{V_{m,1}} + \frac{1}{V_{m,2}} + \frac{1}{V_{m,3}}} \right)^2 + V_{m,2} \left(\frac{\frac{1}{V_{m,2}}}{\frac{1}{V_{m,1}} + \frac{1}{V_{m,2}} + \frac{1}{V_{m,3}}} \right)^2 + V_{m,3} \left(\frac{\frac{1}{V_{m,3}}}{\frac{1}{V_{m,1}} + \frac{1}{V_{m,2}} + \frac{1}{V_{m,3}}} \right)^2$$

Figure 89 – Combination of Kriging estimates from OK with C0 (model 1), OK without C0 (model 2) and with two structures (model 3) for the grasshopper 1993 dataset. This includes differences with other SAKWeb® OK results. Notice that the OK variance difference among the four approaches: the combined version rounds 66% less when compared with the other three.

As Figure 90 depicts, the kb2d.exe GSLib® routine assesses these three OK versions through a PHP® file³⁷³ that uses cookies technology. Cookies are small text files saved by the user browser through the header of an HTML file. Since ASP® and PHP® are able to write³⁷⁴ and read³⁷⁵ them, SAKWeb® creates a cookie called sakwebsessionid which contains the user session identification³⁷⁶. Therefore, the possibility of sharing this crucial information between both technologies becomes real and is central to setting up the parameter file names.

The Global Region Confidence Interval option simulates the minimum and maximum plume based in the rescaled OK variance³⁷⁷ for a particular confidence level (80%, 90% and 95%). This option also determines the largest and the smallest region, for the same confidence level, on the basis of a particular cutoff limit (see Figure 70 in subsection 7.2.3), although the smoothing effect of Kriging estimates can be inadequate for high and low threshold choices. The Local Region Confidence Interval is computed on the basis of the theoretical concepts of subsection 7.2.3.4.

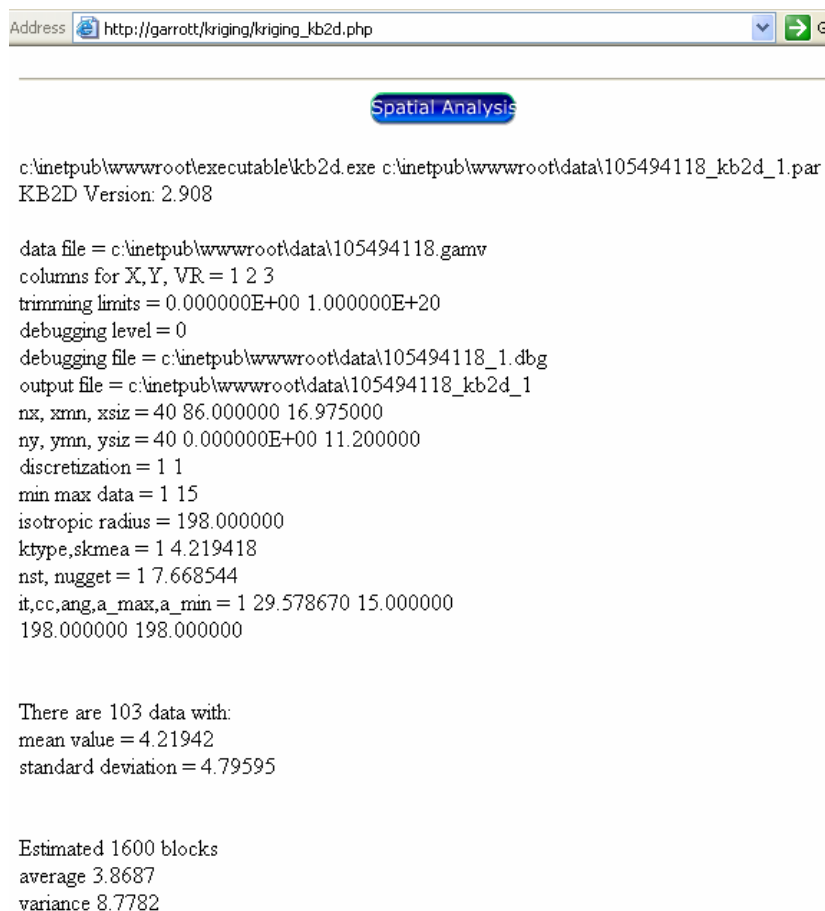


Figure 90 – Partial view of SAKWeb[®] OK assessment based on GSLib[®] routines.

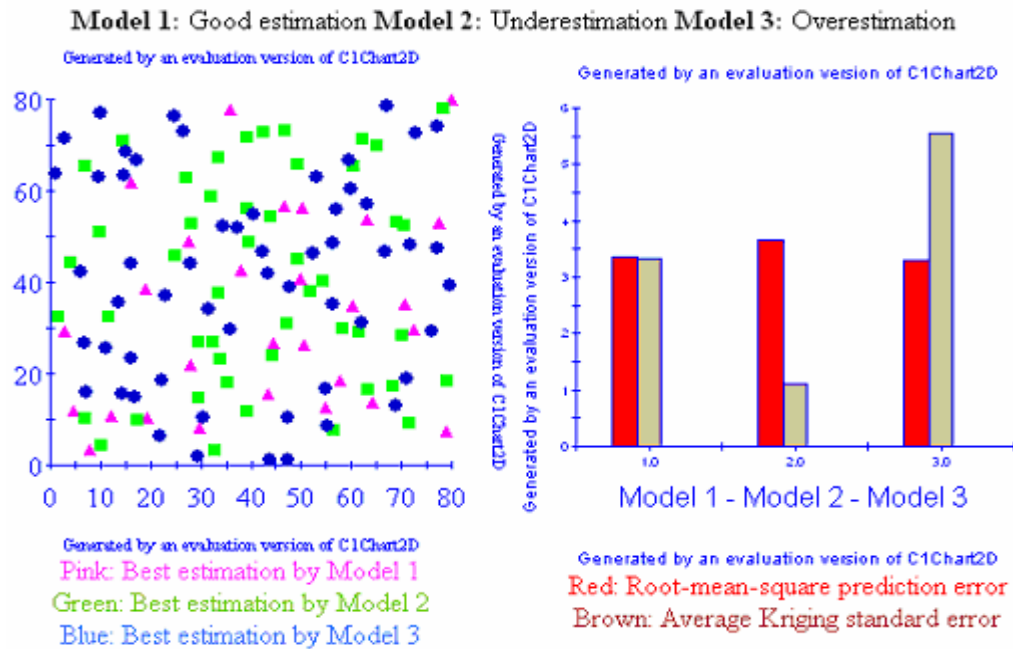
The possibility of validating these methods against an extra dataset or of cross-validating the original samples can be achieved with the Validation With Extra Dataset and Cross-Validation options. The former method requires an additional dataset to test the security of our estimations (see Appendix E) and its input file must show the following pattern for each row: sample ID, coordinate x, coordinate y and true sample value. Fields are separated by a single space and no other type of

lines should be inserted before or after. The positive, negative and total error sum for each OK is also assessed.

The latter approach uses the original samples to check the estimation uncertainty by dropping one observation each time the re-estimated OK process is computed. On the basis of the theory in subsection 3.5.1, SAKWeb[®] includes:

- A comparison of the tables for the estimation error and the estimation variance among the three exact OK models.
- A message for each OK model, according to the root-mean-square prediction error and the average Kriging standard error. If both factors are similar ($\pm 5\%$ difference) then 'good estimation' is displayed. If the first statistic is greater than the second one then 'under-estimation' is shown. Otherwise, 'over-estimation' appears.
- A location map that reveals which model presents the lowest error.
- A bar graph that compares the root-mean-square prediction error and the average Kriging standard error for each model.
- The possibility of visually describing false positive³⁷⁸ (the real value is below the estimated value), false negative³⁷⁹ (the real value is above the estimated value) and true samples with different colors. The number of observations that layout in each quadrant is also given, according to each chosen OK model. Notice that the threshold value is setup by the user.

Sample_id	Coord_x	Coord_y	Sample value	Cross-validation Error Model 1	Cross-validation Error Model 2	Cross-validation Error Model 3	Estimation Variance Model 1	Estimation Variance Model 2	Estimation Variance Model 3
38	28.9	2.1	0.06	.14603490	.15518890	.14644110	.26061500E-01	.84698920E-02	.43098000E-01
40	29.4	14.8	0.07	.11575870	.11217960	.11638490	.23212030E-01	.73051850E-02	.39200360E-01



False positive and false negative detection based on the cross-validation procedure

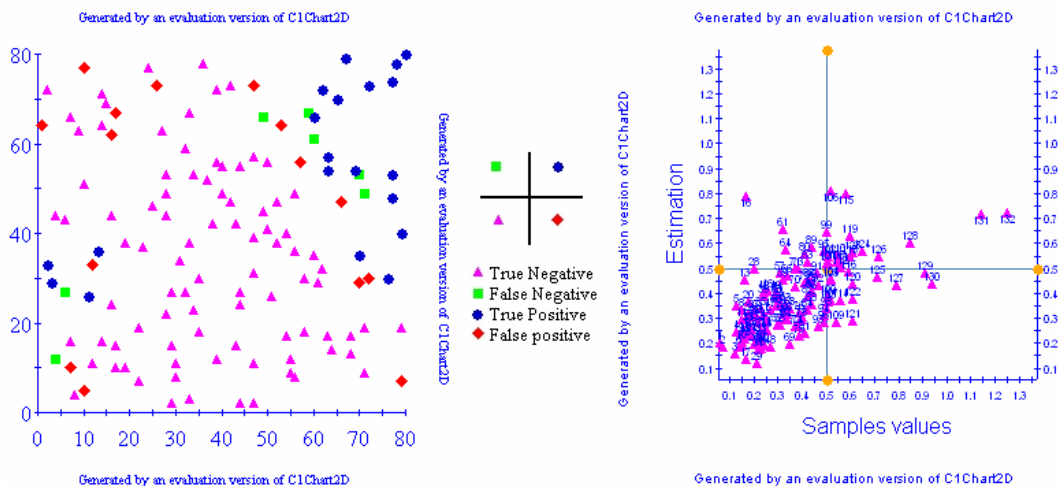
Minimal Value: 0.06 Maximal Value: 1.25

OK Model: OK With C0

Threshold:

Spatial Analysis

SAKWeb Cross-Validation Analysis (OK)

Number of Samples

True Negative: 90

False Negative: 7

True Positive: 20

False Positive: 15

Figure 91 – Partial view of SAKWeb[®] Cross-Validation option with its true-false/positive-negative analysis.

At last, the 3D-2D OK Profile is an attractive graphic option because it allows the user with the 3D surface generated by OK, interactively with the mouse, to outline the 2D profile between 2 points (see figure 92).

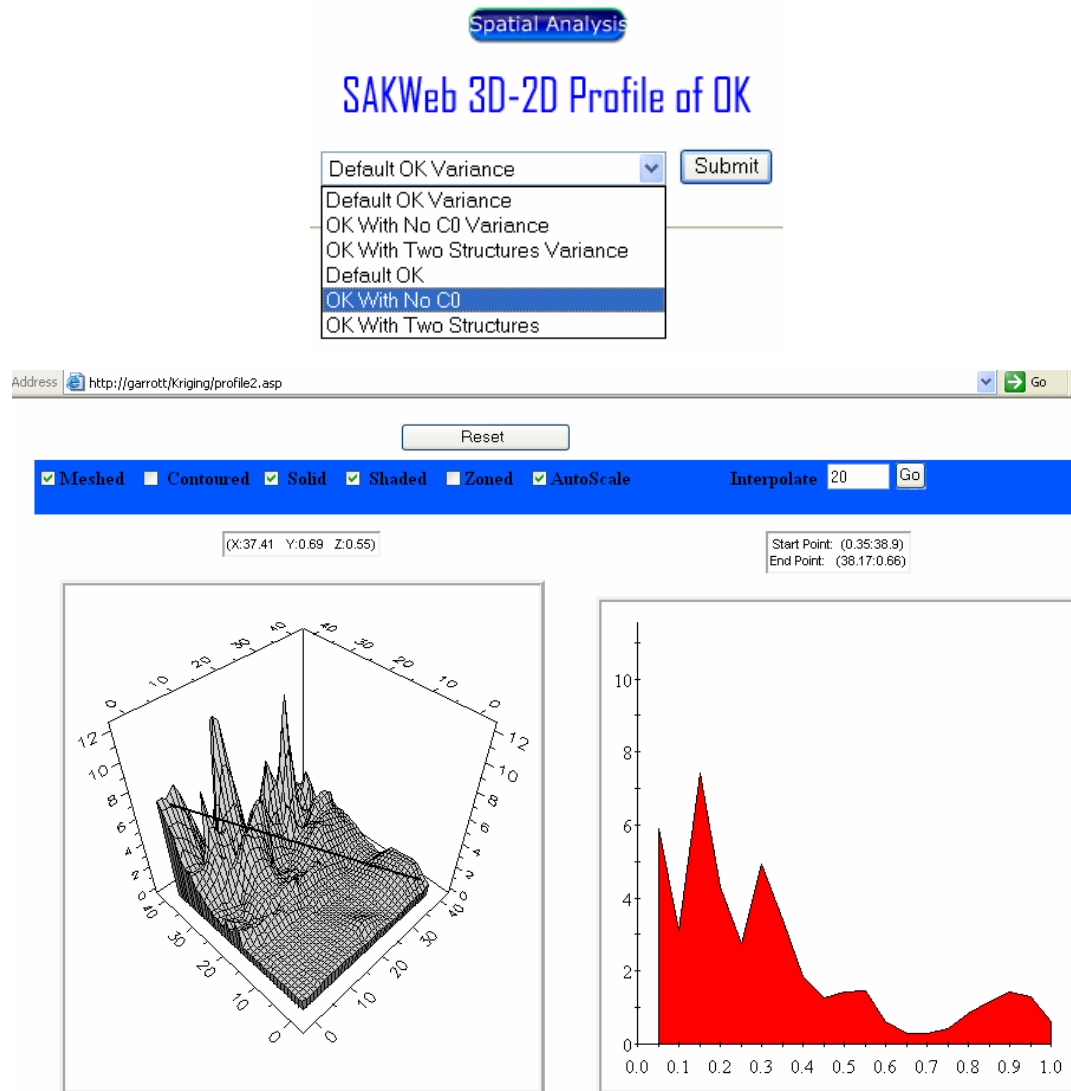


Figure 92 – Notice that the black line in the left image is not drawn by SAKWeb[®].

7.3.4. SAKWeb[®] Help

Flash[®] MX is the SAKWeb[®] Help standard authoring tool and, as Figure 53 (see subsection 7.1.4 for a further review) shows, it is divided into ten menus. The first option is a direct hyperlink to the Macromedia[®] site for users that do not have the

Flash[®] plug-in installed in their computers (necessary to display the following four choices).

The presentation movie tries to capture some screen photographs and spatial analysis keywords to create a storyboard that includes a rough message, the main navigation elements, vector background images and major animations with a consistent design. A pre-loader was developed because it is an excellent device, particularly if used with the `iframeloaded` script that loads all elements into the memory before playing in the computer. Logo animations are brief and simple while text slides waits a few seconds before disappearing. Tweenings with motion guides help bubble effects while audio clip instance creates an audio environment. “Another feature is JavaScript[®] code inclusion in the top browser” [Coelho, 2002], providing announces of GIS keywords. The capability to pause the movie also exists due to the attached script [Speigel, 2003]:

```
on (release, keyPress "") {
  if (paused eq "") {
    paused = "paused";
    stop();
    tellTarget ("_level9.audio") {
      gotoAndStop(7);
    }
  } else if (paused eq "paused") {
    paused = "";
    play();
    tellTarget ("_level9.audio") {
      gotoAndPlay(1);
    }
  }
}
```

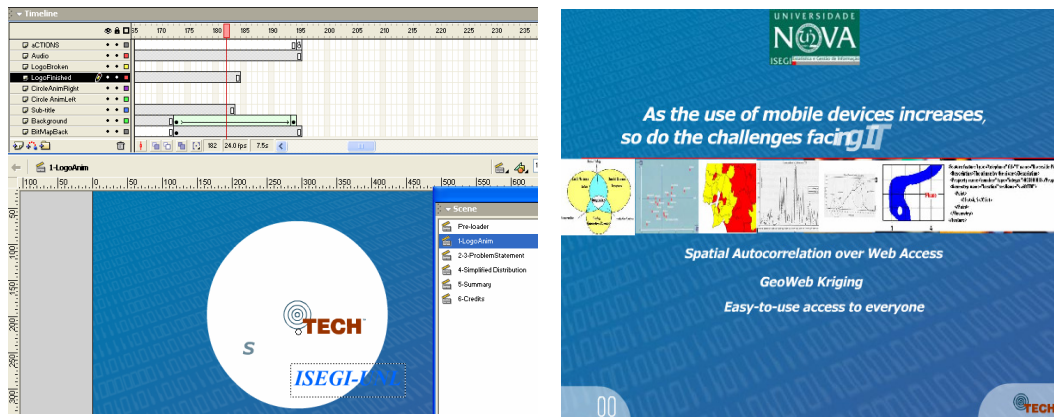


Figure 93 – The SAKWeb[®] presentation movie in terms of its construction (left) and final output (right).

The SAKWeb[®] introduction is also an entire Flash[®] product, though it includes user interaction. The blue spectrum has a powerful impact although the SAKWeb[®] Know How tweening motion represents another Flash[®] feature based on various keyframes. The ability to see different fade rates over the right button is a result of a movie clip with two layers and three frames whose alpha property changes according to the cursor speed. In addition, as confirmed by Street [2002], “one of the true powers of Flash[®] is its ability to dynamically load external information”. To achieve this goal, three steps are taken:

- A single-line dynamic text field is assigned to the variable header.
- A second multi-line dynamic text field called Imported text is created.
- The next frame script loads the appropriate text for each section (the Moran variance scatterplot in this case):

```
setProperty("home", _visible, "1");
header = /:headExpertise; (header equals the variable headExpertise whose reference appears on the
                           text file loaded with the loadVariables(*.txt,/) instruction during pre-loader
                           operation)
Importedtext = /:copyExpertise; (the label copyExpertise is also included on the text file)
tellTarget ("/tracker") {
    gotoAndPlay(1);
}
stop();
```

Conceptually, the summary of SAKWeb[®] help is divided into five major sections:

- SAKWeb[®] Moran scatterplot;
- SAKWeb[®] Moran scatterplot variance;

- SAKWeb[®] nugget-effect approaches including an overview of the role of geography among other sciences;
- SAKWeb[®] hypermedia Web interface;
- SAKWeb[®] sill rescaling and interpolation confidence interval.

SAKWeb[®] Help is divided into two branches, as Figure 94 shows³⁸⁰: SAKWeb[®] Themes and SAKWeb[®] Structure. Concerning the first item, this help corresponds to Chapter 1, while the latter relates to the following chapters and sections. When the user chooses a topic, a definition is displayed and, at the bottom of the browser, one or more hypertext links are highlighted in the relationship structure of SAKWeb[®] Help. This view presents an alternative to major help software that pursues the Contents-Search-Index pattern. In addition, the attention-grabbing cropping, colorizing and blurring eye involving design interactions with a bit of mystery are crucial elements.



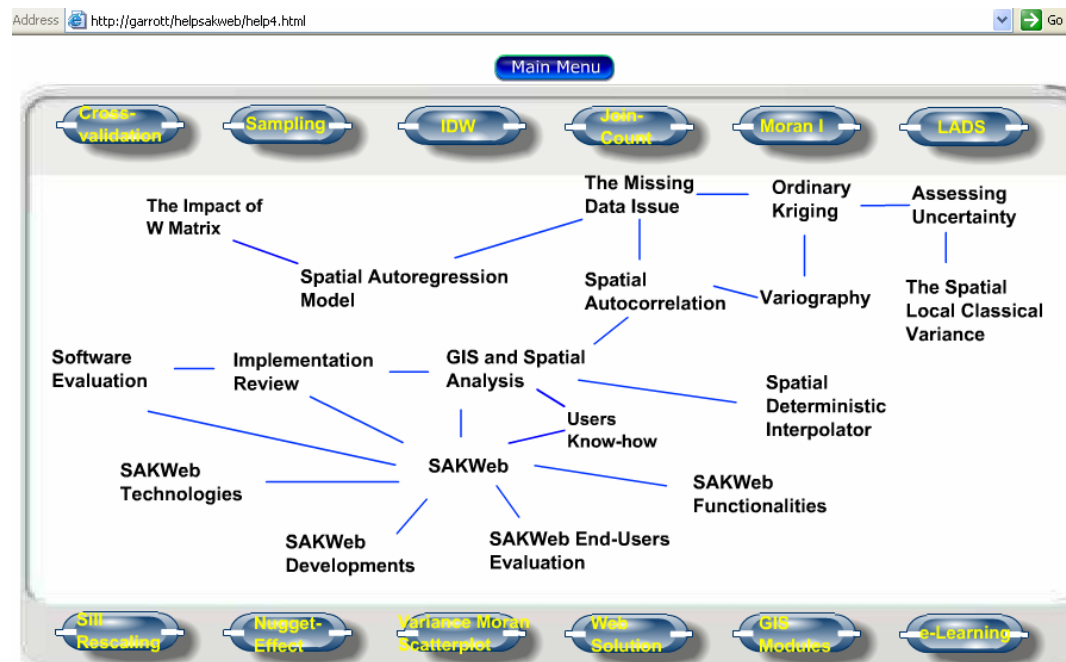


Figure 94 – The SAKWeb[®] Help.

One of SAKWeb Help's achievements is its success in capturing the feel of a living, thinking and interactive solution without being distracting, confusing and gimmicky. Distinctive animation for each section has also been created. For instance, the navigation system is focused in a sensitive way, with every button having a normal, rollover and active state. In fact, any interaction model should pursue a consistent navigation scheme that allows users to understand what a button is without having to think about it. With yahoo.com, for example, its simple interaction model is represented by underlined text links. With SAKWeb[®] help, text link buttons are angular and highlighted when the cursor rolls over them. Color consistency and acid audio in a loop context completes this environment.

Both the components of SAKWeb[®] Helps respect the structure given above with a clean screen layout following the KISS³⁸¹ aphorism, thus providing another solution for future software help. If the help is extensive in terms of contents then hierarchical layer architecture can be built. A third option is presented by SAKWeb[®] appendices and other utilities including Java Applets[®] and messages for the SAKWeb[®] administrator. A further new option is the capacity of the Web master to view, search and delete those messages with a list of options (reached by

the email menu) but protected with a confidential username and password (see figure 95).

Address: <http://garrott/helpsakweb/acesso.asp>

UNIVERSIDADE NOVA ISEG Estatística e Gestão de Informação

Messages Administration of SAKWeb

Username: Password:

Address: <http://garrott/helpsakweb/sucesso.asp>

UNIVERSIDADE NOVA ISEG Estatística e Gestão de Informação

Welcome to the reserve area of SAKWeb® Message Administrator

Address: http://garrott/helpsakweb/email_listar.asp

UNIVERSIDADE NOVA ISEG Estatística e Gestão de Informação

Name: Pierre Goovaerts **Country:** Belgium **Phone:** 7777777777 **Fax:** 7777777777 **Email:** pierr_goo@hotmai.com
Observation: Hi, Your site needs to enclose datasets greater than 500 points.
Browser: Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1) **Remote:** 169.254.201.34 **Username:** GARROTTjoao garrott

Name: pedro cabral **Country:** France **Phone:** 2222222222 **Fax:** 2222222222 **Email:** pcabral@isegi.unl.pt
Observation: Bonjour, ca va? Je veux dire qu'il y a un petit problem avec la routine de variance MS.
Browser: Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1) **Remote:** 169.254.201.34 **Username:** GARROTTjoao garrott

Name: Jorge Sousa **Country:** Portugal **Phone:** 8888888888 **Fax:** 8888888888 **Email:** jsousa@alfa.ist.utl.pt
Observation: Boa tarde, Simulação é o proximo passo para este site.
Browser: Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1) **Remote:** 169.254.201.34 **Username:** GARROTTjoao garrott

Name: Gisela **Country:** Portugal **Phone:** 4444444444 **Fax:** 4444444444 **Email:** gisela@isegi.unl.pt
Observation: Boa tarde, este site está em baixo e os alunos queisam-se que não podem realizar os trabalhos.
Browser: Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1) **Remote:** 169.254.201.34 **Username:** GARROTTjoao garrott

Records 1 to 4 of 5

The figure consists of two screenshots of a web application interface for SAKWeb Administrator Messages.

Top Screenshot: The address bar shows `http://garrott/HelpSakweb/contents_bd2.asp`. The page header includes the "UNIVERSIDADE NOVA" logo and "ISEG - Estatística e Gestão de Informação". A blue button labeled "MNGT Operations" is in the top right. The main heading is "SAKWeb Administrator Messages". The form contains the following fields:

- First Name: Jorge Sousa
- Country/Region: Portugal
- Work Phone: 8888888888
- Fax Number: 8888888888
- Email Address: jsousa@alfa.ist.utl.pt
- Notes: Boa tarde, Simulação é o proximo passo para este site.
- BROWSER: Mozilla/4.0 (compatible; MSIE
- REMOTE: 169.254.201.34
- USERNAME: GARROTT\joao garrott

At the bottom, it says "MailingList 5 of 5" with navigation arrows.

Bottom Screenshot: The address bar shows `http://garrott/HelpSakweb/email_delete.asp`. It features the same header and "MNGT Operations" button. The main section is titled "Search for name:" and includes a dropdown menu with the following options: "—Choose one—", "—Choose one—", "Painho", "Gisela", "Jorge Sousa", "pedro cabral" (highlighted), and "Pierre Goovaerts". A "Delete record" button is located to the right of the dropdown.

Figure 95 – The message administration option of SAKWeb[®].

The Other W3 Spatial Statistics Resources option is a simple HTML page with dynamic layers and a marquee object that have hyperlinks with important Web related resources, including a zip file with Excel[®] datasets for personal downloading and local Adobe Acrobat[®] files regarding the missing data issue³⁸² (see Section 2.6 for further review). Regarding the SAKWeb[®] discussion group for site visitors to communicate with each other, “Microsoft FrontPage[®] carried out this task quite well” [Marques, Mendes, 2001]. Its major features are (see Figure 96):

- A table of contents that contains hyperlinks to articles on the discussion topics;
- A search form that allows visitors to search the articles for a word or phrase³⁸³;
- An entry form in which a visitor types an article to post;

- A reply message to other articles;
- A confirmation page which registers that a visitor's article has been posted;



Figure 96 – The SAKWeb[®] News option.

7.4. SAKWEB[®] END-USER EVALUATION

In January 2003, a third survey was carried out with the same population introduced in Section 6.1. SAKWeb[®] was setup in an IIS[®] server, for ten clients using the intranet the Rainha Dona Leonor college in Lisboa. The Gamma[®] software GS+ Pb dataset³⁸⁴ (see Appendix M) was given to the twenty users (2 for each computer) so that they could test the SAKWeb[®] features. Previously, a 45 minutes refresher overview of spatial autocorrelation, trend surfaces, nearest neighborhoods and OK subjects was presented as most of the users had not reviewed them for almost a year. However, no tips were given on how to use SAKWeb^{®385}. Essentially, this section had three purposes:

- To test the simultaneous access robustness of SAKWeb[®].
- To appraise spatial autocorrelation and Kriging techniques via the Internet.
- To check how users interact with the Web GUI and other SAKWeb[®] features (see Figure 97) since they were only aware of the ESRI[®] Geostatistical Analyst module.

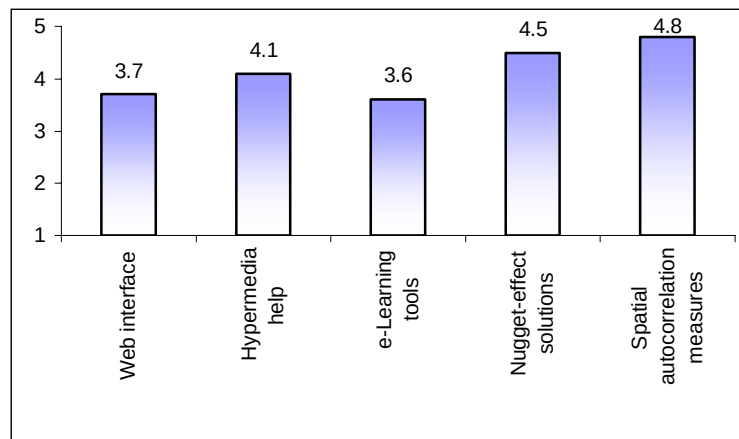


Figure 97 – The average classification for the impact of SAKWeb[®] features.

On the basis of the survey presented in Appendix H and with post-statistical processing, ten interpretations can be drawn:

- The SAKWeb[®] Moran scatterplot variance was too slow to be computed. In fact, 80% of the users complained about this issue, stating that this routine leads to impaired performance, that is, the slowness of the server impacts on concurrent operations of other users. Thus, the ASP[®] script language that underlies this procedure was changed to C[®] in version 2.0³⁸⁶. However, the strength of SAKWeb[®] in concurrent processes did not reveal any other major flaw.
- The capability to brush and link different maps was pointed out as seriously deficient by 60% of the interviewees, a major trend in the ESRI[®] Geostatistical Analyst module (particularly for assessing classical statistics in a moving average context). Accordingly, two new options were setup within SAKWeb[®]: local interactive statistics and 3D-2D OK profile.
- In the survey, the limited number of OK models available, especially when compared with the ESRI[®] Geostatistical Analyst module, is another issue to be taken into consideration. This includes different lag spacing depending on different directions, smaller separation distances for the first lag and other variography tools such as the correlogram, the madogram, the general relative variogram and the pairwise relative variogram. Consequently, these last four features were included after this survey was carried out.
- Although IK, UK, CK, SK and PK were never introduced in detail previously, in the survey 35% of these teachers expect these capabilities in

the near future with WWW links for other geostatistical sites. As a result, SK³⁸⁷ and W3 links for other spatial analysis resources were integrated in version 2.0. Analogous, stochastic simulation is regarded as a future improvement by 20% of the users.

- The questionnaire presented reveals that 45% of the respondents think that it is important to include the Box-Cox, Normal and logarithmic transformations in geostatistical software. Because of the complexity to assess the back unbiased transformation, this option will still not be covered in version 2.0.
- Surprisingly³⁸⁸, 95% of the users demand cross-validation. Hence, this procedure was added in SAKWeb[®] version 2.0.
- This questionnaire indicates that 30% of the users think that topographical and hill shade background maps enhance final Kriging maps. For this reason, image mapping in a new window was added afterwards. Curiously, no users mention the absence of the traditional semivariance/covariance cloud and the QQ-plot in SAKWeb[®].
- When compared with SAKWeb[®], only 20% chooses the ESRI[®] Geostatistical approach with regard to the outlier search issue. The bivariate scatterplot is also a good option. However, the histogram tool in conjunction with the brushing and linkage location map represents a wonderful tool for this purpose.
- All answers reveal that these users support the SAKWeb[®] Moran location scatterplot for verifying trends and local pockets of stationary.
- Figure 98 presents the final ranking for the main advantages of the Web solution.

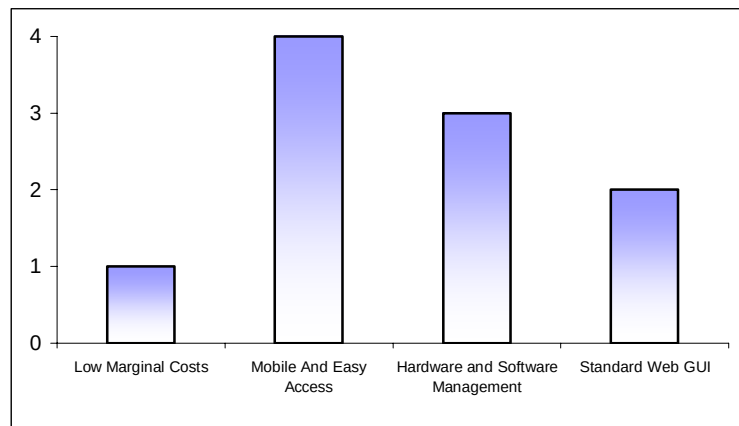


Figure 98 – The ranking for the Web solution.

After the survey was completed, certain issues were discussed in a brief meeting. The respondents confirmed previous suggestions: they really appreciate the capability to generate several Kriging surfaces on-the-fly and compare them without any extra work. It is expected that future software developments will enhance this feature. Therefore, the capability to extend this on-the-fly potential based on different range, model and anisotropy factors could be a remarkable improvement since the present solution is only based on the four nugget-effect strategies. Also, indicator maps for samples to reveal spatial data that are above or below any threshold should not be forgotten³⁸⁹.

With regard to e-Learning, these users agree that these capabilities will require an on-line geostatistic expert to maintain this service, particularly the SAKWeb[®] chat function. Certainly, this extra-cost for the SAKWeb[®] provider (in money and time) must be taken into account. In terms of the interface, it was unanimously agreed that this Web software is extremely easy to use for two main reasons:

- The hints given and the validation check for the input parameters;
- The inability to proceed to the next step if the previous one is not correct (it works like a process wizard).

Two users reclaim the migration from ASP[®] 3.0 to ASP.Net[®] architecture. In technical terms, the user first needs to change the extension of the files mentioned from .asp to .aspx so that they can be executed by the .Net framework. Next, the

user needs to change the application code to make it compatible with the ASP.Net[®]. This includes:

- The directive `<%@language="VBScript"%>` needs to be changed to `<%@Page Language="VB"%>`
- The parameters functions require the ByRef keyword to pass all input by value.
- All ASP.Net[®] functions need to be enclosed in the `<script>` block instead of `<%` and `%>`.
- With ASP.Net[®], it is not possible to render functions with HTML tags. Hence, the user needs to enclose the code to render text in a function and invoke the function when the text needs to be rendered.
- The third step regards the application optimization via the Web.config file and caching implementation.

The introduction of the Moran location scatterplot (instead of the conventional one) and the Moran variance scatterplot represents a major step in spatial autocorrelation measures. According to the questionnaire, the Moran scatterplot was ranked first in terms of usability and visual impact for outliers, trends and patterns search. The Moran I and Geary C took second place while the variogram cloud ranked third. But, this latter measure requires the brushing and linking feature. Another interesting comment made by two users was: “Until now, the emphasis on spatial autocorrelation measures by geosoftware has been limited and almost forgotten. It is hoped that SAKWeb[®] restores in our minds the need to implement them with Kriging software.”

The inclusion of deterministic methods such as IDW, Fourier and B-Splines, particularly for detrending operations and global interpolation, should be considered in the near future too³⁹⁰. Another point confirmed, as already stated in Section 7.1.8, is that the SAKWeb[®] Help approach presents alternative help software to that provided by the conventional Contents-Search-Index pattern. According to most inquiries, laying out the overall help structure with hyperlinks among subjects (see Figure 94 in Section 7.3.4) is really an attractive model. As some users confirmed, “it works like a puzzle where each piece corresponds to

each issue, though at the same time a complete view is given of the structure topics and how they are related to each other.” Future help applications must be encouraged to follow this concept too.

The capability to view the numerical results of OK values and OK variance was also requested by some users, especially when the measurement error was taken in consideration. Thus, the new 2.0 version of Non-Exact OK With Measurement Error already reflects these changes by laying out the OK variance map (see subsection 7.2.4 for a further review) and the observation table with their identification number, sample value, OK estimation and OK variance³⁹¹. Finally, the data grid with the capability to manage the input data, whatever the data source, was also mentioned by users. According to them, “although the present solution works quite well, input data administration is better achieved with an internal data grid since it gives the capability to insert, delete and update the coordinates and observation values quite easily.”

7.5. CONCLUSION

This chapter concentrates on the development and design of SAKWeb[®]. The first section reviews the technical issues of IIS[®], ASP[®], PHP[®], JavaScript[®], ActiveX[®], WebChart[®], ActiveBar[®], Java Applets[®], DHTML and Flash[®] features. References are made to their use within SAKWeb[®]. The differences between server-side and client-side techniques are emphasized too. Section 7.2 is divided into four subsections:

- The SAKWeb[®] variogram - Moran I - Moran scatterplot optimal ESDA;
- The SAKWeb[®] Moran variance scatterplot ESDA;
- SAKWeb[®] variogram sill rescaling including the minimum and maximum plume centered on the OK variance and local variance;
- The SAKWeb[®] nugget-effect.

Previous innovations are highlighted with examples of applications while particular features are examined from a technical point of view. For instance,

subsections 7.2.4.1 and 7.2.4.2 analyze the nugget-effect approaches versus the GS+[®] Pb and the San Diego housing costs datasets. Then, the following inferences are put forward:

- The measurement error only gives different results against the nugget-effect with a sudden jump when the estimation is one of the original samples;
- There is no correlation between the true error and the OK variance;
- The differences among the various approaches are greater for regions with high density values;
- On the basis of the two validation datasets, no neighborhood pattern evidences can be found with regard to which nugget-effect fits best.

Major interfaces are fully dissected in Section 7.3. Finally, a third feedback survey on SAKWeb[®] capabilities and future improvements is presented, followed by the concluding section.

8. SAKWeb[®]: Spatial Analysis of the Grasshopper in Colorado, USA

This chapter has three major objectives: 1) to test the SAKWeb[®] capability to perform true spatial tasks in a W3 environment; 2) to put theory into practice; 3) to make a contribution to the grasshopper outbreak issue by improving the grasshopper monitoring in Colorado. This includes outlining a definition area of infestation that exceeds the U.S. Department of Agriculture threshold and delimiting areas for environmental remediation work³⁹².

Chapter 8 is divided into ten sections. The first section introduces the ecology of grasshoppers while the second focuses on the prediction issue of grasshopper outbreaks. The ecoregions of the study are described in Section 8.3. ESDA is conducted in Section 8.4 to determine whether grasshopper densities exhibit spatial patterning: regional or large-scale versus field or small-scale. The spatial autocorrelation, variography and Moran location scatterplot presented in Sections 8.5, 8.6 and 8.7 determine the patch size of spatial dependence among grasshopper counts. Quoting Chou [1993], “the description and analysis of spatial patterns is essential to understanding the spatial distribution of an ecological phenomenon and the processes that produce that distribution”. If these spatial patterns are found then it is possible to describe grasshopper population growth and to give experts clues about the factors involved in the development of outbreaks (review subsections 2.4.6 and 2.5.1.1 for analogous circumstances). The relationship between the Moran I correlogram and the estimation of variogram parameters is also emphasized. The OK interpolates the map surface in Section 8.8. The next section establishes the results of validation, cross-validation analysis and regional confidence interval procedures. By knowing the estimated and true value at sample points, the overestimation or underestimation biases of the OK technique can be determined, and therefore controlled. The concluding section presents a summary of the current chapter.

8.1. THE GRASSHOPPER ECOLOGY

A grasshopper is an important agricultural pest which can be found in a variety of crops and ecosystems. “As herbivores, they are notorious for their appetites although their destructiveness may be as much due to clipping vegetation as to actually consuming it” [Skinner, 2000]. “Regarding its life cycle, the eggs are usually deposited in soil but some species prefer to lay them among roots of plants, in wood or in cow dung” [University of Wyoming, 2002]. A single female may deposit from 8 to 25 egg masses, each containing 20 to 120 eggs. Yet, females are very selective in oviposition site and commonly insert their ovipositor many times over before locating a satisfactory site. Soil texture, temperature, vegetation conditions and moisture level commonly influence site choice. “Eggs are laid in a packet nearly 0.5 cm long and from 0.25 to 0.75 cm below the soil surface” [Animals, 2002]. “These eggs are mainly deposited during September and begin hatching the next May but a late spring or cool summer may delay their growth and adult feeding so that nymphs are present through the autumn” [BASF Canada, 2001].

There are a hundred of species of grasshopper and, occasionally, they develop into a plague. For instance, major periods of drought are correlated with sunspot cycles. According to the University of Wyoming [2002], “a decrease in the number of sunspots is followed by a fall in precipitation which in turn is followed by grasshopper population increases”. Pfadt [1977] suggests that population densities double annually until the fifth year when densities triple or quadruple.

Since a large numbers of eggs may be deposited in the same location, food can be depleted quickly. “Drought often makes the forage unpalatable and both nymphs and adults walk or fly to more succulent plant material” [Legare, 2000]. Hence, grasshopper migration may occur because of a lack of suitable food.

Because of its negative historical impact on the rangelands, grasshopper habits quite often conflict with human interests. Grasshoppers are most easily controlled

with insecticides when they are still immature and their location is restricted to breeding areas. “When high populations of grasshopper nymphs are detected, generally more than 10 per square yard, prompt treatment is necessary to prevent later problems in adjacent areas” [Cranshaw, Capinera, 2003]. According to Peairs [2002], “the insecticide products available for rangeland grasshopper control are not very effective against adults, if at all”. As expected, entomologists and other experts have looked for other practical ways to prevent grasshopper damage and control the population considering various factors like geographic distribution, life cycles, behavior and environmental influences (weather, fire, sunspots and earthquakes).

A grasshopper outbreak is costly and complicated. In the late 1980s, a computer simulation model called Hopper[®] was developed to improve the decision supporting system for management of the rangeland grasshopper, including treatment options and Kriging to improve the predictive power of the grasshopper maps. “Hopper[®] allows the evaluation of the validity and cost effectiveness of treating outbreaks because treatment decisions are serious business” [Berry, Kemp, Onsager, 2002]. On the one hand, unnecessary treatment wastes money but failure to treat when treatment is needed may damage the local economy.

Major outbreaks occur without any apparent explanation, with extensive damage during outbreak years. Accordingly, several methods have been developed to remove grasshoppers from agricultural land³⁹³ including [Skinner, 1999]:

- Collection with nets and plowing and harrowing to destroy eggs;
- The use of turkeys and other domestic fowl;
- Smudging and concussion;
- Chemical control based on arsenic with bran or sawdust;
- Insecticide sprays such as Malathion and carbaryl;
- Biological control agents like the pathogen *Nosema locustae* Canning.

“The current emphasis of large-scale grasshopper management efforts has shifted from crop protection to rangeland protection” [Skinnner, 1999]. This shift has occurred because farmers can use numerous insecticides to effectively suppress

grasshoppers (see Table 24). “Insecticide applications directed at the myriad of other insect pests associated with crops help keep grasshopper numbers low” [Berry, Kemp, Onsager, 2002]. Certainly, such practices associated with crop production as tillage, ditch burning and weed control also reduce the prevalence of grasshoppers.

Test	Grasshopper Species	% Mortality	Test	Grasshopper Species	% Mortality
Laboratory	<i>Ageneotettix deorum</i>	28	Field	<i>Aeropedellus clavatus</i>	34
	<i>Amphitornus coloradus</i>	4		<i>Ageneotettix deorum</i>	90
	<i>Aulocara ellioti</i>	18		<i>Amphitornus coloradus</i>	0
	<i>Aulocara femoratum</i>	31		<i>Aulocara ellioti</i>	82
	<i>Boopedon nubilum</i>	45		<i>Aulocara femoratum</i>	72
	<i>Melanoplus sanguinipes</i>	50		<i>Camnula pellucida</i>	0
	<i>Metator pardalinus</i>	8		<i>Chorthippus curtipennis</i>	49
	<i>Opeia obscura</i>	0		<i>Encoptolophus sordidus</i>	67
	<i>Phlibostroma quadrimacula</i>	5		<i>Hadrotettix trifasciatus</i>	74
	<i>Phoetaliotes nebrascensis</i>	42		<i>Hesperotettix viridis</i>	0
	<i>Trachyrhachys kiowa</i>	0		<i>Melanoplus spp.</i>	75

Table 24 – Mortality among grasshoppers exposed to insecticide in laboratory and field trials [Onsager, 1978].

Yet, the use of these techniques has serious implications in a long-term perspective. Birds are important grasshopper predators and serve to exemplify some of the problems associated with insecticide use. Materials such as toxaphene and diazinon cause high levels of bird mortality following their application to rangeland. Others such as propoxur and azinphosmethyl cause reduction in bird numbers without direct evidence of bird mortality. Emigration from treated sites is, thus, implied. “Malathion and carbaryl seem to have no direct effect on wildlife populations although emigration has occurred under some circumstances” [University of Wyoming, 2002]. Insecticides used for grasshopper control degrade the environment.

The cost of grasshopper control by means of a reduction in forage is also an important direct expense that ranchers must incur (see Table 25). However, indirect costs must also be included because of the reduction in weight gain by cattle and relocation costs. “Grasshoppers consume from 6% to 12% of the

available forage in the western United States, although in some localities they consume essentially all available forage” [Berry, Kemp, Onsager, 2002]. Grasshoppers eat approximately one-half of their body weight in green forage per day. According to these researchers, with a grasshopper population of seven or eight per square meter in a four hectare field, grasshoppers consume as much forage as a cow.

Treatment	Rate (kg AI/ha)	Forage Saved (kg dry weight/ha)	Forage Saved (Value in \$)	Investment (\$/ha)	Net Return (\$/ha)
Carbaryl	0.56	252	12.30	6.79	5.51
	1.12	297	14.50	9.63	4.87
	0.28	212	10.36	5.37	4.99
	0.56	227	11.11	6.79	4.32
	1.12	242	11.81	9.63	2.18
	0.56	130	6.36	6.79	-0.43
	1.12	132	6.44	9.63	-3.19
Malathion	0.56	125	6.14	5.53	0.61
Untreated		0	0	0	-16.66

Table 25 – Costs and returns associated with control of grasshoppers (forage was valued at \$44 per metric ton and grasshopper density was 10 to 15 per m²) [Onsager, 1978].

8.2. THE GRASSHOPPER PREDICTION ISSUE

“Despite more than 120 years study of rangeland grasshopper ecology and population dynamics, we are still unable to accurately predict when or where the next grasshopper outbreak will occur” [Skinner, 2000]. Several difficulties can be mentioned:

- Control effects frequently use a single treatment to manage all species with varying success;
- The fluctuation in population involves both space and time factors;
- Information gained from laboratory studies may not be applicable under field conditions. For example, natural reproductive rates are much lower than the rates obtained in studies of caged grasshoppers;
- Grasshoppers are managed on various spatial scales. On the field scale, local land owners control grasshoppers by spraying fields and roadsides. At water

shaded scale, changes in species composition and abundance influence the effectiveness of treatment. Over statewide areas, surveys and hazard maps are used to assess the chances of damage, although the probability that an outbreak will develop may change from one state to another. “On a regional scale, rangeland grasshopper management efforts and national agricultural policies are coordinated by national institutions” [Skinner, 1999];

- The grasshopper species is composed of 600 species with particular habitat requirements. Thus, better grasshopper identification and forage biomass estimation techniques are needed.

Can infestations be predicted? “It is surprising that so much remains to be learned about a group of insects which have been so important in American agriculture history” [Skinner, Child, 2000]. This is the legacy of a view that fails to accept the heterogeneity of the grasshopper structure. Although grasshopper dispersion is not uniform, one approach is to look for temporal or spatial patterns in outbreak occurrences instead of the traditional hierarchical levels of biological organization. Another alternative is to look for correlation with auxiliary variables such as drought or sunspot activity in an OLS or spatial autoregressive model.

The research presented here only concerns the former method emphasizing multiple spatial scales, because grasshopper concentration is related to both time and space. The first goal is to identify spatial patterns in grasshopper concentrations in Colorado on the basis of the data available from Skinner’s [1999] dissertation (see Appendix E). ESDA and Kriging techniques are used to characterize spatial dependence among grasshopper counts for the 1993, 1994, 1995 and 1997 datasets (numbers per m²) in the next seven sections.

8.3. THE GRASSHOPPER STUDY REGION

The two ecoregions of interest are the eastern plains of Colorado, USA:

- Ecoregion 25 or Western High Plains – The dominant habitat in this area is a short grass prairie dominated by two low-growing warm-season grasses, blue

grama and buffalo grass. “Western wheat grass also exists along with taller vegetation including widespread prickly-pear cactus and yucca and cholla in the south” [Skinner, 1999]. “The primary agents of large scale disturbances are fire and insect outbreaks” [Hutchings, 2000]. Regarding economical issues, this land is used for dry land crops, grazing and irrigated agriculture³⁹⁴.

- Ecoregion 26 or Southwestern Tablelands – The Colorado Plateau is a high, flat tableland carved into soaring mesas, great cliffs, chasms and canyons. It includes much of northern Arizona, eastern Utah, western Colorado and northwestern New Mexico. It comprises semiarid tablelands with moderate relief where the natural vegetation includes grama, buffalo and bluestem grasses with some sand sage. “Land use includes grazing and dry land agriculture” [Skinner, 1999]. According to Hutchings [2000], the manipulation of water (irrigation and dam building) and other land uses (orchards, farms, industrial and residential) have created major threats to wildlife habitats.

Remark that very few samples are taken in the middle longitudes because that area corresponds to the high mountains of Colorado. The Rocky Mountains cover much of the central region (see Figure 99), with small extensions into southern Wyoming and northern New Mexico. It encompasses the majority of the forested lands in Colorado, except for pinyon-juniper forests. The topography is rugged, with over 50 mountain peaks rising above 4300 meters. “Very large, flat and open mountain basins called parks also occur” [Hutchings, 2000]. Some permanent snowfields and remnant glaciers are found at higher elevations, too.

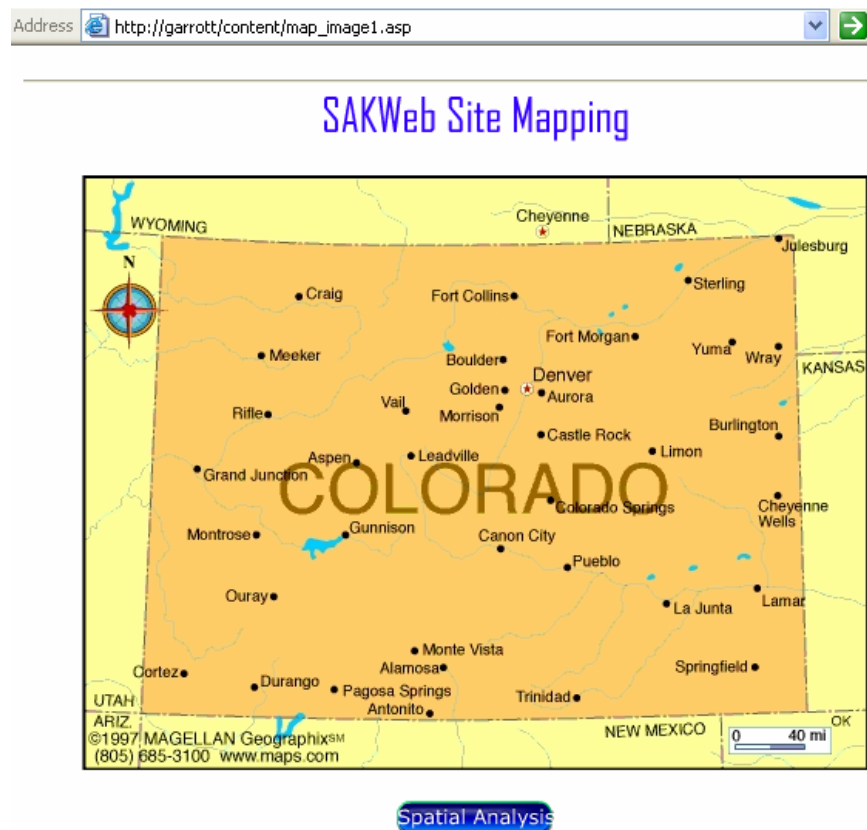


Figure 99 – Map of the cities of Colorado, USA, as shown in the SAKWeb[®] Image Mapping option.

8.4. SAKWEB[®] EXPLORATORY DATA ANALYSIS

The summary aspatial statistics of SAKWeb[®] for grasshoppers densities are depicted in Table 26. The mean grasshopper density differs from year to year: in 1994, it was almost the double of 1997. The U.S. Department of Agriculture suggests that treatment is justified when grasshopper numbers reach approximately 9 per m² which leads to an overall tranquil situation for all years³⁹⁵. Nevertheless, this trend is not followed when a spatial comparison between years and ecoregions are made. For instance, the general linear polynomial trend presented in Appendix E reveals a strong upward drift along the longitude from west to east in all years. In fact, areas of moderate to high densities are somewhat clustered but also scattered among low density sites: densities are low in the western half of Colorado for all years while higher concentrations are dispersed among lower ones in eastern Colorado. This suggests a single pattern for the western section while grasshopper densities exhibit a random pattern for the eastern area.

SAKWeb [®]	1993	1994	1995	1997
Number of points	103	99	97	116
Average (individuals per m ²)	4.21	4.9	3.19	2.44
Standard error of the average	0.47	0.71	0.49	0.23
Variance	23.22	50.79	23.96	6.52
Standard deviation	4.81	7.12	4.89	2.55
Maximum	23.4	32.6	26.8	12.5
Median	2.5	2.1	1.5	1.6
Minimum	0	0	0	0
Mode	0.5	1.1	1.3	0.8
Coefficient of variation	1.14	1.45	1.53	1.04
Absolute average of deviation	3.48	4.61	2.96	1.84
Q1	0.9	1.1	0.7	0.75
Q2	2.5	2.1	1.5	1.6
Q3	5.4	5.8	2.7	3.3
Q3-Q1	4.5	4.7	2	2.55
Skewness	1.98	2.74	3.01	2.01
Pearson	0.77	0.53	0.38	0.64
Kurtosis	4.12	7.73	9.56	4.83

Table 26 – SAKWeb[®] aspatial descriptive results for the four grasshopper datasets.

As expected, the preferential sampling issue is a reality on account of the frequency historical outbreaks or the value of agricultural productivity in those specific regions. Therefore, the EGM and the EGV, based on the nearest neighborhood distance, reveal different values when compared with the conventional average and variance. Yet the density for the region is almost the same, although the total area also changes.

The overall histograms present positive skewness, confirmed by the box-and-whisker plots (see Appendix E). Since SAKWeb[®] does not cover any Box-Cox operations, no logarithm transformation was performed on the original dataset, a situation not pursued in Skinner's work³⁹⁶. The frequency curve is peaked (leptokurtic) for all years with a highly skewness value, a situation followed by many natural phenomenon³⁹⁷. In addition, the highest maximum density was 32.6, in 1994, and the lowest reached 12.5 in 1997. "A possible explanation of this decrease can be closely related with the precipitation of 1997 since moist conditions increase mortality rates" [Skinner, 1999]. Likewise, the eastern plains

of Colorado reveal higher densities because of the drier climate which creates more favorable conditions for the grasshopper population.

The bivariate scatterplots³⁹⁸ reveal a weak relationship between density and latitude-longitude for all years with high densities located mostly between 102°-104° West and 37°-39° North (see Appendix E). On the basis of classical inferential statistics, the parametric t-Student test corroborates the difference between the means of the following populations for a confidence level of 95%³⁹⁹

- 1993 and 1995 (t-Student = 1.699)
- 1993 and 1997 (t-Student = 2.988)
- 1994 and 1995 (t-Student = 1.983)

These inferences lead this analysis to the moving rectangular window for a local summary descriptive assessment including data heteroskedasticity. The need for this SAKWeb[®] capability is that the variogram assumes that the interpolation surface has a constant mean, a situation almost never true⁴⁰⁰. Certainly, there are major differences among results, as Table 27 shows (Regions A, B and C are shown in Figure 100).

Year	Region	Number	Mean	Standard Deviation	Skewness	Kurtosis
1993	A	18	4.01	6.06	2.26	5.58
	B	31	6.16	5.56	1.5	1.82
	C	54	3.17	3.46	1.89	3.38
1994	A	8	0.82	0.73	0.3	-1.49
	B	27	6.49	9.14	2.14	3.74
	C	64	4.74	6.42	2.94	10.11
1995	A	13	0.62	0.5	0.49	-0.93
	B	55	3.4	5.73	2.93	8.1
	C	29	3.94	3.85	1.8	2.9
1997	A	14	0.48	0.51	1.9	3.54
	B	28	3.32	3.61	1.31	1.15
	C	74	2.48	2.08	2.12	6.64

Table 27 – Aspatial descriptive results according to each year and region.

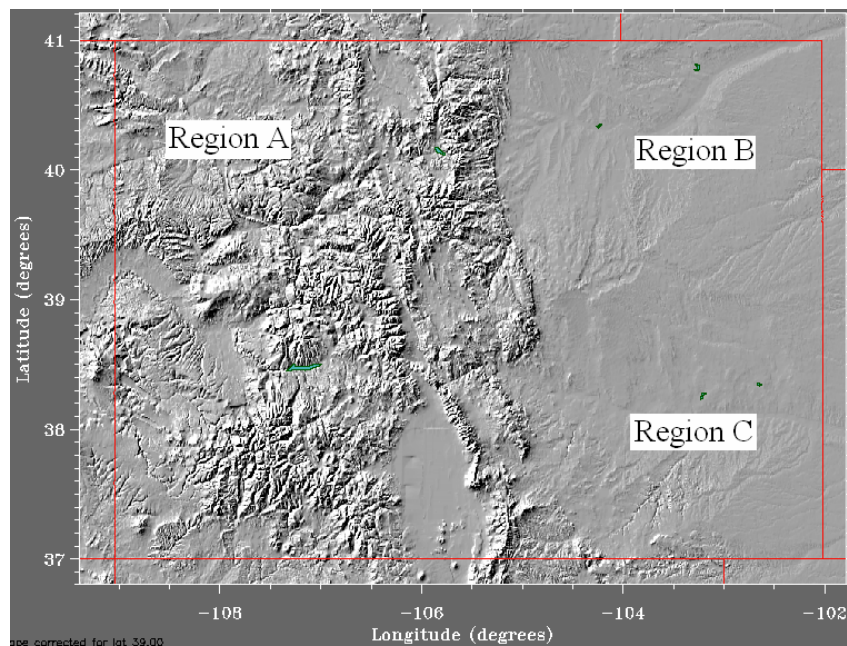


Figure 100 – Topographic map of Colorado, USA.

When dealing with the spatial distribution presented in Table 28, the nearest average distance between samples is 23 km with a standard error of the mean of 1.4 km. For the same area, if all points had been arranged randomly then the expected mean nearest-neighbor distance would be 25.86 km (1994). It is also possible to calculate the expected mean nearest-neighbor distance for these points if they had been arranged in most possible dispersed way possible. So, if all samples have a perfectly uniform point pattern, the expected nearest distance is roughly 57 km, leading to a worse situation with regard to the nugget-effect issue. This occurs because uniform sampling forces a minimum distance frontier among points, the distance interval in which capturing spatial dependency is most important (review subsection 3.3.2 for further discussion). It is also important to stress that these results are very useful for the user when he/she setups the variogram lag, the tolerance lag, horizontal bandwidth and number of lags parameters (see Figure 86 in subsection 7.3.2).

SAKWeb [®] Results	1993	1994	1995	1997
Estimated Global Mean	4.82	5.06	3.29	2.25
Estimated Global Variance	29.57	57.01	28.21	5.82
EGM/Average ratio	1.14	1.03	1.03	0.92
EGV/Variance ratio	1.27	1.12	1.17	0.89

Coord _x of mean center ⁴⁰¹	547.15	683.71	545.58	536.06
Coord _y of mean center	251.56	292.11	272.94	299.2
Coord _x of weighted geographical centroid ⁴⁰²	531.14	738.52	627.63	606.63
Coord _y of weighted geographical centroid	193.2	243.85	256.94	280.58
Standard deviation distance	643.02	772.33	650.26	653.94
Region area	275674	322525	287091	300572
Region density	0.0004	0.0003	0.0003	0.0004
Standard deviation ellipse length ⁴⁰³ (Y axis)	156.62	219.8	489.15	617.83
Standard deviation ellipse length (X axis)	481.35	402.82	389.25	445.46
Standard deviation ellipse angle (clockwise degrees)	66.27	66.6	64.14	61.5
Mean nearest distance	22.46	23.56	22.71	22.97
Standard error of the mean nearest neighbor distance	1.33	1.49	1.44	1.23
Expected mean nearest distance for a random arrangement	25.86	28.53	27.2	25.45
Expected nearest distance for a perfect uniform point pattern situation	55.59	61.33	58.45	54.69
Nearest neighborhood index (0 to 2.15)	0.86	0.82	0.83	0.9

Table 28 – SAKWeb[®] nearest neighborhood analysis of the four grasshoppers datasets.

8.5. SAKWEB[®] SPATIAL AUTOCORRELATION

Variography was performed first (see algorithm depicted in Figure 61 of subsection 7.2.1.7) with the goal to assess the Moran I correlogram leading, hence, to a first attempt of describing spatial continuity. Also, the chosen random function provides the necessary tool for the construction of the **W** matrix of Moran I but respecting Tobler Law.

Most real world patterns are somewhere between a random pattern and a clustered pattern. Grasshopper densities in Colorado follow this tendency. The Moran I correlogram reveals a significant positive spatial autocorrelation for 1993, 1995 and 1997 (see Table 29). This feature is especially important for a neighborhood

distance of 45 km (1995), a situation not confirmed in 1994, for which the highest autocorrelation measure is roughly 70 km. Thus, better results are expected with a lower nugget-effect for 1995 since closer samples have a major impact on the real estimation. In contrast, a wider variogram range⁴⁰⁴ should be estimated in 1994 for two reasons:

- The Moran I correlogram reveals a wider patch area for its highest value;
- The Moran I is lower than in all other years, giving the idea that stronger variation among neighbors is expected and interpolation results will be worse⁴⁰⁵.

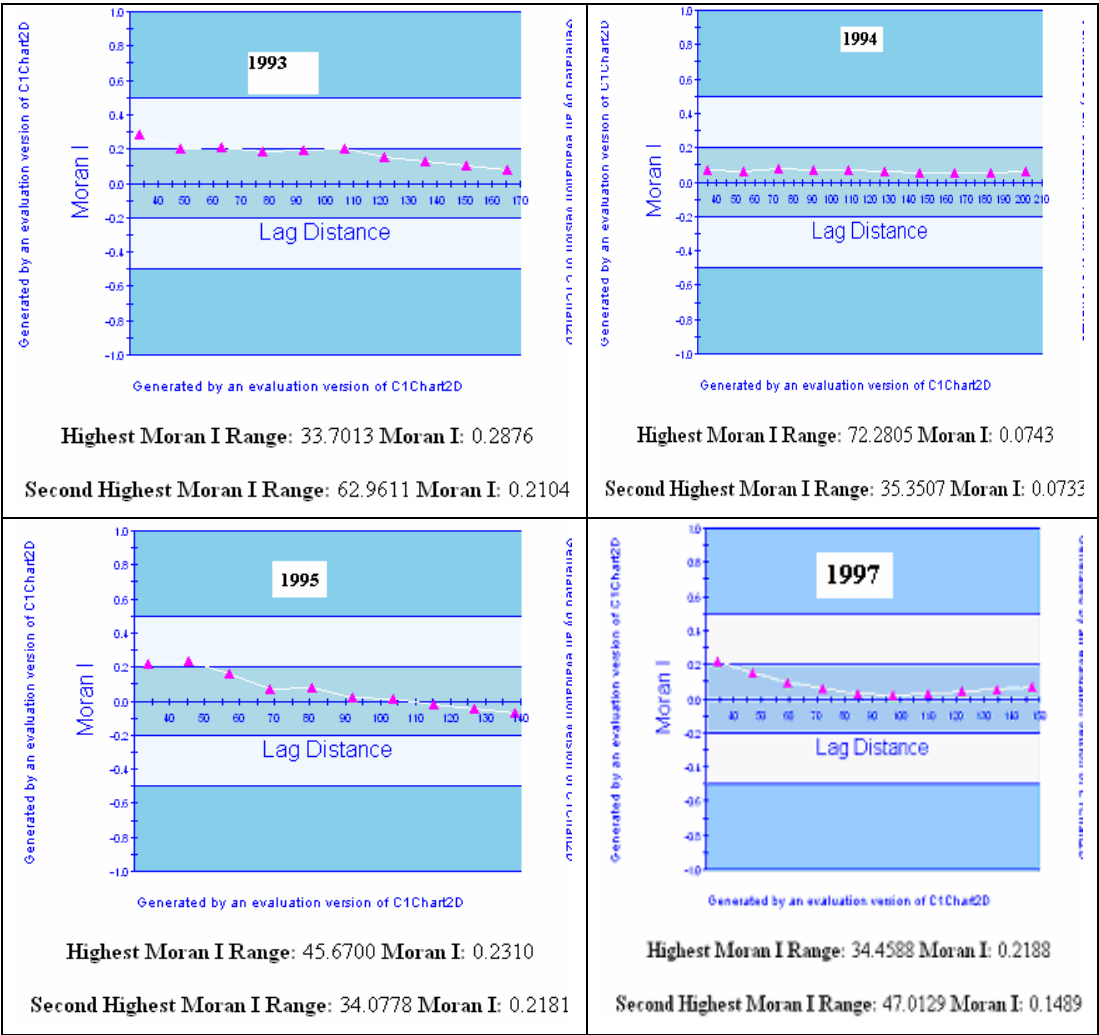


Table 29 – SAKWeb[®] Moran’s I correlogram for the grasshopper dataset.

Regarding 1993, spatial autocorrelation is strong up to a range of 110 km and moderate up to 170 km, leading to the idea that the variogram range should be

close to this distance. Accordingly, the variogram range of the next year is wider, while for 1995 and 1997, it becomes smaller. In terms of grasshopper prevention, the year 1994 is more important for local land owners because of the scatter of epidemics within the state due to the lack of spatial autocorrelation homogeneity.

Confirming the results of Section 2.7, as distance gets larger, the Moran I tendency is zero. Moreover, as spatial autocorrelation gets stronger, the variogram scale (y-axis) gets smaller because the dissimilarity among values is smaller. The variogram of 1993 and 1995 presents a scale range of [0...25] and [0...70], respectively, against the [0...100] of 1994. The standard mean and variance in Table 29 (above) follow this trend. If 1997 does not pursue this tendency, [0...10], it is because the values for the grasshopper densities are rather lower than those in previous years. For instance, the average density and variance for the year of 1997 is 2.44 and 6.52 while for 1994 it reaches an average of 4.9, with a higher variance of 50.79.

Comparing the 1993 distribution with other spatial distributions, that year reveals a more homogeneous grasshopper density distribution while the latest two years reveal more pronounced local spatial patches. In contrast, the year 1994 reveals a more erratic overall picture. This can be corroborated by the slope of the Moran I correlogram: the deeper it is, the clearer the definition of local spatial patterns becomes. Indeed, the Moran I correlogram curves for 1995 and 1997 indicate that there are smaller spatial patterns of density than 1993 and 1994. Accordingly, the grasshopper densities in Colorado's adjacent states are more important for these years than 1995 and 1997 or, in other words, the endemic damage levels are more positively correlated between adjacent states⁴⁰⁶ in the first year than in the last two.

A possible cause of this positive spatial autocorrelation on a larger scale can be a result of the smaller number of observations for a wider spatial area (300,000 km², on average). There may be poor resolution and changes in grasshopper densities on scales of less than 30 km but it is not possible to detect them because the average distance between samples is about 230 km⁴⁰⁷. In fact, the length of the standard deviation ellipse, a sampling bias measure, indicates that this effect is more

pronounced longitudinally than in latitudinally (3 times more so). Hence, different interpolation maps are expected according to the different nugget-effects, particularly in sampling zones. In regions without any observations, the difference among OK approaches is minimal.

This also shows that the current scheme for the surveying the grasshopper population is not adequate for detecting localized infestations. The average distance between samples is too large while the highest scale at which autocorrelation was detected ranged between 33 km and 72 km. If there are small hot spots then it is not possible to detect them.

8.6. SAKWEB[®] VARIOGRAPHY

If spatial autocorrelation is not significant for any distance then variography and Kriging should not be performed because the sample average is the best estimation for any location. Since the four datasets do not exhibit this trend then it makes sense to carry on with the interpolation process. Otherwise, this procedure would have to finish here.

According to the previous section and the algorithm depicted in Figure 61, variogram parameters should be readjusted on the basis of the Moran I correlogram hints. Variography is not a deterministic science but rather a form of stochastic thinking. In this case, the range of the first variogram fitness attempt had to be changed to reflect the inferences of the Moran I correlogram:

- With the 1993 dataset, the isotropic range was modified from 150 to 180;
- With the 1994 dataset, no adjustments were made;
- With the 1995 dataset, the range was shifted from 170 to 150⁴⁰⁸;
- With the 1997 dataset, the range changed from 180 to 160.

The Moran I correlograms were computed again and, in spite of modifications, the inferences made in previous sections continue to be valid. If the results had changed quite dramatically then the process would have had to be re-estimated

from the beginning. This also occurs because the neighborhood definition needed it to assess the Moran I is problematic and based on the variogram itself to embrace Tobler's First Law of Geography. The final variogram parameters are depicted in Table 30.

Year	Model	Major Range	Isotropic	Nugget Effect (C0)	Sill	C0/Sill	Angle Direction
1993	Spherical	180	Yes	7	28	0.25	-30°
1994	Spherical	220	No	10	75	0.13	+45°
1995	Spherical	150	No	0	45	0	+45°
1997	Spherical	160	No	1	9	0.11	-45°

Table 30 – SAKWeb[®] variogram models⁴⁰⁹ for the four datasets.

Note that in 1993, there is a match between the variogram range and the lag distance in the Moran I correlogram⁴¹⁰. Nevertheless, this trend is not followed in the next years because of the existence of a geometric anisotropy (see appendix E to compare the different behaviors according to different directions). It is the minor range (not the major one) that presents a close value to the distance at which Moran I correlogram becomes close to zero. Therefore, the minor ranges for 1994, 1995 and 1997 become 80 km, 90 km and 100 km, respectively.

Another issue relates with the sampling method. If the variogram range changed to small distances then the OK estimation and variance for the Rocky Mountains in Colorado would give non-sense results because of the lack of observations within that range (see Figure 106 of Section 8.9). Looking carefully at the sampling sites, it becomes obvious that the current sampling approach did not cover this wide central region.

However, the most motivating issue with this approach is that the Moran I correlogram can be a useful tool for the determination of the variogram factors. It seems that the neighborhood view of spatial autocorrelation measures can work out as a best parameter guess for the distance view of spatial autocorrelation measures (review Section 2.4 for a further discussion). This happens because the variogram fitness can turn out to be a complicated problem on account of the uncertainty and personal thinking that underlies this procedure. Furthermore, if the author of this

dissertation knew more about grasshoppers' habitats or the Colorado biological environment, for instance, this geographical information could lead to other variogram parameters. This is the main reason why SAKWeb[®] does not contain any automatic procedure for the variogram fitness because the statistical knowledge by itself is not sufficient for high-quality solutions. Automatic procedures can not include this type of valuable information. Recapping Section 6.2, the software must only do the calculation, while the human factor must interpret those results, with the ability to manage the process course because the best solution relies on the best justification model.

The model choice does not differ between years although it could. Nevertheless, during this research, the model considered will be the same in order to compare other parameters with similar conditions. Also, the possibility of splitting data (the coefficient of variation is greater than one) and calculating variograms for the two major ecoregions was not undertaken because of the lack of data. UK, stratified regionalization, detrending, general relative variograms, SK with varying local mean or external drift could be alternatives in a stratification scheme. Definitely, the quasi-stationary assumption is a sampling density compromise⁴¹¹.

The parameters of the overall variogram models differ over time. The nugget-effect and the range, the distance at which data are not spatially autocorrelated, are dissimilar. Hence, grasshopper densities exhibit spatial patterns on multiple scales (possibly nested ones such that autocorrelation on smaller scales is masked by larger patterns). It seems that grasshoppers densities in Colorado exhibit spatial patterning on both local and regional scale and, probably, the two nested scales of SAKWeb[®] may achieve better results than the other methods.

If sill fluctuations correspond to periodic structures then the 1993 spatial distribution reflects this situation. In contrast, 1997 depicts the best variogram structure while the years of 1995 and 1997 confirm that the variogram model has a smaller range (see Table 31). In terms of anisotropy (see Appendix E), 1993 reveals an isotropic situation whose patterns of variation are fairly similar in all directions with a range close to 180 km. The directional variograms of 1994

reveals a more unstructured and periodic continuity: the major axis is 220 km (and also reflected with the omnidirectional one) while the minor equals 80 km, the distance where the Moran I correlogram is closer to zero. A long distance range was, thus, setup as an obscure average of the several directional behaviors. The year of 1995 varies more continuously in the North-West direction with a close range of 150 km and confirmed by the omnidirectional variogram. Also, the minor variogram range matches the 90 km range of the Moran I correlogram, the distance at which the global spatial autocorrelation measure closes to zero. At last, the year of 1997 presents a clear North-South/NorthWest-SouthEast geometric anisotropy with a major continuity of 160 km while the perpendicular direction disclose a shorter range of 100 km. Again, this minor continuity matches the zero value of the Moran I correlogram.

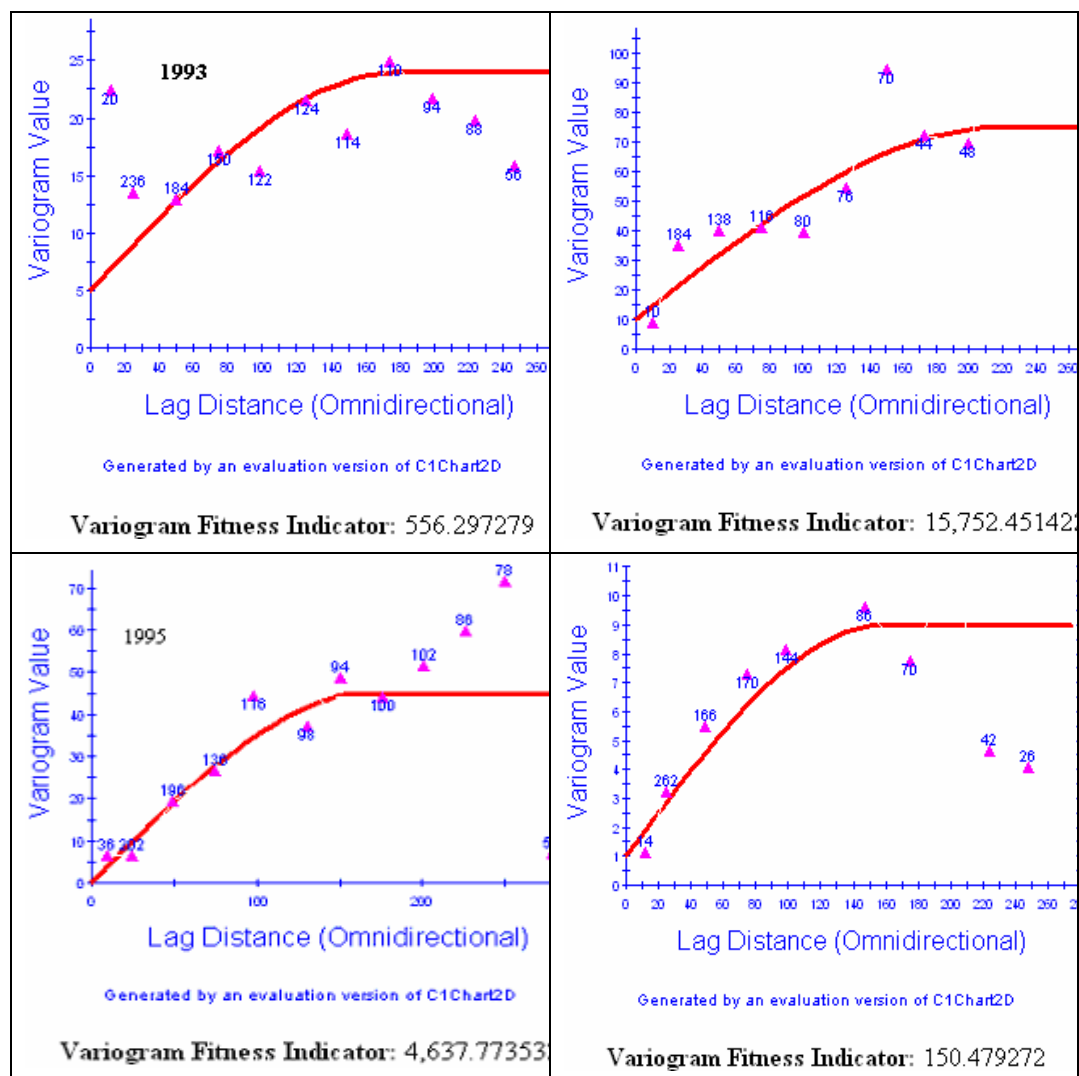


Table 31 – The SAKWeb® overall variogram for the years of 1993, 1994, 1995 and 1997.

8.7. THE SAKWEB[®] MORAN LOCATION SCATTERPLOT AND MORAN VARIANCE SCATTERPLOT

According to Figure 101 and the spatial location of the observations (see Appendix E), it is notable that no grasshopper sampling was undertaken in the Rocky Mountains of Colorado. Generally, there was over-sampling throughout the eastern plains (a region limited by Fort Collins, Boulder, Cripple Creek, Florence, San Luis, Trinidad, Springfield, Cheyenne Wells, Wray and Sterling) and the western area of Craig, Dinosaur, Grand Junction, Delta, Glenwood Springs and Steamboat Springs. In addition, the Telluride and Silverton districts were sampled in 1993 and 1997.

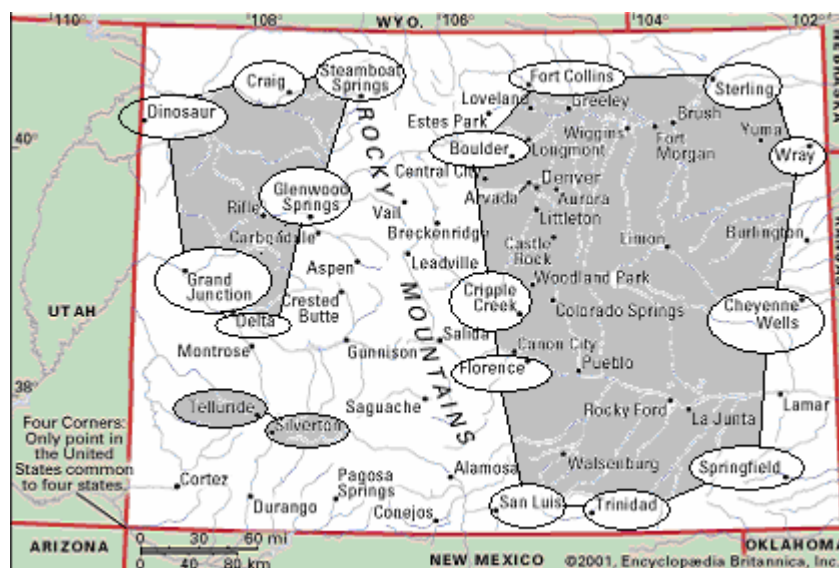


Figure 101 – The map of Colorado, USA.

Two patterns are clear in Figure 102. The positive one lies in the southeast region while the diagonal between the southwest and the northeast does not cause concern about any grasshopper outbreaks. Samples 65, 67, 68, 69, 20 and 45 in quadrant I represent uncertainty observations due to **W** misspecification issue (see subsection 7.2.1.6 for further discussion) and, carefully examination shows that these sites are characterized by a shift region of high and low patterns. The difference between

these points and site 54 (the green symbol in quadrant I) lies in the fact that the latter has a small value (0.7) while the previous ones present high levels of grasshopper concentration (4.3, 4.5, 4.3 and 4.3, respectively). Note the high grasshopper values in the north-west region highlighted by samples 12 (4.8) and 13 (5.4). The green symbols (samples 46, 47 and 49) of quadrant III represents an uncertainty area due to the cutoff mean where densities have a medium value surrounded by low neighbors. As expected, the major negative outliers of quadrant II are situated in the heart of quadrant I, giving the idea that grasshopper infestation is not a continuous process but, rather, the area is sprinkled with high and low spots.

The positive outliers (103, 92, 82 and 11) of quadrant IV follow the previous trend, though these sites are restricted in the negative pattern. However, the high grasshopper densities of sites 43 (21), 48 (5.8), 45 (5.1) and 38 (11.2) are localized in an unstable region surrounded by samples of the other three quadrants. Samples 41, 42 and 43 (the black symbols) of the Moran variance scatterplot show this situation (see Appendix E). The outlier sites 25, 34 and 36 (the yellow symbols) and 19 and 59 (the green symbols) circumscribe this region, too. A further area of medium instability exposed by this measure is in the northwest counties.

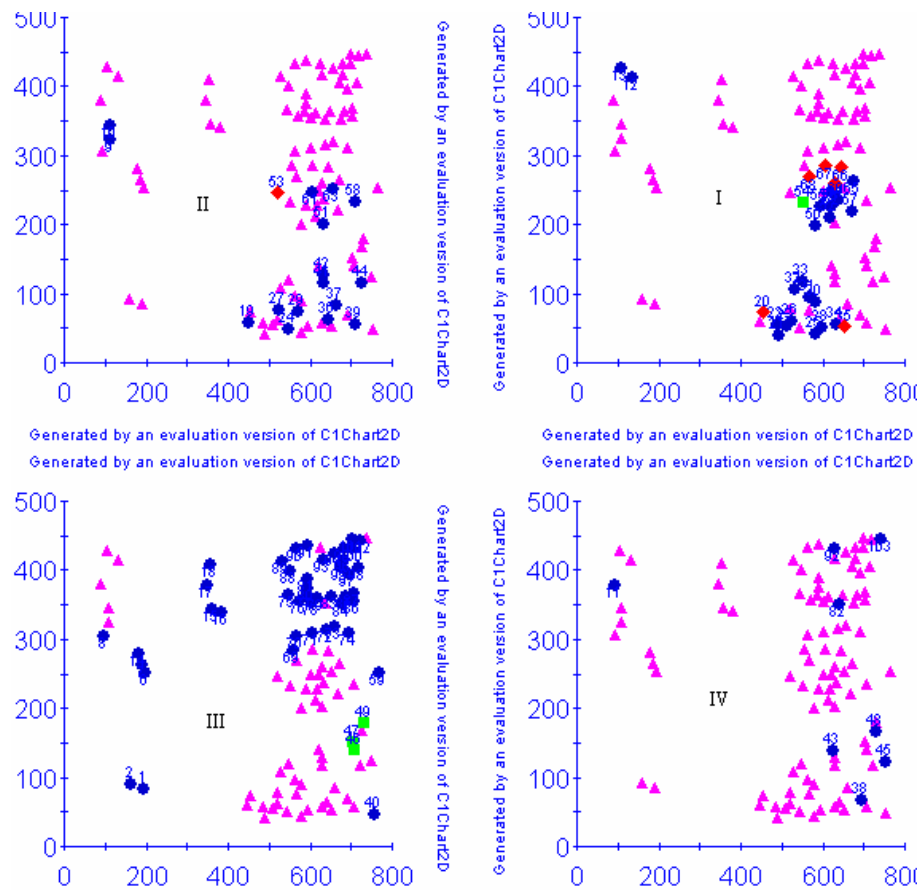


Figure 102 – The SAKWeb[®] Moran location scatterplot of Colorado in 1993.

The unstable 1994 dataset uncovered by the SAKWeb[®] location and Moran variance scatterplots confirms the long range and high nugget-effect of the variogram fitness procedure (notice that this year reveals the highest sill value). There is a major grasshopper infestation in the Cheyenne Wells, Rocky Ford and Lamar districts (central-east region) while a high scarcity of grasshoppers is recorded in three other areas:

1. The Sterling, Fort Collins, Boulder, Cripple Creek, Limon and Wray districts (north-east region);
2. The Steamboat Springs, Craig, Rifle, Grand Junction and Glenwood Springs districts (north-west region);
3. Springfield, Trinidad, San Luis and Walsenburg (south-east region).

In this year, quadrant III highlights two separate regions with the quadrant I locations somewhere in the middle (see Figure 103). Hence, a second order polynomial should be applied for a detrend operation and confirmed by the strange

overall shape of SAKWeb[®] linear regression (see Appendix E). As with the previous year, quadrant III is followed by quadrant I and, therefore, the linear trend regression fitted quite well.

Positive and negative outliers are spread out in the map without any cluster patterning, which confirms the imperfect behavior of the experimental variogram. Since the global average of grasshopper densities in this year is quite high (when compared with the other three years) and the global number of observations considered as a positive pattern is less than half of quadrant III, this suggests that sites in quadrant I are high or sites in quadrant III are low. Nevertheless, the existence of a strong infestation gap between both quadrants and sub-regions is a correct inference: 15.76 against 2.27.

Site 15 (0.3) of quadrant II represent uncertainty points (as a result of the **W** misspecification issue presented in subsection 7.2.1.6) with a low intensity but with neighbors that register a medium grasshopper density. If this situation is not considered then these points will be shown as true negative outliers (low grasshopper concentrations surrounded by high values). Furthermore, the false outlier 35 (4.8) represents an average infestation with close high densities (a true outlier classification if a conventional Moran scatterplot is created).

Note that although the Moran variance scatterplot highlights major positive and negative outliers, some of them are not emphasized on account of the threshold setup for both axes (in SAKWeb[®], both cutoff limits are 14.3% of the maximum values reached on each axis)⁴¹².

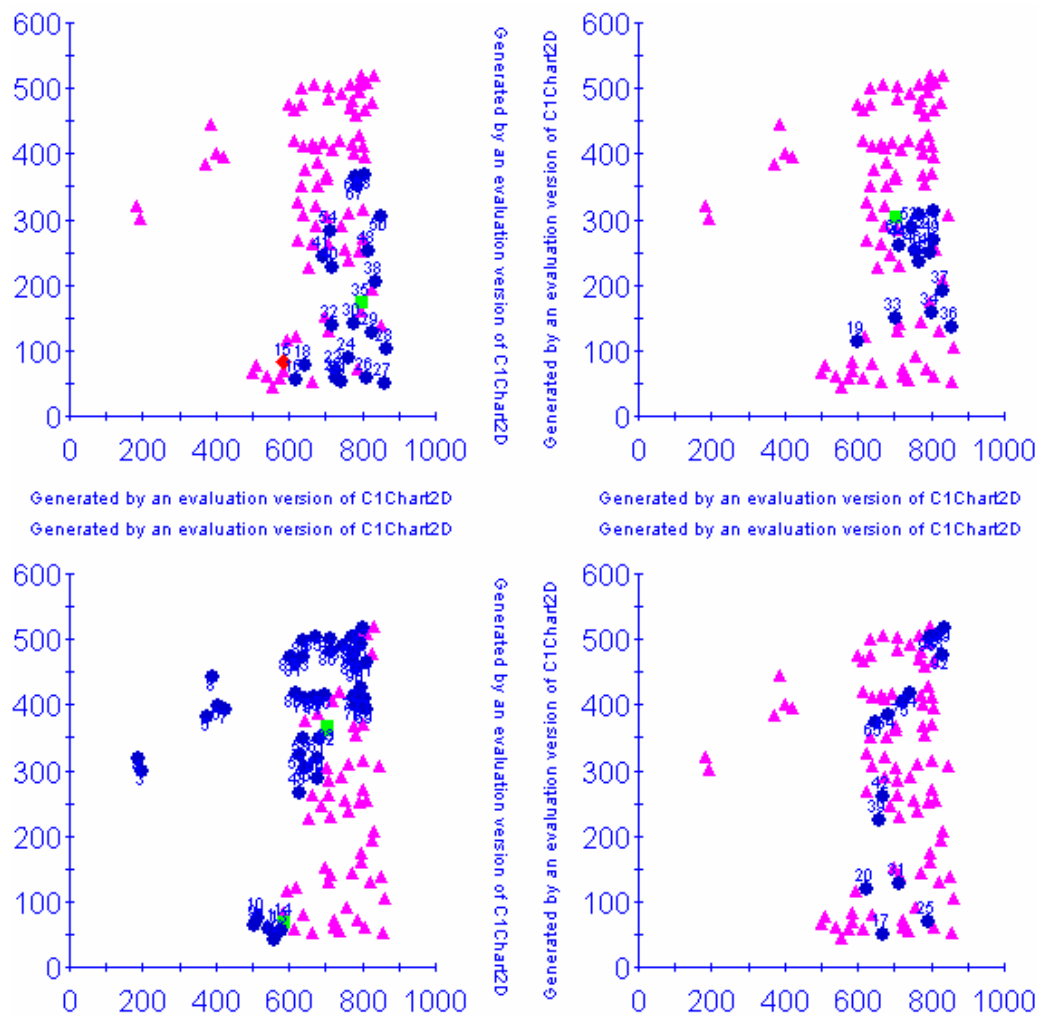


Figure 103 – The SAKWeb[®] Moran location scatterplot of Colorado for 1994.

According to the 1995 dataset, two outbreaks were triggered: in the Burlington region, with an average of 17.68 grasshoppers per m², and the Lamar-La Junta-Springfield districts, with an average of 9.7 grasshoppers per m². Notice that the overall average is 3.15⁴¹³. Fourteen negative outliers and five positive outliers emerge from quadrant II and IV and the need to recheck their original values can be a requirement: sample 19 (4), sample 42 (13.4), sample 47 (19.2), sample 88 (3.7) and sample 97 (9).

The main difference between this Moran location scatterplot and 1993 is that minor local infestation patches are more pronounced than in 1995 then in 1993 because of the existence of a long range structure (see Table 32 for the relationship between the variogram and the Moran I correlogram). This trend is confirmed by the smaller dissimilarity values of the variogram for the first year.

With isotropic variograms, there is a direct relationship between the lag distance of the Moran I correlogram and the variogram range (x-axis).
With geometric anisotropy, there is a close match between the minor range of the variogram and the distance of the Moran I correlogram at which reaches zero value.
There is an inverse relationship between the positive value of the Moran I correlogram and the variogram sill (y-axis).
Although it can be hard to detect spatial autocorrelation for shorter lag distances due to the lack of observations, a strong Moran I correlogram for small distances suggests a lower nugget-effect.
Unlike the Moran I correlogram ⁴¹⁴ , the comparison between variogram sill is valid if and only if the sample sites have similar values. Quite often, lower densities lead to lower dissimilarity values.
No variography or Kriging should be carried out for a zero and negative Moran I correlogram for all lag distances since it does not respect Tobler's Law ⁴¹⁵ .
Deeper slopes in the Moran I correlogram indicate a shift from a Spherical to an Exponential model while flat profiles with a high spatial autocorrelation reveal a shift from a Spherical to a Gaussian one ⁴¹⁶ .

Table 32 – Main trends in the Moran I correlogram and the variogram parameters.

If a closer look is taken at the Moran variance scatterplot, six observations in the south and eight in the north-center are to be considered outliers. It becomes clear that those samples are located within low and high local spatial clusters, respectively. However, this spatial measure does not discriminate positive and negative outliers unless the sample value is also displayed. Also, samples 67 and 42 are stressed as a high variability region because their neighboring sites are quite different. For instance, sample 66 (9), sample 68 (19.2), sample 69 (23.4), sample 62 (1.1) and sample 63 (1.3) are the neighbors of site 67 that has a grasshopper infestation of 26.8.

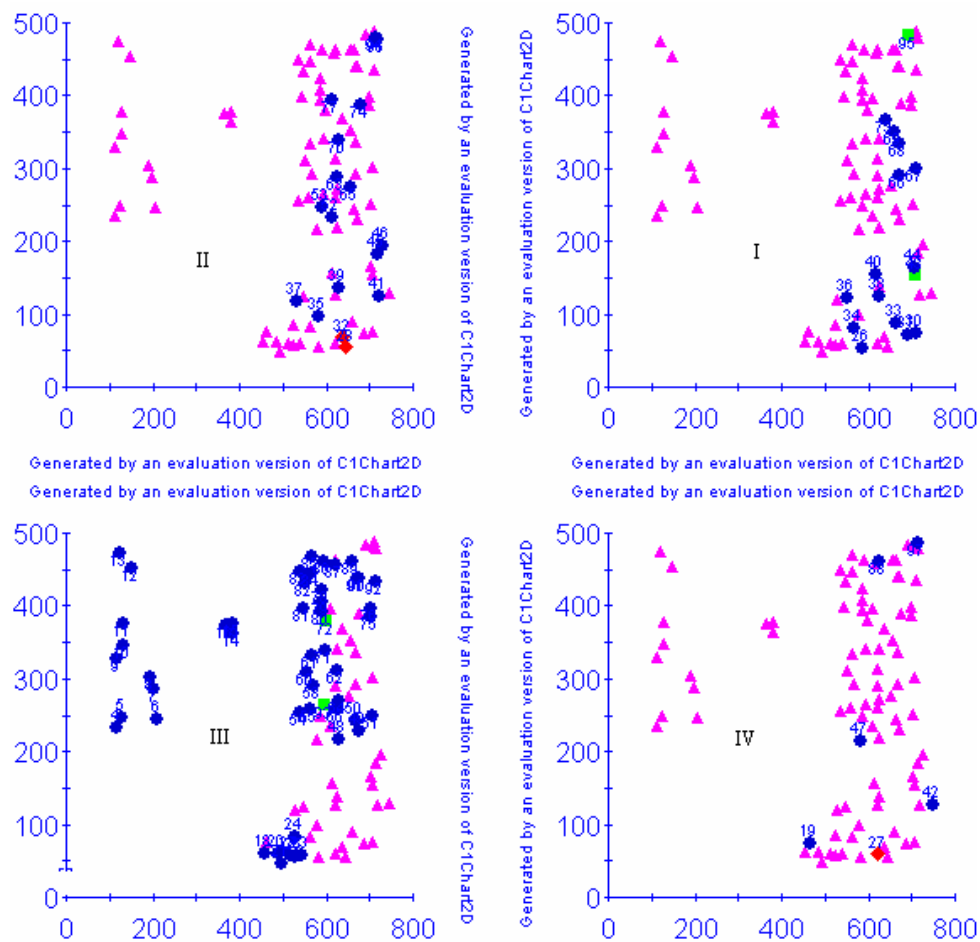


Figure 104 – The SAKWeb® Moran location scatterplot of Colorado in 1995.

The year 1997 can be described as one with small spatial patterns and big concerns for local land owners. According to Figure 105, several negative and positive local spots in a scatter arrangement can be found, leading to a experimental variogram structure with a short range. Furthermore, the central-east region of Wray, Limon and Cheyenne Wells presents some grasshopper infestation fluctuations which have a negative impact on the accuracy of the final estimates. Hence, according to Isaaks and Srivastava [1989], “these transitional regions are more disposed to misclassification for a particular cutoff limit, with major economic consequences, than regions classified as exhibiting positive or negative patterns”.

However, the low global mean makes this year the most tranquil one for farmers and local state funding. In fact, the positive outliers in quadrant IV are lower than in the previous years with a range of [2.9, 7.6] and a global average of 4.19 grasshoppers per m^2 . Compared with the 1994 dataset, for instance, the range

equals [5.4, 32.6] while the positive outlier average equals 11.12 grasshoppers per m².

As in previous years, the central and east region of the Burlington and Cheyenne Wells districts registers high levels of grasshopper infestation. In addition, the west side of the Rocky Mountains in Colorado and the northern and west districts of Sterling, Yuma, Fort Morgan, Wiggins and Fort Collins do not suffer from any grasshopper outbreak in these four years. It is the cold wind off the permanent ice in the Aspen district that is responsible for this pattern? Is there any other local pocket to be taken, though it is not shown by this approach because of the sampling design? What are the disparities in the physical and biotic factors in the eastern and western region that may influence the grasshopper's outbreaks in the east? Does it rain more or is it warmer in the eastern region than in the west? Should more samples be taken in the eastern districts to confirm the permanent low densities of grasshoppers? In terms of soil attributes such as salinity, clay content or organic matter, how does oviposition differ?⁴¹⁷ Since there is an enormous variety of grasshoppers, do these results reflect all of them? Comparing levels of grasshopper infestations with those in other countries, is there any spatial characteristic to be revealed? Can I extrapolate the grasshopper infestation levels in adjacent states for this year and next? Is there any preferred route for grasshopper plagues? Does the direction of the prevailing wind correspond to those routes? What are the links between food preference location and these spatial patterns shown in Figure 102, 103, 104 and 105? Is there any long-term cycle of grasshopper infestations that cannot be captured by the present dataset? According to the University of Wyoming [2002], "major epidemics occurred in 1936-38, 1957-58 and 1980-82, at approximately 22 year intervals". Smaller epidemics occurred midway between major outbreaks. This is simply a confirmation that major droughts in the western United States tend to occur at 22 year intervals, followed by minor ones at 11-year intervals. To sum up, recapping Section 2.1, the added value of Geography is to discuss spatial autocorrelation and spatial interpolation processes to find space rules for other science experts in the context of a process of detection.

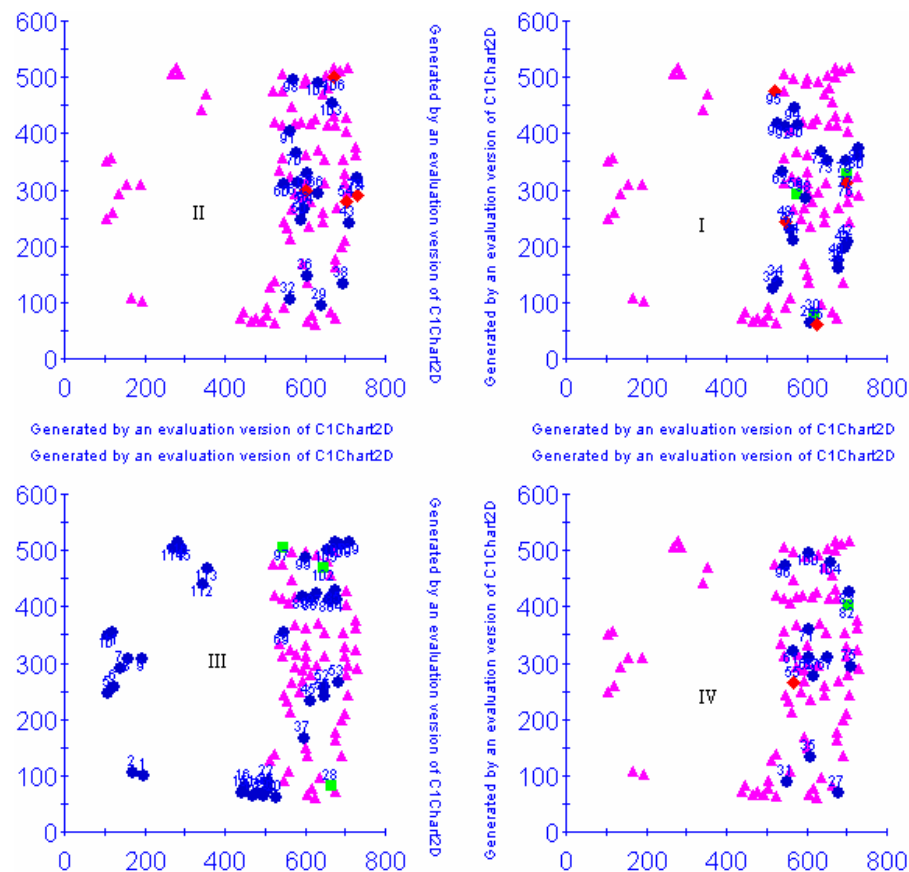


Figure 105 – The SAKWeb[®] Moran location scatterplot of Colorado for 1997.

8.8. SAKWEB[®] ORDINARY KRIGING

Although OK variance is an excessively spatial operation (see Section 3.6 for further discussion), this uncertainty measure is considered essential for most users to assign confidence intervals. Nevertheless, as stated in subsection 7.2.3, some type of transformation must be applied to take into account the discrepancy between the sample variance and the variogram sill. This happens because most times, the variogram sill overestimates the global sample variance. With SAKWeb[®], each sample weight⁴¹⁸ relies on the nearest neighbor distance between all samples and the estimated one. The estimated global mean, EGM, and the estimated global variance, EGV, are computed on the basis of this rule. Naturally, if the EGV is greater or less than the sill then variogram rescaling must be setup⁴¹⁹.

According to Table 33, this rescale operation is important for all years. With the exception of 1993, the same trend is followed by the higher sill of the variogram versus the EGV that confirms the Cressie [1993] statements in subsection 7.2.3. Therefore, the following rescale factors were applied to the original variograms:

- 1.232 in 1993;
- 0.76 in 1994;
- 0.627 in 1995;
- 0.647 in 1997.

	1993	1994	1995	1997
EGV	29.57	57.01	28.21	5.82
Classical Variance	23.22	50.79	23.96	6.52
Variogram Sill	24	75	45	9

Table 33 – Comparison of EGV, classical variance and variogram sill results for the four years.

Another closely related issue for OK variance regards the inclusion of measurement error. A lower prediction of the Kriging estimations becomes real with the exception of the samples available (it is zero when no measurement error is included). SAKWeb[®] has this capability too⁴²⁰.

It became clear in subsection 7.2.4 that the nugget-effect has a major impact on the OK estimates. The maps presented in Tables 34, 35, 36 and 37 confirm this. For each year, four final maps will be displayed (note the simultaneous 2D and 3D views obtained by SAKWeb[®]). However, the measurement error is not plotted since its result does not differ much when compared with the exact OK model with the nugget-effect (model 1). Only at sample locations, the results differ according to the following two rules:

- With the measurement error approach, the OK estimation can be lower or higher when compared with the sample itself (see Figure 75 in subsection 7.2.4);
- With the exact OK nugget-effect, the OK estimation equals the sample value.

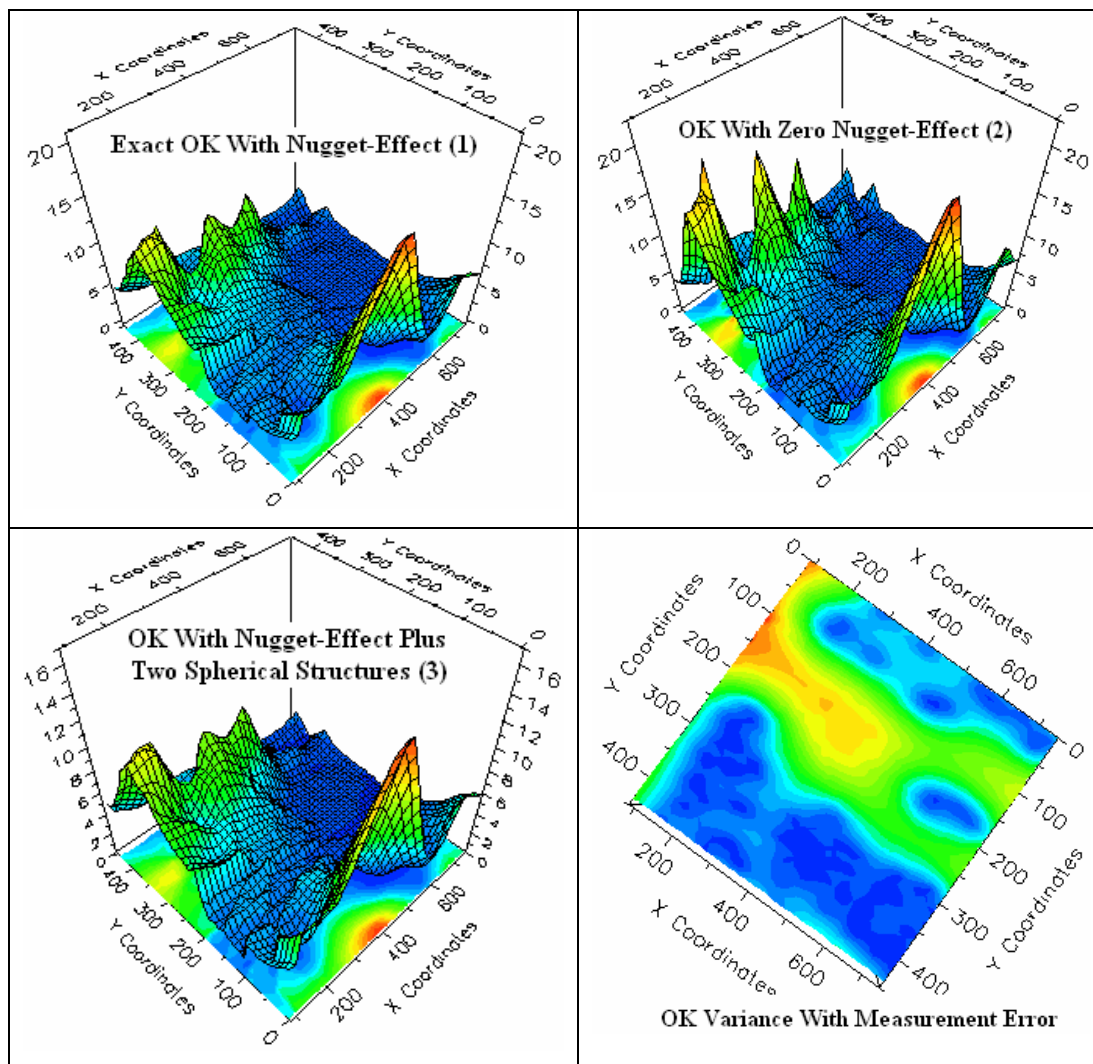


Table 34 – SAKWeb® OK and OK variance with measurement error (30%) of grasshopper Colorado (1993).

As expected, higher OK variances are expected where there is a lack of observations (particularly true along the Rocky Mountains in Colorado). The major exception relates to the 1994 spatial dataset because of the lack of observations on the western side of Colorado, thus leading to a higher OK variance (see Table 35). The differences in the OK estimates are also considerably smaller in those same areas, giving a flat layout on the estimate map⁴²¹. Furthermore, the existence of a nugget-effect reflects a higher OK variance, on average, against OK without C0. Finally, the average for the OK variance that uses OK with two variogram structures is always higher than that for the other methods because of the higher sill (note that this overall sill is the sill sum of the micro and the long-range variograms).

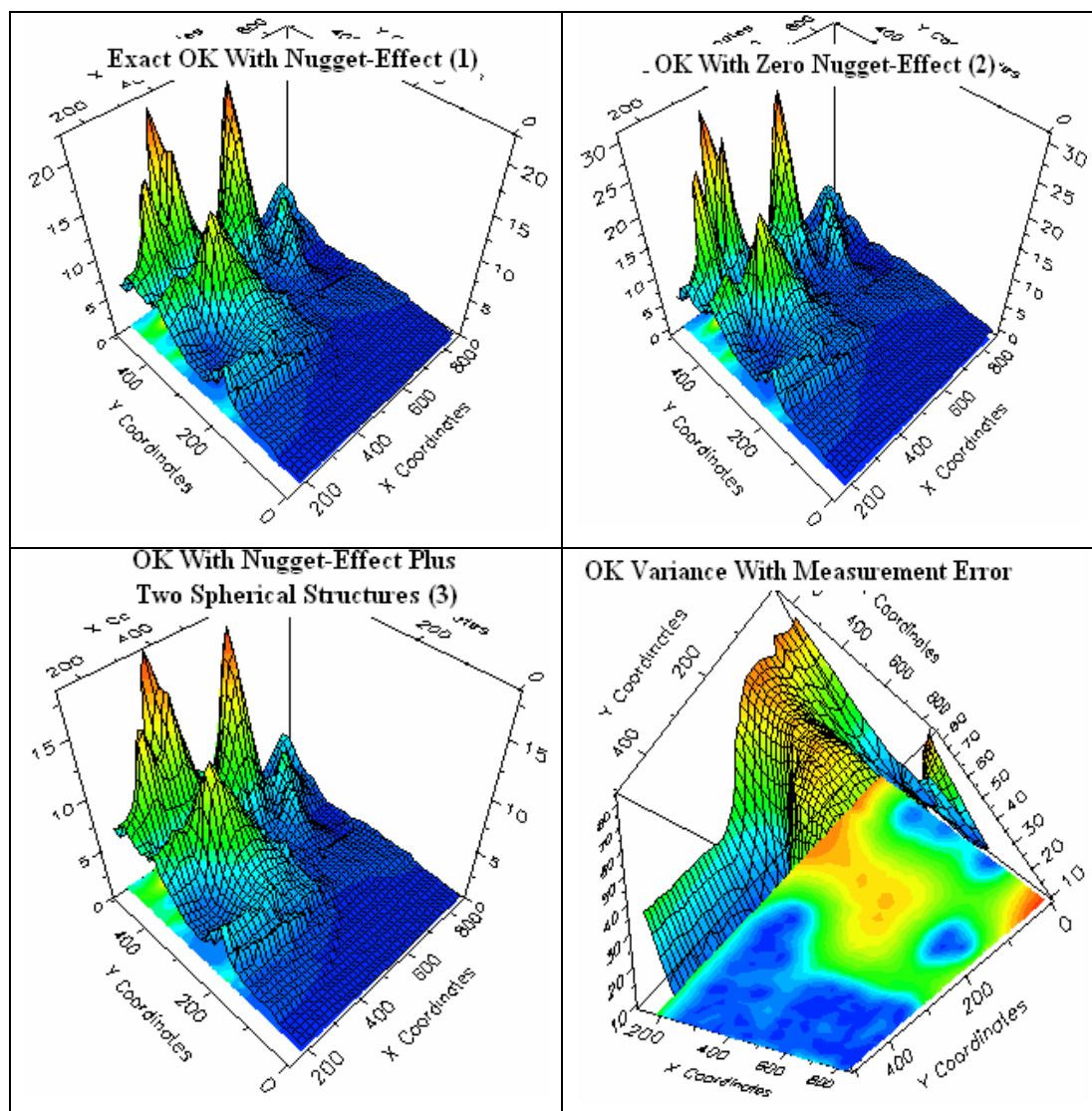


Table 35 – SAKWeb® OK and OK variance with measurement error (20% with a rescaled variogram sill factor of 0.7602) for the grasshopper dataset, Colorado (1994).

Another pattern with these four tables is that OK with zero nugget-effect always achieves a wider range for the final interpolation. In confirmation of previous conclusions, as the nugget-effect gets higher, the interpolation values become smoother and closer to the overall mean. Thus, the difference between exact OK with a nugget-effect and OK with the two structures is smaller than when no nugget-effect is taken into account.

These previous trends are not followed in the 1995 dataset because the original variogram does not have C0. According to Appendix E, the discrepancy arising from the presence or absence of C0 is nil and the variation among the other

approaches hardly exists. Because all methods present the same variogram for a lag of less than the SSI⁴²² distance, these minor differences correspond to the power weights of further distance samples. In fact, major Kriging weights rely on the estimation itself.

According to Section 8.1, the general trend reveals a strong upward drift longitudinally from west to east, in all years. Nevertheless, two exceptions must be made:

- The zero densities of the western half of Colorado leads to a small flat surface in the linear polynomial for the 1994, 1995 and 1997 datasets;
- The 1993 OK interpolation reveals that a second order polynomial trend is more suitable than the linear one.

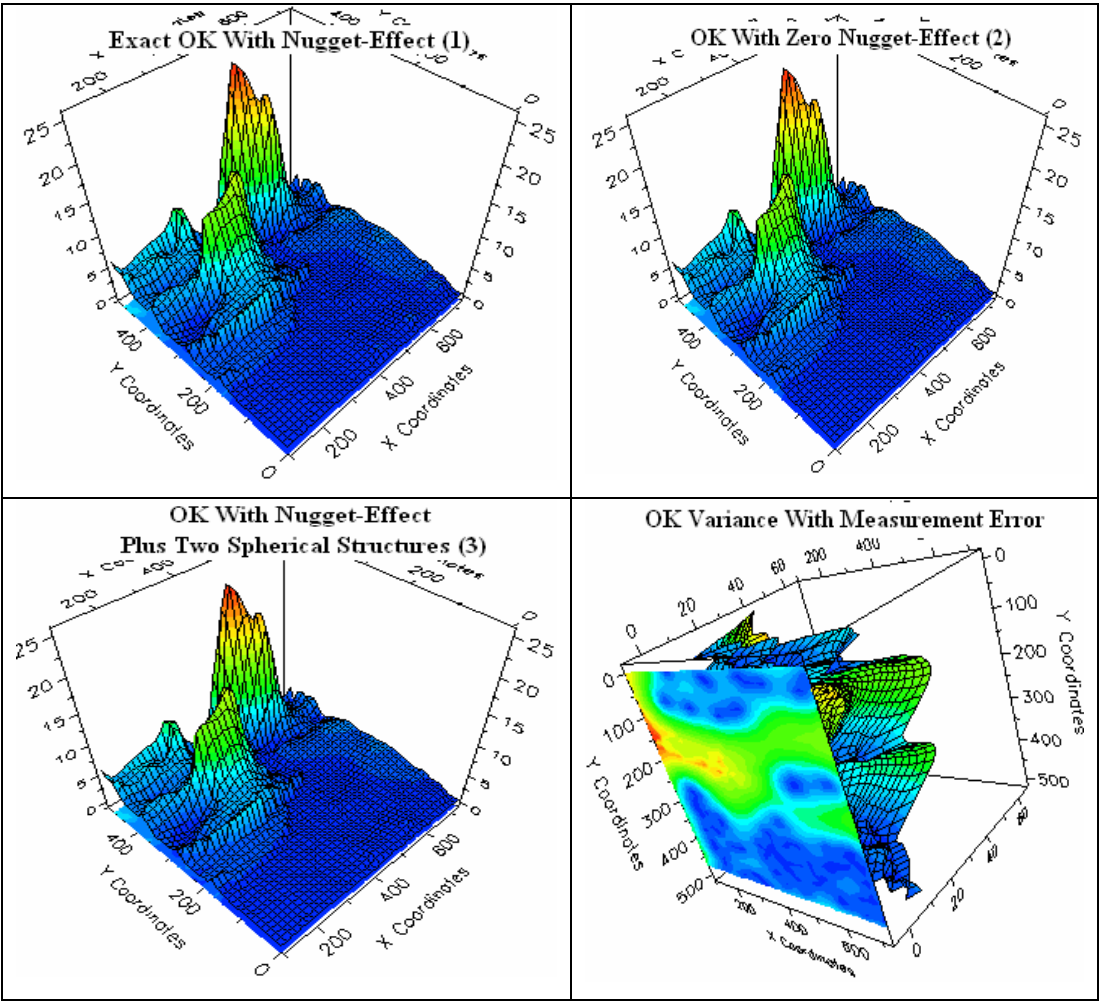


Table 36 – SAKWeb[®] OK and OK variance with measurement error (20%) for the grasshopper, Colorado (1995).

Another inference concerns the 1995 interpolation. Besides the fact that it has the greatest grasshopper outbreak for all years, the SAKWeb[®] layout calls for particular attention because the infestation level is concentrated in a single place. While all other years present a sprinkle pattern of high and low values in the eastern lands of Colorado, this year registers a unique cluster of grasshopper infestation.

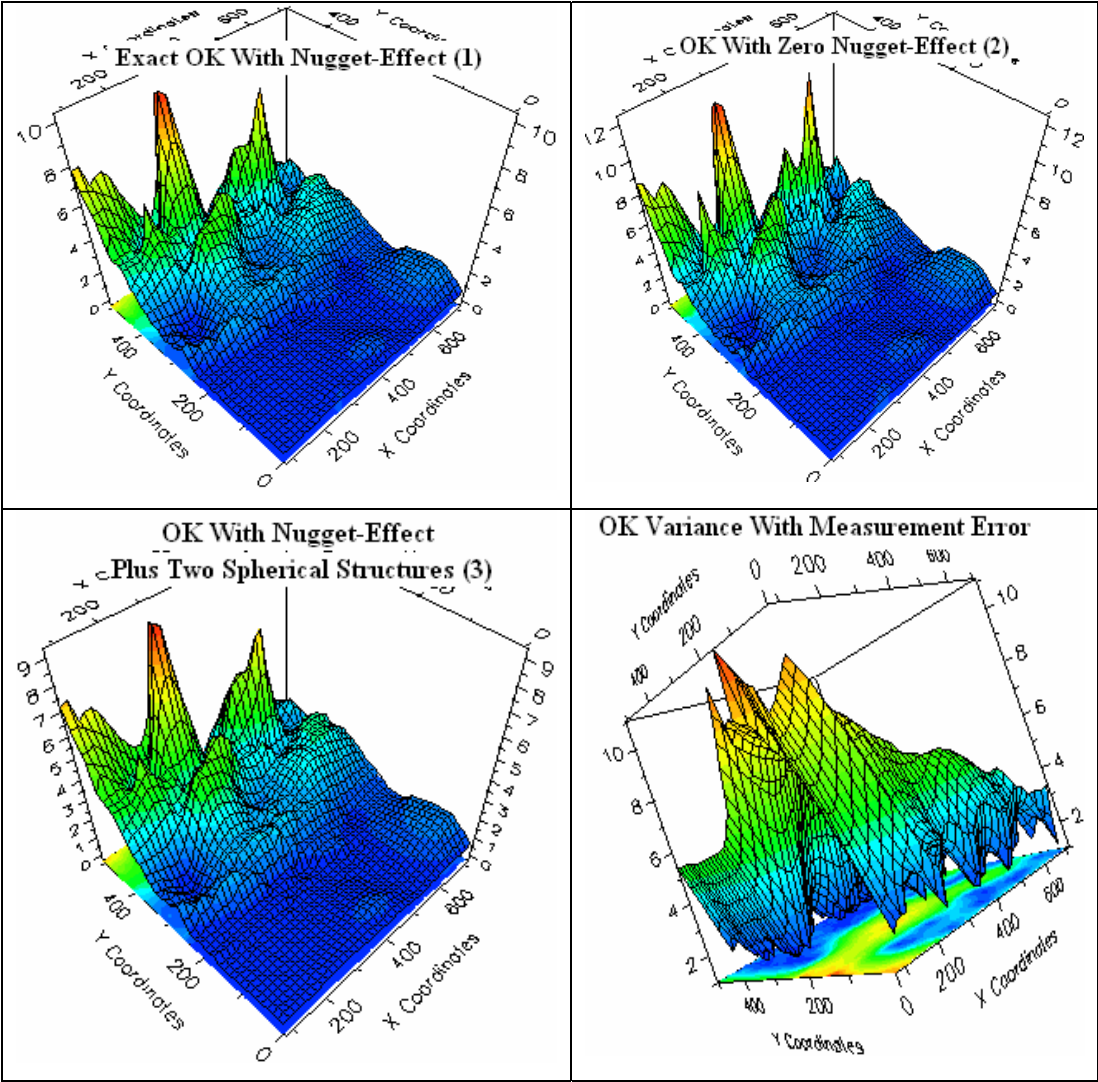


Table 37 – SAKWeb[®] OK and OK variance with measurement error (20%) for the grasshopper, Colorado (1997).

8.9. SAKWEB[®] VALIDATION PROCEDURES

On the basis of four additional validation datasets, the SAKWeb[®] results confirm previous conclusions in a validation procedure (see Appendix E):

- Contrary to Hypothesis 5, once again, no evidence of neighborhood patterns can be found in relation to which nugget-effect fits best.
- Statistically, the default nugget-effect strategy of the major software presents worst outcomes against the zero and two-structure models. In fact, OK with C0 only records better results with validation sample 1 for the 1993 and 1997 datasets.
- In 1995, OK with no C0, OK with measurement error and OK with C0 register the same outcome because the original variogram does not have any nugget-effect.
- In 1993 and 1997, there is a clear superiority of OK with two structures over the other options. The assumption of Section 8.3 regarding the existence of local and regional scales is confirmed. Nevertheless, this trend is not fully pursued in 1994 and 1995 where OK with two structures and OK without C0 share the best results.
- Two reasons were pointed out in Section 8.2 regarding the need for a wider variogram range for the 1994 grasshopper dataset. If the range is 100 km⁴²³ then a strange and non-sense outcome is given by SAKWeb[®] (see Figure 106) because of the lack of observations within the OK system⁴²⁴.

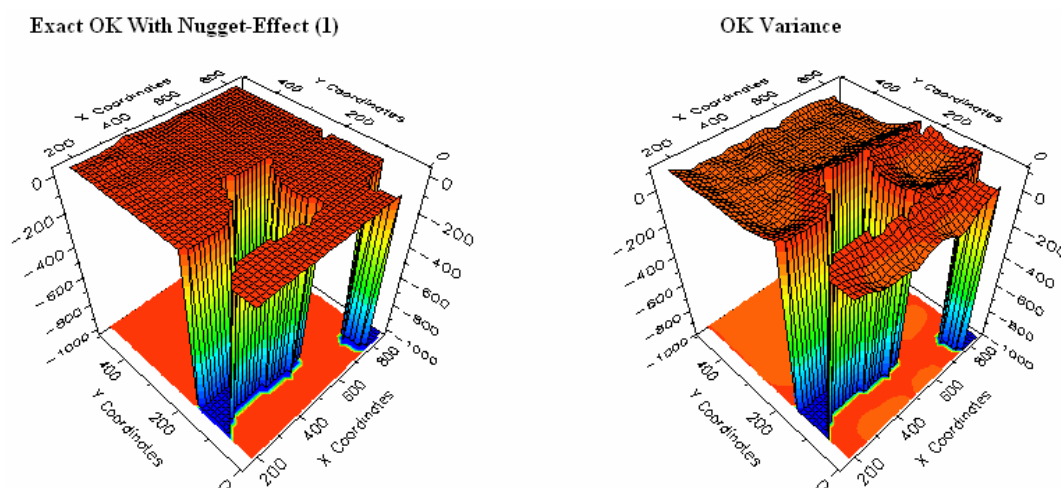


Figure 106 – The SAKWeb[®] OK estimation and OK variance for a 100 km range.

- A non-appropriate variogram range leads to a larger validation error as Table 38 confirms⁴²⁵. Since the 1993 and 1994 datasets reveal a positive Moran I for the default variogram range (180 and 220, respectively), the number of lower error estimations for these validation datasets are higher when compared with other ranges (150 and 200, respectively). Hence, the information given by the SAKWeb[®] Moran I correlogram can turn out to be very useful for the estimation of the variogram range.

Year	Variogram Range	Number Of Estimations With Higher Error	Number Of Estimations With The Same Error	Number Of Estimations With Lower Error
1993	Range=180 (versus Range=150)	8	4	15
1994	Range=220 (versus Range=200)	6	3	15

Table 38 – According to the two variogram ranges, the error comparison for the 1993 and 1994 grasshopper datasets clearly demonstrates the appropriate hints of the Moran I correlogram.

- The cross-validation of SAKWeb[®] (see Table 39) backs up previous results whose root-mean-square prediction error is lower⁴²⁶ for model 3⁴²⁷ in 1993, 1994 and 1997. In 1995, model 1⁴²⁸ and model 2⁴²⁹ register the same results because the variogram nugget-effect equals zero. Oddly, model 3 has a

similar outcome, though always slightly lower when compared with the previous models.

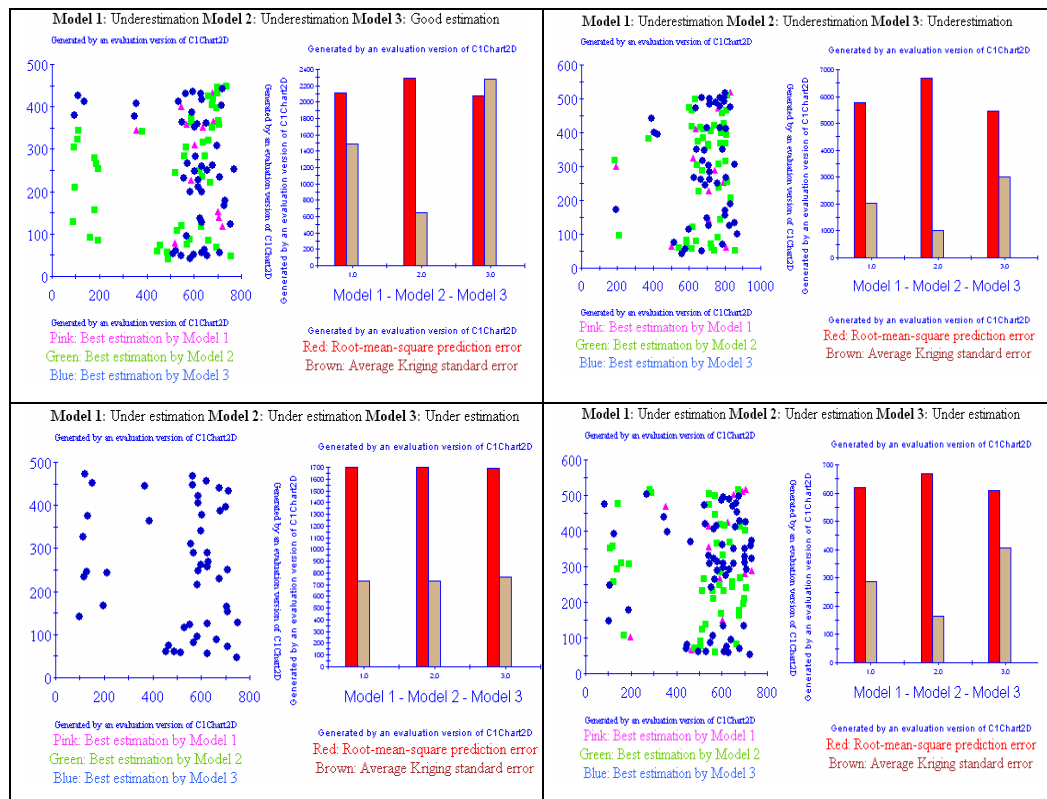
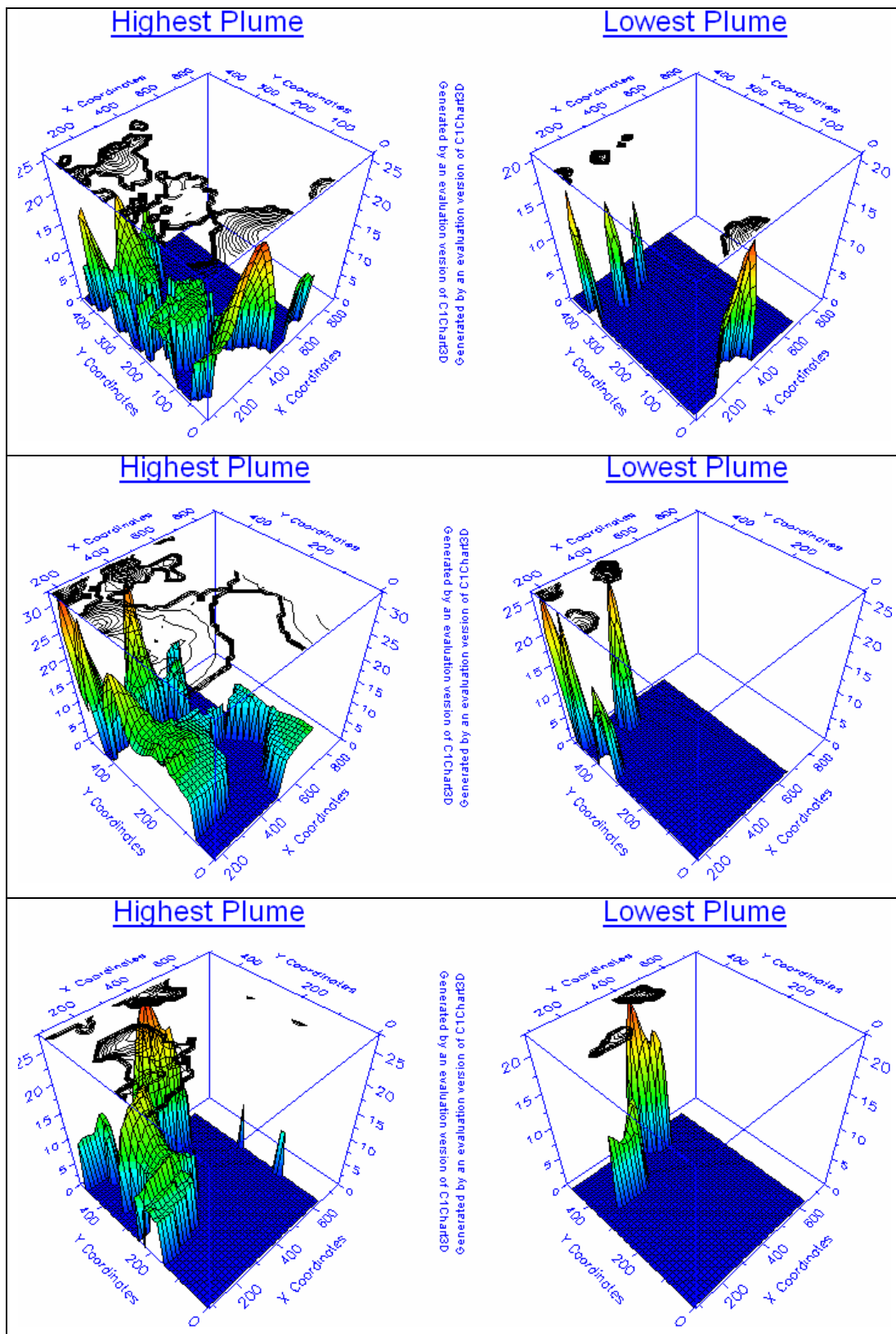


Table 39 – The grasshopper cross-validation for 1993, 1994 (upper row), 1995 and 1997 (lower row).

- On the basis of the cross-validation procedure, the nugget-effect suggestion given in Table 32 is also confirmed. The nugget-effect of the 1993 variogram was changed from 5 to 12 (a moderately positive Moran I correlogram) and, as expected, worse results emerged: 1) The difference between the root-mean-square prediction error and the average Kriging standard error becomes higher for the three models; 2) In general, the root-mean-square prediction error becomes higher too.
- “The U.S. Department of Agriculture (1977) suggests that treatment is justified when grasshopper numbers reach approximately 9 per m²” [Skinner, 1999]. Table 40 presents the global confidence interval assessed by SAKWeb[®] for the 1993, 1994, 1995 and 1997 datasets, respectively.



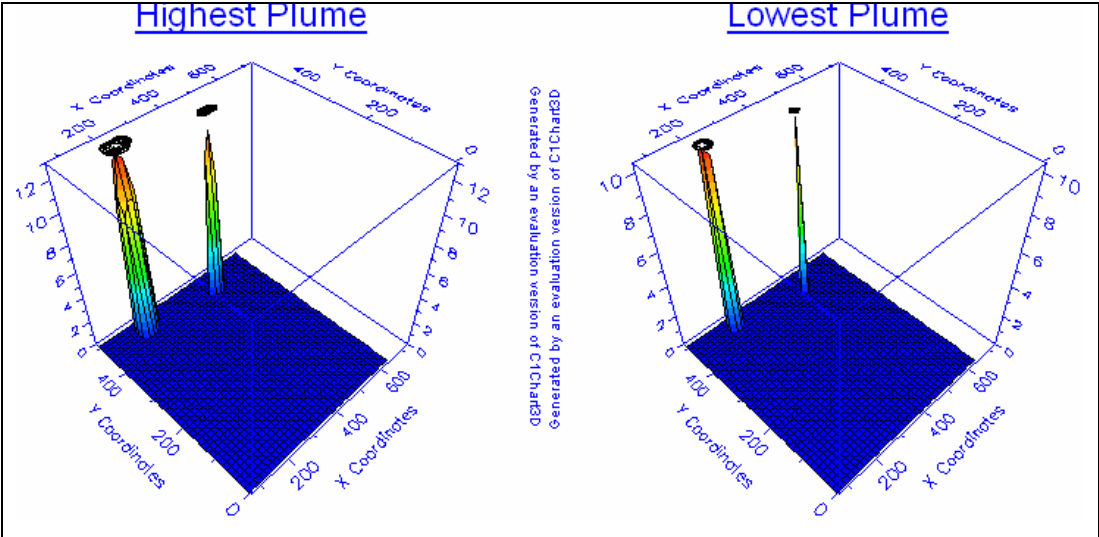
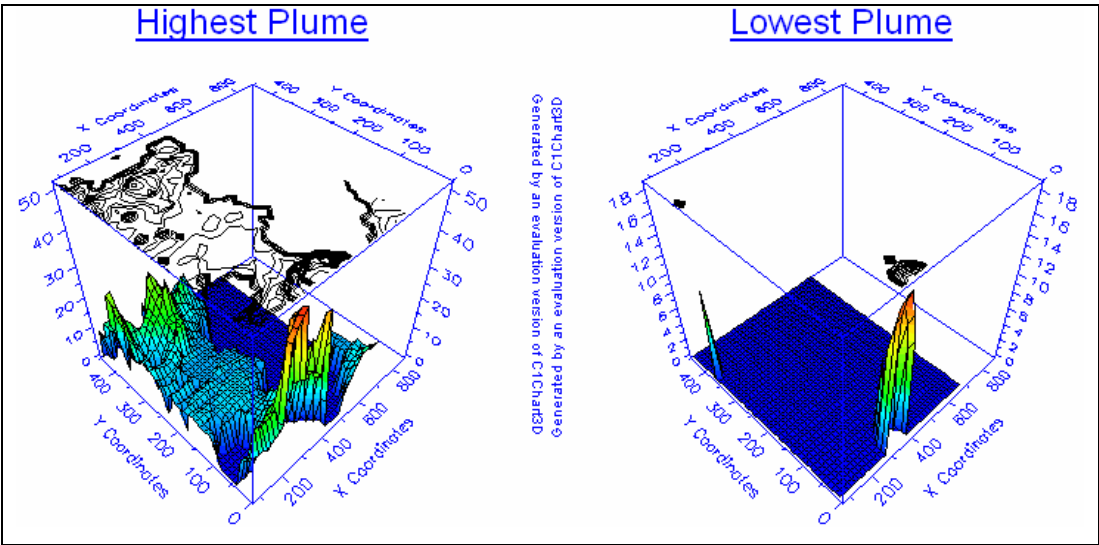


Table 40 – The global confidence interval for a level of 80% and a cutoff value of 9 with an OK model without C0.

- Table 41 presents the local confidence interval assessed by SAKWeb[®] for the 1993, 1994, 1995 and 1997 datasets, respectively.



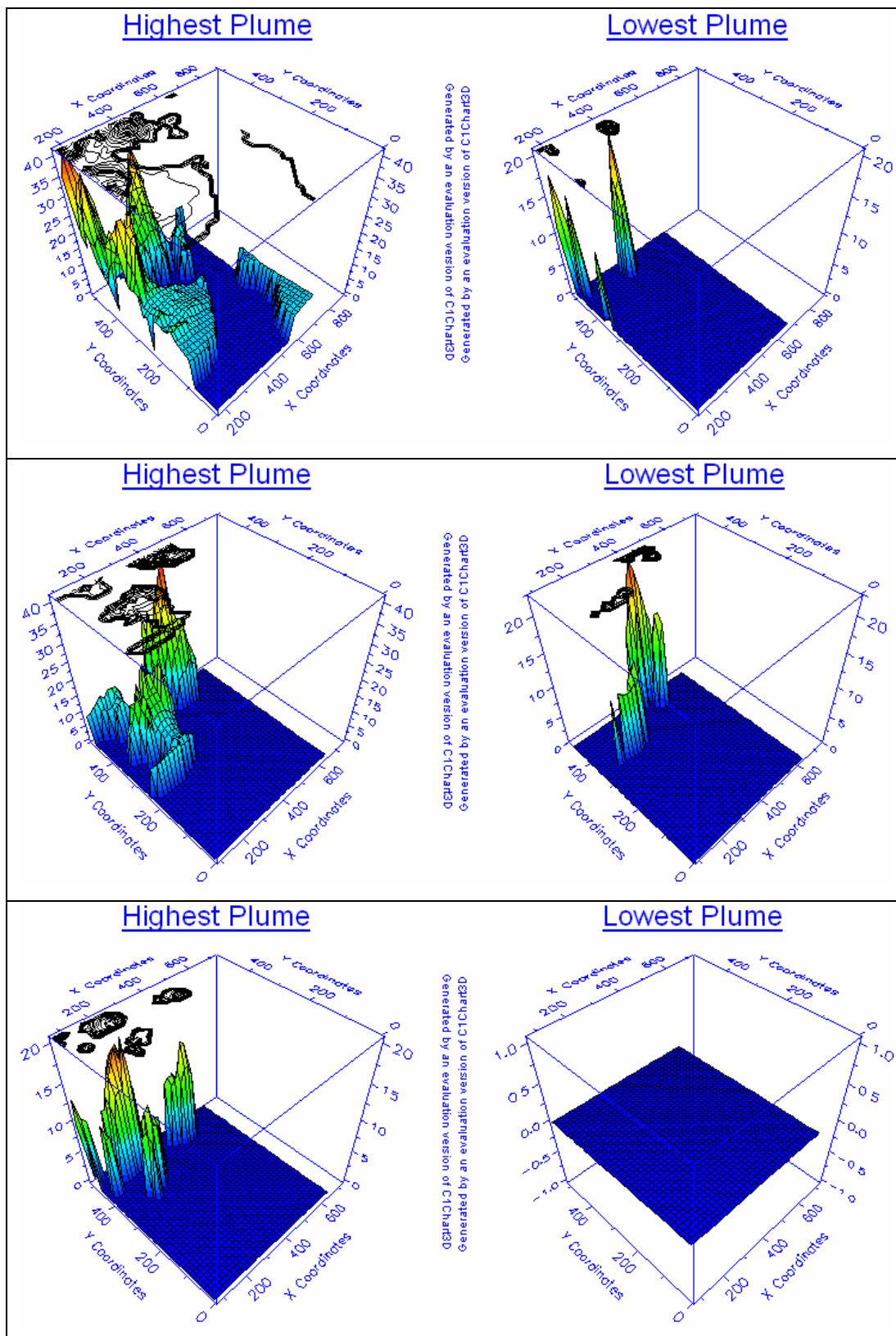
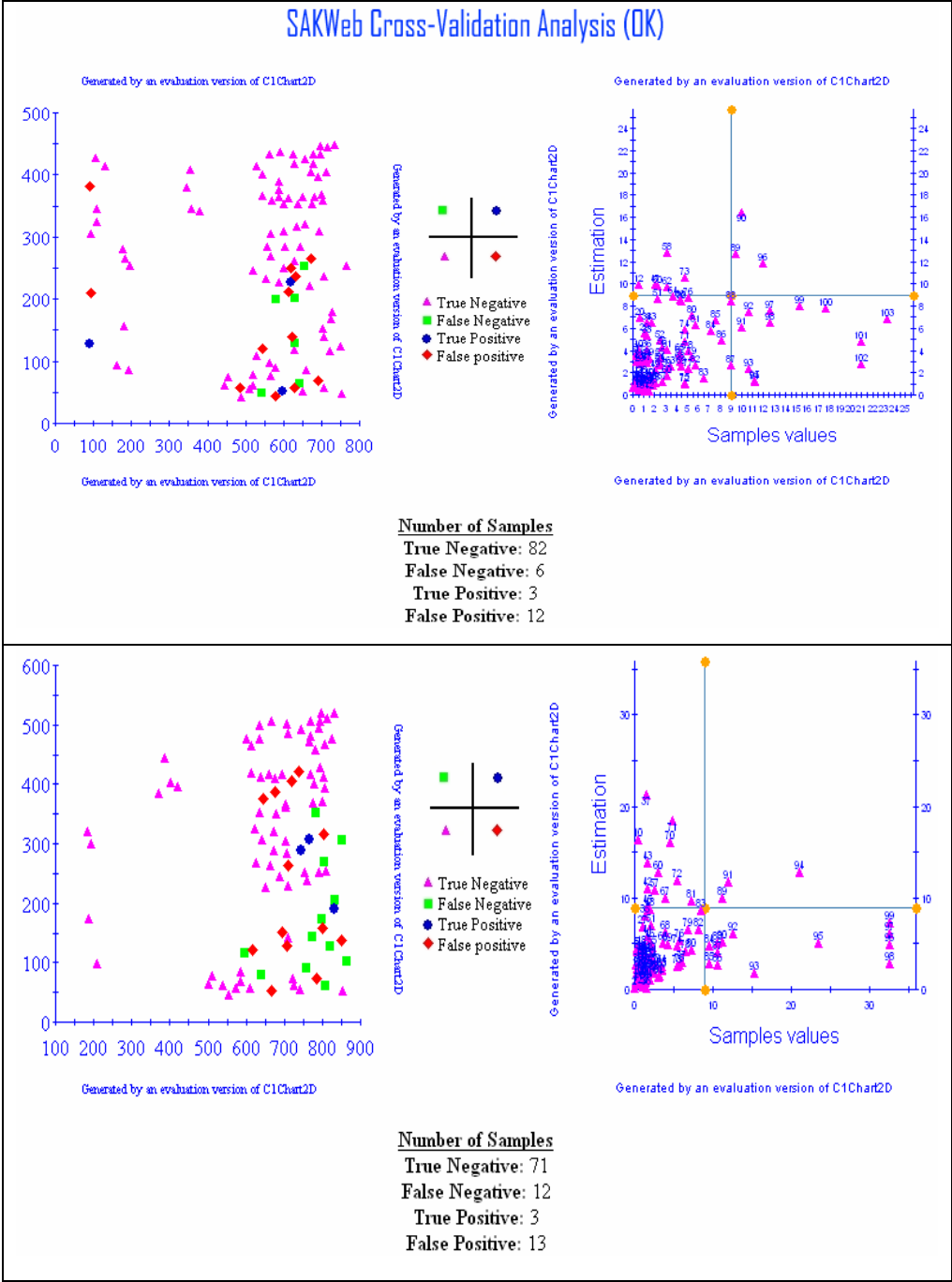


Table 41 – The local confidence interval for a level of 80% and a cutoff value of 9 with an OK model without C0.

- Based on the previous threshold treatment, Table 42 depicts the cross-validation extension to reveal the positive-negative/true-false samples for a subsequent analysis of costs⁴³⁰.



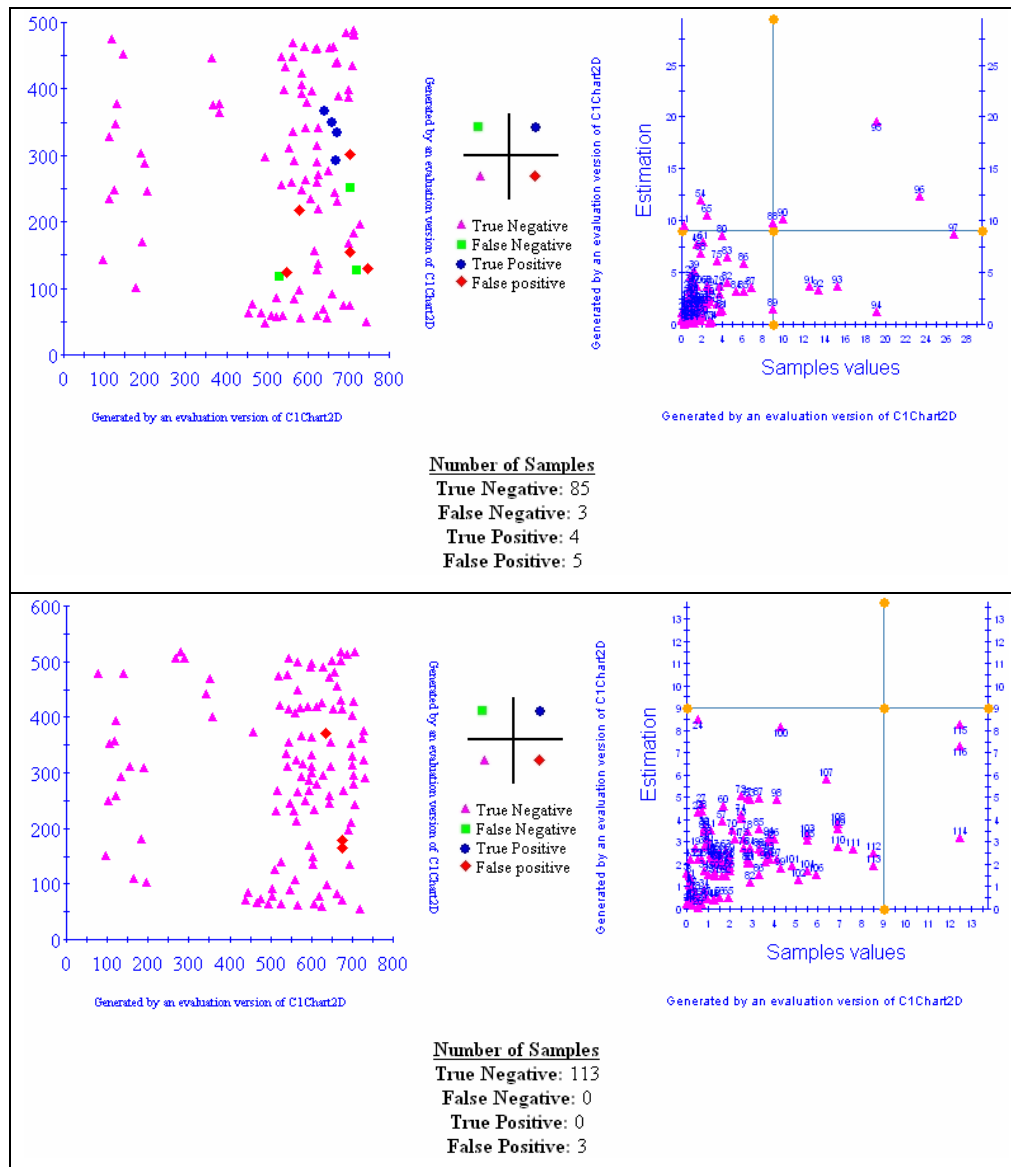


Table 42 – The grasshopper analysis of cross-validation for 1993, 1994, 1995 and 1997 (OK model with C0).

8.10. CONCLUSION

It is unrealistic to expect that only one factor totally accounts for trends in grasshopper abundance but, according to specialized sources such as Animals [2002], “weather appears to be a major driving force behind increases in grasshopper numbers”. Grasshopper density in Colorado exhibits spatial patterning on two spatial scales, implying the influence of at least two processes in the structuring of their density in space. The question is, then, “What are the factors

that create and influence both scales?” [Skinner, 1999]. In essence, there are two major categories:

- The biotic factor, on local scale, including plant community composition, competition, density, natural enemies and population dynamics and predation and parasitism.
- The abiotic factor, on global scale, covering habitat characteristics such as slope, aspect, soil properties, topography, precipitation and vegetation.

The efforts to model grasshopper dynamics on any scale should incorporate both factors. Yet, these factors are likely to differ in scale where processes at lower levels are included in higher ones in a hierarchical model:

- The lack of food in a particular area creates pressure on its local neighbors;
- Soil, slope aspect and vegetation have a major influence on an intermediate scale;
- Climatic events have a greater impact on a larger one.

A permanent observation and warning system throughout the effected region, to provide regular reports on these factors, may help individual land owners and county, state and government authorities. In addition, the greater the positive spatial autocorrelation identified, the greater the intervention of the institutions mentioned above should be. Although no cost comparisons are made here, the benefits of annual grasshopper surveys would most likely far outweigh the costs of the devastation the grasshoppers cause.

Standardization of sampling methods, frequency and timing must be achieved to investigate its temporal dynamics. With the present four datasets, positive spatial autocorrelation was detected on scales up to 220 km although this scale differed from year to year. However, the current grasshopper survey for local infestations is likely to be missed. Therefore, the sampling for monitoring purposes should be done at small spatial lags to capture the variability in abundance on both the field and regional scales.

Finally, any predicted model must incorporate a spatial-temporal scale, though its design is outside the scope of this dissertation. For instance, grasshopper populations reach their annual peak in early July when they are in the adult stage and sharply decrease in the fall at the time of oviposition. Thus, any factor that influences the grasshopper during this period will have an obvious effect on the outbreak of the next year. However, the goal of spatial autocorrelation is to explore, like detective looking for evidence and clues, and not to describe, an analyst evaluating the clues. The latter ambition belongs to the ecology experts.

9. Summary and General Conclusion

“...Promote the diffusion of analysis based on GIS throughout the scientific community and provide a conduit for disseminating information regarding GIS research, teaching and applications.”

NCGIA Annual Report, 1995

Users now have the ability to collect and to explore large amount of geo-referenced data. With the advent of Web technology and modern wireless computing, it has become necessary to develop a W3 software for interpolation to understand the often complex spatial autocorrelation that exist among the samples collected in space. The Internet geostatistical solution and, particularly, the close linkage between spatial autocorrelation and Kriging were never considered. The inclusion of these two major components provided the inspiration for SAKWeb[®] version 2.0.

In this dissertation, several new analytical methodologies are introduced and technologies are merged for the purpose of analyzing spatial data. All of the research objectives described in Section 1.2, and summarized in the six spatial analysis themes given below, were accomplished:

- The missing data issue that relates spatial autocorrelation, spatial autoregression and Kriging;
- The background analysis of deterministic and stochastic interpolation;
- An implementation review and software evaluation;
- The current spatial analysis know-how of Portuguese users;
- SAKWeb[®] developments and the technology involved;
- A detailed analysis of the grasshopper infestations in Colorado, USA.

9.1. SUMMARY

The present study began with the role of Geography among the sciences⁴³¹ and the lack of true spatial analysis methods within GIS. Besides environmental and race factors that have influenced the evolution of Mankind (“every kind of soil makes its men, its trees and its flowers”)⁴³², new applied mathematical space formulations are becoming crucial. However, mathematics will always have limitations because the earth has single and complex terrain processes⁴³³. Spatial procedures are necessary to combat this lack of statistical methods for GIS as a computational science, particularly if deductive⁴³⁴ and inductive⁴³⁵ approaches are considered. According to Ferreira and Simões [1993], “the kernel of geography is to think geographically, that is, to study spatial distribution phenomena and their correlations”.

Traditional statistics must be reformulated to properly account for spatial correlation and spatial heterogeneity within georeferenced data. Spatial autocorrelation is a reality but also a requirement to carry out spatial interpolation and spatial autoregressive modeling. For instance, if the autoregressive residuals reveal a medium-strong spatial autocorrelation then any missing variable within the initial regression model can be significant.

For Kriging, classical statistics are viewed as an aspatial diagnostic tool. Yet, the neighborhood view of spatial autocorrelation measures can provide a better ESDA tool for the recognition of atypical locations, trends, transition regions and spatial patterns. Regrettably, none of the Kriging software available nowadays follows this development. Accordingly, SAKWeb[®] introduces an interactive computation process for the chosen vicinity range based on the highest global Moran I where the **W** matrix is defined in proportion to the covariogram contiguity. After the range is defined, the Moran scatterplot is presented in two different forms: the conventional and the location. In addition, SAKWeb[®] presents the Moran variance scatterplot because the conventional one is not sensitive to the neighbor variance. Generally, this analysis reveals outliers and stable and changeable regions.

“Diagnostics are playing an increasingly vital role in spatial analysis” [Griffith, 1993] and Kriging describes the best linear unbiased estimator in the sense of least variance. Kriging is B.L.U.E. (best linear unbiased estimator) and B.U.E. (best unbiased estimator if data respects the ‘bell’ curve). When Kriging is compared with deterministic interpolators, there are major differences, e.g., the former provides uncertainty assessment, anisotropy detection or methodology assumptions. Also, to fulfill the Gaussian assumption, it is common to apply Box-Cox and logarithm transformations. Yet, two major topics should be stressed:

- These transformations focus solely on attribute information and completely ignore location which may create biased results.
- The direct exponential back-transform is biased too.

It became clear in Chapter 5 that the other individual features that were missing in the geosoftware available had become the cornerstone of SAKWeb[®]:

- No software includes the variability factors that underlie the Moran scatterplot;
- No statistical packages that cover spatial autocorrelation issues regard simple observations with no defined boundaries;
- No software considers variogram rescaling⁴³⁶ and the regional confidence interval;
- Measurement error is only undertaken by the ESRI[®] Geostatistical module;
- No software covers the declustering operation based on the nearest neighborhood analysis;
- No software compares final estimations on the fly according to several nugget-effect solutions;
- All software requires an installation procedure, on each computer, that depends on the operating system;
- No software uses the browser interface;
- No software considers the Web as a valuable spatial analysis platform;
- No software presents an e-Learning framework for education purposes. In addition, SAKWeb[®] Help does not follow the traditional Contents-Search-Index pattern of major help software. Instead, it lays out the overall help

structure with hyperlinks among subjects, giving a complete view of the matters in the structure and how they are related to each other in an alternative and attractive framework. Future help applications must be encouraged by this concept.

In terms of implementation technologies, nineteen different software packages were used by SAKWeb[®]: ASP[®], IIS[®] with Server Extensions, PHP[®], FrontPage[®], VBScript[®], ActiveX[®], Dreamweaver[®], Ultradev[®], Flash[®], Director[®], Fireworks[®], WebChart[®], ActiveBar[®], Java Applets[®], JavaScript[®], HTML, DHTML, Fortran and C language. This list does not include the Excel[®] format input, a major deficiency of SAKWeb[®] since no other forms of interoperability are accepted. The present and the future of SAKWeb[®] capabilities is based upon the criticism of the third feedback survey. Nevertheless, SAKWeb[®] was very well accepted by users and substantiated by the grasshopper results of the Colorado case.

9.2. DISCUSSION OF THE HYPHOTESES

The current e-Learning trend illustrates a strong development in the distribution of brain-ware know-how, supported by the time and cost savings of the Web infrastructure. Also, statistical comparisons between traditional and distance learning reveal no major differences in terms of student success and failure. Although some companies still show resistance to recruiting employees that had this type of education, the fact is that the opinions of business executives are shifting and accepting this new culture⁴³⁷. In conjunction with the need to share educational contents relating to spatial analysis, SAKWeb[®] follows this trend by creating virtual-school features like chatting in a virtual room, clarifying questions via e-mail, retrieving notes and other historical messages or accessing hypermedia resources.

The present geography curriculum at universities needs to upgrade these quantitative matters. Moreover, as with Macromedia[®] or Cisco[®], international accreditation to put conventional users and experts on the same level is significant.

As a result of the present situation, conventional users and experts have different software preferences and different goals. While conventional users prefer to use GIS modules because of the trademark that underlies the software or the direct need of topology access, experts tend to use statistical products and independent geosoftware where interfaces are a non-concern issue⁴³⁸. Also, conventional users continue to manage large volumes of spatial data with very deterministic spatial operations⁴³⁹ while experts are more concerned with stochastic issues such as MAUP and fuzzier boundaries⁴⁴⁰. If a high quality solution depends on geostatistical and geographical knowledge then the WWW with hypermedia help⁴⁴¹ is the best way to fill this knowledge gap between both communities. Accordingly, the first and second hypotheses are met (see Section 1.4 for a further review).

Spatial autocorrelation is characterized by innumerable meanings: a descriptor of map pattern types, an index of the information content latent in geo-referenced data⁴⁴², a diagnostic tool for spatial model misspecification, a spatial process mechanism attributable to the geographical ordering of data, a spatial spillover effect and a search for unobserved variables. Hence, another objective was to review global and local spatial autocorrelation measures, spatial autoregression and the importance of residuals within these models. An important constraint on this numerical aspect is the use of LADS and the join-count statistics.

A further important focus is the association of spatial autocorrelation, spatial autoregression and Kriging, also known as the missing data issue, a requirement not emphasized by geostatistic literature. A particular aspect is the existence of the **W** matrix, needed for the assessment of spatial autocorrelation and commonly setup by the binary contiguity or IDW. However, it was demonstrated in Section 2.7 that if the former approach fulfills **W** specification quite well, the latter method is blinded to its natural neighbors because the exponent of $1/d^{\text{exp}}$ already indicates the number of neighbors to take into consideration. Therefore, the third hypothesis is confirmed.

Another relevant issue lies in the implementation philosophy of theoretical research papers produced by SAKWeb[®], e.g., for the variogram rescaling operation. Quite often, the research papers end up on a library shelf without any application for the common GIS user. It is essential for theoretical research to be reflected in practical outcomes. For instance, OK variance is an operation that does not depend on data although most users demand the use of this feature within geosoftware. A new classical local spatial variance that is both data and geometrically dependent was introduced and, according to the results shown in subsection 3.6.1, some promising results were achieved (fourth hypothesis). Curiously, the variogram range is based on the distance at which the global Moran I is higher, giving the idea that the scale neighborhood dependency is a factor to take into account. Although further investigation and confirmation must pursue this topic in order to reach an improved uncertainty indicator, this index must also be converted into a software tool to reach the conventional GIS user, since any theory without a functional use will always have a limited success.

Kriging seems to achieve better interpolation results than other methods although the sampling distribution can turn out to be a problem, especially when the support is not the same or uniform sampling is applied. This latter issue leads to another topic identified as the nugget-effect, the distance interval in which capturing spatial dependency is most important. Independently of conventional geostatistics software, SAKWeb[®] developments have been made using four approaches to this factor: with and without nugget-effect, measurement error and micro-macro scale structures. Although no spatial patterns can be found regarding which nugget-effect fits best, contrary to the fifth hypothesis, the possibility of simulating different on-the-fly models and comparing discrepancy among them is already significant. Parallel to Regard[®], which enhanced the brushing and linkage feature at the beginning of the nineties, it is expected that future software progress will also develop this aspect, with the inclusion of range, model and anisotropy factors.

Uncertainty in the OK variance was taken into consideration via a variogram rescale procedure based on the estimated global variance against the original sill. Although the OK interpolation procedure is not affected, the OK variance surface

and global and local confidence interval⁴⁴³ can be improved when compared with traditional software, especially because the sample weights are given by the nearest neighborhood analysis (sixth hypothesis).

The revised Moran scatterplot for conventional sampling relates to the last two hypotheses. Subsection 7.2.1, as an ESDA tool, introduces the Moran location scatterplot based on the following combination:

- The covariogram setups the **W** matrix to reflect Tobler's Law;
- The Moran I correlogram indicates the best range for the covariogram;
- The conventional Moran scatterplot is replaced by the location one, where true-false outliers and transition regions are highlighted.
- This interactive procedure also leads to several key suggestions for the variogram definition, as the grasshopper example in Colorado demonstrates. Seven relationships between the Moran I correlogram and variogram parameters are shown in Table 32 as guiding principles to avoid the blind automatic fitness procedure of major geosoftware. The variogram fitness is a skill science like the **W** matrix specification.

The Moran variance scatterplot deals with the homogeneity/heterogeneity of observations and their neighborhood and, therefore, it underlies the second moment of the Moran scatterplot. Specifically, it becomes beneficial to detect heteroskedasticity among residuals in a detrended procedure. As the end-users of the third survey confirmed, spatial autocorrelation measures with Kriging must be encouraged in new software.

9.3. LIMITATIONS AND FUTURE RESEARCH

The first technical topic for a future SAKWeb[®] version 3.0 is to migrate this ASP[®] site to an ASP.Net[®] architecture, “although the two types of application can coexist together” [Pandey, 2002]. Therefore, it is possible to run ASP.Net[®] and ASP[®] 3.0 Web page applications on the same host. Migration of a Web application

is a three step procedure as explained in Section 7.4. Also, ASP.Net[®] already incorporates the WAP technology for wireless connections.

An international domain name for the future server machine, like www.sakweb.pt or sakweb.isegi.unl.pt, is another requirement in order to fulfill the interaction of this service around the globe, including SCSI disks and dual processors. Yet, the limitations of today provide the improvements of tomorrow:

- A non-uniform interpolation grid;
- Different data posting;
- Quadrant and octant search for IDW;
- Datasets with UTM coordinates;
- Interoperability with other format software like GS+[®], Variowin[®], MS-Access[®] and Surfer[®];
- Procedures to avoid missing and negative values;
- Data transformations with unbiased back-transformations;
- More types of variogram models;
- Data grid inclusion for data management;
- A cross cross-validation procedure;
- A brushing operation for the Moran variance scatterplot.

The future of SAKWeb[®] can be bright. Already beyond of the scope of this dissertation, the present SAKWeb[®] infrastructure can be easily be applied as a WWW interface with GSLib[®] routines to avoid the reinvention of the wheel⁴⁴⁴, a situation required by some respondents in the third survey. The Next Kriging Menu selection has this aim by leading the user to a new choice of options that include SK, UK, IK and CK. In effect, SK has already been implemented with the following options: SK Calculation, Default SK, SK With No C0, SK With Two Structures, Exact SK Differences, SK Versus OK, Validation With Extra Dataset and Cross-Validation (as with OK, this selection includes the location of positive-negative/true-false samples). A third menu screen has already been sketched for simulation and spatial autoregression procedures. As Section 6.3 suggested, the respectability of the academic research concerning implementation should be a

reality, where the results are reflected in appealing pictures while the arduous work is taking place beyond-stage in a W3 environment.

Literature Cited

Albaredes, G., GIS Usability In Question, Proceedings 15th European Urban Data Management Symposium, Lyon, 1994.

Amrhein, C., Reynolds, H., Using The Getis Statistic To Explore Aggregation Effects In Metropolitan Toronto Census Data, The Canadian Geographer, 41, 1998.

Animals, Porto Editora, 2002.

Anselin, L., Exploratory Spatial Data Analysis In Geocomputational Environment in Geocomputation A Primer, John Wiley & Sons, 1998.

Anselin, L, The Moran scatterplot As An ESDA Tool To Assess Local Instability In Spatial Association in Spatial Analytical Perspectives On GIS, Taylor & Francis, 1996.

Anselin, L., Local Indicators of Spatial Association – LISA, GISDATA Specialist Meeting On GIS And Spatial Analysis, Amsterdam, 1995.

Anselin, L., Exploratory Spatial Data Analysis And Geographical Information Systems, New Tools For Spatial Analysis, Eurostat, 1993.

Anselin, L., SpaceStat Tutorial - A Workbook For Using SpaceStat In The Analysis Of Spatial Data, University Of Illinois, Urbana-Champaign, 1992.

Anselin, L., O'Loughlin, J., Geography Of International Conflict And Cooperation: Spatial Dependence And Regional Context In Africa, 1991.

Anselin, L., Spatial Econometrics, Kluwer, 1988.

ArcNews, Newspaper, Fall 2000.

ArcUser, Magazine, Spring 2000.

Arentze, T, Borgers, A., Timmermans, H., A Generic Spatial Decision Support System For Planning Retail Facilities in Geographical Information Research, Taylor & Francis, 1998.

Arlinghaus, S., Nystuen, J., The Importance Of Position In Space in Practical Handbook Of Spatial Statistics, 1996.

Armstrong, M., Basic Linear Geostatistics, Springer, 1998.

Aronof, Geographical Information Systems: A Management Perspective, WDL Publications, Canada, 1995.

Atkinson, P., A Method For Describing Quantitatively The Information, Redundancy And Error In Digital Spatial Data in Innovations In GIS 3, Taylor & Francis, 1997.

Aunón, J., Hernandez, J., Dual Kriging With Local Neighborhoods: Application To The Representation Of Surfaces, Mathematical Geology, 32, 2000.

Baço, F., Os Sistemas De Informação Geográficos E As Empresas, Dissertação ISEGI-UNL, 1997.

Baddeley, A., Orive, L., The Rao-Blackwell Theorem In Stereology And Some Counter-Examples, Advanced Applications Probabilities, 27, 1995.

Bailey, T., A Review Of Statistical Spatial Analysis In GIS in Spatial Analysis And GIS, Taylor & Francis, 1996.

Baklanov, A., Makarova, T., Nazarenko, L., GIS Applications In The Artic Environment in Geographical Information Research, Taylor & Francis, 1998.

Baptista, C., Fundamental Dos Sistemas Digitais, FCA-Lidel, 2002.

BASF Canada, <http://www.agsolutions.ca>, 2001.

Batty, M., Xie, Y., Urban Analysis In A GIS Environment: Population Density Modeling Using ArcInfo in Spatial Analysis And GIS, Taylor & Francis, 1996.

Batty, M., Dodge, M., Doyle, S., Smith, A., Modeling Virtual Environments in Geocomputation A Primer, John Willey & Sons, 1998.

Beckers, J., Modeling The Oro Moraine Multi-Aquifer System: Role Of Geology, Numerical Model, Parameter Estimation And Uncertainty, University Of Waterloo, Canada, Ph.D. Thesis, 1999.

Berry, B., Baker, A., Geographical Sampling, Prentice Hall, NJ, 1968.

Berry, J., Kemp, W., Onsager, J., Hopper[®], Version 4.0 Users Guide, <http://www.usda.gov>, 2002.

Bivand, R., Scripting And Toolbox Approaches To Spatial Analysis In A GIS Context in Spatial Analytical Perspectives On GIS, Taylor & Francis, 1996.

Bonham-Carter, G., Agterberg, F., Wright, D., Integration Of Geological Datasets For Gold Exploration In Nova Scotia in Introductory Readings in GIS, Taylor & Francis, 1992.

Bonham-Carter, G., Geographical Information Systems For Geoscientists: Modeling With GIS, Pergamon, 1996.

Boykova, A., Moho Discontinuity In Central Balkan Peninsula In The Light Of The Geostatistical Structural Analysis, Physics Of The Earth And Planetary Interiors, 114, 1999.

Bryan, C., http://filebox.vt.edu/users/anieto/web/thesis_csm/chapt6.pdf, 1994.

Brown, D., Internal Migration In Rwanda in Practical Handbook Of Spatial Statistics, CRC Press, 1996.

Bullen, N., Linking GIS And Multilevel Modeling In The Spatial Analysis Of House Prices in Geographical Information Research, Taylor & Francis, 1998.

Burrough, P., McDonnell, R., Principles Of Geographical Information Systems, Oxford University Press, 1998.

Burrough, P., Principles Of Geographical Information Systems For Land Resources Assessment, Clarendon Press-Oxford, 1991.

Butenfield, B., Mackaness, W., Visualization, GIS Volume I: Principles, London, 427-443, 1993.

Byte Magazine, October 1999.

Calkins, H., Creating Large Digital Files From Mapped Data in Introductory Readings In GIS, Taylor & Francis, 1992.

Câmara, A., Planos de Negócio, GASA, 2000.

Câmara, A., Ferreira, F., Castro, P., Spatial Simulation Modeling in Spatial Analytical Perspectives On GIS, Taylor & Francis, 1996.

Can, A, Griffith, D., Spatial Statistical/Econometric Versions Of Simple Urban Population Density Models in Practical Handbook Of Spatial Statistics, CRC Press, 1996.

Capper, B., Watson, S., Towards A Corporate GIS Strategy Or Corporate GIS In Local Government - Fact Or Fiction?, London, 1993.

Cassettari, S., Introduction To Integrated GeoInformation Management, Chapman & Hall, 1994.

Ceroli, A, Riani, M., The Ordering Of Spatial Data And The Detection Of Multiple Outliers, Journal Of Computational And Graphic Statistics, 8, 1999.

Chainey, S., Stuart, N., Stochastic Simulation: An Alternative Interpolation Technique For Digital Geographical Information in Innovations In GIS 3, Taylor & Francis, 1997.

Chrisman, N., The Error Component In Spatial Data, GIS Volume I: Principles, Longman Scientific & Technical, 165-173, 1992.

CIIST, Introdução Ao Sistema Operativo VAX/VMS[®], 1992.

Clark, I., Harper, W., Practical Geostatistics, Ecosse North America, 2000.

Clark, I., Harper, W., Practical Geostatistics 2000 Answers To The Exercices, Ecosse North America, 2001.

Clark, W., Hosking, P., Statistical Methods For Geographers, John Wiley & Sons, 1986.

Clarke, K., Visualizing Different Geofutures in Geocomputation A Primer, John Wiley & Sons, 1998.

Coelho, P., JavaScript, FCA-Lidel, 2002.

Coelho, P., Programação Em Java 2, FCA-Lidel, 2000.

Coelho, P., Pinheiro, J., Spatial Variability In The General Linear Mixed Model in Revista De Estatística, 1998.

Collins, S., Modeling Urban Air Pollution Using GIS in Geographical Information Research, Taylor & Francis, 1998.

Couclelis, H., Geocomputation In Context in Geocomputation A Primer, John Wiley & Sons, 1998.

Cowen, D., GIS Versus CAD Versus DBMS: What Are The Differences? in Introductory Readings In GIS, Taylor & Francis, 1992.

Cranshaw, W. S., Capinera, J. L., Colorado State University Cooperative Extension, <http://www.ext.colostate.edu>, 2003.

Cressie, N, Statistics For Spatial Data, John Wiley & Sons, New York, 1993.

Cressie, N., Wikle, C., A Dimension Reduced Approach To Space-Time Kalman Filtering, Biometrika, 86, 1999.

Chou, Y., Critical Issues In The Evaluation Of Spatial Autocorrelation, Lecture Notes In Computer Science 716, 1993.

Christensen, W., Analysis Of Multivariate Spatial Data Using Latent Variables, Iowa State University, Ph.D. Thesis, 1999.

Côrtés, P., ASP Caixa De Ferramentas para Web, Érica, 2000.

Cross, D., Fast Fourier Transforms, 2000.

Curran, P., Milton, E., Atkinson, P., Foody, G., Remote Sensing: From Data To Understanding in Geocomputation A Primer, John Wiley & Sons, 1998.

Dalthorp, D., Nyrop, J., Villani, M., Estimation Of Local Mean Population Densities Of Japanese Beetle Grubs (Scarabaeidae: Coleoptera), Entomological Society Of America, 1999.

Dana, P., Coordinate System Overview, University of Texas At Austin, 2000.

Dangermond, J., A Classification Of Software Components Commonly Used In GIS in Introductory Readings In GIS, Taylor & Francis, 1992.

Dangermond, J., A Review Of Digital Data Commonly Available And Some Of The Practical Problems Of Entering Them Into A GIS, ESRI, 1991.

Davis, J., Whigham, P., Grant, I., Representing And Applying Knowledge About Spatial Process In Environmental Management in Introductory Readings In GIS, Taylor & Francis, 1992.

Davis, R., Skold, M., Berry, J., Kemp, W., The Economic Threshold For Grasshopper Control On Public Rangelands, Journal & Resource Economics 17, 1992.

Diário Notícias (DN), Gota De Água Inteligente, 22/Nov/2001.

Ding, Y., Fotheringham, A., The Integration Of Spatial Analysis And GIS, State University of New York, 1992.

Drill Guide, Drill Guide Description, 2000.

Dubois, G., A FAQ On GIS & Geostatistics: 2nd Issue, http://www.ai-geostats.org/Geostats_Faq, [2002].

Dutilleul, P., Modifying The T-test For Assessing The Correlation Between Two Spatial Processes, *Biometrics* 49, 1994.

Ebdon, D., *Statistics In Geography*, Blackwell, 1998.

Editora, Porto, *Dicionário de Língua Portuguesa*, 2001.

Egenhofer, M. J., Why Not SQL?, *International Journal Of GIS*, Vol 6, No 2, 71-85, 1994.

Ella, V., *Spatial Analysis And Simulation Of Selected Hydrogeologic And Groundwater Quality Parameters In A Glacial Till Acquitter*, Ph.D. Thesis, Iowa State University, 2000.

Englund, E., Sparks, A., *GEO-EAS 1.2.1 Users Guide*, 1991.

Ertl, G., Gleixner, G., Ranzinger, M., *Move-X: A System For Combining Video Films, Computer Animations And GIS Data*, *Proceedings 15th European Urban Data Management Symposium*, Vol 1, 1993.

Espinosa, R., Olmo, M., *Application Of Geostatistics To Identify Gold Rich Areas In The Finisterre-Fervenza Region, NW Spain*, *Applied Geochemistry*, 14, 1999.

ESRI, *Using AVENUE*, USA, 1997.

ESRI, *Using ArcGIS Geostatistical Analyst*, USA, 2001.

ESRI, <http://www.esri.com/education>, 2002.

Evans, P., Wurster, T., *Getting Real About Virtual Commerce*, *Harvard Business Review*, Nov-Dec 1999.

FDTI (Fundação Divulgação Tecnologias Informação), Macromedia Director[®] 7, 2000.

FDTI (Fundação Divulgação Tecnologias Informação), Macromedia Flash[®] 5, 2001.

Feng, H., Spatial Statistical Modeling Of Explicit And Implicit Determinants Of Regional Fertility Rates: A Case Study Of He-Nan Province, China in Practical Handbook Of Spatial Statistics, CRC Press, 1996.

Ferreira, C., Simões, N., A Evolução Do Pensamento Geográfico, Gradiva, 1993.

Ferreira, C., Simões, N., Tratamento Estatístico E Gráfico Em Geografia, Gradiva, 1994.

Ferreira, P., Macromedia Director 7 & Lingo, FCA-Lidel, 2000.

Ferreira, P., Azevedo, E., Flash 5 Conceitos E Prática, FCA-Lidel, 2001.

Figueira, J., Geografia 12º, Editorial O Livro, 2000.

Fischer, P., Concepts And Paradigms Of Spatial Data in Geographical Information Research, Taylor & Francis, 1998.

Fischer, P., Nijkamp, P., GIS And Spatial Analysis, The Annals Of Regional Science, 1992.

Flash[®] 5 Lite, Makron Books, 2001.

Florax, R, Rey, S., The Impacts Of Misspecified Spatial Interaction In Linear Regression Models in New Directions In Spatial Econometrics, Springer-Verlag, 1996.

Flowerdew, R., Green, M., Areal Interpolation And Types Of Data in Spatial Analysis And GIS, Taylor & Francis, 1996.

Fonseca, J., Perfis Urbanos – Duas Escalas De Observação No Concelho de Lisboa, Dissertação ISEGI-UNL, 1998.

Fonseca, J., Torres, D., Exercícios De Estatística, .Edições Sílabo, 2000.

Foote, K., Huebner, D., Error, Accuracy And Precision, University of Texas At Austin, 2000.

Foote, K., Lynch, M., Data Sources For GIS, University of Texas At Austin, 2000.

Fotheringham, A., Charlton, M., Brunsdon, C., The Geography Of Parameter Space: An Investigation Of Spatial Non-Stationarity in Innovations In GIS 3, Taylor & Francis, 1997.

Fotheringham, A., Ding, Y., The Integration Of Spatial Analysis And GIS in Computer, Environment and Urban Systems, Vol 16, 1992.

Frank, J., Reet, E., Jackson, W., Combining Data Helps Pinpoint Infill Drilling Targets In Texas Field, Oil & Gas Journal, 1999.

Froidevaux, R., Constrained Kriging As An Estimator Of Local Distributions Functions in Proceedings Of The International Workshop On Statistics of Spatial Processes, Italy, 1993.

Gabrosek, J., The Effect Of Local Uncertainty In Geostatistics, Ph.D. Thesis, Iowa State University, 1999.

Gaudart, M, Karson, M., Linder, E., Sinha, D., Bayesian Spatial Prediction, Environmental And Ecological Statistics, 6, 1999.

GASA, <http://gasa3.dcea.fct.unl.pt/tig>, FCT-UNL, 2001.

Gates, B., Rumo Ao Futuro, 1995.

Gaudard, M., Karson, M., Linder, E., Sinha, D., Bayesian Spatial Prediction, Environmental And Ecological Statistics, 6, 1999.

Geostatistics FAQ, <http://curie.ei.jrc.it/faq/index.html>, 1997.

Getis, A., Ord, J., The Analysis Of Spatial Association By Use Of Distance Statistics in Geographical Analysis, 1992.

Getis, A., Spatial Dependence And Heterogeneity And Proximal Databases in Spatial Analysis And GIS, Taylor & Francis, 1996.

Gold, C., Surface Interpolation, Spatial Adjacency And GIS in Innovations In GIS 3, Taylor & Francis, 1997.

Gonçalves, V., Escolas Profissionais E A Envolverte Local, Dissertation ISEGI-UNL, 1997.

Goodchild, M, Spatial Autocorrelation, Norwich: GeoBooks, 47, 1987.

Goodchild, M., Different Data Sources And Diverse Data Structures: Metadata And Other Solutions in Geocomputation A Primer, John Wiley & Sons, 1998.

Goovaerts, P., Geostatistical Analysis Of Environmental Data, ISEGI Conference, 2002.

Goovaerts, P., Geostatistical Approaches For Incorporating Elevation Into The Spatial Interpolation Of Rainfall, Journal Of Hydrology, Elsevier Science, 2000.

Goovaerts, P., Using Elevation To Aid The Geostatistical Mapping Of Rainfall Erosity, Catena, 34, 1999.

Goovaerts, P., Geostatistics In Soil Science: State-Of-The-Art And Perspectives, Geoderma, 89, 1998.

Goovaerts, P., Geostatistics For Natural Resources Evaluation, Oxford University Press, 1997.

Gotway, C., Ferguson, R., Herget, G., Peterson, T., Comparison Of Kriging And IDW methods For Mapping Soil Parameters, Soil Science Society, 60, 1996.

Green, D., Bossomaier, T., Online GIS And Spatial Metadata, Taylor & Francis, 2002.

Griffith, D., Some Guidelines For Specifying The Geographical Weights Matrix Contained In Spatial Statistical Models in Practical Handbook Of Spatial Statistics, CRC Press, 1996.

Griffith, D., Spatial Regression Analysis On The PC: Spatial Statistics Using SAS, Association Of American Geographers, 1994.

Griffith, D., What Is Spatial Autocorrelation? Reflections On The Past 25 Years Of Spatial Statistics, L'Éspace Géographique, 3, 1993.

Griffith, D., Advanced Spatial Statistics For Analysing And Visualizing Geo-Referenced Data, Int. Journal GIS, Vol.7, 1993.

Griffith, D., Which Spatial Statistics Techniques Should Be Converted To GIS Functions?, Syracuse University, 1992.

Griffith, D., Csillag, F., Exploring Relationships Between Semivariogram And Spatial Autoregressive Models, The Journal Of The RSAI, 72, 1993.

Groenigen, J., Siderius, W., Stein, A., Constrained Optimization Of Soil Sampling For Minimization Of The Kriging Variance, *Geoderma*, 87, 1999.

GSLib, <http://www.gslib.com> [2002].

Gunnink, J., Burrough, P., Interactive Spatial Analysis Of Soil Attribute Patterns Using Exploratory Data Analysis And GIS in Spatial Analytical Perspectives On GIS, Taylor & Francis, 1996.

Gunnink, J., Hazelhoff, L., Linking Tools For Exploratory Analysis Of Spatial Data With GIS, EGIS, 1994.

Haining, R., Designing A Health Needs GIS With Spatial Analysis Capability in Spatial Analytical Perspectives On GIS, Taylor & Francis, 1996.

Haining, R., Ma, J., Wise, S., Design Of A Software For Interactive Statistical Analysis Linked To A GIS, *Computational Statistics* 11, 1997.

Hart, B., Geology Plays Key Role In Seismic Attribute Studies, *Oil & Gas Journal*, 1999.

Haslett, J., Bradley, R., Craig, P., Unwin, A., Dynamic Graphics For Exploring Spatial Data With Application To Locating Global And Local Anomalies, *The American Statistician*, Vol. 45, 3, 1992.

Haslett, J., Wills, G., Unwin, A., SPIDER, An Interactive Statistical Tool For The Analysis Of Spatial Distributed Data, *International Journal GIS*, 3, 1991.

Hear, D., Baker, M., Computer Graphics, Prentice-Hall Inc, Englewood Cliffs, N.J, 1999.

Hock, R., Jensen, H., Application Of Kriging Interpolation For Glacier Mass Balance Computations, *Geografiska Annaler*, 81, 1999.

Hopkins, L., Ensor, K., Rifai, H., Empirical Evaluation Of Ambient Ozone Interpolation Procedures To Support Exposure Models, *Air & Waste Management*, 49, 1999.

Horsssen, P., Schot, P., Barendregt, A., A GIS Based Plant Prediction Model For Wetland Ecosystems, *Landscape Ecology*, 14, 1999.

Hunsburger, B., *Unix Quick Reference*, Que Corporation, 1998.

Hutchings, S., Land Bird Conservation Plan, <http://www.rmbo.org>, 2000.

Instituto Superior Técnico, *Sistemas De Comunicação Integrada I, Parte II*, AEIST, 1998.

Isaaks, E., Srivastava, R., *An Introduction To Applied Geostatistics*, Oxford University Press, New York, 1989.

Isaías, P., Sousa, I., *Concepção E Pesquisa De Informação Na Internet*, Universidade Aberta, 1999.

Journel, A., Huijbregts, C., *Mining Geostatistics*, Academic Press, 1978.

Journel, A., *Modeling Uncertainty And Spatial Dependence: Stochastic Imaging*, *Int. Journal GIS*, Vol. 10, 1997.

Juang, K., Lee, D., Comparison Of Three Nonparametric Kriging Methods For Delineating Heavy-Metal Contaminated Soils, *Journal Of Environmental And Quality*, 29, 2000.

Kalkomey, C., Potential Risks When Using Seismic Attributes As Predictors Of Reservoir Properties, The Leading Edge, Vol 16, pp. 247-251, 1999.

Kansas Geological Survey, <http://www.kgs.ukans.edu>, 2000.

Kelk, 3-D GIS For Geosciences, Computers And Geosciences, 17, 1994.

Kelly, M., Spatial Analysis And GIS, Taylor & Francis, 1996.

Kollias, V., Kalivas, D., Yassoglou, N., Mapping The Soil Resources Of A Recent Alluvial Plain In Greece Using Fuzzy Sets In A GIS Environment, European Journal Of Soil Science, 50, 1999.

Korth, H., Silberschatz, A., Database System Concepts, McGraw-Hill International Editions, 1994.

Kravchenko, A., Bullock, D., A Comparative Study Of Interpolation Methods For Mapping Soil Properties, Agronomy Journal, 91, 1999.

Krivoruchko, K., GIS And Geostatistics: Spatial Analysis Of Chernobyl's Consequences In Belarus, ESRI, 2000.

Kriging Description, <http://www.fortner.com/docs>, 2000.

Lang, C., Kriging Interpolation, <http://www.cs.cornell.edu>, 1997.

Langran, G., Time In GIS, Taylor And Francis, 1992.

Laudon, K, Laudon, J, Management Information Systems, Prentice-hall International, 2002.

Leenaers, H, Burrough, P., Okx, J., Effective Mapping Of Heavy Metal Pollution On Floodplains By CK From Elevation Data in Innovations In GIS 3, Taylor & Francis, 1997.

Legare, S., Grasshoppers, <http://www.lams.losalamos.k12.nm.us>, 2000.

Legendre, P., Spatial Autocorrelation: Trouble Or New Paradigm, Ecology, 74, 1994.

Lele, S., Curriero, F., A Composite Likelihood Approach Variogram Estimation, Journal Of Agricultural, Biological And Environmental Statistics, 4, 1999.

Levine, N., Spatial Statistics And GIS Software Tools To Quantify Spatial Patterns, APA Journal, 1997.

Li, B., Implementing Spatial Statistics On Parallel Computers in Practical Handbook Of Spatial Statistics, CRC Press, 1996.

Liebholt, A., Geostatistics And Geographical Information Systems In Applied Insect Ecology, Annual Rev., 1995.

Limaïem, A., Elmaraghy, H., Curve And Surface Modeling With Uncertainties Using Dual Kriging, Journal Of Mechanical Design, 121, 1999.

Long, D., Spatial Statistics For Analysis Of Variance Of Agronomic Field Trials in Practical Handbook Of Spatial Statistics, CRC Press, 1996.

Longley, P., Foundations in Geocomputation A Primer, John Wiley & Sons, 1998.

Longley, P., Goodchild, M., Maguire, D., Rhind, D., Geographical Information Systems And Science, John Wiley & Sons, 2001.

Loureiro, P., TCP/IP Em Redes Microsoft, FCA-Lidel, 2000.

MacDonald, M., A Spatial Decision Support System For Collaborative Solid Waste Planning in Geographical Information Research, Taylor & Francis, 1998.

MacDougal, E., The Accuracy Of Map Overlays, Landscaping Planning 2, 1978.

Macmillan, W., Pierce, T., Optimization In A GIS Framework: The Problem Of Political Redistricting in Spatial Analysis In GIS, Taylor & Francis, 1996.

Maguire, D. J., An Overview And Definition Of GIS, GIS Volume I: Principles, Longman, 1992.

Marble, D., Geographical Information Systems: An Overview in Introductory Readings in GIS, Taylor & Francis, 1993.

Marques, F., Mendes, A., FrontPage 2000, FCA-Lidel, 2001.

Marques, F., Mendes, A., Henriques, M., Sena, R., Programação De Base De Dados Com Visual Basic 6, FCA-Lidel, 2001.

Marques, J., Atlas Do Ambiente – Aplicações De Suporte À Informação, Dissertação ISEGI-UNL, 1997.

Marx, R., The TIGER System: Automating The Geographical Structure Of The US Census in Introductory Readings In GIS, Taylor & Francis, 1993.

Maryland, State of, MAGI: Maryland Automated Geographical Information System in Introductory Readings In GIS, Taylor & Francis, 1993.

Matos, J., Fundamentos De Informação Geográfica, FCA-Lidel, 2001.

Mayhew, S., Dictionary Of Geography, Oxford, 1999.

McBride, P., Iniciação Ao Java, Editorial Presença, 1998.

McMaster, R., Modeling Community Vulnerability To Hazardous Materials Using GIS in Introductory Readings In GIS, Taylor & Francis, 1993.

Miller, E., Towards A 4D GIS: Four-Dimensional Interpolation Utilizing Kriging, Innovations in GIS 3, Taylor & Francis, 1997.

Monteiro, A., Camargo, E., Felgueiras, C., Fuks, S., Integração De Geoestatística E SIGs: Uma Necessidade, Instituto Nacional De Pesquisas Espaciais-Brazil, 2001.

Monteiro, E., Boavida, F., Engenharia De Redes Informáticas, FCA-Lidel, 2001.

Morgan, C., Tudo Sobre Xenix, Pioneira Informática, 1996.

Murteira, B., Análise Exploratória Dos Dados – Estatística Descritiva, McGraw Hill De Portugal, 1993.

Nalder, I., Wein, R., Spatial Interpolation Of Climatic Normals: Test Of A New Method In The Canadian Boreal Forest, Agriculture And Forest Meteorology, 92, 1998.

Nass, C., Garfinkle, D., Localized Autocorrelation Diagnostic Statistic (LADS) for Spatial Models: Conceptualization, Utilization And Computation, Regional Science And Urban Economics 22: 333/346, 1992.

NCGIA (National Center For Geographic Information And Analysis), Annual Report, 1995.

National Portuguese Personal Data Committee, Protecção Dos Dados Na União Europeia, 2001.

Negreiros, J., Apontamentos De Programação Por Objectos E Java Applets, COCITE, 2001.

Negreiros, J., ASPs, IIS, Access, SQL Server: Uma Solução Web, CIIST, 2001.

Negreiros, J., Bases de Dados Relacionais: Conceitos Teóricos, CIIST, 1997.

Negreiros, J., Conceitos Teóricos De Base De Dados, CIIST, 1998.

Negreiros, J., Internet, HTML E FrontPage, CIIST, 1999.

Negreiros, J., Introdução Aos Expert Systems, COCITE, 1998.

Negreiros, J., Manual De Operação De VMS[®], CIIST, 1995.

Negreiros, J., Manual de Word, Excel E Access, Fernave, 2000.

Negreiros, J., Sistemas De Informação De Gestão I, COCITE, 1996.

Negreiros, J., Amador, J., Castro, J., Unix-Curso Completo, FCA-Lidel, 1998.

Negreiros, J., Ferreira, A., Programação Na World Wide Web Com CGIs, FCA-Lidel, 1999.

Negreiros, J., Lopo, C., Beidatsch, H., Conceitos Teóricos de SAP R/3, ISEGI-UNL, 1999.

Negreiros, J., Orfão, R., Manual de Hardware, CIIST, 1997.

Olivier, M., Geostatistics, Rare Disease And The Environment in Spatial Analytical On GIS, Taylor & Francis, 1996.

Olson, D., A Study Of Voter Turnout In New York City Elections Using Spatial Statistics in Geographical Information Research, Taylor & Francis, 1998.

Openshaw, S., A Framework For Research On Spatial Analysis Relevant To GeoStatistical Information Systems in New Tools For Spatial Analysis, Eurostat, 1993.

Openshaw, S., What Is GISable Spatial Analysis, New Tools For Spatial Analysis, Eurostat, 1994.

Openshaw, S., Two Exploratory Space-Time-Attribute Pattern Analyzers Relevant To GIS in Spatial Analysis And GIS, Taylor & Francis, 1996.

Openshaw, S., Building Automated Geographical Analysis And Explanation Machines in Geocomputation A Primer, John Wiley & Sons, 1998.

Openshaw, S., Brunsdon, C., Charlton, M., A Spatial Analysis Toolkit For Analysis in Conference On Geografical Systems, 1992.

Openshaw, S., Clarke, G., Developing Spatial Analysis Functions Relevant To GIS Environments in Spatial Analytical Perspectives On GIS, Taylor & Francis, 1996.

Painho, M., Geographical Information Systems, ADISEGI, 1999.

Painho, M., Modeling Errors In Digital Landuse/Landcover Maps, Ph.D. Thesis, UCSB, California, 1992.

Painho, M., Cabral, P., Peixoto, M., Pires, P., , E-teaching and GIS: ISEGI-UNL learning experience, Third European GIS Education Seminar EUGISES,, 2002.

Pandey, N., ASP.Net Fast & Easy Web Development, Premier Press, 2002.

Passarella, G., Greene, E., Vurro, M., Spatial And Temporal Study Of Nitrate Concentration In Groundwater By Means Of Co-regionalization, Environmental Geology, 36, 1999.

Peairs, F., Grasshoppers May Be Troublesome This Year, <http://www.ext.colostate.edu>, 2002.

Peuquet, D., Marble, D., ARCINFO: An Example Of A Contemporary GIS in Introductory Readings In GIS, Taylor & Francis, 1993.

Pinto, C., Curto, D., Estatística Para Economia E Gestão, Edição Silabo, 1999.

Poeter, E., Wingle, W., McKenna, S., UNCERT: Geostatistics, Uncertainty Analysis And Visualization Software Applied To Groundwater Flow And Contaminant Transport Modeling, Computers & Geosciences, 25, 1999.

Pritsker, A., Introduction To Simulation And SLAM II, John Wiley & Sons, 1993.

Ralston, B., Object Oriented Spatial Analysis in Spatial Analysis And GIS, Taylor & Francis, 1996.

Raper, J., Bundock, M., UGIX: A GIS Independent User Interface Environment, Proceedings Autocarto 10, Baltimore, 1993.

Remy, N., GSTL: The Geostatistical Template Library In C++, Master Thesis, Stanford, USA, 2001.

Rogerson, P., Fotheringham, S., GIS And Spatial Analysis: Introduction And Overview in Spatial Analysis And GIS, Taylor & Francis, 1996.

Rosário, D., Interface Para Divulgação De Informação Estatística Usando A Internet, Dissertação ISEGI-UNL, 1997.

Rossiter, D., Geographical Information Systems, Cornell University, 1999.

Royle, J., Berliner, L., A Hierarchical Approach To Multivariate Spatial Modeling And Prediction, Journal Of Agriculture, Biological And Environmental Statistics, 4, 1999.

Rowley, J., A Strategy For GIS Standards, Proceedings Association For GIS Conference, Birmingham, 1994.

Rowlingson, B., Gatrell, A., Spatial Point Process Modeling in Spatial Analysis And GIS, Taylor & Francis, 1996.

Santos, A., Ensino À Distancia & Tecnologias De Informação E-Learning, FCA, 2000.

Santos, M., Por Uma Nova Geografia, 1978.

Sarkozy, F., GIS Function Interpolation, Technical University Budapest, 1999.

SAS Institution, Introduction To Time Series Forecasting Using SAS/ETS Software, Course Notes, 2001.

Schildt, H., Teaching Yourself C, Osborne, .McGraw-Hill, 1998.

Schultz, P., Seismic Guided Estimation Of Reservoir Properties, SPE 28386, SPE 69th Annual Technical Conference and Exhibition, 1997.

Schultz, C., Myers, S., Hipp, J., Young, C., Nonstationary Bayesian Kriging: A Predictive Technique To Generate Spatial Corrections For Seismic Detection, Location And Identification, Physics Of The Earth And Planetary Interiors, 113, 1999.

Semana Informática, N° 595, 2002.

Scientific Software Group, <http://www.scisoftware.com>, 2002.

Schwartz, J., M. McGuinness, Einstein Para Principiantes, Publicações Dom Quixote, 1982.

Silva, A., Agentes De Software Na Internet, Centro Atlantico, 1999.

Silva, R., Automatização Da Generalização Cartográfica E Aquisição De Conhecimentos: Uma Abordagem Teórica E Experimental, Dissertação ISEGI-UNL, 1995.

Skinner, K., Spatial And Multivariate Analysis Of Colorado Rangeland Grasshopper Abundances: Pattern And Process, Ph.D. Thesis, Colorado State University, 1999.

Skinner, K., The Past, Present And Future Of Rangeland Grasshopper Management, Rangelands, 22, 2000.

Skinner, K., Child, R., Multivariate Analysis Of The Factors Influencing Changes In Colorado Grasshopper Abundance, Rangelands, 22, 2000.

Soares, A., Geoestatística Para As Ciências Da Terra E Do Ambiente, IST Press, 2000.

Sousa, J., Apontamentos De Estatística Espacial, Departamento De Minas Do IST, 1998.

SpaceStat, <http://www.spacestat.com>, 2002.

Speigel, J., Flash MX Application Design And Development, New Riders Publishing, 2003.

Spotlight: The Economic Impact Of E-Commerce, Distribution Savings On E-Goods Is Dramatic, May 1999.

Spring, <http://www.dpi.inpe.br>, 2000.

Stehman, S., Overton, W., Spatial Sampling in Practical Handbook Of Spatial Statistics, CRC Press, 1990.

Stein, A., Zaadnoordijk, W., Improved Parameter Estimation For Hydrological Models Using Weighted Object Functions, Hydrological Processes, 13, 1999.

Stephens, R., Visual Basic Graphics Programming, Second Edition, Wiley, 2001.

Street, S., Flash[®] For The Real World, SAMS, 2002.

Super Interessante Magazine, 41, 2001.

Tang, A., Adams, T., Usery, E., A Spatial Model Design For Feature-Based GIS in Innovations In GIS 3, Taylor & Francis, 1997.

Tercan, A., Importance Of Orthogonalization Algorithm In Modeling Conditional Distributions By Orthogonal Transformed Indicators Methods, Mathematical Geology, 31, 1999.

Tereso, C., Redes Locais em Windows, FCA-Lidel, 2000.

Tobler, W., Resolution Resampling And All That In Building DataBases For Global Science, Taylor & Francis, 1988.

Tomlinson, R., Current And Potential Uses Of GIS: The North American Experience in Introductory Readings In GIS, Taylor & Francis, 1992.

Tomlinson, R., GIS – A New Frontier in Introductory Readings In GIS, Taylor & Francis, 1992.

Trezentos, P., Cardoso, A., Fundamental Do Linux, FCA-Lidel, 2000.

Trocado, P., Operações De Análise Espacial, IST, 2001.

United Nations Year Book, 2000.

University of Wyoming, <http://www.sdvc.uwyo.edu>, 2002.

Unwin, D., GIS And Spatial Statistical Analysis in Geographical Information Research, Taylor & Francis, 399-409, 1998.

Upton, G., Information From Regional Data in Spatial Statistics: Past, Present and Future, Ann Arbor, 1992.

Upton, G., Fingleton, B., Spatial Data Analysis By Example, Volume I, John Wiley & Sons, 1995.

US Bureau of the Census, Technical Description Of The DIME System in Introductory Readings In GIS, Taylor & Francis, 1993.

Vale, M., Construção De Um SIG Para Avaliação E Implementação Do Plano De Ordenamento Da Albufeira Do Castelo de Bode, Dissertação Mestrado, ISEGI-UNL, 1994.

Vale, M., Discussão De Dissertação De Doutoramento, *ISEGI-UNL*, 2001.

Vasiliev, I., Visualization Of Spatial Dependence: An Elementary View Of Spatial Autocorrelation in Practical Handbook Of Spatial Statistics, CRC Press, 1996.

Vieira, J., Programação Web Com ASPs, Centro Atlântico, 2000.

Vilar, J., Programação Em Access 2000 Com VBA, FCA-Lidel, 2001.

Vilares, M., Sistemas De Informação Estatística, *ISEGI-UNL*, 1993.

Walford, N., Geographical Data Analysis, John Wiley & Sons, 1995.

Walter, N., Spatial Interpolation, University Of Calgary, Canada, 2001.

Wang, F., Hall, G., Fuzzy Representation Of Geographical Boundaries In GIS, *Int. Journal GIS*, Vol. 10, 1997.

Wang, X., Zhang, Z., A Comparison Of Conditional Simulation, Kriging And Trend Surface Analysis For Soil Heavy Metal Pollution Pattern Analysis, *Journal Of Environmental Sciences And Health*, Mercel Dekker, Inc., 1999.

Webster, R, Olivier, M., Kriging: A Method Of Interpolation For GIS, *International Journal Of GIS*, VOL. 4, No. 3, 1992.

Wiecek, M., Multiobjective Models In Identification Of Random Fields, *European Journal Of Operational Research*, 32, 1988.

Wingle, W., Poeter, E., McKenna, S., UNCERT: Geostatistics, Uncertainty Analysis And Visualization Software Applied To Groundwater Flow And Contaminant Transport Modeling, *Computers & Geostatistics*, 25, 1999.

Wong, D., Aggregation Effects In GeoReferenced Data in *Practical Handbook Of Spatial Statistics*, CRC Press, 1996.

Wong, D., Lee, J., Statistical Analysis With ArcView[®] GIS, John Wily & Sons, Inc., 2001.

Wood, G., Webster, R., Olivier, M., Estimating Soil Salinity By Disjunctive Kriging In Soil Use And Management, 6, 97-104, 1992.

Worboys, M., GIS: A Computing Perspective, Taylor & Francis, 1996.

XGobi, <http://www.public.iastate.edu/~arcview-xgobi>, 2002.

Yao, T., Nonparametric Cross-covariance Modeling As Exemplified By Soil Heavy Metal Concentrations From The Swiss Jura, Geoderma, 88, 1999.

Yates, S., Yates, M., Geostatistics For Waste Management, U.S. Environmental Protection Agency, 1991.

Yates, S., Shouse, P., Zhang, R., Estimates Of Soil Nitrate Distributions Using CK With Pseudo-Cross-variograms, Journal Of Environmental And Quality, 428, 1999.

Yeh, T., Guzmán, J., Sequential Kriging And CK: Two Powerful Geostatistical Approaches, Stochastic Environmental And Risk Assessment, 13, 1999.

Yu, K., Ecologists, Farmers, Tourists - GIS Support Planning Of Red Stone Park, China in Geographical Information Research, Taylor & Francis, 1998.

Zhang, Z., Griffith, D., Developing User-Friendly Spatial Statistical Analysis Modules For GIS: An Example Using ArcView, Computers, Environment And Urban Systems, Vol. 21, 1998.

Zimmerman, D., Pavlik, C., Ruggles, A., Armstrong, M., An Experimental Comparison Of Ordinary And UK And Inverse Distance Weighting, Mathematical Geology, Vol. 31, 1999.

Appendix A

Operating System	Author	Software	Type	Major Applications	Source Coding; Pricing; Home Page	Year
MS-Windows [®]	Yves Gratton	MATLAB [®] Kriging Toolbox	Statistical Package (MATLAB [®] is mandatory)	Variogram, rodogram, correlogram, 2D and 3D modeling, point and block Kriging, point and block CK, universal CK, normality test (Kolmogorov-Smirnov) and image processing functions.	MATLAB [®] code available; Free; http://www.inrs-eau.quebec.ca	1996
MS-Windows [®]	Geostat Systems International Inc. [®]	GeoStat [®]	Independent Geosoftware	Interactive variogram analysis, OK and conditional simulation.	Not available; Not available; http://www.geostat.com	1997
MS-Windows [®]	Correlations Company [®]	GViz [®]	Independent Geosoftware	Interactive variogram analysis, nearest neighbor maps, fractal map, Kriging, stochastic simulation and Jackknife for validation.	Not available, US\$15000, http://www.correlations.com	1997
MS-Windows [®] and Unix [®]	University of Adelaide, Australia	Spatial functions	Statistical Package	Simulations of K-functions, simulation of binomial spatial point process, Sequential and Strauss spatial point process, correlograms and	Both SPlus [®] and R [®] codes are	1997

		for Kriging and point pattern analysis for SPlus [®]	(SPlus [®] is a requirement)	variograms, spatial covariance, pseudo-likelihood of Strauss, trend surface analysis, trend surface over a grid, Kriging with its standard error of prediction.	available; Free; http://www.stats.ox.ac.uk	
MS-Windows [®]	University of Ottawa	Rooks Case [®] (Requires Excel 97 [®])	Independent Geosoftware	Regular Lattice Data: Moran's I and Geary's C with Rook's, Queen's or Bishop's Adjacency; Join-Count Statistics for presence/absence data with Rook's, Queen's or Bishop's Adjacency, reports variance, expected values and summary statistics. Irregular Lattice Data: Moran's I and Geary's C; single lags; correlograms; reports variance, nearest neighbor distribution statistics, min, 1st quartile, mean, median, 3rd Quartile, max nearest neighbor distance, interactive base correlogram, local spatial autocorrelation statistics, $G_i(d)$, significance tests based on Bonferroni-like adjustments for up to 1000 points. Monte-Carlo simulation module reporting the Monte-Carlo p-value and providing the distribution for post-analysis.	Not Available; Free; http://www.uottawa.ca	1998
MS-Windows [®]	GeoSoft [®]	Isomap [®]	Independent Geosoftware	OK, hydro-geological and geo-physical applications.	Not available; Not available; http://www.geoandsoft.com	1999
Unix [®]	Frieder Tonn	SGS [®]	Independent Geosoftware	Trend analysis and modeling (3 rd order), spatial correlation analysis	C [®] available; Free;	1999

				(covariance, correlogram, variogram), sequential Gaussian simulation, SK and OK, Kriging with local anisotropy, output in 3D, isolines and data histograms.	http://www.frieder-tonn.de	
MS-Windows [®]	Louisiana State University and University of Alaska	MATLAB [®] Spatial Statistics Toolbox 1.1	Statistical Package (MATLAB [®] is mandatory)	This software can estimate simultaneous spatial autoregression (SAR), conditional spatial autoregression (CAR) and mixed regressive spatially autoregressive estimates. Spatial W matrix can be calculated for very large datasets (>100000 points). It can be based upon nearest neighbors (symmetric or asymmetric) and Delaunay triangles (symmetric). The Delaunay spatial W leads to a variance-covariance matrix that depends upon the autoregressive parameter. In contrast, the nearest neighbor concentration matrices or variance-covariance matrices depend upon two parameters: the autoregressive parameter and the number of neighbors. Computation of the log-determinants for a grid of autoregressive parameters can also be done.	MATLAB [®] codes available; Free; http://www.spatial-statistics.com	1999
Unix [®] and MS-Windows [®]	Insightful [®]	Splus [®] for Arcview [®] and S+Spatial Stats [®]	GIS Module	3D point clouds, nearest neighbor search, Geary and Moran spatial autocorrelation coefficients, directional variograms and correlograms for exploring anisotropy, spatial regression models including conditional and simultaneous autoregressive models, spatial randomness tests and simulation of spatial random processes, Ripley's K-functions, local intensity estimation, OK and UK, Kriging prediction at arbitrary locations with standard errors, parametric and nonparametric trend surfaces.	Not available; US\$800; http://www.insightful.com	2000
DEC Alpha [®] workstations	Iowa State University	Arcview [®] XGobi [®]	GIS Module	The link between ArcView [®] and XGobi [®] is designed to combine the strengths of both packages in the exploration and analysis of data	Free; http://www.publ	2000

and Unix [®] .	(Arcview for Unix [®])	collected in a spatial framework. ArcView [®] is a GIS that allows spatial data in many forms to be displayed, queried and manipulated. XGobi [®] is a dynamic graphics program that allows multivariate data to be explored through the manipulation of scatterplots. XGobi [®] provides tools for the exploration of multivariate data such as linked brushing, identification, the grand tour and projection pursuit. The link allows data collected at spatial locations (stored in ArcView [®]) to be dynamically passed to XGobi [®] and explored. The link between the data points in XGobi [®] and the locations from which they were collected is maintained through "linked brushing". Linked brushing is the ability to change the glyph/size/color of the points in either ArcView [®] or XGobi [®] and to see the corresponding points in the other application simultaneously including CDF plots, variogram cloud, lagged-scatterplots and multivariate variogram clouds.				ic.iastate.edu/~a rcview-xgobi
OS/2 [®] , Unix [®] , SAS [®] VMS [®] and MS- Windows [®]		SAS/STAT [®]	Statistical Package	Variogram (isotropic or anisotropic) and OK.	Not available; Unknown; http://www.SAS.com	2000
MS-Windows [®] and Unix [®]	Applied Computer Engineering	GrisStat [®]	Independent Geosoftware	Variograms, covariograms, 3D Kriging, CK, conditional simulation and 3D visualization.	Not available; US\$495; http://www.gridstat.com	2001
MS-Windows [®] , Linux [®] and	Dezhang Chu	EasyKrig [®]	Statistical Package	Variogram, rodogram, point and block Kriging, point and block CK, panels and polygonal areas and cross-validation.	MATLAB [®] codes available;	2001

Macs [®]			(MATLAB [®] is mandatory)		Free, http://globec.whoi.edu	
MS-Windows [®]	The Institute for Environmental Modeling at the University of Tennessee, Knoxville	SADA [®]	Independent Geosoftware	2D and 3D visualisation tools, variogram analysis and modeling, nearest neighbors, OK, IK, inverse distance weighted and Shepard's interpolation. As well, SADA [®] can accept map layers from a GIS saved in DXF.	Not available; Free; http://www.tiem.utk.edu/~sada	2002
Unix [®] and MS-Windows [®]	Ecole des Mines de Paris (Centre de Geostatistique)	ISATIS [®]	Independent Geosoftware	Map of proportional symbols, fully multivariate for any number of variables, variogram, covariance, correlogram, madogram, rodogram, variogram clouds and surface analysis, interactive or automatic variogram modeling (nugget-effect, exponential, spherical, Gaussian, cardinal sine, stable, gamma, J and K Bessel), generalized Cauchy, factorized exponential, power, first order generalized covariance, automatic drift identification using the IRF-k theory (Intrinsic Random Functions of order k), PCA, declustering and regularization, polygons of influence, moving average, moving median and moving dip average, inverse of a powered distance, least squares polynomial fit, discrete splines, Kriging and CK, Bundled external drift simulations , ...), non conditional simulations (multigaussian, truncated gaussian, simulations of random sets, ...) and cross-validation.	Not available; Not available; http://www.geovariances.fr	2002
MS-Windows [®]	Ned Levine and Associates	CrimeStat II [®]	Independent Geosoftware	The spatial statistics in CrimeStat [®] are subdivided into four categories: 1) Statistics describing the spatial distribution of incidents (such as the mean center, center of minimum distance, standard deviational ellipse,	Not available; Free; http://www.icpsr	2002

				Moran's I or angular mean). 2) Statistics for describing properties of distances between incidents including nearest neighbor analysis, linear nearest neighbor analysis and Ripley's K statistic. 3) Three routines for conducting hot spot analysis: hierarchical nearest neighbor clustering, K-means clustering and local Moran statistics. 4) Single kernel density routine for producing a surface or contour estimate of the density of incidents and a dual-variable kernel density estimation for comparing the density of incidents to the density of an underlying baseline.	.umich.edu	
MS-Windows [®]	Technical University of Denmark	DACE [®]	Statistical Package (MATLAB [®] is mandatory)	Variogram (7 models implemented), SK, OK, Kriging with external drift, regression model up to 2 nd order. Parameters of the covariance function are estimated by maximum likelihood.	Matlab [®] codes available, Free, http://www.imm.dtu.dk/~hbn	2002
Linux [®] , MS-Windows [®] and Unix [®]	Paulo Justiniano Ribeiro, Jr and Petter Diggle	GeoR [®]	Statistical Package	General statistics, transformed Box-Cox models, variogram analysis with OLS/WLS fitting, least squares fitting and likelihood based methods, Bayesian methods, SK, OK, UK, Kriging with external drift and Gaussian simulations.	R [®] codes available; Free; http://www.est.ufpr.br/geoR	2003

Appendix B



First Questionnaire

Name of University or Institution:

Gender (Always make a circle on the right answer):

Male

Female

Age:

<=22

>22 and <=32

> 32 and <=42

>42 and <=52

>52

Current Professional Status:

Undergraduate Student

Graduate Student

Doctoral Student

Professor

Employee or Researcher

Bachelorlato or Licenciatura Name:

Main Subjects of Studies (You may choose more than one option):

Maths

Statistic

Geographical Information Systems

Geography

Geology

Environmental Issues

Computers

Social Sciences and Politics

Economics, Management, Business and Administration

Other

Basics Knowledge

Relational DataBases	Yes	No
Computer Programming	Yes	No
Geographical Information Systems	Yes	No

Theoretical Knowledge of Statistical Exploratory Data Analysis

Mean, median, variance, standard deviation	Yes	No
Hypothesis Tests	Yes	No
Linear Relationships ($y=ax+b$)	Yes	No

Theoretical Knowledge of Spatial Interpolation Methods

Fourier	Yes	No
B-Splines	Yes	No
Inverse Distance Weighted	Yes	No
Polynomial Trend Surface	Yes	No
Kriging	Yes	No

Knowledge of GIS Software

ESRI (ArcView [®] or ArcInfo [®])	Yes	No
Intergraph [®]	Yes	No
Others, if any:		

Knowledge of GIS Modules Software

Network Analysis	Yes	No
Spatial Analysis	Yes	No
Others, if any:		

Knowledge of Deterministic Spatial Functions

Buffering	Yes	No
Overlay	Yes	No
Path Finding	Yes	No

Knowledge of Statistical Software

SAS [®]	Yes	No
SPSS [®]	Yes	No
MINITAB [®]	Yes	No

GLIM[®] Yes No

MATLAB[®] Yes No

Others, if any: _____

Have you ever used Statistical Software with spatial data including topological information (location, connectivity and adjacency of geographical features)? Yes No

Have you ever used Kriging Spatial Interpolation approach within...

Statistical Software? Yes No

GIS Software? Yes No

Knowledge of Kriging Software

VarioWin[®] Yes No

GeoMS[®] Yes No

Regard[®] Yes No

GS+[®] Yes No

Others, if any: _____

Have you ever used any Kriging or Spatial Autocorrelation Software coupled to GIS Software?

Yes No

Do you know Tobler's First Law or First Law of Geography?

Yes No

Theoretical Knowledge of Measures of Spatial Autocorrelation

Moran I Yes No

Geary C Yes No

Variogram Cloud Yes No

G_i(d) Yes No

Have you ever measured Spatial Autocorrelation with any type of Software? Yes No

If yes, which ones? (You may choose more than one option)

GIS Software

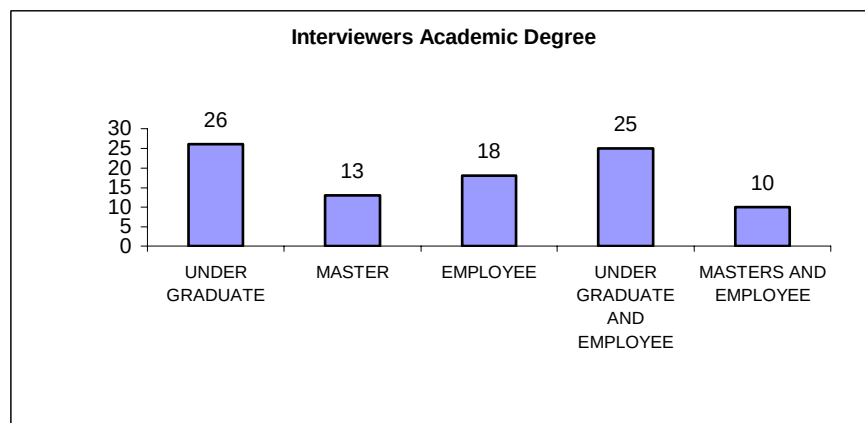
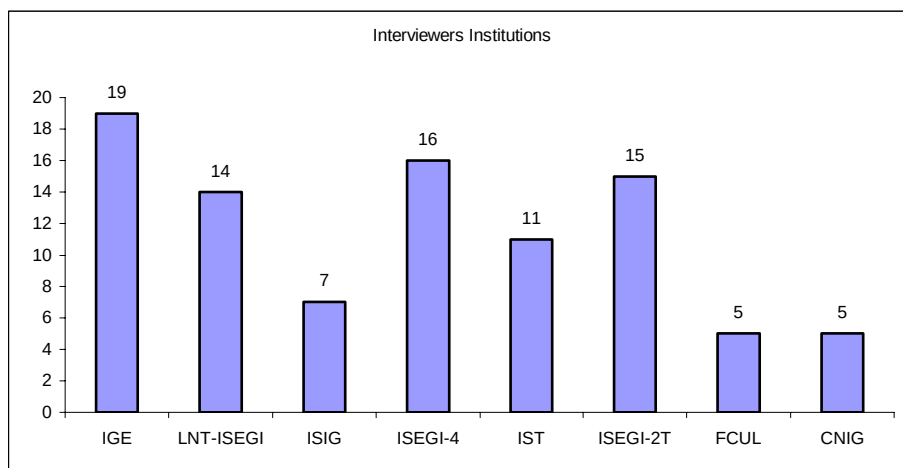
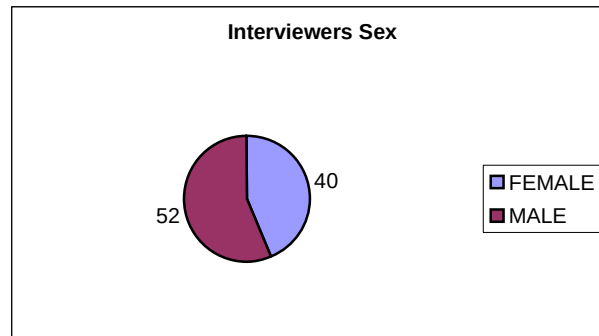
Statistical Software

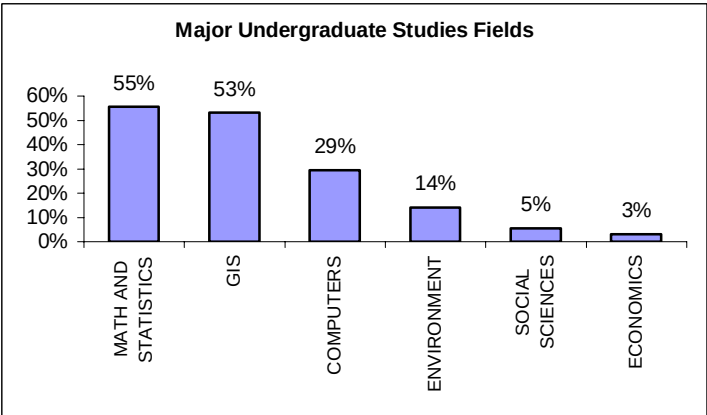
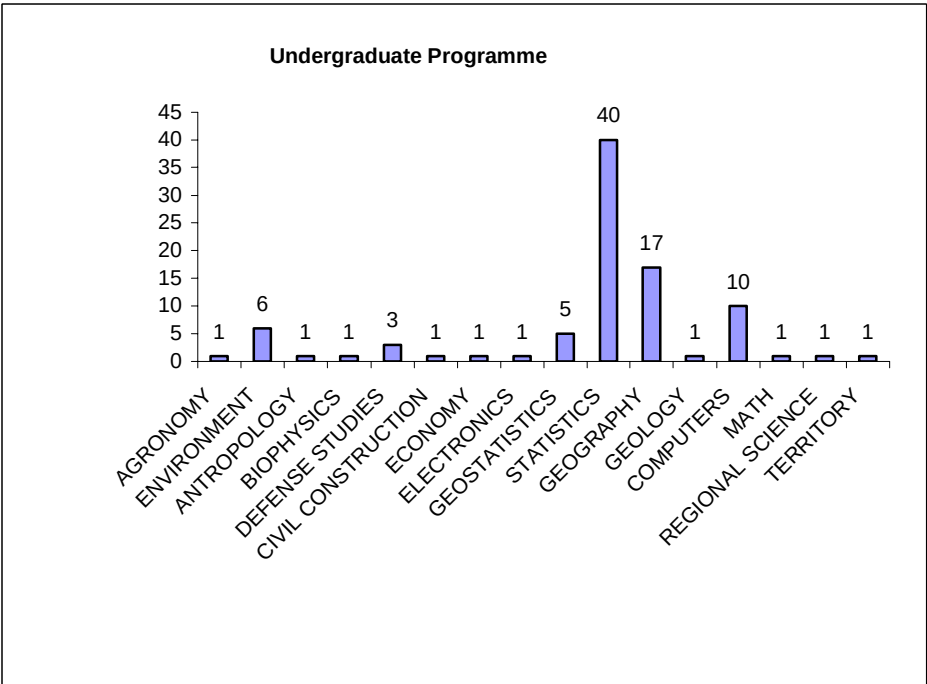
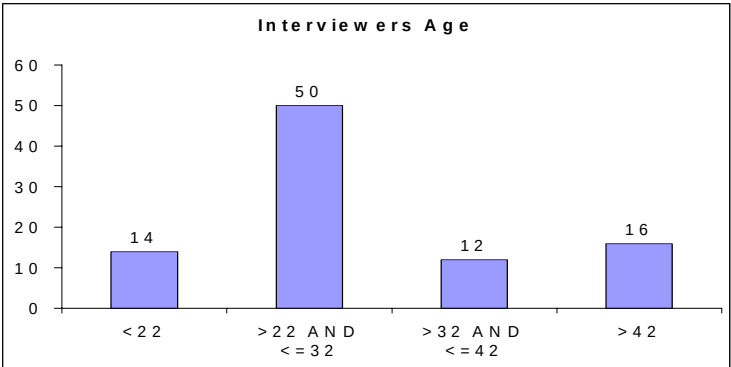
Geostatistical Software

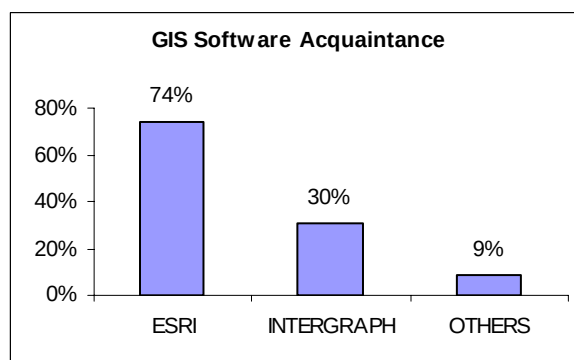
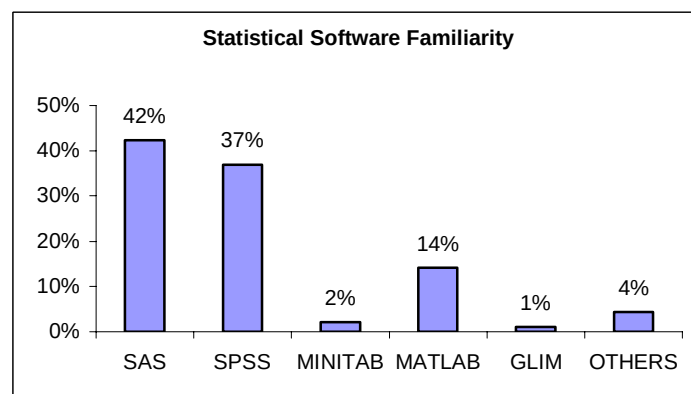
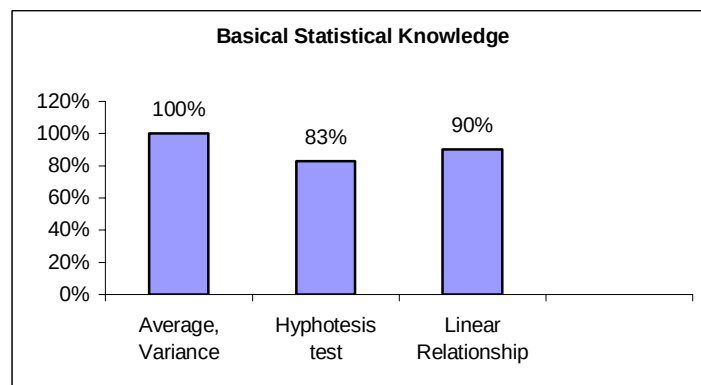
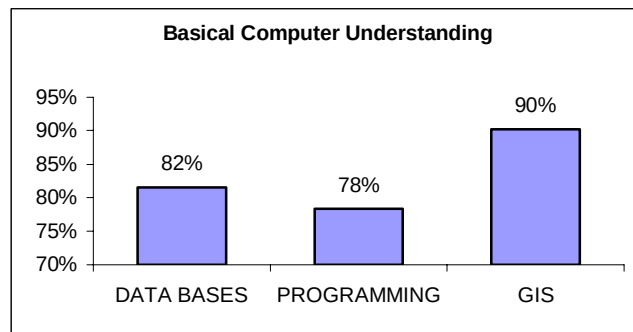
Thanks, Dude!!!!!!

Appendix C

The Statistical Results Of The First Survey







Appendix E

Grasshoppers Datasets of 1993, 1994, 1995 and 1997

SampleID	Coord X	Coord Y	GrassHoppers 1993	SampleID	Coord X	Coord Y	GrassHoppers 1993
1	192	85	0.7	53	521	246	0.7
2	161	93	0.3	54	553	233	5.4
3	86	130	11.9	55	589	228	7.6
4	93	210	23.4	56	616	229	9.5
5	180	157	7.2	57	669	222	5.1
6	195	254	0.5	58	706	236	1.1
7	185	266	0.7	59	765	254	3.2
8	92	306	1.3	60	632	237	15.3
9	108	324	0.5	61	600	250	3.7
10	110	345	0.3	62	621	250	12.6
11	90	381	10.6	63	654	253	0.5
12	132	414	4.8	64	628	261	4.3
13	105	428	5.4	65	673	265	11.2
14	177	280	0	66	645	284	4.3
15	357	346	1.5	67	604	285	4.5
16	380	342	1.7	68	568	269	4.3
17	345	380	0.2	69	556	285	2.1
18	354	409	1.1	70	565	306	0.9
19	445	61	1.7	71	601	310	0.9
20	456	74	4.3	72	636	316	2.5
21	488	42	9	73	658	321	1.5
22	485	57	10	74	693	310	3.2
23	509	56	5.1	75	545	366	1.9
24	542	50	3.2	76	569	358	0.9
25	580	44	17.7	77	588	364	0.5
26	521	62	9	78	600	354	3.2
27	521	78	0.7	79	612	362	2.5
28	594	53	10	80	589	376	0.3
29	568	77	1.3	81	589	389	1.5
30	580	89	4.8	82	638	353	6.5
31	564	97	8.1	83	650	364	0.5
32	528	109	5.8	84	674	353	1.5
33	546	120	10.6	85	682	365	1.3
34	628	57	21	86	704	358	0.7
35	652	52	4.5	87	701	369	0.3
36	642	65	2.1	88	546	401	1.3
37	661	85	0.7	89	530	414	0.9
38	692	69	11.2	90	562	434	1.9
39	706	58	2.3	91	590	437	1.3

40	753	48	2.5	92	626	433	4.8
41	628	118	1.3	93	628	418	0.3
42	629	129	2.3	94	658	426	2.7
43	622	140	21	95	673	405	0.5
44	722	117	1.3	96	680	418	0.5
45	749	124	5.1	97	692	397	0.9
46	706	140	3.5	98	713	405	2.1
47	702	152	2.7	99	678	434	0.5
48	726	168	5.8	100	698	434	0.5
49	729	180	2.5	101	698	446	1.1
50	580	201	4.8	102	717	444	0.7
51	629	202	3.2	103	736	448	4.8
52	613	212	12.6				

SampleID	Coord X	Coord Y	GrassHoppers 1994	SampleID	Coord X	Coord Y	GrassHoppers 1994
1	208	98	1.1	51	804	316	32.6
2	188	174	0.3	52	764	309	21
3	192	301	0	53	742	290	11.9
4	184	320	0	54	708	285	1.3
5	372	384	0.5	55	673	289	1.1
6	401	402	1.1	56	641	306	1.3
7	420	396	1.9	57	622	326	2.1
8	386	444	1.7	58	636	352	2.3
9	501	65	2.7	59	673	320	3.5
10	512	78	3.5	60	703	305	7.2
11	540	62	1.1	61	680	351	0.3
12	554	45	2.3	62	705	362	1.1
13	573	58	3.2	63	704	367	1.1
14	585	69	1.7	64	677	386	10.6
15	586	84	0.3	65	646	375	9.5
16	612	57	0.9	66	775	369	2.3
17	665	53	15.3	67	782	353	1.7
18	640	81	1.7	68	801	370	0.9
19	594	117	5.4	69	807	395	0.3
20	617	121	23.4	70	779	401	2.1
21	740	54	1.1	71	802	413	1.3
22	725	61	1.5	72	793	429	0.7
23	722	73	1.5	73	770	416	2.1
24	757	92	2.5	74	739	421	10.6
25	785	73	32.6	75	720	406	11.2
26	808	62	4.5	76	694	417	1.9
27	854	53	3.7	77	675	410	0.9
28	863	104	0.5	78	661	416	0.9
29	820	129	1.5	79	637	412	0.2
30	771	145	3	80	613	420	1.1
31	706	129	9.5	81	614	466	0.2
32	711	141	1.9	82	600	476	0.5
33	696	151	12.6	83	634	476	1.5
34	799	159	32.6	84	634	500	1.3
35	797	173	4.8	85	666	505	0.3

36	850	137	32.6	86	710	485	0.7
37	828	193	11.2	87	708	502	3
38	830	207	4	88	744	492	3.5
39	652	227	6.5	89	765	471	1.3
40	711	229	4	90	781	458	0.7
41	688	246	1.7	91	808	468	1.1
42	663	263	5.4	92	826	477	5.8
43	625	269	1.1	93	770	481	2.5
44	710	263	10.6	94	768	506	2.1
45	761	239	8.1	95	791	495	0.9
46	754	253	8.5	96	793	506	6.1
47	791	252	6.8	97	796	520	4.3
48	810	255	2.1	98	813	510	5.4
49	803	270	7.2	99	832	520	5.8
50	849	307	1.7				

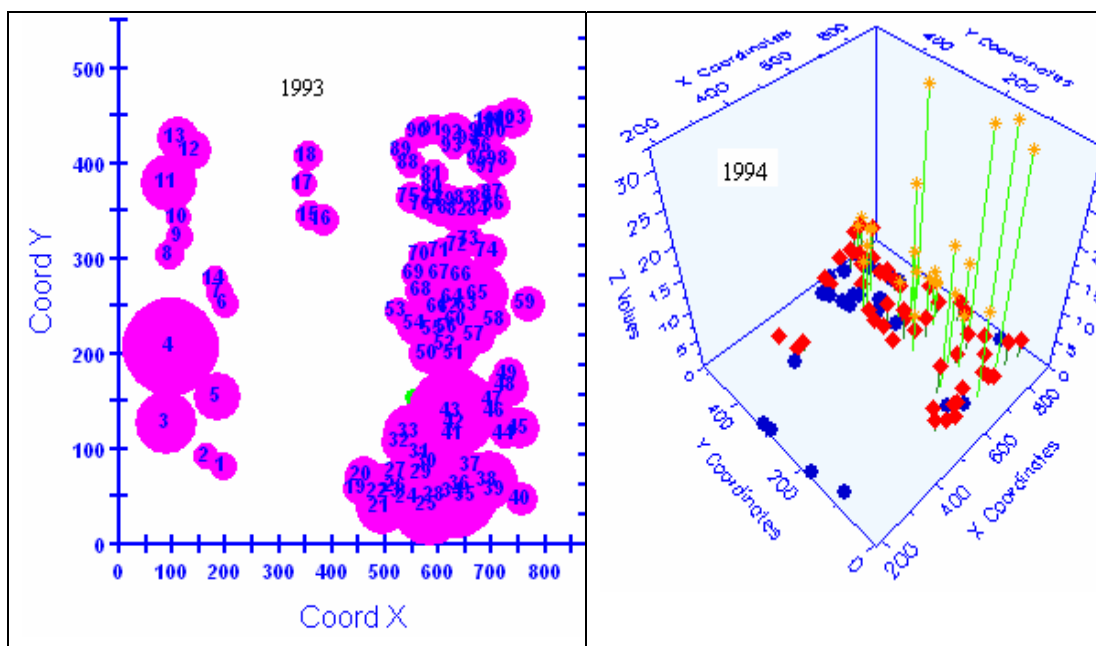
SampleID	Coord X	Coord Y	GrassHoppers 1995	SampleID	Coord X	Coord Y	GrassHoppers 1995
1	176	102	0.9	49	671	231	1.9
2	96	144	0.7	50	665	244	1.5
3	193	170	0.2	51	704	252	0.3
4	112	235	0.3	52	608	235	0.9
5	126	248	1.3	53	585	249	1.3
6	207	246	0.7	54	534	255	0.9
7	198	288	0	55	560	260	0.6
8	189	303	0.2	56	622	260	2.1
9	111	329	0.5	57	596	264	3
10	127	348	1.5	58	566	292	0.3
11	130	377	0.5	59	496	298	0.7
12	147	453	1.3	60	553	312	0.7
13	119	475	0	61	562	335	2.7
14	382	365	0.3	62	623	314	1.1
15	382	377	0	63	622	291	1.3
16	366	376	0.2	64	625	271	1.3
17	363	446	2.7	65	651	277	1.3
18	453	63	1.5	66	667	293	9
19	464	77	4	67	705	302	26.8
20	485	63	1.3	68	669	336	19.2
21	495	48	0.3	69	656	352	23.4
22	511	60	0.5	70	625	341	1.5
23	538	60	1.1	71	596	342	0.3
24	524	85	0.5	72	598	380	2.1
25	524	57	0.3	73	638	369	10
26	583	55	5.4	74	676	389	1.9
27	622	59	3.5	75	700	387	1.3
28	646	56	1.1	76	699	398	0.7
29	745	49	6.1	77	609	396	2.7
30	705	75	4.5	78	585	423	1.3
31	688	74	6.8	79	585	407	1.7
32	638	69	1.1	80	585	394	0.2
33	659	91	3.7	81	542	398	1.9

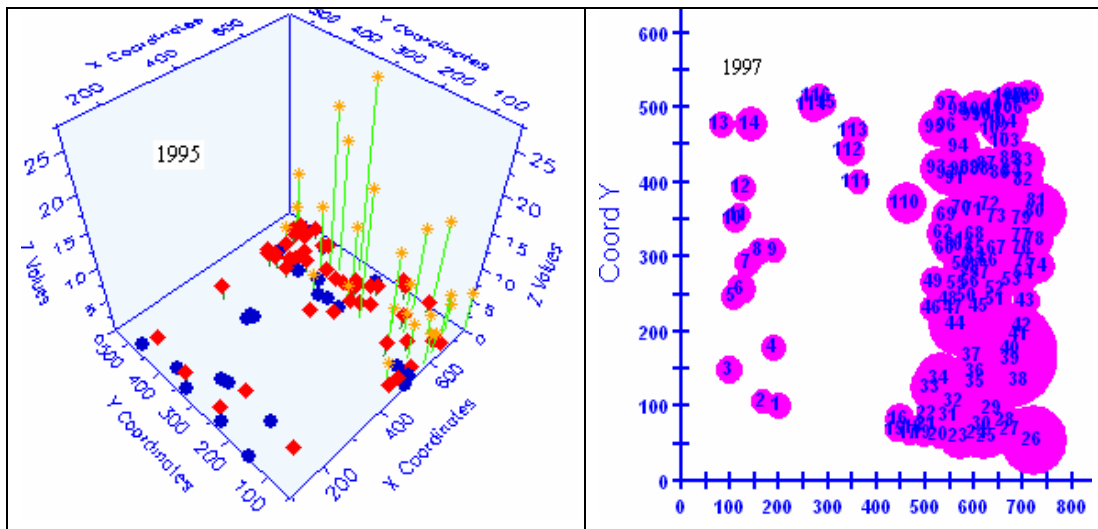
34	565	84	6.1	82	546	434	2.3
35	579	98	2.1	83	536	449	2.7
36	548	125	15.3	84	563	448	1.9
37	528	119	2.5	85	562	470	1.5
38	622	127	3.7	86	591	463	0.5
39	625	138	2.7	87	619	459	1.1
40	615	157	3.5	88	623	462	3.7
41	718	127	1.9	89	655	462	2.5
42	747	130	13.4	90	671	441	1.7
43	705	155	12.6	91	669	439	2.1
44	701	167	4	92	710	436	1.9
45	714	184	2.5	93	664	464	1.3
46	727	196	2.5	94	712	482	0.9
47	580	217	19.2	95	693	484	4.5
48	626	220	0.7	96	714	480	0.5
				97	712	489	9

SampleID	Coord X	Coord Y	GrassHoppers 1997	SampleID	Coord X	Coord Y	GrassHoppers 1997
1	196	103	0.3	59	576	292	3.4
2	166	109	0.1	60	542	312	0.8
3	97	151	0.7	61	562	324	5.9
4	184	180	0.3	62	537	334	2.8
5	103	249	0.2	63	578	316	1.1
6	121	259	1.2	64	602	300	2.2
7	135	294	0.8	65	601	312	3.6
8	157	311	0.2	66	630	295	1.7
9	190	309	0.1	67	647	312	7.6
10	105	352	0.2	68	601	332	0.8
11	118	357	0.1	69	544	356	2
12	122	395	0.5	70	575	367	0.4
13	79	479	0.2	71	600	363	3.3
14	139	479	1.9	72	634	370	12.5
15	438	71	0.1	73	647	354	4.3
16	445	84	1.3	74	731	290	1.9
17	467	66	0	75	705	295	3.8
18	476	73	0	76	699	313	2.5
19	494	65	0.5	77	699	330	5.5
20	526	63	0.8	78	727	324	1.6
21	505	77	0.1	79	697	353	3.3
22	504	92	0.1	80	725	361	6.9
23	568	62	4	81	728	376	2.8
24	606	65	4.1	82	701	403	2.6
25	625	60	2.5	83	703	429	2.8
26	720	56	8.5	84	677	415	1.2
27	675	71	3.8	85	673	431	1.8
28	663	82	0.9	86	655	414	1
29	637	98	0.5	87	627	425	2
30	618	78	8.5	88	612	418	1.5
31	547	90	4.3	89	591	420	1.4
32	559	107	0.7	90	573	417	2.9

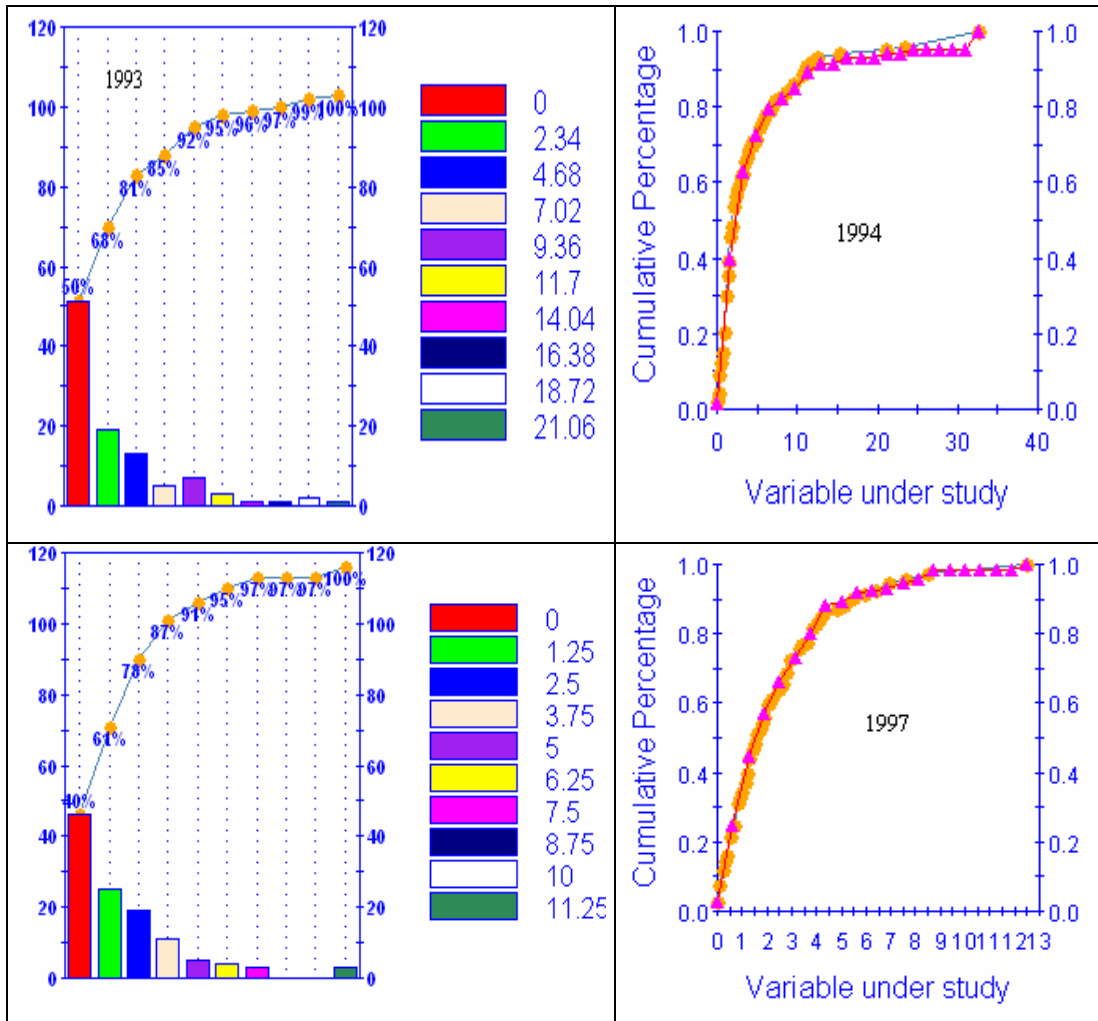
33	511	126	3.8	91	561	407	1.8
34	526	139	5.5	92	544	414	3.8
35	604	135	5.5	93	524	421	2.8
36	603	149	0.7	94	567	448	3.3
37	595	170	2.1	95	519	475	2.6
38	693	136	0.5	96	543	476	4.1
39	675	164	12.5	97	544	505	1.4
40	675	178	12.5	98	568	499	0.5
41	692	198	6.4	99	597	490	1.6
42	698	210	2.9	100	601	498	2.9
43	706	243	0.7	101	628	491	0.5
44	562	212	6.9	102	645	471	1.1
45	608	234	0.2	103	663	456	1.2
46	513	232	0.8	104	658	481	5.1
47	557	232	6.9	105	651	502	2
48	548	244	2.5	106	672	501	1.3
49	518	267	1.3	107	672	517	1
50	586	249	0.8	108	689	512	1.3
51	644	244	1.4	109	706	517	1.6
52	644	259	1.7	110	458	373	3.3
53	679	269	1.1	111	358	401	0
54	702	280	2	112	342	443	0.9
55	566	266	3.8	113	352	470	0.8
56	593	268	1.3	114	267	506	1.5
57	613	279	4.8	115	288	507	0.4
58	593	287	2.9	116	279	517	0.3

SAKWeb[®] 2D And 3D Data Postings

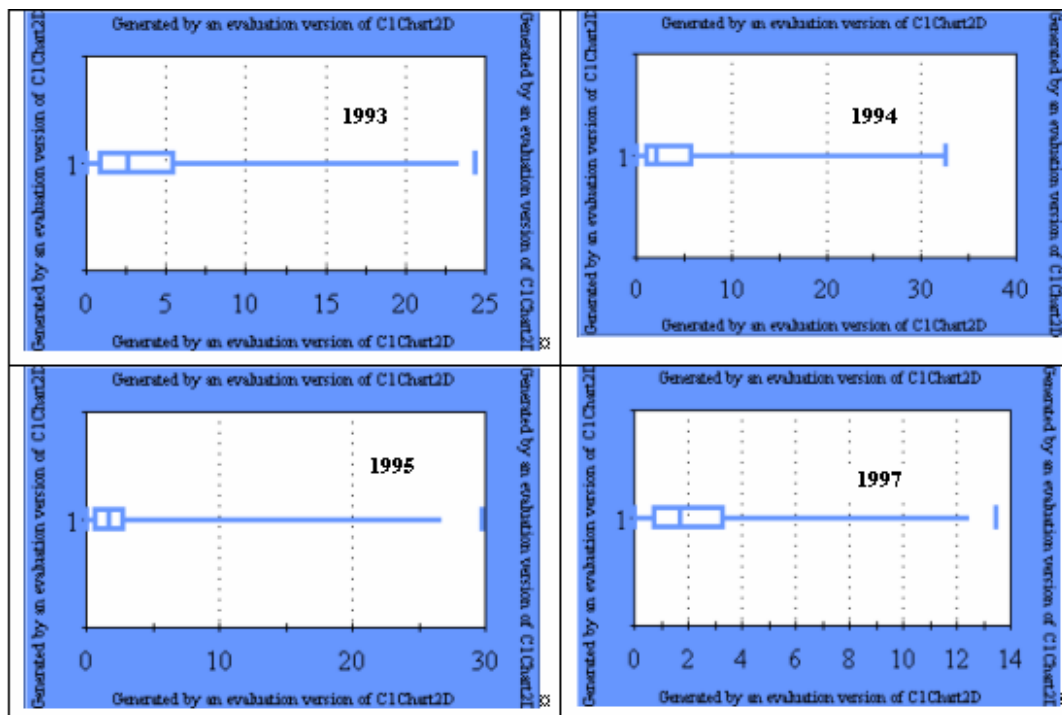




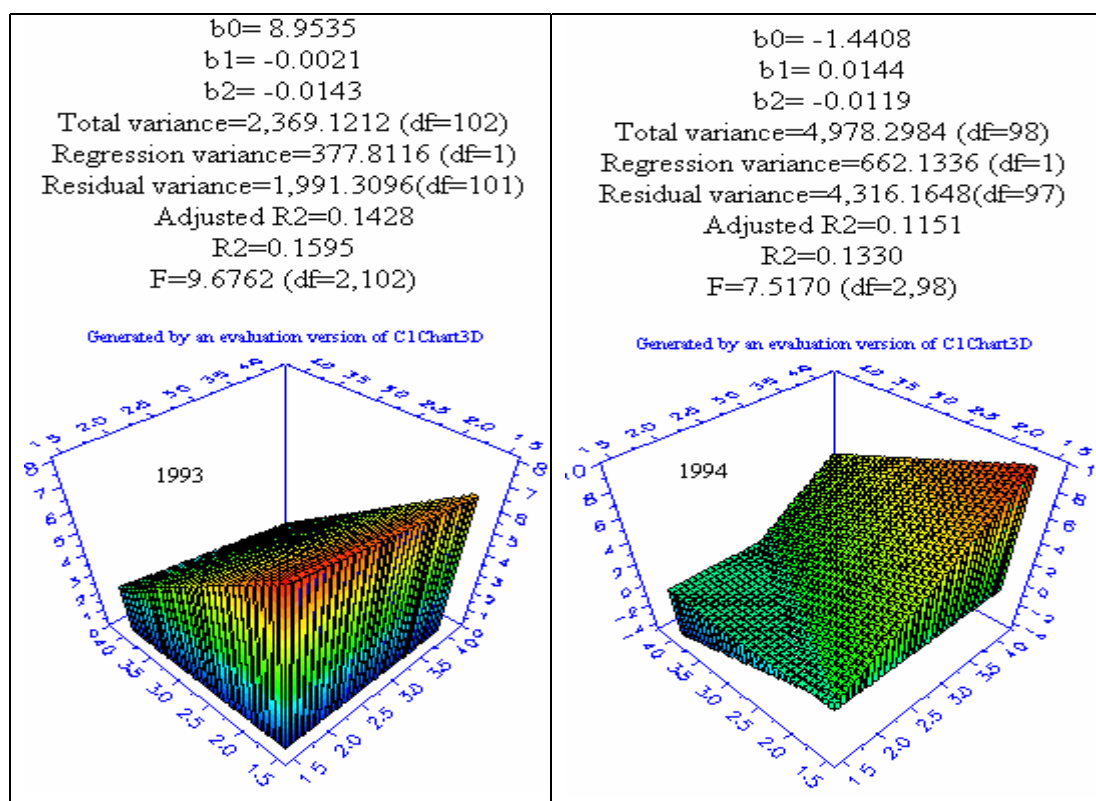
SAKWeb® Cumulative Histograms

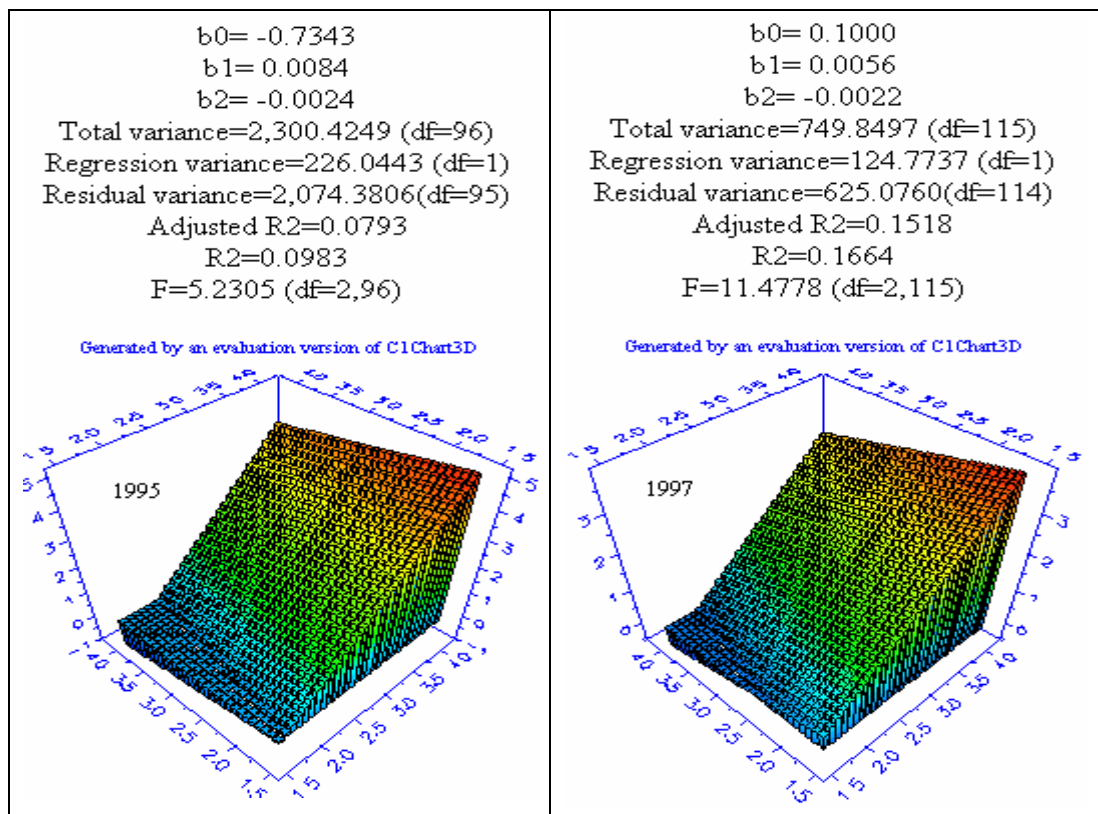


SAKWeb[®] Box-And-Whisker Plot

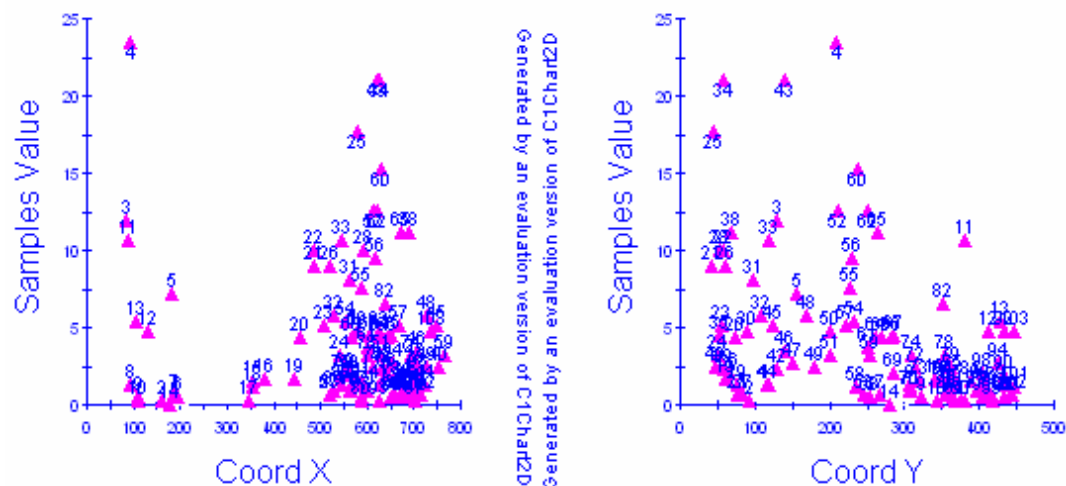


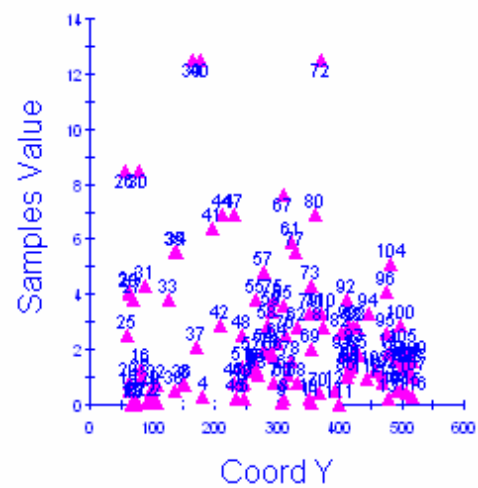
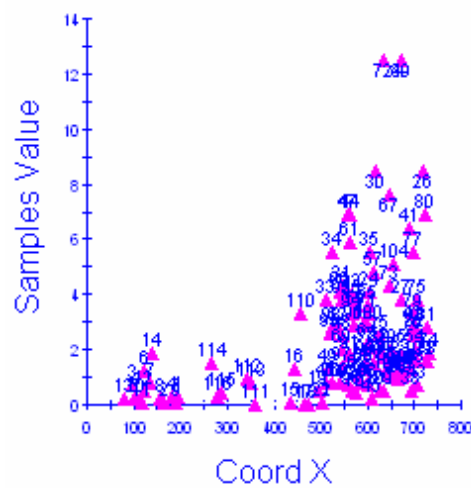
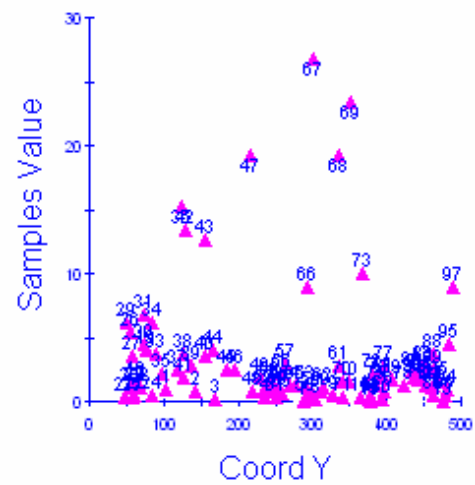
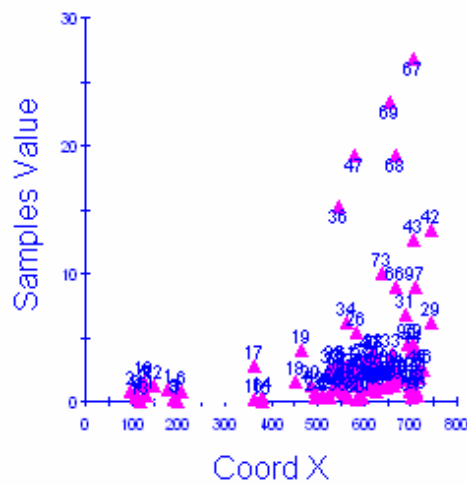
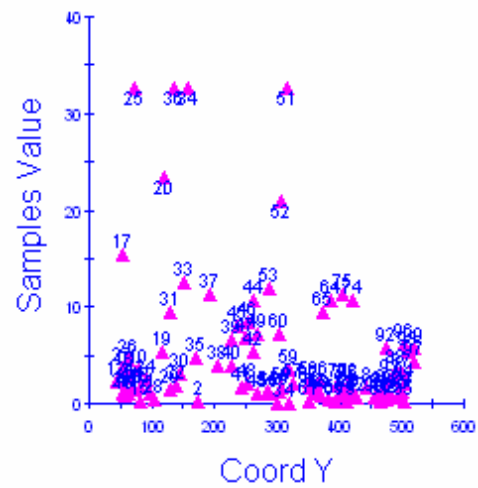
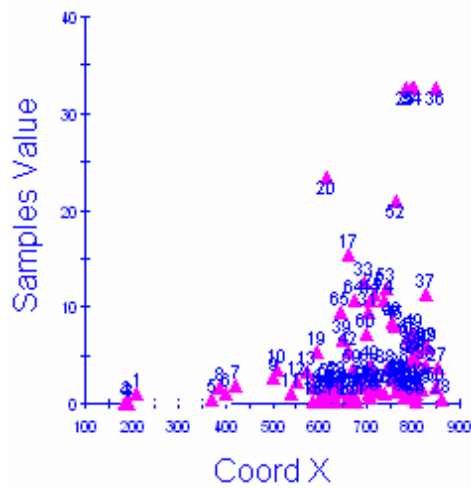
SAKWeb[®] Linear Polynomial Trend Model



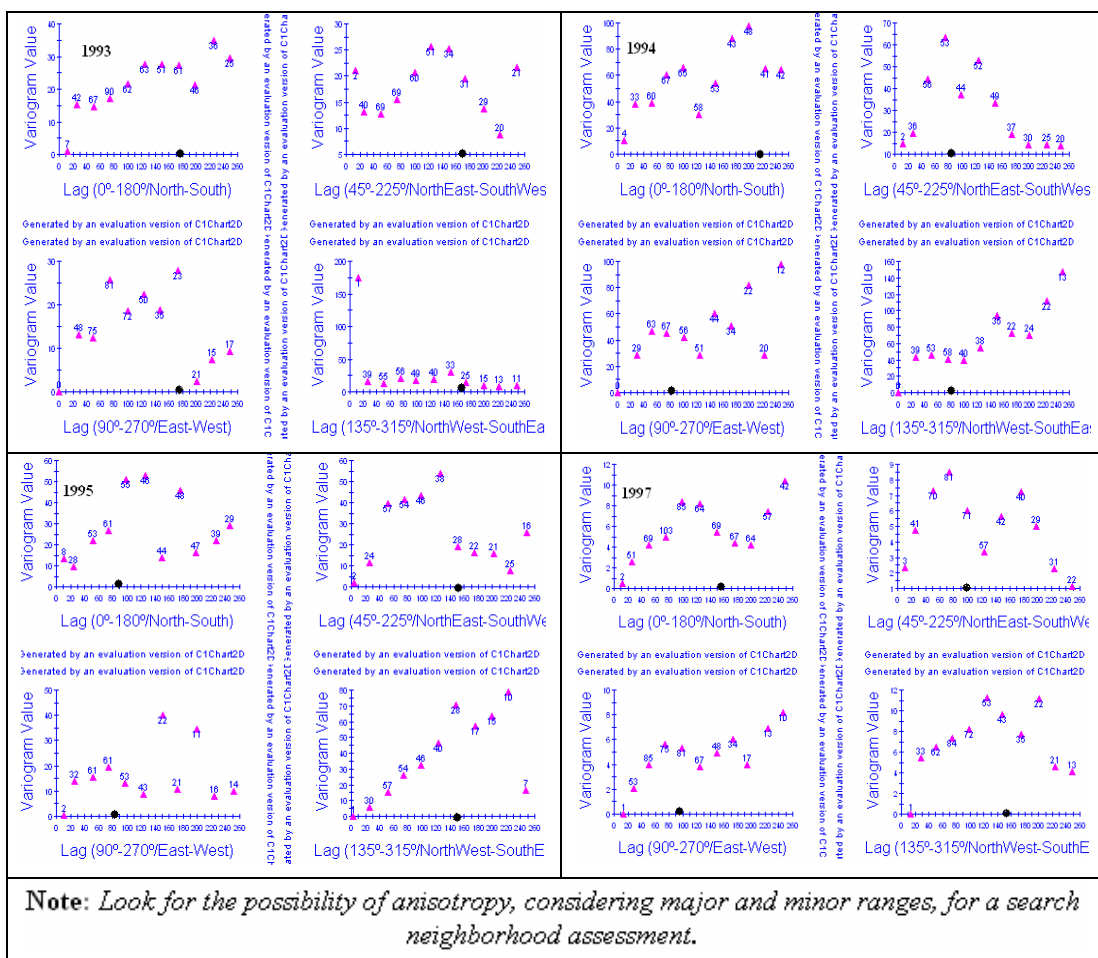


The Bivariate Scatterplot (density versus longitude and latitude) Of Grasshoppers Datasets (1993, 1994, 1995 and 1997, respectively)

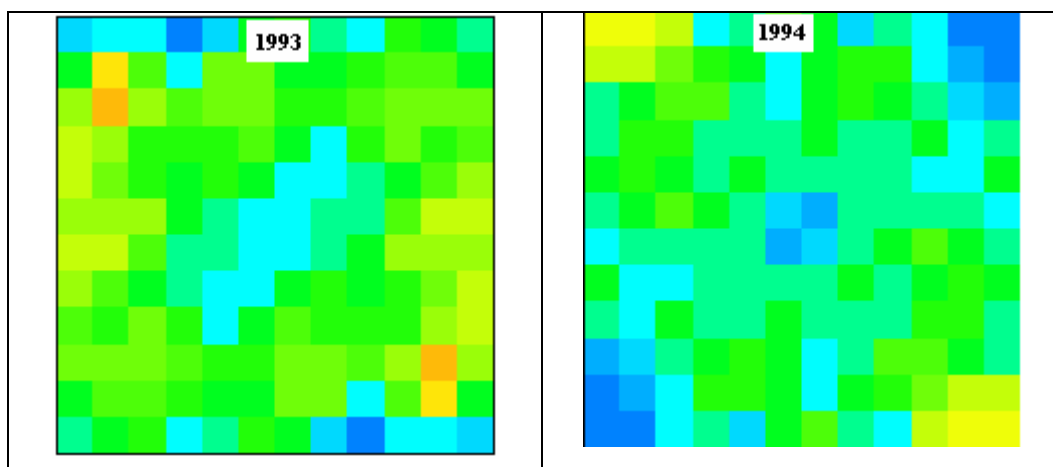


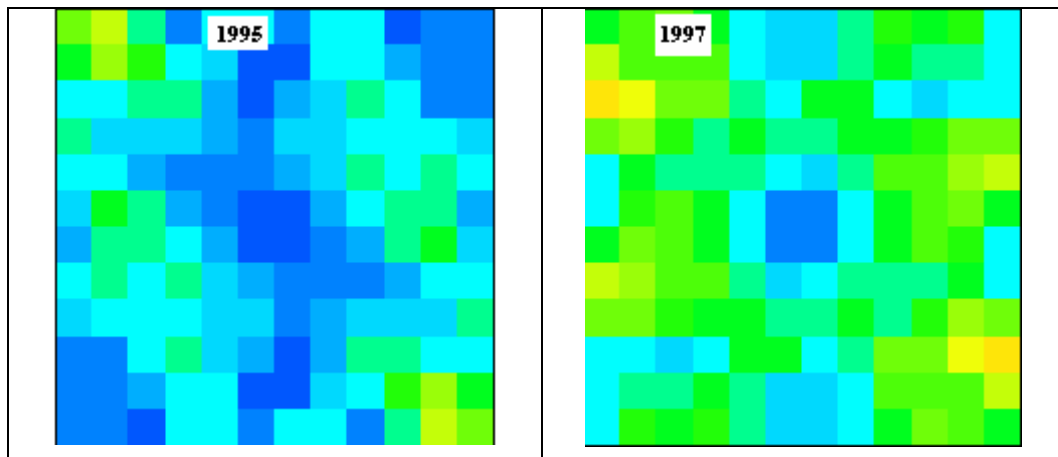


SAKWeb[®] Variogram According To Four Directions

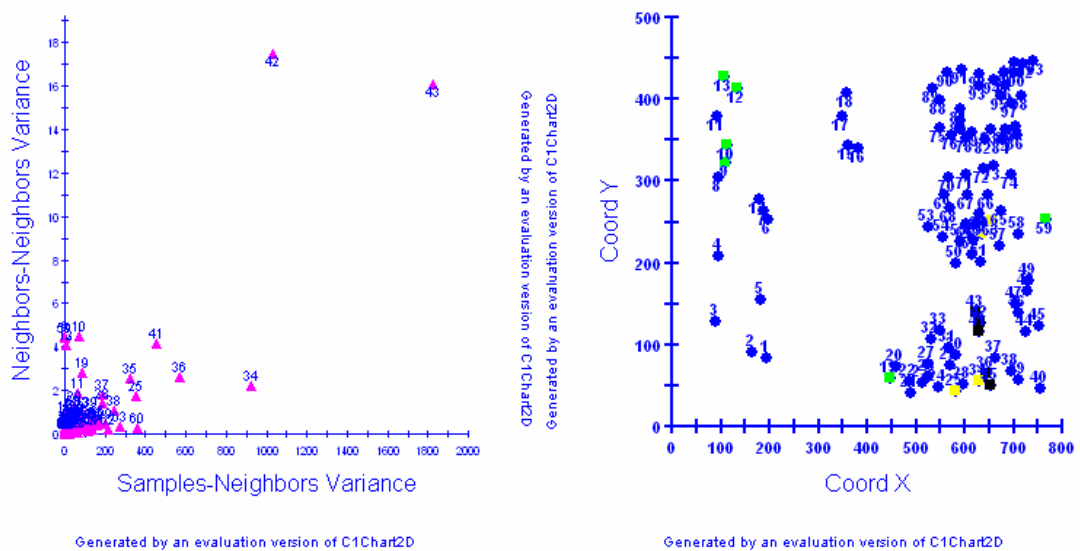


SAKWeb[®] Anisotropy Map





The SAKWeb[®] Location And Variance Of The 1993 Moran Scatterplot



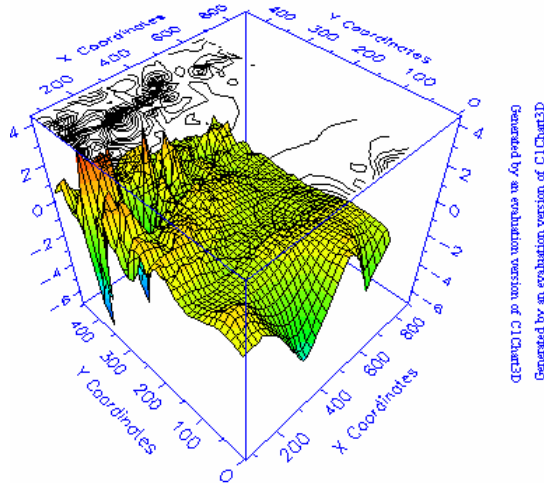
Black - Highly variability region
 Green - Medium central-neighbors versus low central-neighbors variance
 Yellow - Outliers
 Blue - Stable region

The SAKWeb[®] Exact OK Differences Of 1993

SAKWeb Exact OK Differences

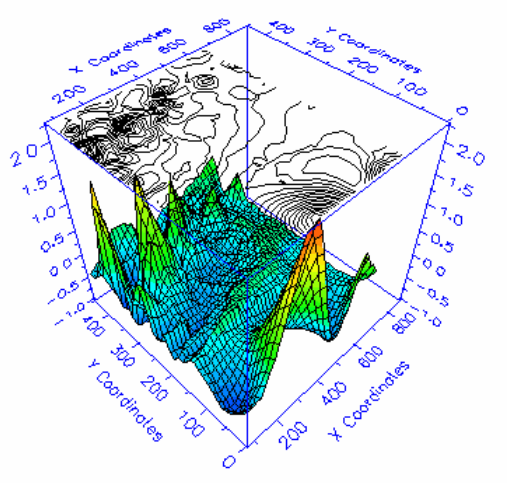
OK difference between First and Second approach

Generated by an evaluation version of C1Chart3D



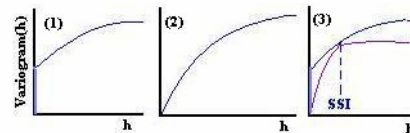
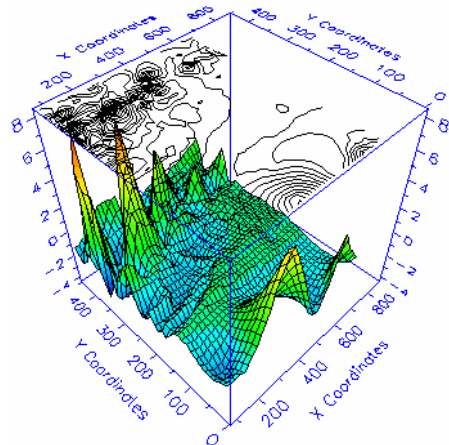
OK difference between First and Third approach

Generated by an evaluation version of C1Chart3D



OK difference between Second and Third approach

Generated by an evaluation version of C1Chart3D



OK With C0 (difference average value): 3.85234

(Maximum Difference: 18.75887)

(Minimum Difference: -0.15475)

OK Variance Average: 21.97120

OK Without C0 (difference average value): 3.75503

(Maximum Difference: 22.24619)

(Minimum Difference: -0.88978)

OK Variance Average: 13.72206

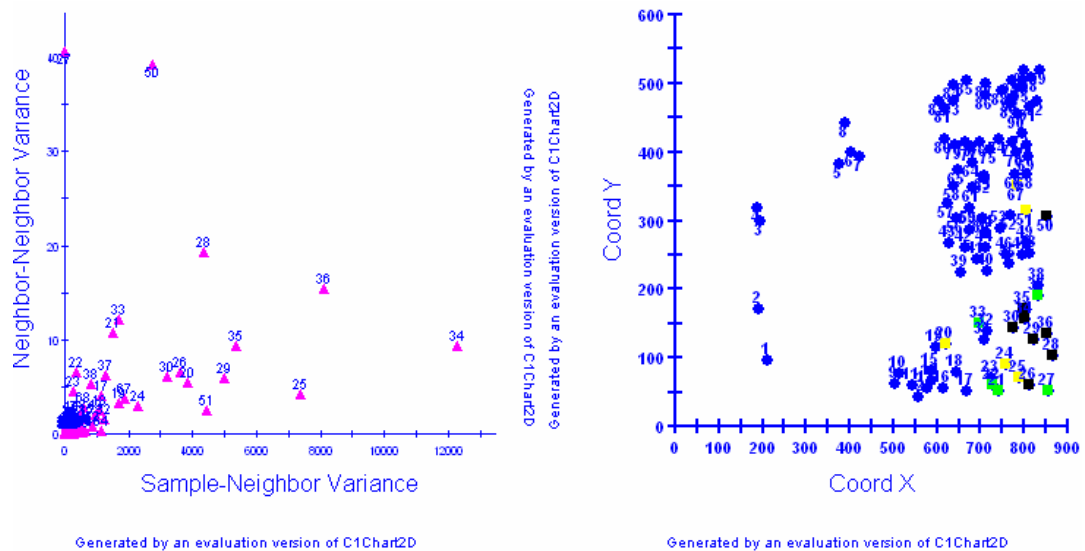
OK With C0 Plus Two Structures (difference average value): 3.91091

(Maximum Difference: 16.51323)

(Minimum Difference: 0.40153)

OK Variance Average: 29.76893

The SAKWeb[®] Variance Of The 1994 Moran Scatterplot



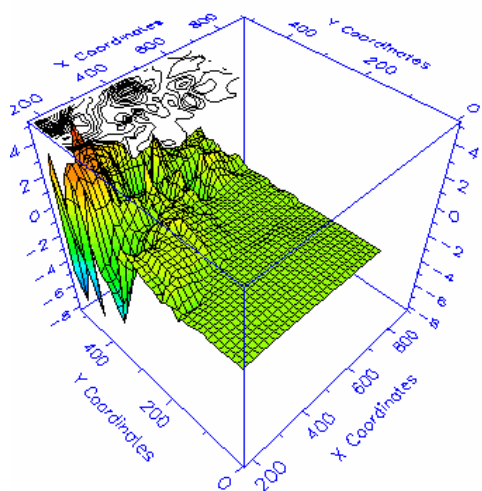
Black - High variability region
 Green - Medium central-neighbors versus low central-neighbors variance
 Yellow - Outliers
 Blue - Stable region

The SAKWeb[®] Exact OK Differences Of 1994

SAKWeb Exact OK Differences

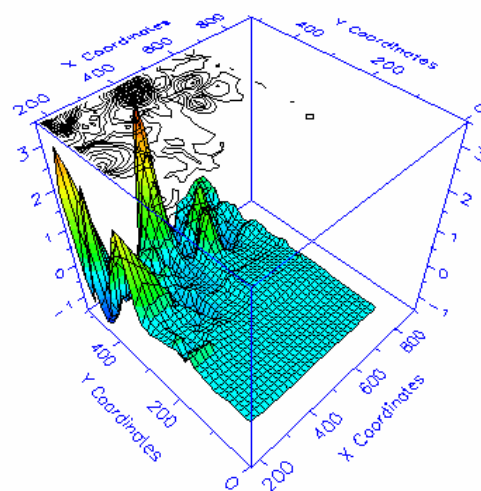
OK difference between First and Second approach

Generated by an evaluation version of C1Chart3D

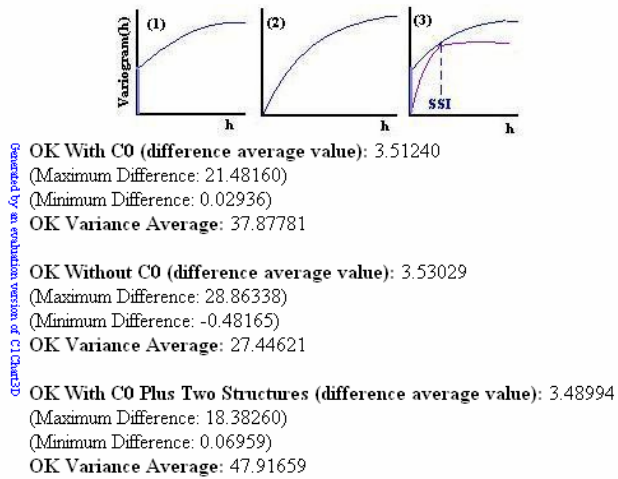
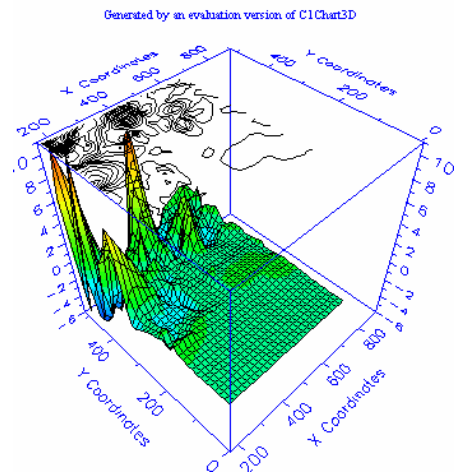


OK difference between First and Third approach

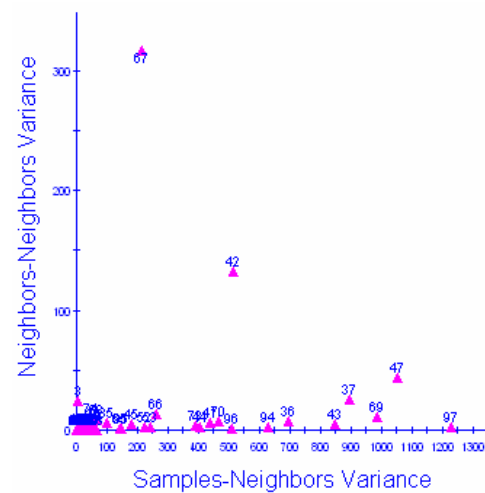
Generated by an evaluation version of C1Chart3D



OK difference between Second and Third approach

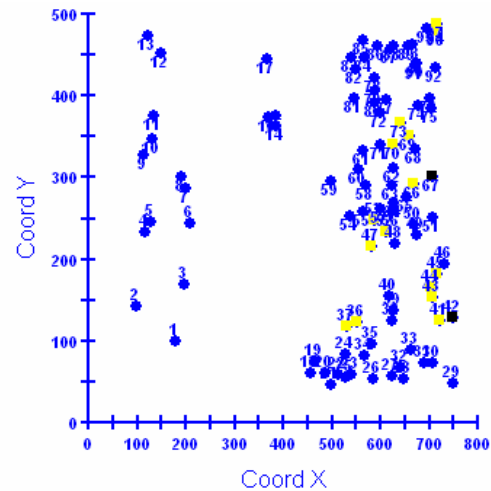


The SAKWeb® Variance Of The 1995 Moran Scatterplot



Generated by an evaluation version of C1Chart2D

Generated by an evaluation version of C1Chart2D



Generated by an evaluation version of C1Chart2D

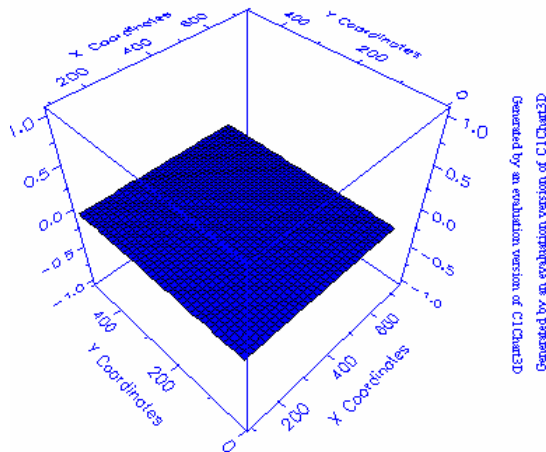
Black - Highly variability region
Green - Medium central-neighbors versus low central-neighbors variance
Yellow - Outliers
Blue - Stable region

The SAKWeb[®] Exact OK Differences Of 1995

SAKWeb Exact OK Differences

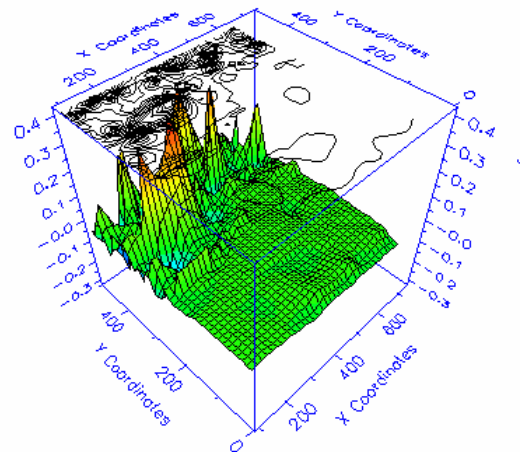
OK difference between First and Second approach

Generated by an evaluation version of C1Chart3D



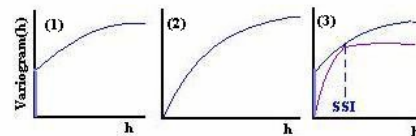
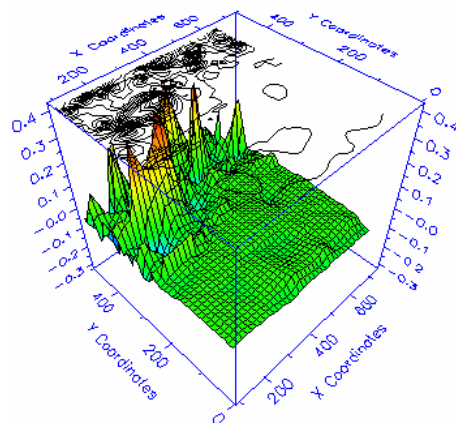
OK difference between First and Third approach

Generated by an evaluation version of C1Chart3D



OK difference between Second and Third approach

Generated by an evaluation version of C1Chart3D

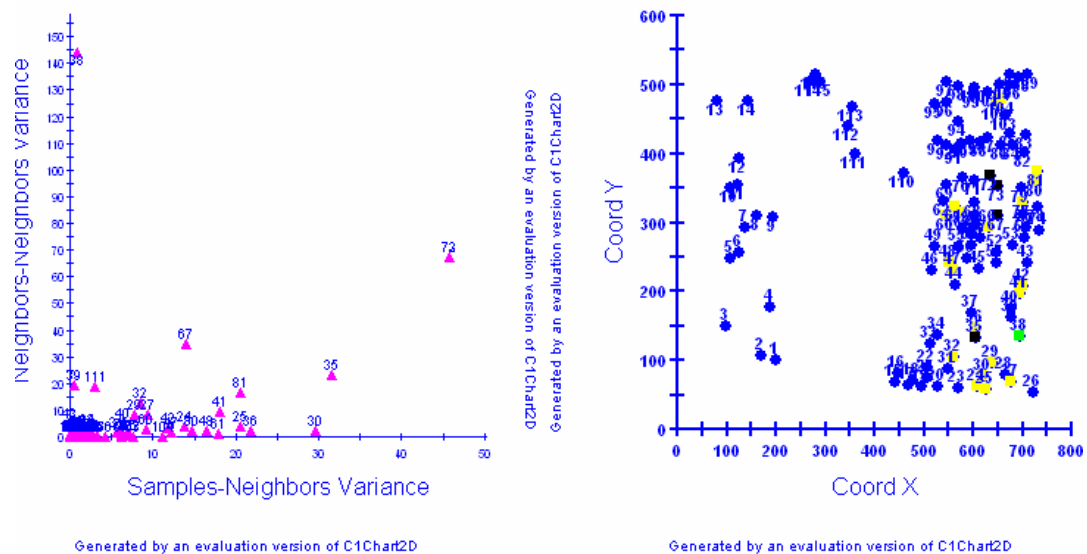


OK With C0 (difference average value): 2.53458
(Maximum Difference: 24.28161)
(Minimum Difference: -0.14040)
OK Variance Average: 16.67592

OK Without C0 (difference average value): 2.53458
(Maximum Difference: 24.28161)
(Minimum Difference: -0.14040)
OK Variance Average: 16.67592

OK With C0 Plus Two Structures (difference average value): 2.532
(Maximum Difference: 23.97388)
(Minimum Difference: -0.09034)
OK Variance Average: 17.09188

The SAKWeb® Variance Of The 1997 Moran Scatterplot



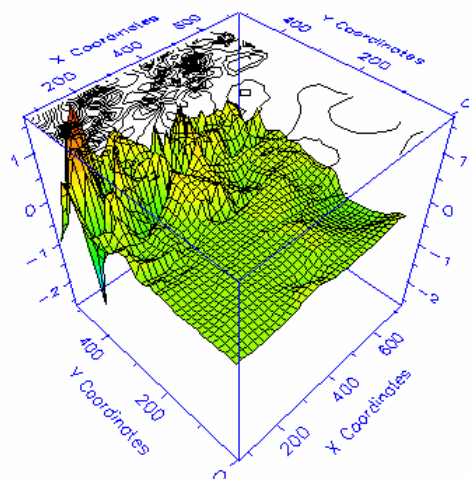
Black - Highly variability region
 Green - Medium central-neighbors versus low central-neighbors variance
 Yellow - Outliers
 Blue - Stable region

The SAKWeb® Exact OK Differences Of 1997

SAKWeb Exact OK Differences

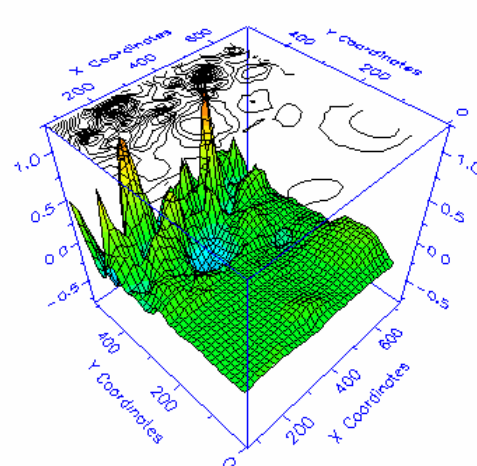
OK difference between First and Second approach

Generated by an evaluation version of C1Chart3D

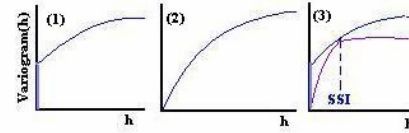
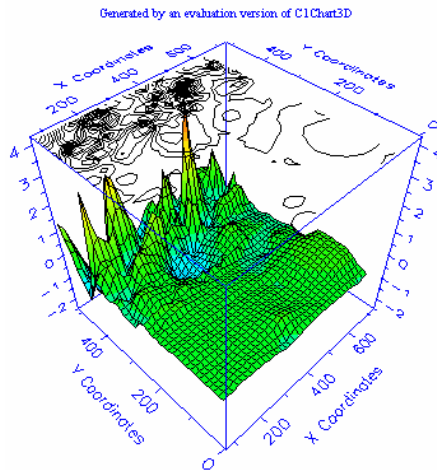


OK difference between First and Third approach

Generated by an evaluation version of C1Chart3D



OK difference between Second and Third approach



OK With C0 (difference average value): 1.70039

(Maximum Difference: 9.98653)

(Minimum Difference: 0.08674)

OK Variance Average: 4.11661

OK Without C0 (difference average value): 1.66771

(Maximum Difference: 11.56023)

(Minimum Difference: -0.00009)

OK Variance Average: 3.04476

OK With C0 Plus Two Structures (difference average value): 1.70954

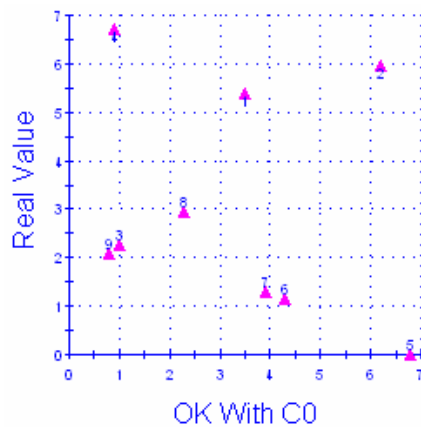
(Maximum Difference: 8.99647)

(Minimum Difference: 0.13047)

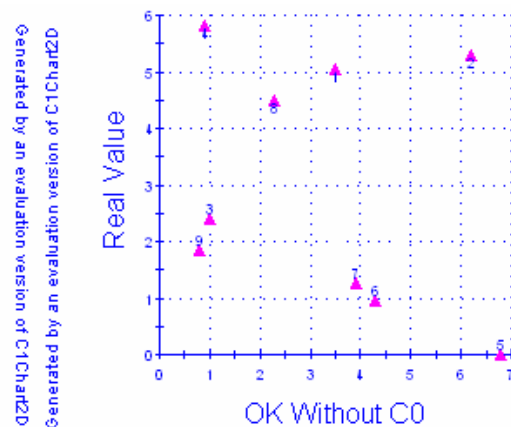
OK Variance Average: 5.17393

The SAKWeb® Validation Procedure Of 1993

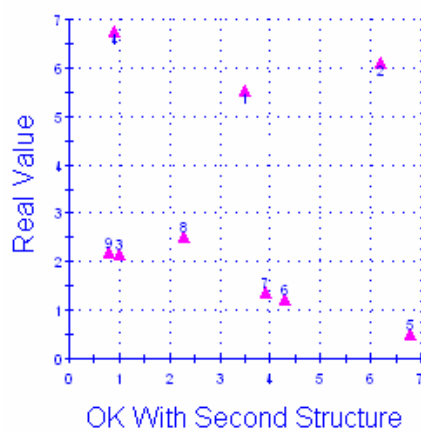
Sample ID	Coord X	Coord Y	Sample Value	OK With C0 (Difference against real)	OK Without C0 (Difference against real)	OK With Second Structure (Difference against real)
1	185	200	3.5	5.3736580 (-1.873658)	5.0506320 (-1.550632)	5.5147920 (-2.014792)
2	110	400	6.2	5.9736990 (0.226301)	5.3055390 (0.8944610000000001)	6.0874250 (0.1125750000000001)
3	360	150	1	2.2504730 (-1.250473)	2.3879540 (-1.387954)	2.1402510 (-1.140251)
4	360	370	0.9	6.7022150 (-5.802215)	5.8127300 (-4.91273)	6.7595130 (-5.859513)
5	540	70	6.8	0 (6.8)	0 (6.8)	.50884640 (6.2911536)
6	650	130	4.3	1.1409650 (3.159035)	.94979410 (3.3502059)	1.2184010 (3.081599)
7	590	280	3.9	1.3026790 (2.597321)	1.2526460 (2.647354)	1.3538390 (2.546161)
8	730	360	2.3	2.9455270 (-0.645527)	4.5010910 (-2.201091)	2.5133620 (-0.213362)
9	580	420	0.8	2.0650220 (-1.265022)	1.8314380 (-1.031438)	2.2013900 (-1.40139)



Generated by an evaluation version of C1Chart2D
Generated by an evaluation version of C1Chart2D



Generated by an evaluation version of C1Chart2D



Generated by an evaluation version of C1Chart2D

Total Error Sum

OK With C0: 23.619552
OK Without C0: 24.7758659
OK With Two Structures: 22.6607966

Positive Error Sum

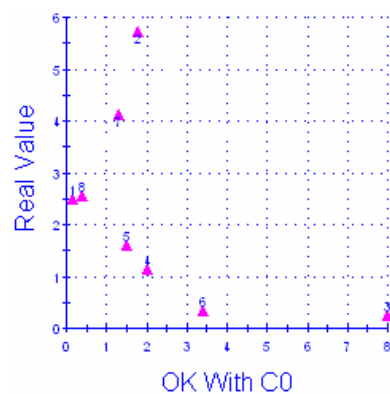
OK With C0: 12.782657
OK Without C0: 13.6920209
OK With Two Structures: 12.0314886

Negative Error Sum

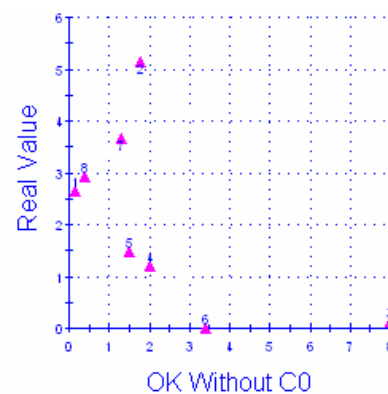
OK With C0: -10.836895
OK Without C0: -11.083845
OK With Two Structures: -10.629308

The SAKWeb® Validation Procedure Of 1994

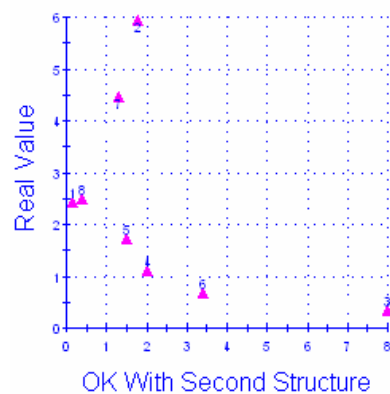
Sample ID	Coord X	Coord Y	Sample Value	OK With C0 (Difference against real)	OK Without C0 (Difference against real)	OK With Second Structure (Difference against real)
1	190	250	0.15	2.4856470 (-2.335647)	2.6574080 (-2.507408)	2.4418040 (-2.291804)
2	431	410	1.8	5.7335730 (-3.933573)	5.1305470 (-3.330547)	5.9338550 (-4.133855)
3	620	70	8	.24414240 (7.7558576)	.12022060 (7.8797794)	.32393790 (7.6760621)
4	792	120	2	1.1438620 (0.856138)	1.2044250 (0.795575)	1.1100600 (0.88994)
5	610	280	1.5	1.5956750 (-0.095675)	1.4879710 (1.202900000000001E-02)	1.7134290 (-0.213429)
6	755	347	3.4	.33986390 (3.0601361)	0 (3.4)	.67177750 (2.7282225)
7	644	481	1.3	4.1249220 (-2.824922)	3.6660800 (-2.36608)	4.4754210 (-3.175421)
8	830	415	0.4	2.5470700 (-2.14707)	2.9313060 (-2.531306)	2.4827400 (-2.08274)



Generated by an evaluation version of C1Chart2D
Generated by an evaluation version of C1Chart2D



Generated by an evaluation version of C1Chart2D



Generated by an evaluation version of C1Chart2D

Total Error Sum

OK With C0: 23.0090187
OK Without C0: 22.8227244
OK With Two Structures: 23.1914736

Positive Error Sum

OK With C0: 11.6721317
OK Without C0: 12.0873834
OK With Two Structures: 11.2942246

Negative Error Sum

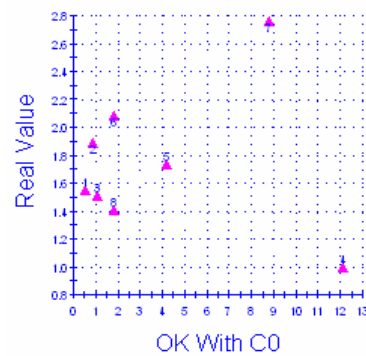
OK With C0: -11.336887
OK Without C0: -10.735341
OK With Two Structures: -11.897249

The SAKWeb® Validation Procedure Of 1995

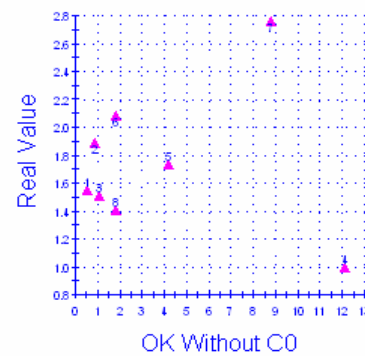
SAKWeb Validation Dataset

Sample ID	Coord X	Coord Y	Sample Value	OK With C0 (Difference against real)	OK Without C0 (Difference against real)	OK With Second Structure (Difference against real)
1	145	180	0.55	1.5507030 (-1.000703)	1.5507030 (-1.000703)	1.5695350 (-1.019535)
2	161	402	0.9	1.8921250 (-0.992125)	1.8921250 (-0.992125)	1.9193000 (-1.0193)
3	360	422	1.1	1.5110330 (-0.411033)	1.5110330 (-0.411033)	1.5571690 (-0.457169)
4	597	154	12.1	.99569990 (11.1043001)	.99569990 (11.1043001)	.98826180 (11.1117382)
5	695	135	4.2	1.7330300 (2.46697)	1.7330300 (2.46697)	1.7144630 (2.485537)
6	601	312	1.8	2.0806340 (-0.280634)	2.0806340 (-0.280634)	2.0684660 (-0.268466)
7	650	400	8.8	2.7593220 (6.040678)	2.7593220 (6.040678)	2.7692510 (6.030749)
8	700	427	1.8	1.4073010 (0.392699)	1.4073010 (0.392699)	1.5033010 (0.296699)

Generated by an evaluation version of C1Chart2D

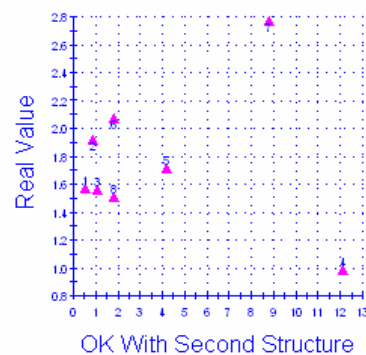


Generated by an evaluation version of C1Chart2D



Generated by an evaluation version of C1Chart2D

Generated by an evaluation version of C1Chart2D



Generated by an evaluation version of C1Chart2D

Generated by an evaluation version of C1Chart2D

Total Error Sum

OK With C0: 22.6891421
 OK Without C0: 22.6891421
 OK With Two Structures: 22.6891932

Positive Error Sum

OK With C0: 20.0046471
 OK Without C0: 20.0046471
 OK With Two Structures: 19.9247232

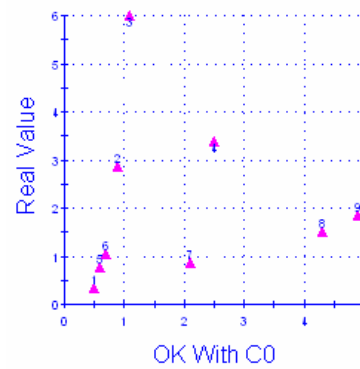
Negative Error Sum

OK With C0: -2.684495
 OK Without C0: -2.684495
 OK With Two Structures: -2.76447

Spatial Analysis**The SAKWeb® Validation Procedure Of 1997****SAKWeb Validation Dataset**

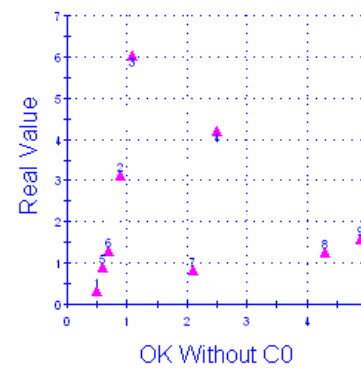
Sample ID	Coord X	Coord Y	Sample Value	OK With C0 (Difference against real)	OK Without C0 (Difference against real)	OK With Second Structure (Difference against real)
1	171	179	0.5	.35158400 (0.148416)	.33861310 (0.1613869)	.34372140 (0.1562786)
2	140	370	0.9	2.8605860 (-1.960586)	3.1346080 (-2.234608)	2.7705910 (-1.870591)
3	300	490	1.1	5.9901700 (-4.89017)	6.0342240 (-4.934224)	5.9137840 (-4.813784)
4	472	388	2.5	3.3871510 (-0.887151)	4.2142490 (-1.714249)	3.0966040 (-0.596604)
5	569	106	0.6	.78370060 (-0.1837006)	.90986290 (-0.3098629)	.72880490 (-0.1288049)
6	631	145	0.7	1.0461380 (-0.346138)	1.2791490 (-0.579149)	.95095830 (-0.2509583)
7	659	219	2.1	.84976290 (1.2502371)	.81592910 (1.2840709)	.86690890 (1.2330911)
8	703	379	4.3	1.5058440 (2.794156)	1.2395410 (3.060459)	1.6623960 (2.637604)
9	582	476	4.9	1.8503220 (3.049678)	1.5650700 (3.33493)	2.0369560 (2.863044)

Generated by an evaluation version of C1Chart2D



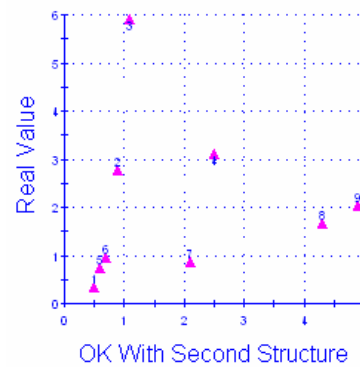
Generated by an evaluation version of C1Chart2D

Generated by an evaluation version of C1Chart2D



Generated by an evaluation version of C1Chart2D

Generated by an evaluation version of C1Chart2D



Generated by an evaluation version of C1Chart2D

Generated by an evaluation version of C1Chart2D

Total Error Sum

OK With C0: 15.5102327
 OK Without C0: 17.6129397
 OK With Two Structures: 14.5507599

Positive Error Sum

OK With C0: 7.2424871
 OK Without C0: 7.8408468
 OK With Two Structures: 6.8900177

Negative Error Sum

OK With C0: -8.2677456
 OK Without C0: -9.7720929
 OK With Two Structures: -7.6607422

Generated by an evaluation version of C1Chart2D

Spatial Analysis

Appendix F

The Sampling Price at *Direcção Geral Do Ambiente* [2001]

Parâmetros Analisados no Sector de Metais

PARÂMETROS	MATRIZ	MÉTODO	PREÇO
ALUMÍNIO	Água superficial, subterrânea, rede	EAAFG ICP	2.500\$00 12.47 €
	Água residual	ICP	5.000\$00
	Amostra sólida	FRX	24.94 €
ARSÉNIO	Água superficial, subterrânea, rede	EAAGH	3.500\$00 17.46 €
	Água residual	EAAGH	7.000\$00
	Amostra sólida		34.92 €
BÁRIO	Água superficial, subterrânea, rede	ICP	2.500\$00 12.47 €
	Água residual		5.000\$00
	Amostra sólida		24.94 €
BORO	Água superficial, subterrânea, rede	ICP	2.500\$00 12.47 €
	Água residual		5.000\$00
	Amostra sólida		24.94 €
CÁDMIO	Água superficial, subterrânea, rede	EAAFG	2.500\$00 12.47 €
	Água residual	ICP	5.000\$00
	Amostra sólida	FRX	24.94 €

Tabela de Preços (Matriz sólida e águas residuais)

DETERMINAÇÃO	PREÇO
<i>FITOPLÂNCION</i> - Análise quantitativa	44.000\$00 219.47 €
<i>FITOPLÂNCION</i> - Análise qualitativa	24.000\$00 119.71 €
<i>ECOTOXICOLOGIA</i> - Teste de inibição da bioluminescência de <i>Vibrio fischeri</i>	40.000\$00 199.52 €

Tabela de Preços (Matriz sólida e águas residuais)

PARÂMETRO	MÉTODO	PREÇO
INSECTICIDAS ORGANOCLORADOS	Cromatografia	20.000\$00 99.76 €
PCB's	Cromatografia	20.000\$00 99.76 €
HERBICIDAS	Cromatografia	20.000\$00 99.76 €
INSECTICIDAS ORGANOFOSFORADOS	Cromatografia	20.000\$00 99.76 €
TOC	Espectrofotometria de Infravermelho	16.000\$00 79.81 €
ANÁLISE QUALITATIVA DE COMPOSTOS ORGÂNICOS	Cromatografia	40.000\$00 199.52 €

Appendix G



Second Survey

School Name: _____

Name: _____

Gender (Always make a circle on the right answer): Male Female

Age: <=30 >30 and <=40 > 40 and <=50 >50 and <=60 >60

Bachelorato or Licenciatura Name: _____

University Name: _____

Main Subjects of Studies (You may choose more than one option):

- Maths
- Statistic
- Geographical Information Systems
- Geography
- Environmental Issues
- Computers
- Economics, Management, Business and Administration

Basics Knowledge

Relational DataBases	Yes	No
Computer Programming	Yes	No
Geographical Information Systems	Yes	No

Theoretical Knowledge of Statistical Exploratory Data Analysis

Mean, median, variance, standard deviation	Yes	No
Hypothesis Tests	Yes	No
Linear Relationships ($y=ax+b$)	Yes	No

Theoretical Knowledge of Spatial Interpolation Methods

Fourier	Yes	No
B-Splines	Yes	No
Inverse Distance Weighted	Yes	No
Polynomial Trend Surface	Yes	No
Kriging	Yes	No

Knowledge of GIS Software

ESRI (ArcView [®] or ArcInfo [®])	Yes	No
Intergraph [®]	Yes	No

Others, if any: _____

Do you know Tobler's First Law or First Law of Geography? Yes No

ArcView GIS Program [Classification from 1 (-) to 5 (+) and Yes/No]

Module I - Introduction to GIS

(Internal Organization, Visualization, Alphanumerical Analysis) – 8 hours

Difficulty:_____ Time Sufficiency:_____

Module II – Spatial Analyst

(Matrix Manipulations, Spatial Models, Areal Assessment, Surfaces Studies, Spatial Analysis, Reclassification) – 12 hours

Difficulty:_____ Time Sufficiency:_____

Module III – Network Analyst

(Short Cost Distance, Network Construction) – 4 hours

Difficulty:_____ Time Sufficiency:_____

Module IV – 3D Analyst

(TIN, 3D Spatial Analysis, Visibilities) – 4 hours

Difficulty:_____ Time Sufficiency:_____

Module V – Geostatistical Analyst

(Theoretical Concepts, Deterministic Interpolators, Exploratory Spatial Data Analysis, Geostatistical Techniques, Geostatistical Layers Displaying) – 22 hours

Difficulty:_____ Time Sufficiency:_____

From 1(-) to 5(+), indicate the difficulty level to acquire Geostatistics Analyst[®] from ESRI:_____

Concerning software implementation, three basical solutions exist at present (Statistical products, GIS modules, Independent software). Do you think that spatial analysis should be implemented over the Web? _____ If your answer is yes, why? _____

What are the major features that Web spatial analysis software should include? _____

Thanks, Dude!!!!!!

Appendix H



Third Survey

School Name: _____

Name: _____

Comments

Web interface: _____

Hypermedia help: _____

E-Learning tools: _____

Nugget-effect solutions: _____

Spatial autocorrelation measures: _____

SAKWeb[®] [Classification from 1 (-) to 5 (+)]

Web interface	_____
Hypermedia help	_____
E-learning tools	_____
Nugget-effect solutions	_____
Spatial autocorrelation measures	_____

SAKWeb[®] another attractive features: _____

Other SAKWeb[®] future improvements: _____

Concerning the Web solution given by SAKWeb[®], rank the following features in terms of importance:

Low Costs	_____
Mobile and Easy Access	_____
Hardware and Software Management	_____
Web Standard GUI	_____

Concerning global and local spatial autocorrelation measures, rank the following ones in terms of importance:

Moran I and Geary C	_____
The Spatial Lag Pies	_____
Gi(d)	_____
Moran scatterplot	_____
The Variogram Cloud	_____

Do you think these previous measures should be included with all Kriging software as an ESDA tool?

Yes No

Although OK variance is not a good uncertainty measure, do you think that this feature should be banned from Kriging software?

Yes No

Although cross-validation is not regarded by SAKWeb[®], do you think it should be included?

Yes No

Although topographical and hill shade maps are not covered by SAKWeb[®], do you think it should be integrated?

Yes No

Do you think that the cross-validation procedure is an important feature to be considered in SAKWeb[®]?

Yes

No

Regarding the outlier search issue, which software do you prefer (circle one)?

1 - SAKWeb[®] (Moran I- Moran scatterplot – Variogram and Moran scatterplot variance options)

2 - ESRI[®] Geostatistical module (Semivariogram/covariance cloud and Voronoi options)

Regarding the pattern search issue, which software do you prefer (circle one)?

1 - SAKWeb[®] (Moran I- Moran scatterplot – Variogram and Moran scatterplot variance options)

2 - ESRI[®] Geostatistical module (Trend analysis option)

In spite of the possible bias of back-transformations towards the original data, do you think that the Box-cox and the logarithmic transformations must be incorporate within SAKWeb[®]?

Yes

No

Concerning software trademarks such as Adobe[®], ESRI[®], Microsoft[®], SPSS[®] and Macromedia[®], do you think that their products are trustfull and reliable?

Yes

No

If you have two softwares whose goal is the same but one produced by a known company and another by a research institute or university then which ones do you prefer?

Comercial software

☐

Research institute or university products

☐

If a spatial dataset was given to produce an OK surface, which software would you choose?

Statistical packages such as SAS-GIS[®] or SPSS-X[®]

☐

GIS modules such a ESRI[®] Geostatitital Analyst module

☐

Independent software such as GSLib[®] or Regard[®]

☐

Why? _____

Concerning operating systems, which ones do you have knowledge of?

Linux[®]

☐

Sun[®] Unix

☐

MS-Windows[®]

☐

Mac[®]

☐

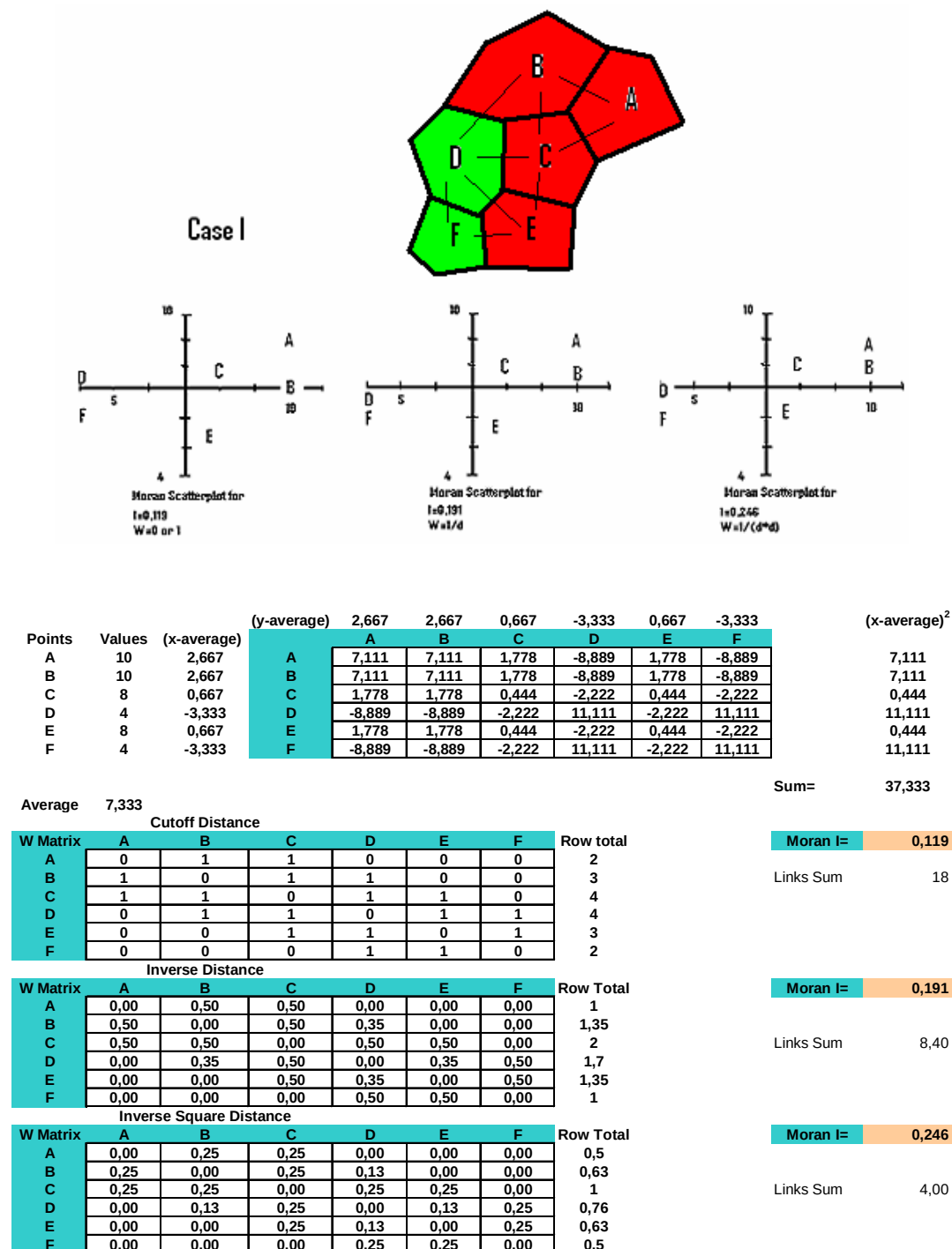
IBM[®]

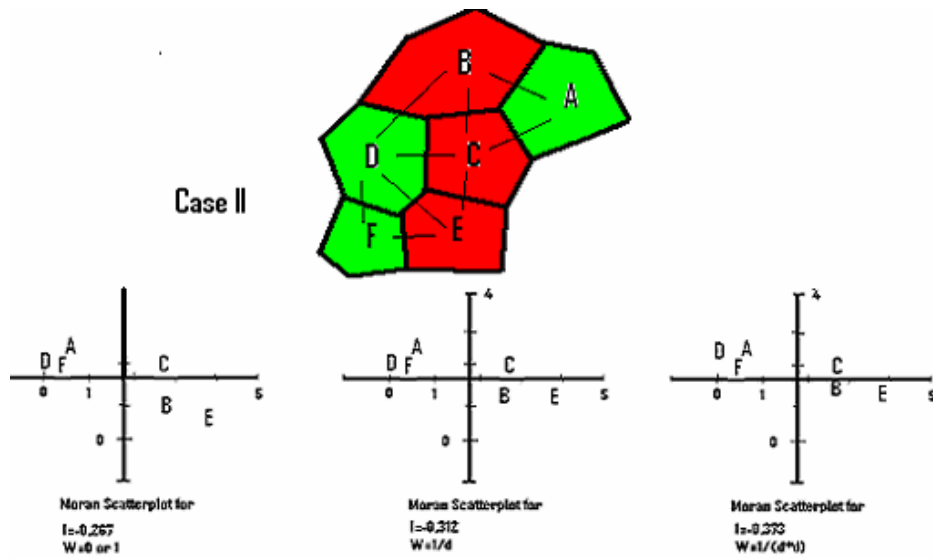
☐

Thanks, Dude!!!!!!

Appendix I

The Moran Results According To The Three Weights Vicinity Matrices





Points	Values	(x-average)	(y-average)	-1,200	0,900	0,700	-1,400	2,100	-1,100	(x-average) ²
A	0,5	-1,200	A	1,440	-1,080	-0,840	1,680	-2,520	1,320	1,440
B	2,6	0,900	B	-1,080	0,810	0,630	-1,260	1,890	-0,990	0,810
C	2,4	0,700	C	-0,840	0,630	0,490	-0,980	1,470	-0,770	0,490
D	0,3	-1,400	D	1,680	-1,260	-0,980	1,960	-2,940	1,540	1,960
E	3,8	2,100	E	-2,520	1,890	1,470	-2,940	4,410	-2,310	4,410
F	0,6	-1,100	F	1,320	-0,990	-0,770	1,540	-2,310	1,210	1,210
										Sum=
										10,320

Average 1,700

Cutoff Distance

W Matrix	A	B	C	D	E	F	Row total
A	0	1	1	0	0	0	2
B	1	0	1	1	0	0	3
C	1	1	0	1	1	0	4
D	0	1	1	0	1	1	4
E	0	0	1	1	0	1	3
F	0	0	0	1	1	0	2

Moran I= -0,373

Links Sum 18

Inverse Distance

W Matrix	A	B	C	D	E	F	Row Total
A	0,00	0,50	0,50	0,00	0,00	0,00	1
B	0,50	0,00	0,50	0,35	0,00	0,00	1,35
C	0,50	0,50	0,00	0,50	0,50	0,00	2
D	0,00	0,35	0,50	0,00	0,35	0,50	1,7
E	0,00	0,00	0,50	0,35	0,00	0,50	1,35
F	0,00	0,00	0,00	0,50	0,50	0,00	1

Moran I= -0,312

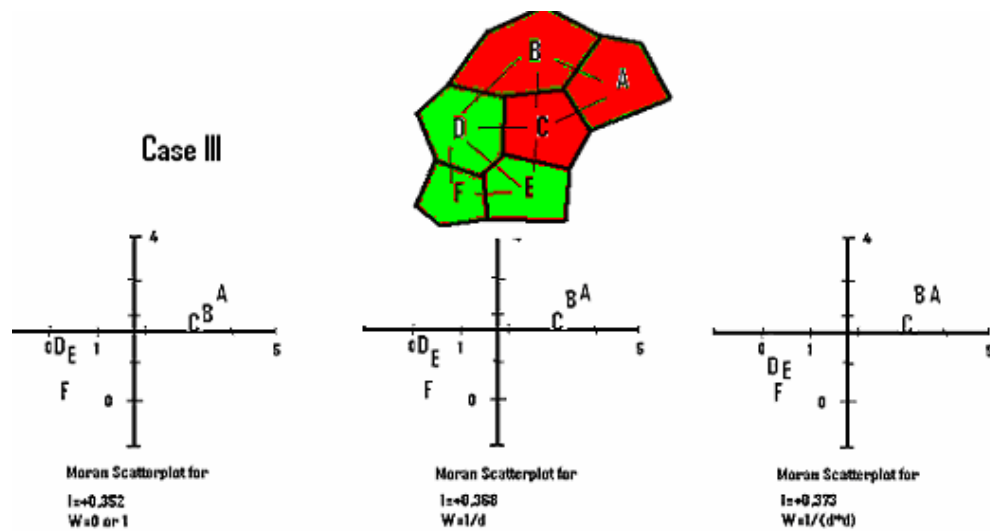
Links Sum 8,40

Inverse Square Distance

W Matrix	A	B	C	D	E	F	Row Total
A	0,00	0,25	0,25	0,00	0,00	0,00	0,5
B	0,25	0,00	0,25	0,13	0,00	0,00	0,63
C	0,25	0,25	0,00	0,25	0,25	0,00	1
D	0,00	0,13	0,25	0,00	0,13	0,25	0,76
E	0,00	0,00	0,25	0,13	0,00	0,25	0,63
F	0,00	0,00	0,00	0,25	0,25	0,00	0,5

Moran I= -0,267

Links Sum 4,00



Points	Values	(x-average)	(y-average)	2,100	0,900	0,700	-1,400	-1,200	-1,100	(x-average) ²
			A	B	C	D	E	F		
A	3,8	2,100	A	4,410	1,890	1,470	-2,940	-2,520	-2,310	4,410
B	2,6	0,900	B	1,890	0,810	0,630	-1,260	-1,080	-0,990	0,810
C	2,4	0,700	C	1,470	0,630	0,490	-0,980	-0,840	-0,770	0,490
D	0,3	-1,400	D	-2,940	-1,260	-0,980	1,960	1,680	1,540	1,960
E	0,5	-1,200	E	-2,520	-1,080	-0,840	1,680	1,440	1,320	1,440
F	0,6	-1,100	F	-2,310	-0,990	-0,770	1,540	1,320	1,210	1,210
										Sum= 10,320

Average 1,700

Cutoff Distance

W Matrix	A	B	C	D	E	F	Row total
A	0	1	1	0	0	0	2
B	1	0	1	1	0	0	3
C	1	1	0	1	1	0	4
D	0	1	1	0	1	1	4
E	0	0	1	1	0	1	3
F	0	0	0	1	1	0	2

Moran I= 0,352

Links Sum 18

W Matrix	A	B	C	D	E	F	Row Total
A	0,00	0,50	0,50	0,00	0,00	0,00	1
B	0,50	0,00	0,50	0,35	0,00	0,00	1,35
C	0,50	0,50	0,00	0,50	0,50	0,00	2
D	0,00	0,35	0,50	0,00	0,35	0,50	1,7
E	0,00	0,00	0,50	0,35	0,00	0,50	1,35
F	0,00	0,00	0,00	0,50	0,50	0,00	1

Moran I= 0,368

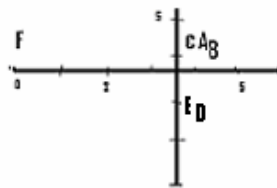
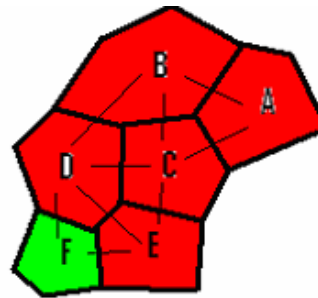
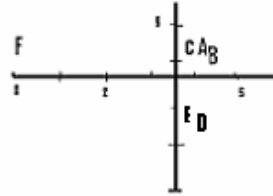
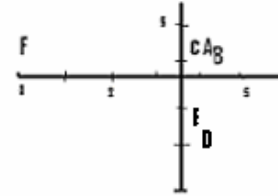
Links Sum 8,40

W Matrix	A	B	C	D	E	F	Row Total
A	0,00	0,25	0,25	0,00	0,00	0,00	0,5
B	0,25	0,00	0,25	0,13	0,00	0,00	0,63
C	0,25	0,25	0,00	0,25	0,25	0,00	1
D	0,00	0,13	0,25	0,00	0,13	0,25	0,76
E	0,00	0,00	0,25	0,13	0,00	0,25	0,63
F	0,00	0,00	0,00	0,25	0,25	0,00	0,5

Moran I= 0,381

Links Sum 4,00

Case IV


Moran Scatterplot for
I=-0,144
W=0 or 1

Moran Scatterplot for
I=-0,174
W=1/d

Moran Scatterplot for
I=-0,196
W=1/(d*d)

Points	Values	(x-average)	(y-average)	0,500	0,600	0,400	0,700	0,900	-3,100	(x-average) ²
A	3,8	0,500	A	0,250	0,300	0,200	0,350	0,450	-1,550	0,250
B	3,9	0,600	B	0,300	0,360	0,240	0,420	0,540	-1,860	0,360
C	3,7	0,400	C	0,200	0,240	0,160	0,280	0,360	-1,240	0,160
D	4	0,700	D	0,350	0,420	0,280	0,490	0,630	-2,170	0,490
E	4,2	0,900	E	0,450	0,540	0,360	0,630	0,810	-2,790	0,810
F	0,2	-3,100	F	-1,550	-1,860	-1,240	-2,170	-2,790	9,610	9,610

Sum= 11,680

Average 3,300

Cutoff Distance

W Matrix	A	B	C	D	E	F	Row total
A	0	1	1	0	0	0	2
B	1	0	1	1	0	0	3
C	1	1	0	1	1	0	4
D	0	1	1	0	1	1	4
E	0	0	1	1	0	1	3
F	0	0	0	1	1	0	2

Moran I= -0,144

Links Sum 18

Inverse Distance

W Matrix	A	B	C	D	E	F	Row Total
A	0,00	0,50	0,50	0,00	0,00	0,00	1
B	0,50	0,00	0,50	0,35	0,00	0,00	1,35
C	0,50	0,50	0,00	0,50	0,50	0,00	2
D	0,00	0,35	0,50	0,00	0,35	0,50	1,7
E	0,00	0,00	0,50	0,35	0,00	0,50	1,35
F	0,00	0,00	0,00	0,50	0,50	0,00	1

Moran I= -0,174

Links Sum 8,40

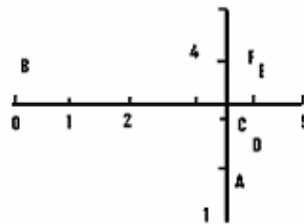
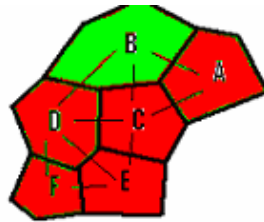
Inverse Square Distance

W Matrix	A	B	C	D	E	F	Row Total
A	0,00	0,25	0,25	0,00	0,00	0,00	0,5
B	0,25	0,00	0,25	0,13	0,00	0,00	0,63
C	0,25	0,25	0,00	0,25	0,25	0,00	1
D	0,00	0,13	0,25	0,00	0,13	0,25	0,76
E	0,00	0,00	0,25	0,13	0,00	0,25	0,63
F	0,00	0,00	0,00	0,25	0,25	0,00	0,5

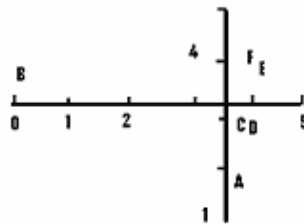
Moran I= -0,196

Links Sum 4,00

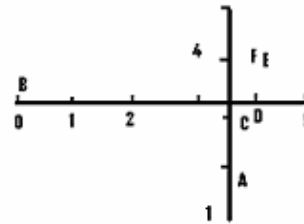
Case V



Moran Scatterplot for
 $I = -0,144$
 $W = 1 \text{ or } 0$



Moran Scatterplot for
 $I = -0,126$
 $W = 1/d$



Moran Scatterplot for
 $I = -0,113$
 $W = 1/(d*d)$

Points	Values	(x-average)	(y-average)	0,500	-3,100	0,400	0,700	0,900	0,600	(x-average) ²
A	3,8	0,500	A	0,250	-1,550	0,200	0,350	0,450	0,300	0,250
B	0,2	-3,100	B	-1,550	9,610	-1,240	-2,170	-2,790	-1,860	9,610
C	3,7	0,400	C	0,200	-1,240	0,160	0,280	0,360	0,240	0,160
D	4	0,700	D	0,350	-2,170	0,280	0,490	0,630	0,420	0,490
E	4,2	0,900	E	0,450	-2,790	0,360	0,630	0,810	0,540	0,810
F	3,9	0,600	F	0,300	-1,860	0,240	0,420	0,540	0,360	0,360
										Sum=
										11,680

Average 3,300

Cutoff Distance

W Matrix	A	B	C	D	E	F	Row total
A	0	1	1	0	0	0	2
B	1	0	1	1	0	0	3
C	1	1	0	1	1	0	4
D	0	1	1	0	1	1	4
E	0	0	1	1	0	1	3
F	0	0	0	1	1	0	2

Moran I= -0,144

Links Sum 18

Inverse Distance

W Matrix	A	B	C	D	E	F	Row Total
A	0,00	0,50	0,50	0,00	0,00	0,00	1
B	0,50	0,00	0,50	0,35	0,00	0,00	1,35
C	0,50	0,50	0,00	0,50	0,50	0,00	2
D	0,00	0,35	0,50	0,00	0,35	0,50	1,7
E	0,00	0,00	0,50	0,35	0,00	0,50	1,35
F	0,00	0,00	0,00	0,50	0,50	0,00	1

Moran I= -0,126

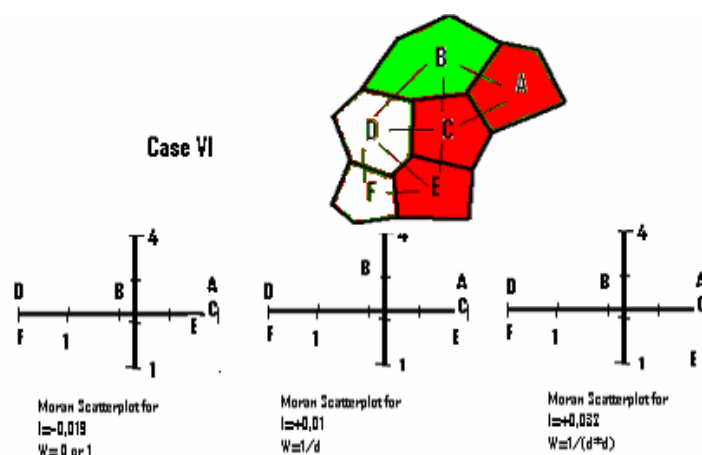
Links Sum 8,40

Inverse Square Distance

W Matrix	A	B	C	D	E	F	Row Total
A	0,00	0,25	0,25	0,00	0,00	0,00	0,5
B	0,25	0,00	0,25	0,13	0,00	0,00	0,63
C	0,25	0,25	0,00	0,25	0,25	0,00	1
D	0,00	0,13	0,25	0,00	0,13	0,25	0,76
E	0,00	0,00	0,25	0,13	0,00	0,25	0,63
F	0,00	0,00	0,00	0,25	0,25	0,00	0,5

Moran I= -0,113

Links Sum 4,00



Points	Values	(x-average)	(y-average)	1,567	-0,333	1,767	-2,233	1,467	-2,233	(x-average) ²
A	3,8	1,567	A	2,454	-0,522	2,768	-3,499	2,298	-3,499	2,454
B	1,9	-0,333	B	-0,522	0,111	-0,589	0,744	-0,489	0,744	0,111
C	4	1,767	C	2,768	-0,589	3,121	-3,946	2,591	-3,946	3,121
D	0	-2,233	D	-3,499	0,744	-3,946	4,988	-3,276	4,988	4,988
E	3,7	1,467	E	2,298	-0,489	2,591	-3,276	2,151	-3,276	2,151
F	0	-2,233	F	-3,499	0,744	-3,946	4,988	-3,276	4,988	4,988
			Sum=							17,813

Average 2,233

Cutoff Distance							Row total
W Matrix	A	B	C	D	E	F	
A	0	1	1	0	0	0	2
B	1	0	1	1	0	0	3
C	1	1	0	1	1	0	4
D	0	1	1	0	1	1	4
E	0	0	1	1	0	1	3
F	0	0	0	1	1	0	2

Moran I= -0,019

Links Sum 18

Inverse Distance							Row Total
W Matrix	A	B	C	D	E	F	
A	0,00	0,50	0,50	0,00	0,00	0,00	1
B	0,50	0,00	0,50	0,35	0,00	0,00	1,35
C	0,50	0,50	0,00	0,50	0,50	0,00	2
D	0,00	0,35	0,50	0,00	0,35	0,50	1,7
E	0,00	0,00	0,50	0,35	0,00	0,50	1,35
F	0,00	0,00	0,00	0,50	0,50	0,00	1

Moran I= 0,010

Links Sum 8,40

Inverse Square Distance							Row Total
W Matrix	A	B	C	D	E	F	
A	0,00	0,25	0,25	0,00	0,00	0,00	0,5
B	0,25	0,00	0,25	0,13	0,00	0,00	0,63
C	0,25	0,25	0,00	0,25	0,25	0,00	1
D	0,00	0,13	0,25	0,00	0,13	0,25	0,76
E	0,00	0,00	0,25	0,13	0,00	0,25	0,63
F	0,00	0,00	0,00	0,25	0,25	0,00	0,5

Moran I= 0,032

Links Sum 4,00

Appendix J

The Average House Prices Dataset Of The San Diego, CA, County (September 1989)

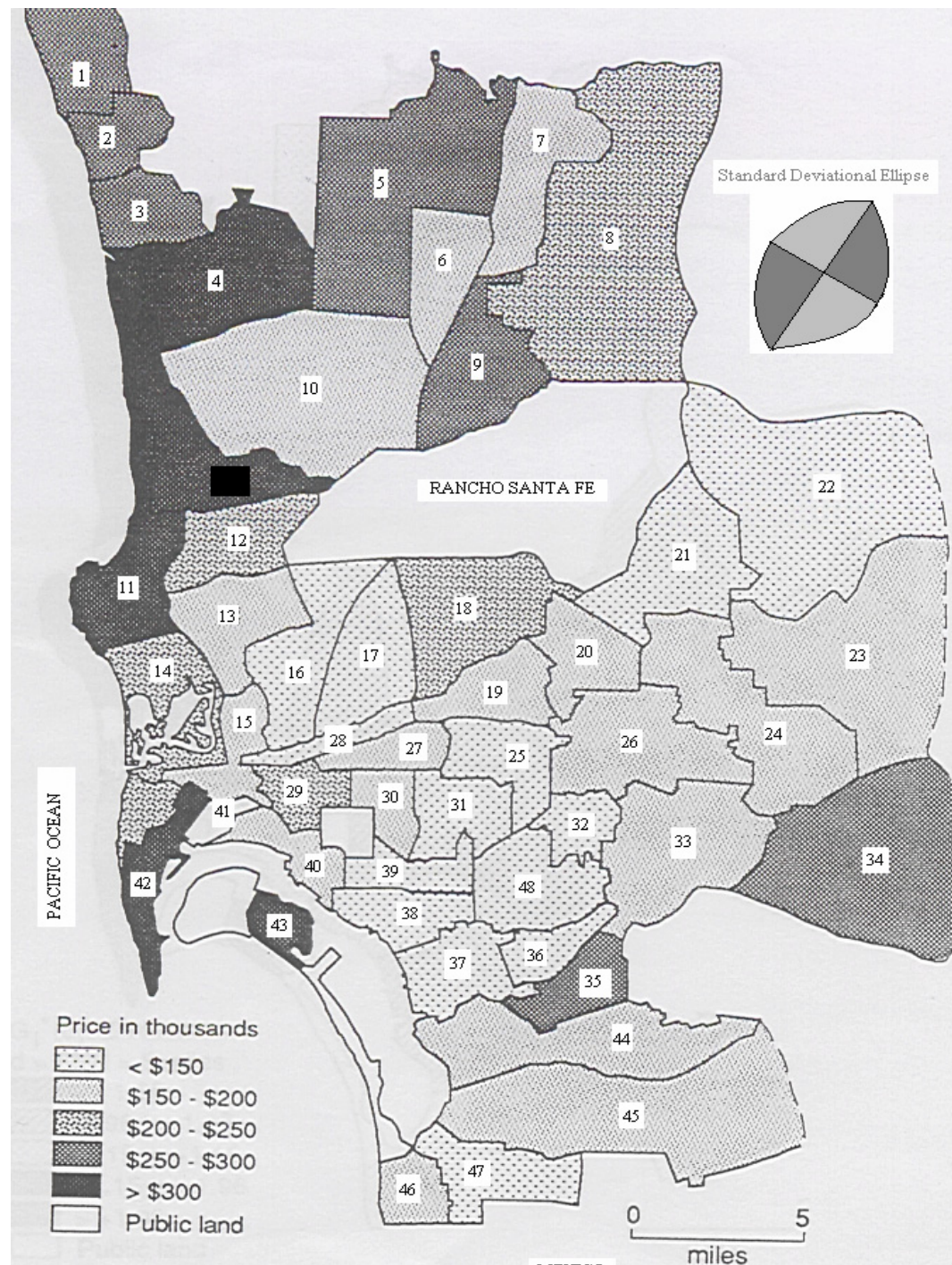
County Number	County Name	Price (in thousands)	County Number	County Name	Price (in thousands)
1	Encinitas	264	25	La Mesa	169
2	Cardiff	260	26	College	138
3	Solana Beach	261	27	Kensington	192
4	Del Mar	309	28	Mission Valley	89
5	Lake Hodges	265	29	Hillcrest	225
6	Rancho Penasquitos	194	30	North Park	152
7	Rancho Bernardo	191	31	East San Diego	111
8	Poway	236	32	Lemon Grove	137
9	Scripps Ranch	270	33	Spring Valley	150
10	Mira Mesa	162	34	Jamul	291
11	La Jolla	398	35	Bonita	297
12	University City	201	36	Paradise Hills	117
13	Clairemont	192	37	National City	99
14	Beaches	249	38	Logan Heights	84
15	Bay Park	152	39	West San Diego	88
16	Kearny Mesa	138	40	Downtown	175
17	Mission Valley	131	41	Ocean Beach	229
18	Tierrasanta	221	42	Point Loma	338
19	Del Cerro	187	43	Coronado	374
20	San Carlos	182	44	Chula Vista East	165
21	Santee	124	45	Chula Vista West	184
22	Lakeside	147	46	Imperial Beach	164
23	El Cajon East	151	47	Otay Mesa	126
24	El Cajon West	150	48	Center San Diego	126

The Counties Comparison Regarding The W Matrices For Patterns And Outliers

County	Value	The W Matrix Choice Over The Moran Scatterplot		Neighborhood Average				
		Pattern	Outlier	1 or 0 (I)	1/d (II)	1/d ² (III)	1/d ³ (IV)	1/d ⁴ (V)
1	264	I or V	II	260	212	242	257	260
2	260	I or V	II	262,5	218	251	261	262
3	261	III	II	284,5	222	254	280	283
4	309	I	II	271,5	209	235	252	257
5	265	I	V	214	204	210	204	198
6	194	I	V	208	210	236	252	259
7	191	II	I	241,25	205	220	228	230
8	236	I	II	230,5	196	201	202	200
9	270	V	I or II	195,75	195	200	202	203
10	162	II	I	287,5	207	222	234	242
11	398	I	II	222,6	191	194	196	198
12	201	II	I	299,66	201	217	237	254
13	192	II	I	227,6	198	205	208	206
14	249	I	II or V	261,8	194	197	196	194
15	152	I	V	179,4	193	198	203	208
16	138	I	II	141	187	178	168	163
17	131	V	II	158,75	184	172	162	154
18	221	V	I	147,25	179	171	170	180
19	187	V	I	160,28	177	170	172	178
20	182	V	IV or V	170,2	178	169	167	179
21	124	V	I	194,25	187	172	164	158
22	147	V or IV	II	137,5	184	164	145	143
23	151	V	II	178,25	179	163	153	150
24	150	V	I or II	177	177	166	159	154
25	169	V	IV	176	178	162	157	163
26	138	I or V	I or V	153	183	161	155	153
27	192	I	V	173	172	151	138	131
28	89	V	II or IV	191,55	178	179	178	177
29	225	II	V	167,4	178	165	154	147
30	152	I	V	154	174	149	136	129
31	111	I	II	139,2	174	152	144	142
32	137	V	II	145,75	171	155	144	138
33	150	V	I	183,85	178	172	270	168
34	291	II	I	150,33	174	163	157	154
35	297	II	V	132,75	161	135	124	120
36	117	I	V	168	186	219	260	281
37	99	V	II	160,8	172	151	137	130
38	84	V	II	122	174	150	129	114
39	88	V	II	129,6	171	148	132	124
40	175	I	V	190	191	204	227	248
41	229	III	V	204,66	203	233	269	294
42	338	V	V	217,66	193	208	222	228

43	374	II	III		178	175	176	177
44	165	II or III	III	193	176	174	180	185
45	184	II	I	145,5	173	159	152	149
46	164	III	I	160	180	168	156	146
47	126	I	II or V	174	181	176	178	181
48	126	V	II	115,5	168	149	139	134

The San Diego Map Including The Standard Deviational Ellipse Layout



The Neighborhood Counties According To The Standard Deviational Ellipse

County	W Matrix For The Standard Deviational Ellipse		County	W Matrix For The Standard Deviational Ellipse	
	Neighbors Number	Vicinity Mean		Neighbors Number	Vicinity Mean
1	2	260	25	20,32,33	156,3
2	1,3	262,5	26	19,20,27,31,32,48	154,3
3	2,4	284,5	27	17,18,19,26,28,30,31,36	143,2
4	3,10	211,5	28	15,16,17,27,29,30,40	166,4
5	6	194	29	15,16,28,30,41,43	189
6	5,7,9	242	30	18,27,28,29,31,38,39,40	148,1
7	6,8	215	31	19,26,27,30,38,39,40,48	142,7
8	7	191	32	25,33,36	145,33
9	6	194	33	24,25,26,32	148,5
10	4	309	34		
11	12,13,14	214	35	36,44	141
12	11,13	295	36	27,32,35,37,44,48	169,3
13	11,12,14,15,16	227,6	37	36,38,39,48	148,7
14	15	152	38	30,31,37,39,40	125
15	13,14,16,18,28,29,41	191,8	39	30,31,37,38	111,5
16	13,15,17,18,28,29,41	177	40	18,28,29,30,31,41,43	200,1
17	16,18,27,28,	160	41	15,16,29,40,42	205,6
18	15,16,17,19,27,30,40	182,7	42	41	229
19	18,20,26,27,31	168,8	43	29,40	200
20	19,21,25,26	154,5	44	35,36,45	199,3
21	20,22	164,5	45	44,47	145,5
22	21	124	46	47	126
23	24	150	47	45, 46	174
24	23,33	150,5	48	26,31,36,37,47	118,2

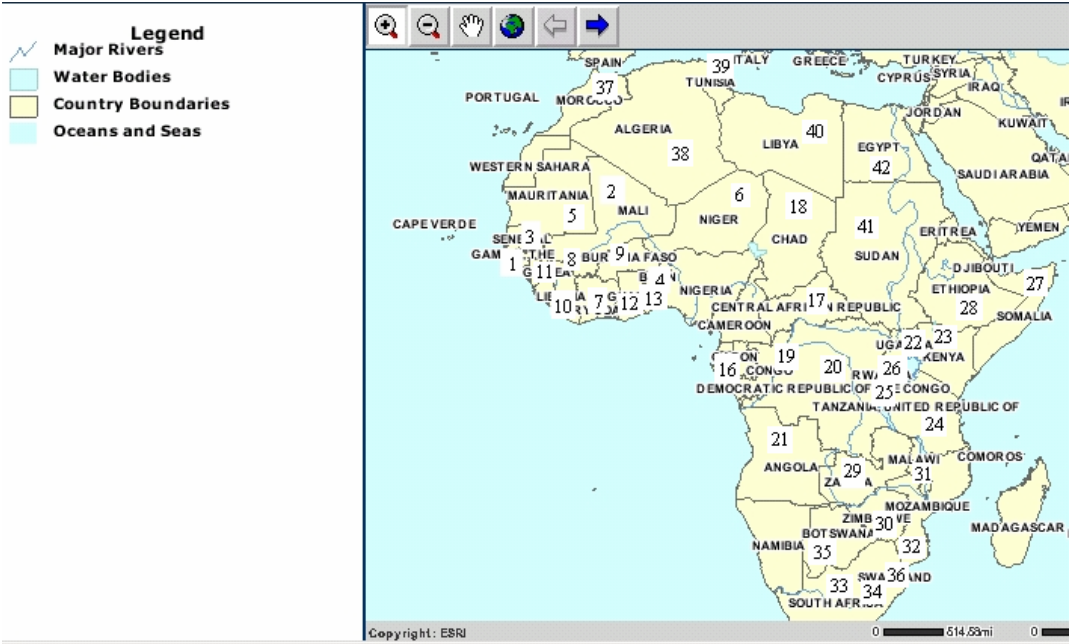
Appendix K

The First Fifteen Neighborhood Weights Counties By The Exponential Covariogram (Range=5.91)

Central District	Neighbor District	Euclidean Distance	Covariogram Value	Assigned Weight	Vicinity Mean
1 - Encinitas	2	3.16	3804.82	0,59	260.41
	3	5.39	2612.08	0,41	
2 - Cardiff	1	3.16	3804.82	0,34	274.15
	3	2.24	4450.37	0,40	
	4	5.00	2787.98	0,25	
3 - Solana Beach	2	2.24	4450.37	0,53	283.27
	4	2.83	4025.94	0,47	
4 - Del Mar	2	5.00	2787.98	0,23	239.03
	3	2.83	4025.94	0,33	
	5	5.39	2612.08	0,21	
	10	5.00	2787.98	0,23	
5 - Lake Hodges	6	2.83	4025.94	0,44	213.14
	7	5.10	2741.66	0,30	
	9	5.83	2422.30	0,26	
6 - Rancho Penasquitos	5	2.83	4025.94	0,25	231.02
	7	4.24	3169.16	0,19	
	8	5.00	2787.98	0,17	
	9	3.16	3804.82	0,23	
	10	5.66	2494.71	0,15	
7 - Rancho Bernardo	5	5.10	2741.66	0,29	230.32
	6	4.24	3169.16	0,34	
	8	3.61	3529.88	0,37	
8 - Poway	6	5.00	2787.98	0,31	216.11
	7	3.61	3529.88	0,39	
	9	5.00	2787.98	0,31	
9 - Scripps Ranch	5	5.83	2422.30	0,21	211.13
	6	3.16	3804.82	0,32	
	8	5.00	2787.98	0,24	

	10	5.10	2741.66	0,23	
10 - Mira Mesa	4	5.00	2787.98	0,26	245.45
	6	5.66	2494.71	0,23	
	9	5.10	2741.66	0,26	
	12	5.39	2612.08	0,25	
11 - La Jolla	12	3.16	3804.82	0,29	201.37
	13	3.61	3529.88	0,27	
	14	4.12	3233.91	0,25	
	15	5.83	2422.30	0,19	
12 - University City	11	3.16	3804.82	0,24	230.81
	13	3.00	3910.74	0,25	
	14	5.39	2612.08	0,16	
	16	4.47	3048.45	0,19	
	17	5.66	2494.71	0,16	
13 - Clairemont	11	3.61	3529.88	0,14	198.29
	12	3.00	3910.74	0,16	
	14	2.83	4025.94	0,16	
	15	5.00	2787.98	0,11	
	16	2.24	4450.37	0,18	
	17	4.12	3233.91	0,13	
	28	5.00	2787.98	0,11	
14 - Beaches	11	4.12	3233.91	0,13	204.28
	12	5.39	2612.08	0,10	
	13	2.83	4025.94	0,16	
	15	3.61	3529.88	0,14	
	16	4.12	3233.91	0,13	
	28	5.39	2612.08	0,10	
	29	5.66	2494.71	0,10	
	41	4.12	3233.91	0,13	
15 - Bay Park	11	5.83	2422.30	0,06	217.96
	13	5.00	2787.98	0,07	
	14	3.61	3529.88	0,09	
	16	4.47	3048.45	0,07	
	17	5.66	2494.71	0,06	
	27	5.10	2741.66	0,07	
	28	3.16	3804.82	0,09	
	29	2.24	4450.37	0,11	
	30	5.10	2741.66	0,07	
	40	3.61	3529.88	0,09	
	41	3.16	3804.82	0,09	
	42	4.24	3169.16	0,08	
	43	5.10	2741.66	0,07	

Appendix L

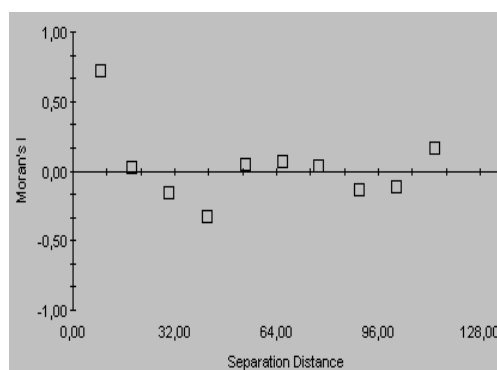


The Africa Dataset

Number	X Coord	Y Coord	Country	Area	Population	GDP Per Person	Pop. Density
1	5	125	Gambia	11295	750000	1453	66
2	37	130	Mali	1240000	7800000	681	6
3	10	125	Senegal	196192	6700000	1307	34
4	45	115	Benin	112622	4000000	867	36
5	20	140	Mauritania	1030700	1800000	1563	2
6	60	130	Niger	1267000	6400000	739	5
7	28	110	Côte D' Ivory	322462	9500000	1598	29
8	15	117	Guinea	245857	5700000	1782	23
9	35	120	Burkinia Faso	274200	6700000	870	24
10	18	107	Liberia	111369	2200000	910	20
11	12	112	Serra Leone	71740	3700000	458	52
12	35	110	Ghana	238537	14000000	1735	59
13	42	110	Togo	56785	3000000	1234	53
14	70	103	Cameroon	470200	9700000	810	21
15	60	110	Nigeria	923758	92000000	795	100

16	65	90	Gabon	227667	1200000	890	5
17	90	105	Centro-African Rep.	622984	2600000	1118	4
18	85	128	Chad	1284000	5000000	856	4
19	75	90	Congo	342000	1800000	995	5
20	95	85	Zaire	2345409	32000000	822	14
21	80	60	Angola	1246700	9300000	1821	7
22	115	95	Uganda	241139	16000000	1074	66
23	130	90	Kenya	582646	20000000	980	34
24	123	78	Tanzania	939828	21500000	480	23
25	110	83	Burundi	27834	5000000	570	180
26	110	82	Rwanda	26338	6000000	660	228
27	150	105	Somalia	637657	6000000	460	9
28	131	110	Ethiopia	1221900	42000000	574	34
29	115	55	Zambia	752614	7000000	719	9
30	110	45	Zimbabwe	390580	8200000	2669	21
31	118	62	Malawi	118484	7000000	523	59
32	118	45	Mozambique	784961	14000000	782	18
33	100	25	South Africa	1125800	32000000	7380	28
34	105	20	Lesoto	30355	1500000	1266	49
35	95	40	Botswana	600372	1100000	6103	2
36	113	28	Swaziland	17363	650000	3816	37
37	35	165	Marroco	458730	24000000	3305	52
38	50	160	Algeria	2381741	21000000	4792	9
39	67	170	Tunisia	163610	7200000	5704	44
40	80	155	Libya	1759540	3700000	6697	2
41	110	120	Sudao	2505813	22000000	467	9
42	110	150	Egypt	1001449	48000000	3401	48

The Africa Map Posting For The Moran I Correlogram (Population Density Dataset)



Appendix M

The Pb Dataset From Gamma[®] Design Software

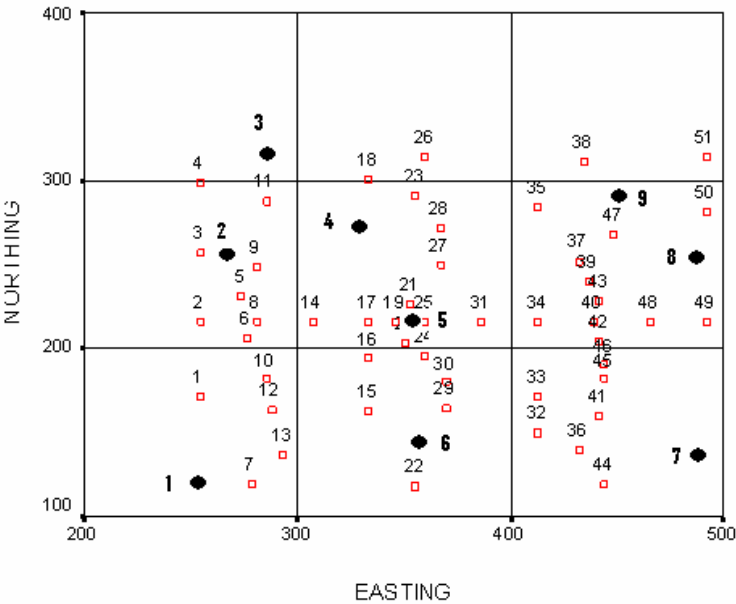
Sample	East	North	Pb	Sample	East	North	Pb	Sample	East	North	Pb
1	4,50	11,90	0,42	44	27,70	48,90	0,37	87	51,90	46,60	0,31
2	2,70	29,40	0,60	45	28,00	53,10	0,24	88	50,10	56,20	0,32
3	1,60	32,60	0,60	46	27,00	62,90	0,44	89	52,80	63,50	0,94
4	4,10	44,50	0,43	47	26,20	73,30	0,54	90	56,50	7,60	0,17
5	0,60	64,00	0,51	48	32,60	3,20	0,21	91	57,70	18,50	0,18
6	2,40	71,80	0,34	49	30,30	10,70	0,17	92	58,00	29,90	0,25
7	7,80	3,50	0,37	50	33,60	23,40	0,15	93	56,10	35,50	0,30
8	6,70	10,20	0,61	51	31,10	34,30	0,27	94	56,10	48,80	0,20
9	6,70	16,30	0,46	52	33,20	37,80	0,16	95	56,60	56,20	0,71
10	6,50	27,20	0,44	53	34,10	52,60	0,15	96	59,70	60,90	0,32
11	5,70	42,70	0,38	54	32,10	58,80	0,13	97	59,30	67,00	0,17
12	9,70	51,10	0,47	55	33,40	67,40	0,42	98	64,20	13,70	0,15
13	9,40	63,30	0,44	56	39,10	11,90	0,23	99	63,00	16,70	0,19
14	6,80	65,60	0,37	57	35,30	18,20	0,14	100	61,20	29,20	0,18
15	9,80	77,40	0,61	58	35,40	29,90	0,15	101	61,60	31,50	0,33
16	10,20	4,60	0,56	59	38,20	42,50	0,18	102	60,20	35,00	0,33
17	12,20	10,80	0,44	60	39,60	48,80	0,34	103	63,10	53,70	0,61
18	13,90	15,80	0,50	61	37,00	52,40	0,47	104	62,70	57,30	0,72
19	10,80	26,00	0,50	62	39,10	56,40	0,39	105	60,40	65,50	0,65
20	11,60	32,70	0,91	63	39,00	71,70	0,29	106	62,00	71,60	1,14
21	13,20	35,80	0,55	64	36,00	77,60	0,23	107	68,50	13,30	0,49
22	14,50	63,60	0,31	65	43,50	1,60	0,15	108	68,00	17,30	0,40
23	14,60	68,80	0,30	66	43,40	15,40	0,23	109	69,80	28,70	0,52
24	14,40	71,00	0,24	67	44,10	23,90	0,41	110	66,20	47,00	0,79
25	17,10	10,00	0,22	68	44,40	26,80	0,31	111	68,80	53,50	0,56
26	19,40	10,30	0,13	69	43,10	42,30	0,38	112	65,00	70,00	1,25
27	16,60	15,10	0,18	70	41,80	47,00	0,27	113	66,60	79,00	0,52
28	15,90	23,80	0,41	71	42,40	72,90	0,30	114	71,30	9,40	0,30
29	19,10	38,40	0,24	72	47,00	1,50	0,22	115	70,70	19,10	0,26
30	15,70	44,50	0,42	73	47,10	10,90	0,23	116	72,30	29,60	0,59
31	16,30	62,00	0,61	74	47,10	31,00	0,25	117	70,50	35,30	0,56
32	16,90	66,90	0,52	75	47,40	39,20	0,20	118	71,30	48,70	0,41
33	21,60	6,60	0,23	76	49,80	40,90	0,40	119	70,30	52,70	0,33
34	21,80	18,90	0,39	77	49,20	45,20	0,49	120	72,50	73,10	0,58
35	22,70	37,40	0,21	78	46,80	56,70	0,29	121	79,10	7,30	0,51
36	24,60	46,00	0,34	79	49,30	66,00	0,49	122	79,00	18,70	0,44

37	24,30	76,60	0,16	80	46,80	73,40	0,52	123	75,80	29,60	0,51
38	28,90	2,10	0,06	81	54,80	8,80	0,24	124	79,30	39,60	0,51
39	29,70	8,20	0,12	82	55,00	12,50	0,19	125	76,60	47,90	0,56
40	29,40	14,80	0,07	83	54,50	17,00	0,35	126	77,40	53,00	0,54
41	28,00	21,80	0,16	84	50,70	26,40	0,18	127	76,90	74,50	0,85
42	29,30	27,20	0,23	85	51,80	38,00	0,20	128	78,30	78,10	0,60
43	27,70	44,30	0,25	86	54,00	40,30	0,32				

The Lead Dataset From GeoEAS[®] Software

Easting	Northing	Lead	Easting	Northing	Lead	Easting	Northing	Lead
254.4	128	4.5	333.6	271	36.25	412.8	285	32.25
254.4	172	19.75	333.6	301	25.5	432	140	12.25
254.4	216	146.7	345.6	210	59	432	252	16.25
254.4	257	16.5	346	216	40.75	434.4	312	24
254.4	299	14.5	350.4	203	25.75	436.8	240	12.25
273.6	231	151.5	352.8	226	127.7	439.2	216	70
273.6	269	20	355.2	118	9.5	441.6	160	31
276	206	37.5	355.2	291	16.75	441.6	204	56.5
278.4	119	13	357.6	139	4.25	441.6	228	18
280.8	216	268	360	195	33.25	444	119	1
280.8	249	19.25	360	216	94.25	444	182	118
285.6	182	80	360	315	21.25	444	190	56.5
285.6	288	30.25	367.2	250	122.2	448.8	268	3.5
288	164	46	367.2	272	55	451.2	295	9.5
288	311	18.25	369.6	165	4	465.6	216	25
292.8	137	10	369.6	180	21.5	492	150	42.5
307.2	216	98	386.4	216	68	492	216	33
333.6	163	36	412.8	150	49.75	492	249	23.5
333.6	194	37.5	412.8	172	48.5	492	282	14.25
333.6	216	44	412.8	216	60.75	492	315	302.5

The 51 Lead Map Posting Used For Kriging In Conjunction With The 9 Validation Ones

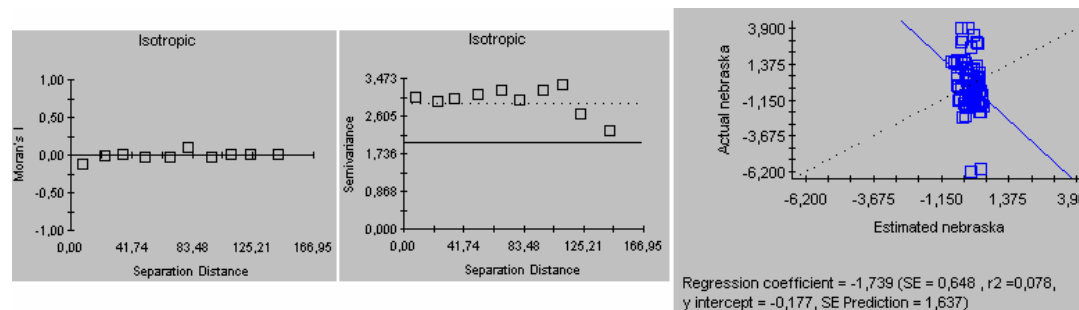


Appendix N

The Soil Organic Matter Of North Platte, Nebraska, USA (Clark and Harper, 2000)

East	North	Soil Org Matter (g/kg)	East	North	Soil Org Matter (g/kg)	East	North	Soil Org Matter (g/kg)
3	180	13.8	39	84	12.3	84	180	17.6
3	156	12.8	39	60	12.8	84	156	16.8
3	132	9.7	39	36	13.8	84	132	14.4
3	108	10	39	12	11.2	84	108	14.2
3	84	10.5	48	180	12.3	84	84	15.2
3	60	14.8	48	156	11.9	84	60	17.6
3	36	10.1	48	132	12.7	84	36	15.2
3	12	9.5	48	108	13.2	84	12	13.6
12	180	12.8	48	84	10.5	93	180	17
12	156	14.7	48	60	12.3	93	156	16.8
12	132	13.1	48	36	14.4	93	132	18.8
12	108	12	48	12	10.5	93	108	20.1
12	84	11.5	57	180	13.1	93	84	18
12	60	11.2	57	156	12	93	60	14.4
12	36	10.7	57	132	12	93	36	12.7
12	12	10.5	57	108	14.6	93	12	11.5
21	180	11.2	57	84	12.8	102	180	14.4
21	156	11.2	57	60	14.8	102	156	14.8
21	132	12.3	57	36	15.2	102	132	16.4
21	108	13.1	57	12	12	102	108	17.6
21	84	12.8	66	180	12.4	102	84	20.1
21	60	12.7	66	156	14.8	102	60	15.6
21	36	12	66	132	14.6	102	36	12.6
21	12	9.5	66	108	15.4	102	12	11.2
30	180	10.7	66	84	12.8	111	180	16.8
30	156	10.7	66	60	15.6	111	156	17.6
30	132	10.6	66	36	17.6	111	132	11.6
30	108	12	66	12	12	111	108	16.8
30	84	15	75	180	12.8	111	84	18.4
30	60	13	75	156	13.2	111	60	15.4
30	36	15.2	75	132	15.8	111	36	14.8
30	12	10	75	108	13.6	111	12	11.1
39	180	13.4	75	84	14.4	111	60	15.4
39	156	12.4	75	60	14.6	111	36	14.8
39	132	12.7	75	36	14.6	111	12	11.1
39	108	6.5	75	12	11.6			

The Moran I Correlogram, Variogram And Cross-Validation Based On The Previous Data

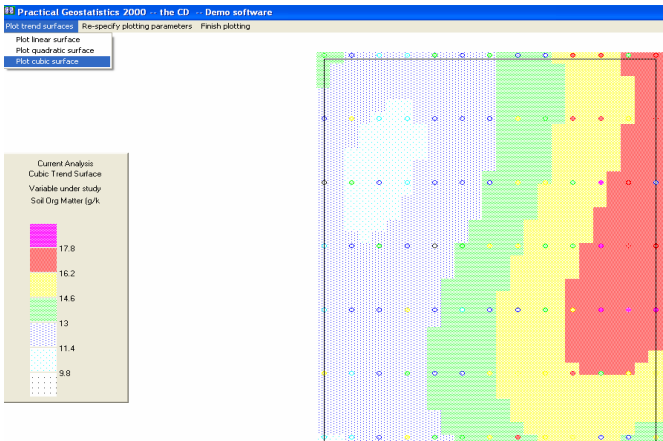


The Anova And Cubic Trend Surface

Analysis of Variance table				
Please note that Analysis of Variance is only technically valid for Normally distributed residuals				
Mean and standard deviation of values: 13.5000 2.4994				
	Sum of Squares	Deg of freedom	Mean square	F ratio
Linear Trend	256.6345	2	128.3173	33.51
Linear Residuals	386.7854	101	3.8296	
Quadratic Trend	300.1534	5	60.0307	17.14
Diff (Linear and Quadratic)	43.5189	3	14.5063	4.14
Quadratic Residuals	343.2665	98	3.5027	
Cubic Trend	384.9889	9	42.7765	15.56
Diff (Quadratic and Cubic)	84.8355	4	21.2089	7.71
Cubic Residuals	258.431	94	2.7493	
Total variation	643.4199	103		

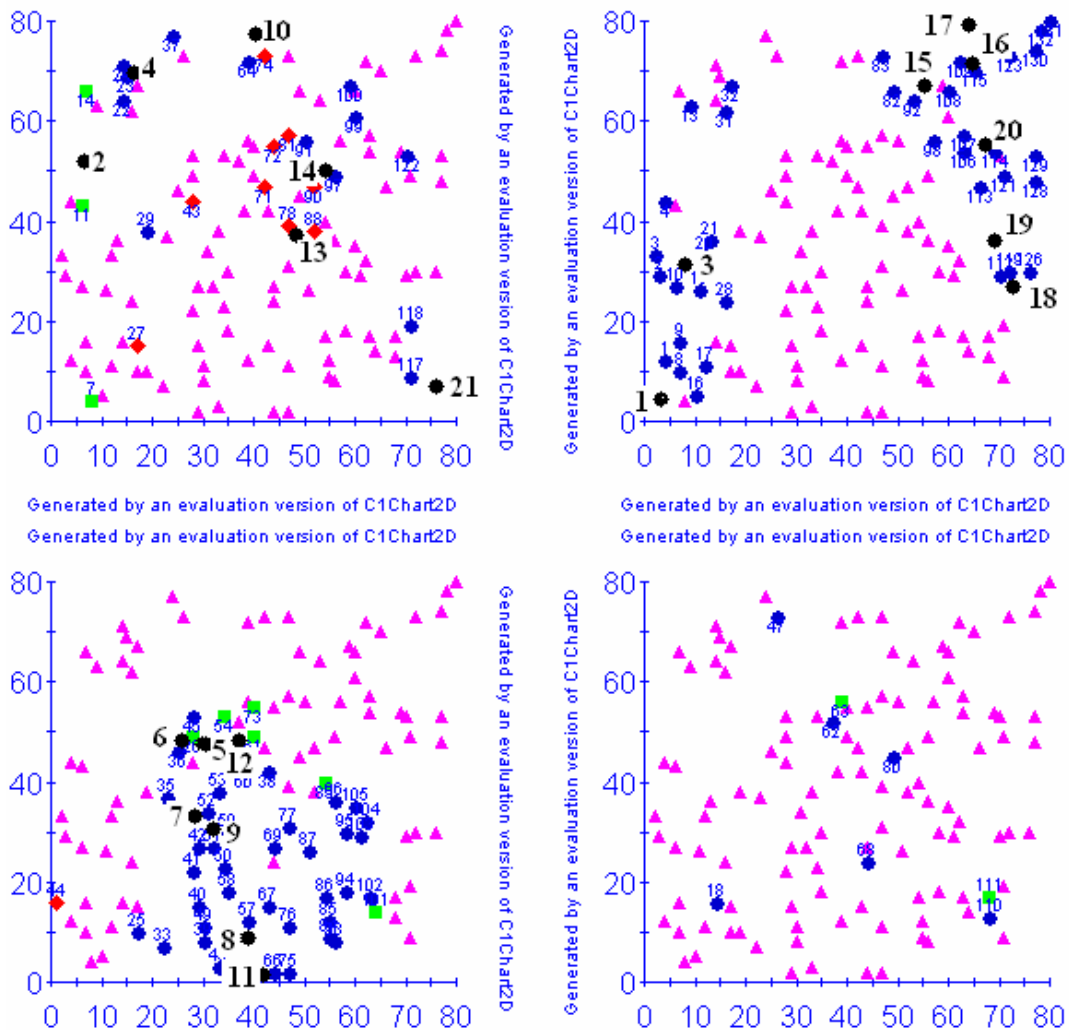
Percentage of Total Sum Of Squares:

Linear Component	39.89
Quadratic Compon...	46.65
Cubic Component	59.83

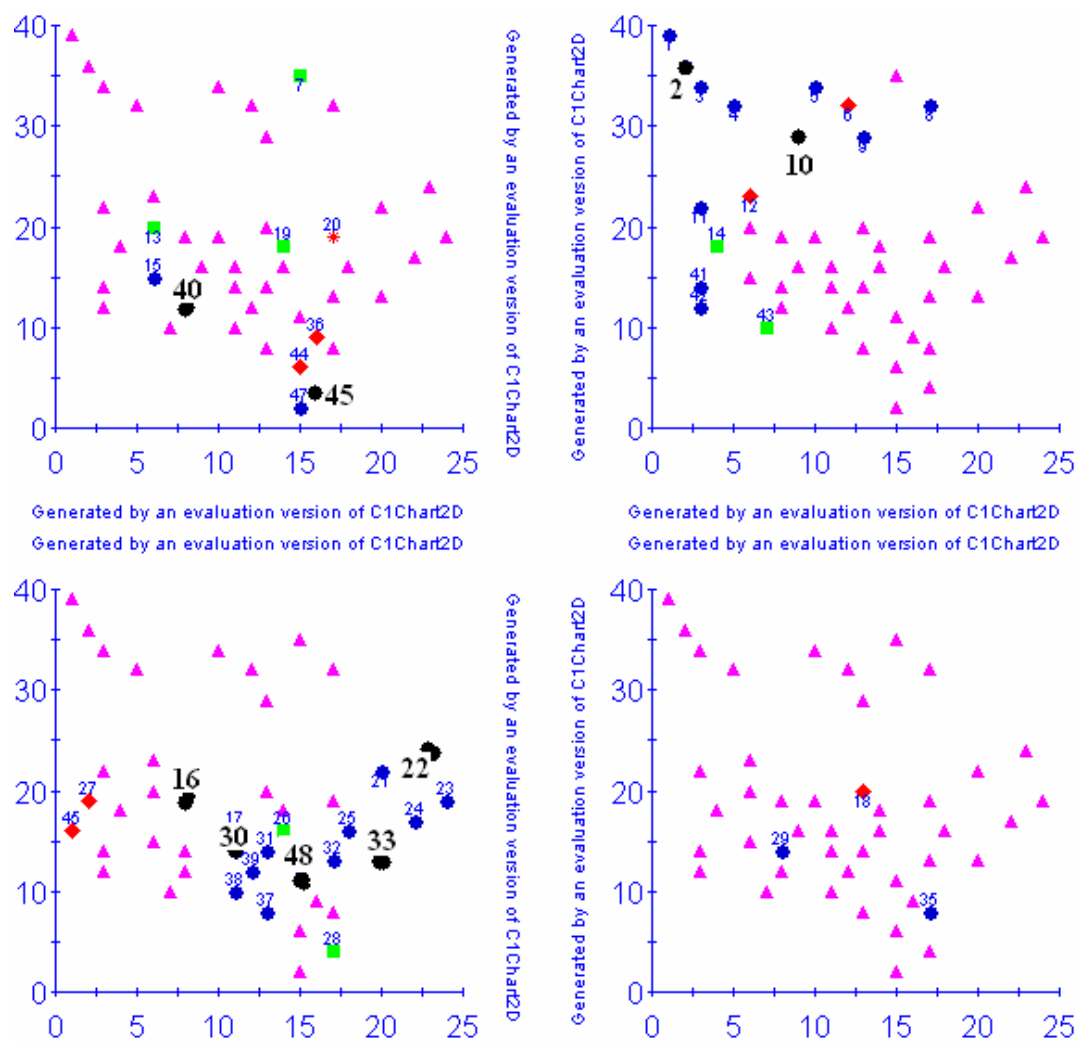


Appendix O

The Locations Of The 21 Validation Samples Of GS+ Pb Dataset



The Locations Of The 12 Validation Samples Of San Diego, CA, Housing Costs Dataset



Acronyms and Mathematical Symbols

.COM – Component Object Model	KI – Knowledge Information
.PCF – Pair Comparison File	LAN – Local Area Network
ABS() – Absolute Value	LIK - Log Likelihood
AI – Artificial Intelligence	LIS – Land Information System
AIC - Akaike Information Criterion	LK – Lognormal Kriging
AM/FM – Automated Mapping/Facilities Management	M-Learning – Mobile Learning
AMC – Arithmetic Mean Center	MAGI - Maryland Automated GIS
AML – Advanced Macro Language	MAUP – Modifiable Areal Unit Problem
ANSI – American National Standard Institute	MIS – Management Information Systems
AR – AutoRegressive Model	MLE – Maximum Likelihood Estimation
ASCII – American Standard Code for Information Exchange	MSE – Mean Square Error
ASP – Active Server Pages®	MSS – Mean Square Sum
BK – Block Kriging	NAA – Node-Arc-Area
CAD/CAM – Computer Assisted Design/Computer Aided Mapping	NTF – National Transfer Format
CAR – Conditional AutoRegressive Model	NYC – New York City
CASE – Computer Aided Software Engineering	OK – Ordinary Kriging
CBD – Central Business District	OLS – Ordinary Least Squares
CCDF – Conditional Cumulative Distribution Function	OO – Object Oriented
CGI – Common Gateway Interface	OODB - Object Oriented DataBase
CIS – Cadastral Information Systems	OOP - Object Oriented Programming
CK - CoKriging	PC – Personal Computer
CLS – Conditional Least Squares	PDA – Personal Digital Assistant
COGO – Coordinate Geometry	PK – Probability Kriging
CORBA – Common Object Request Broker Architecture	pNEM – EPA’s National Ambient Air Quality Standards Exposure Model
COTS – Customer-Of-The-Shelf	PROLOG – Programming in Logic
CRM – Customer Relationship Management	RCTS – Rede Ciência Tecnologia e da Sociedade
CV – Coefficient of Variation	RDBMS – Relation DataBase Management System
DBMS – DataBase Management System	RISC – Reduced Instruction Set Computer
DCL – Doubly Connected List.	RK – Rank Uniform Kriging
DEM – Digital Elevation Model	RML – Restricted Maximum Likelihood
DF – Degrees of Freedom	ROI – Return Of Investment

DK – Disjunctive Kriging	RPC – Remote Procedure Call
DLG – Digital Line Graphic	RS – Remote Sensing
DSS – Decision Support System	RSS – Relative Sum of Squares
DTF – Document Type Definition	RSSE – Relative Sum of Squares Errors
DTM – Digital Terrain Model	SAM - Spatial Analysis Module
DXF – Digital Exchange Format	SAP – Software Application Software
EDI – Electronic Data Interchanged	SAR – Spatial AutoRegressive Model
EDP – Electricidade De Portugal	SC - Schwartz Criterion
EER – Extended Entity Relationship	SDSS – Spatial Decision Support System
ESDA – Exploratory Spatial Data Analysis	SGML – Standard Graphic Markup Language
EGM – Estimated Global Mean	SIS – Soil Information System
EGV – Estimated Global Variance	SK – Simple Kriging
GAL – Geographical Algorithms Library	SQL – Structured Query Language
GAM – Geographical Analysis Machine	SVG – Scalable Vector Graphics
GDF – Geographical Data File	SWOT – Strengths, Weaknesses, Opportunities and Threats
GDSS – Group Decision Support System	TCP/IP – Transmission Control Protocol/Internet Protocol
GIDS - Gradient Inverse Distance Square	TGS – Tourism Geographical System
GIGO – Garbage In Garbage Out	TIGER – Topologically Integrated Geographical Encoding and Referencing
GIS – Geographical Information Systems	TIN – Triangulated Irregular Network
GLS – Generalized Least Squares	UK – Universal Kriging
GML – Geographical Markup Language	UNL – Universidade Nova de Lisboa
GNIS – Gas Network Information Systems	USGS – United States Geological Survey
GPS – Global Position System	UTL – Universidade Técnica de Lisboa
HTML – HyperText Markup Language	UTM – Universal Transverse Mercator
I/O – Input/Output	VRML – Virtual Reality Markup Language
I² – Integration Implementation	W3 – World Wide Web
IC – Information Criteria	W3C – WWW Consortium
IDW – Inverse Distance Weight	WAN – Wide Area Network
IE[®] – Internet Explorer [®]	WIMP – Windows Icons Mouse Pointer interface
IIS[®] – Internet Information Server [®]	WLS – Weight Least Squares
ISEGI – Instituto Superior de Estatística e Gestão de Informação	WWW – World Wide Web
IST - Instituto Superior Técnico	WYSIWYG – What You See Is What You Get
IT – Information Technology	XML – Extensible Markup Language
KEE – Knowledge Engineering Environment, Intillcorp	XQL – Extended Query Language

$\gamma_n(v)$ - Bias correction factor based on natural logarithms variance and n samples	G_{pi} – Goodness of prediction index
χ^2_{lower} - Lowest χ^2 distribution value for a particular confidence level and degrees of freedom	χ^2_{upper} - Upper χ^2 distribution value for a particular confidence level and degrees of freedom
x'_i - Transposition of coordinate x of sample i^{th} to the mean center	y'_i - Transposition of coordinate y of sample i^{th} to the mean center
E_j - Expected value of j class	O_j – Observed value of j class
$\gamma^*(h)$ - Experimental variogram	σ^* – Expected standard deviation
σ^2_{OK} - OK variance	σ – Standard deviation
$\bar{X}$ - Sampling mean	μ – Distribution mean
ξ - Residuals vector	Y – Dependent observations vector
C_{me} or $W(.)$ – Measurement error of nugget-effect	S – Sampling standard deviation
e - The neperian constant	C_{ms} or $\eta(.)$ – Micro-scale variation of nugget-effect
h – Value of lag distance	Ψ - Lagrange
H_0 – Statistical initial hypothesis	h – Directional lag distance
h_1 – Lag distance of direction 1	ε_i – Error of sample i
I_m - Identity matrix	h_2 - Lag distance of direction 2
L_i - Lisa index	ρ - Spatial autoregressive of autocorrelation coefficient
n – Number of observations or samples	$m_2 - \sum_i x_i^2/n$
$N(h)$ - Indicates the number of pairs available for a specific h lag distance	$C1$ or C_1 - Partial Sill
O_{ij} – Observed number of i region of j group	$\gamma(h)$ – Theoretical variogram
P_n - Predictor number variables	E_{ij} - Expected number of i region and j group
R^2 – Regression fitness index	$\ln$ – Neperian logarithm
$R_{nearest}$ – A nearest neighborhood analysis index	df – Degrees of freedom
$S_0 = \sum_i \sum_j w_{ij}$ - The sum of the spatial weights	nc – Number of classes
W - n -by- n matrix of connectivity	σ^2 – Distribution variance
W_{ij} - Weight assigned between i site and j neighbor sample	β or B – Regression estimators vector
WY – Neighborhood impact vector	Z – Number of groups
x_j - Represents i 's neighbor	C_0 or $\gamma(0)$ – Nugget-effect
$Z(NN)$ – Standard value of negative joins	N - Number of regions
$Z(PP)$ - Standard value of positive joins	$Z(PN)$ - Standard value of positive-negative joins
$z(x)$ – Sample value at x location	θ – Clockwise angle in radians
$z(x+h)$ - Sample value at $x+h$ location	$E()$ – Expected value of a particular parameter
$z^*(x)$ or x^* - Estimated value at x location	K – Constant

Z_G – Standard value of $G_i(d)$	r_{ani} – Anisotropy ratio
α – Exponent	S^2 – Sampling variance
$\rho(\mathbf{h})$ – Correlogram	d_{ij} – Distance between i and j
σ_{ij}^2 – Variance between i^{th} and j^{th}	$C(\mathbf{h})$ or $Cov(\mathbf{h})$ – Spatial covariance for two samples that are $\mathbf{h}$ lag apart
χ^2 – Chi-Square distribution	$C(i,j)$ or $Cov(i,j)$ – Covariance between samples i and j
$\mathbf{X}$ – Independent observations matrix	$\mu(s)$ – Trend value for a particular s site using a coordinate regression system, for instance
Δl – Lag tolerance	θ – Amplitude or tolerance angle

¹ “Everything is related to everything, but closer things tend to be more related”.

² Lag distance, lag interval, principal axis, offset tolerance and anisotropy.

³ Minimum of 15 points.

⁴ If spatial distance is less than the variogram range then the $\mathbf{W}$ matrix depends on the $c(h)=sill-\gamma(h)$ model. Otherwise, its value is zero.

⁵ A non-concern for most GIS users.

⁶ Particularly with spatial autocorrelation and autoregressive models.

⁷ But not undertaken as complete black box software.

⁸ In order to avoid reading hundred of pages of documentation.

⁹ Spatial autocorrelation and data configuration.

¹⁰ Token Ring, FDDI, Ethernet, Wireless LANs, Personal Area Networks, Distributed Queue Dual Bus MAN, phone lines, X.25, Frame Relay, ISDN, ATM and Wavelength Division Multiplexing.

¹¹ Bibliographies, electronic newsgroups, W3 sites and learning software.

¹² From the preserve of the few to the right of the many.

¹³ When compared with the binary or the standard deviational ellipse.

¹⁴ Error measurement plus microstructure.

¹⁵ It is possible to attain it by multiple observations per location.

¹⁶ Model, tolerance angle, lag distance, continuity angle, sill, range and nugget-effect.

¹⁷ This includes tools such as e-mail, chat rooms, news, games, calendars, calculator, credits and Web search.

¹⁸ Particularly if Web access and efficiency of time use are considered.

¹⁹ Parish level.

²⁰ Municipality level.

²¹ County level.

²² Two truly geographical space issues.

²³ A sign of the regular process working.

²⁴ Connectivity or arc-node, area definition or polygon-arc, contiguity or right-left.

²⁵ *Geodata Analysis and Display System.*

²⁶ *Computing practitioners have some times been guilty of viewing MIS as being devoid of people.*

²⁷ *The summary measures used for describing linear distributions are defined with reference to the system of statistical moments like the mean and the standard deviation.*

²⁸ <http://www.ems-i.com/gms/index.html>.

²⁹ *One of the five GIS M's: mapping, measurement, monitoring, modeling and management.*

³⁰ *The absolute one (coordinates) and the relative (topology).*

³¹ *The expression Tobler's Law is not completely accurate (probably it should be called Tobler's empirical postulation) since a law, by definition, is an invariable relationship among variables [Porto Editora, 2001], a geographical situation that may break down in Mother Nature. Nevertheless, this expression will be used in this thesis.*

³² *For instance, nonwhite populations are often clustered in similar districts of a metropolitan area for reason of history, discrimination and limited housing access.*

³³ *For instance, the high concentration of liquor stores in low-income communities encourages crime.*

$$\bar{x} - t_{n-1, 2.5} \frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{n-1, 2.5} \frac{s}{\sqrt{n}}.$$

$$\chi^2_{n-1, 97.5} < \frac{(n-1) s^2}{\sigma^2} < \chi^2_{n-1, 2.5}.$$

³⁶ *Degrees of freedom (df) that equals (num_regions-1)×(num_groups-1), that is, df=(N-1)(Z-1).*

³⁷ *With n_1 and n_2 observations but $n_1 \approx n_2$.*

$$\hat{\sigma}_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \frac{(n_1 - 1)\sigma_1^2 + (n_2 - 1)\sigma_2^2}{n_1 + n_2 - 2}.$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\hat{\sigma}_{\bar{x}_1 - \bar{x}_2}} \text{ with } df = n_1 + n_2 - 2.$$

⁴⁰ *MSR/MSE with $df = (\text{var\_num}, \text{observations\_num} - \text{var\_num})$.*

⁴¹ *β_j/σ_j with $df = \text{observations\_num} - \text{var\_num} - 1$.*

⁴² *It creates a new spatial lag variable within the model.*

⁴³ *The lower the AIC and SC global tests, the better for the overall model.*

⁴⁴ *As a rule of thumb, values of the condition number larger than 20 or 30 are considered to be suspected.*

⁴⁵ *If a rejection of the null hypothesis of a normal error becomes true then the tests for heteroskedasticity and spatial dependence should be interpreted with caution since they are based on the normal assumption.*

⁴⁶ *A redundancy issue.*

⁴⁷ *The values become similar to the mean.*

⁴⁸ *The first Openshaw criteria [1992].*

⁴⁹ *Moran I decreases from +1 to 0 or increases from -1 to 0.*

$$^{50} \text{Cov}(i,j) = \sigma^2 \times \rho^{|i-j|}.$$

⁵¹ With a $\rho=0.26$ correlation factor and $n=10$.

⁵² While neighboring procedures are limited to assessing the closest adjacent values, the variogram overcomes this weakness by inspecting the spatial dependence at all scales.

⁵³ Closer distances, higher variability.

⁵⁴ A neighborhood weighted average.

⁵⁵ The product sum of each observation with its corresponding weight from the i^{th} row of W matrix.

⁵⁶ Factorial of n .

$$^{57} Z_i = \frac{(I - E(I))}{\sigma^2(I)}.$$

⁵⁸ It only takes the particular sites within the study region, a non-free sampling situation.

⁵⁹ The distance at which the curve levels off indicates the scale of the spatial pattern. Otherwise, this cumulative plot will reach an asymptote, a situation where patches are aggregated or randomly distributed.

⁶⁰ The similarity of attributes i_s and j_s are compared with each other directly.

$$^{61} Z_G = (G_i(d) - E(G_i(d))) / \sqrt{\sigma^2 G_i(d)}.$$

⁶² Values below the mean x_j .

⁶³ Values above the mean x_j .

⁶⁴ $y_i = b \times x_i$ yields the same scores as x_i .

⁶⁵ $y_i = b + x_i$ does not yield the same result as x_i .

⁶⁶ Global Geary C correlation coefficient for a specific distance d .

⁶⁷ Global Moran I correlation coefficient for a specific distance d .

⁶⁸ A weighted average of the neighboring values.

⁶⁹ It is based on the principle of selecting values for the parameters that maximize the probability of the observed data.

⁷⁰ The Spatial AR option within SpaceStat[®] (Normal distribution and homoskedasticity might not be satisfied) where the spatial error dependent term is modeled as a global trend (together with a stationary local covariance structure).

⁷¹ Degrees of freedom equal the number of exploratory variables.

⁷² Degrees of freedom equal one.

⁷³ Programmed in SAS[®].

⁷⁴ Mean square error (non-explained variance).

⁷⁵ Mean square sum (total variance).

⁷⁶ Like the Multi Instruction Single Data CM5.

⁷⁷ ρ =spatial autocorrelation and WY =neighborhood_crime.

⁷⁸ The spatially lagged dependent variable, WY , is considered an explanatory variable as a way to assess the degree of spatial dependence (ρ).

⁷⁹ $df=5$.

-
- ⁸⁰ Kelegian-Robinson equals 11.55 with $df=3$.
- ⁸¹ $C_{ij}=8.32+0.65Y_{ij}$ with $AIC=-544.8$ and $SBC=-546.5$.
- ⁸² Notice that the mean is basically the same.
- ⁸³ Not suitable for agricultural production.
- ⁸⁴ LADS gives the probability for a given topology with N units in a randomly selected subset of size E (smaller values increase the likelihood that LADS is below the critical value) that contains a contiguous C block (the larger the contiguous region, the small the LADS value).
- ⁸⁵ Significantly higher in these seven states than in the remaining 43.
- ⁸⁶ Assessed by SAKWeb[®].
- ⁸⁷ If similar locations are clustered then there will be few NP joins. If locations are dispersed then there would be relatively many.
- ⁸⁸ The residual probability of being positive or negative is limited by the total number of positive and negative in the research area, a less rigorous assumption than with the free sampling approach.
- ⁸⁹ Usually, it is zero but if and only if there is no measurement error, C_{me} , since the nugget-effect, C_0 , is also caused by micro-scale variations, C_{ms} , that cannot be resolved by the sample spacing.
- ⁹⁰ In a chess context, it corresponds to horizontal/vertical and diagonal, in both directions, respectively.
- ⁹¹ Omitting true spatial dependence adjacencies.
- ⁹² Including false spatial dependence adjacencies.
- ⁹³ The independent observations case of conventional statistics.
- ⁹⁴ See subsection 7.2.1.6 for further details.
- ⁹⁵ The point is inside a given ellipsoid.
- ⁹⁶ A centrographic measure.
- ⁹⁷ With its long axis representing the maximum dispersion path while the short one represents the minimal direction
- ⁹⁸ Note that boundary length or others secondary vicinity rings are not considered, a restrictive constraint.
- ⁹⁹ The Kiefer-Salmon asymptotic regression can test residuals, a statistic that follows a χ^2 distribution with two degrees of freedom for Normal errors. A low probability indicates a rejection of the null hypothesis of Gaussian error.
- ¹⁰⁰ The pair line linkage of neighboring points is bisected at right angles through the polygon boundary.
- ¹⁰¹ In terms of the mean absolute error.
- ¹⁰² Land cover like buildings, grass or wood.
- ¹⁰³ With the Algarve Portugal dataset, the erosivity linear equation equals $968.2+3.8087 \times \text{elevation}$ with $\rho=0.75$.
- ¹⁰⁴ An inexact global basis model based on spatial coordinates.

-
- ¹⁰⁵ The trend (regional features) and the non-systematic white noise fluctuation (local features).
- ¹⁰⁶ In terms of the equation, the least-square solution equals $\mathbf{B}=(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y})$.
- ¹⁰⁷ A constant dip in a single direction.
- ¹⁰⁸ A bowl or dome shape.
- ¹⁰⁹ A very cost-effective means of data vector capture.
- ¹¹⁰ The conditional bias issue.
- ¹¹¹ Smoothing out of the dips and the humps.
- ¹¹² A good estimation method will produce estimated values whose distribution is similar to the distribution of true values.
- ¹¹³ IDW works better if samples are evenly distributed.
- ¹¹⁴ For most software, the observed values are copied over forcing this technique to be an exact interpolator.
- ¹¹⁵ At first, SAKWeb[®] uses the second highest Moran I range to look for new neighbors. If no neighbors are founded then the overall mean is used to full field these cells.
- ¹¹⁶ $(\sum (x^* - \bar{x})^2 - \sum (x - \bar{x})^2) / (n-2)$
- ¹¹⁷ Regression equation for pyretic sulphur coal, created by Gomez & Hazen (1970).
- ¹¹⁸ It represents a systematic increase, continuation or decrease in the regionalized variable for a particular direction.
- ¹¹⁹ Deterministic large-scale modeling plus small-scale probabilistic variability.
- ¹²⁰ The science of inaccuracy, ambiguity, vagueness, fuzziness, error, imprecision and replication [Frank et al., 1999].
- ¹²¹ If space behavior is regarded as isotropic, direction is an irrelevant factor and notification of h is considered for statistical formulas. Otherwise, vector $\mathbf{h}$ (distance and direction) is contemplated.
- ¹²² It simply means that we are ignorant of the spatial process [Isaaks and Srivastava, 1989].
- ¹²³ The mean and the variance.
- ¹²⁴ The expected difference between values at two locations separated by distance $\mathbf{h}$ is zero. If there were a trend or drift over the area then the expected value would change and the mean difference between any two samples would not necessarily be zero.
- ¹²⁵ The measurements can be viewed as a realization drawn by God of the random function model [Geostatistics FAQ, 1997]. Thus, it is implicit that the stationary probabilistic model is closely related with representativeness of the samples and homogeneity of the true spatial distribution.
- ¹²⁶ Stationary of order two.
- ¹²⁷ A characteristic for any spatial pattern.
- ¹²⁸ This distance is determined by the degree of spatial autocorrelation presented in the spatial data.
- ¹²⁹ Non-ergodic relationships.
- ¹³⁰ isotropics $\subset$ stationaries $\subset$ intrinsic.
- ¹³¹ SAKWeb[®] only supports bounded variograms.

¹³² $C(\mathbf{h}) = E[z(\mathbf{x} + \mathbf{h}) \times z(\mathbf{x})]$.

¹³³ SAKWeb[®] covers this factor severely.

¹³⁴ White noise or random dataset.

¹³⁵ It computes the trend and the residual local estimate at the same time, taking into consideration the restricted neighborhood of Kriging stationary.

¹³⁶ Remark that this parameter does not include the scale factor. Further, this test requires statistical independence among the data within each sample dataset.

¹³⁷ Both univariate and bivariate analysis treat data as a set of numbers and ignore its spatial nature.

¹³⁸ These lead to a study of the global trend, the location of severe values and levels of continuity (features not accessible with uni and bivariate measures).

¹³⁹ As many spatial variables do.

¹⁴⁰ The Anderson-Darling, the Ryan-Joiner, the Shapiro-Wilk or the Kolmogorov-Smirnov statistics are other possibilities.

¹⁴¹ This may be indicative of different generating events.

¹⁴² According to the article, this measure does not follow any particular distribution.

¹⁴³ It works with its equivalent ranks from a Normal distribution. After making predictions on the transformed scale, a back-transformation is needed it. According to ESRI [2002], the goal of NST is to make all random errors for the whole population normally distributed. Therefore, the NST transformation for DK and SK must occur after detrending since the variogram is calculated on residuals after trend correction. Yet, with Box-Cox and log transformations, these are applied on spatial data before the trend or spatial autocorrelation is removed and before, quite often, normal residuals emerge.

¹⁴⁴ These variograms consider the available data at each time while the cross-variograms intersect the datasets from two different months.

¹⁴⁵ $\gamma_{2,2}(h)$ for $2m \times 2m$ and $\gamma_{3,3}(h)$ for $3m \times 3m$.

¹⁴⁶ Oversamples and undersamples for different sub-regions.

¹⁴⁷ This effect can happen for three reasons: natural cyclicity, small local high-low patterns and undesirable noise.

¹⁴⁸ For stationary variables, as $h \rightarrow \infty$, $\gamma = \sigma^2 = C_0(\text{Nugget-effect}) + C_1(\text{Partial Sill}) = \text{Total Sill}$. The percentage ratio between the nugget-effect and total sill is called the relative nugget-effect.

¹⁴⁹ How large the search window should be (the local neighborhood stationarity issue).

¹⁵⁰ The closest data tend to cover the influence of those further away.

¹⁵¹ The variogram as a spatial search pattern process.

¹⁵² A non-zero nugget-effect indicates repeated measurements at the same point with different outcomes, a precision issue.

¹⁵³ The higher the nugget-effect, the higher the Kriging variance.

¹⁵⁴ Qualitative maps, for instance.

$$^{155} \gamma(\mathbf{h}) = C(0) - C(\mathbf{h}) = \sigma^2 - C(\mathbf{h}).$$

$$^{156} \gamma(\mathbf{h}) = \rho(0) - \rho(\mathbf{h}) = 1 - \rho(\mathbf{h}) = 1 - C(\mathbf{h})/C(0).$$

¹⁵⁷ The $\gamma(\mathbf{h})$ is not a resilient estimator.

¹⁵⁸ Knowledge of phenomena genesis can be helpful to detect this constraint.

¹⁵⁹ Mainly used for a stratified phenomenon as Figure 20 (right) shows.

¹⁶⁰ The combination of two spherical models is possible as long as each one stops as it reaches its range of influence [Beckers, 2000].

¹⁶¹ This variogram initially rises more rapidly, leading to fuzzier images (less structured) than Gaussian ones because they are linear near the origin.

¹⁶² This model reaches about 66% of its height at the square-root of its range [Lang, 1997]. Yet, according to Englund and Sparks [1991], this model does not work well, particularly when it has a low nugget-effect and/or closely spaced data, a situation not confirmed by Goovaerts [1997].

¹⁶³ Between 50 to 100 for Burrough and McDonnell [1998] and at least 30 for Journel and Huijbregts [1978].

¹⁶⁴ Additional sampling is always welcome as long as the stationary assumption of the search window matches the reality perspective.

¹⁶⁵ Implemented within EcoSse[®], for instance.

¹⁶⁶ The confidence measure.

¹⁶⁷ According to Soares [2000], the OK variance equals $-\sum w_i C(x_i, x_0) - \psi + C(x_0, x_0)$.

¹⁶⁸ Row k of matrix $\mathbf{B}$ changes to zero if the nugget-effect is zero.

¹⁶⁹ The Kriging variance can be rewritten as $\sigma_{OK}^2 = C(0) - \mathbf{W}^T \times \mathbf{B}$ while $\mathbf{x}^* = \mathbf{B}^T \times \mathbf{A}^{-1} \times \mathbf{x}_i$.

¹⁷⁰ The underlying idea is to work with increments rather than with the variable itself.

¹⁷¹ It incorporates anisotropic distance, direction, spatial autocorrelation and clustering information.

$$^{172} \frac{\sum (x_i^* - x_i)}{n}.$$

$$^{173} \sqrt{\frac{\sum (x_i^* - x_i)^2}{n-1}}.$$

$$^{174} \sqrt{\frac{\sum \sigma_{OK}(x_i)}{n-1}}.$$

¹⁷⁵ Under SAKWeb[®], this equality range equals $\pm 10\%$ of the root-mean square prediction error.

¹⁷⁶ Risk β or health cost.

¹⁷⁷ Risk α or remediation cost.

¹⁷⁸ Additional sampling at those locations might be worth to avoid risk α and β costs.

¹⁷⁹ Unlike OK variance, IK is data dependent and avoids a strong assumption about distribution shape.

¹⁸⁰ Original values are ranked in a $[0..1]$ domain.

¹⁸¹ 1, if $z(x) \leq z_{cutoff}$; 0, otherwise.

¹⁸² According to Goovaerts [1997], threshold values beyond the first and ninth quartile of the sample cdf may be inappropriate.

¹⁸³ In contrast, Kriged surface determines the most likely value at a particular location, an accuracy issue.

¹⁸⁴ Concerning the variogram model choice, according to Englund and Sparks [1991], the OK variance is more sensitive than the OK estimate.

¹⁸⁵ As the variable becomes more erratic, this term increases in magnitude.

¹⁸⁶ As samples get closer together, a clustering issue, the average covariance increases.

¹⁸⁷ Uncertainty decreases as samples are closer to the value being estimated.

$$\sup{188} E[\sum w_i x_i - \bar{x}] = \sum w_i \bar{x} - \bar{x} = \bar{x} [\sum w_i - 1].$$

¹⁸⁹ This implies symmetry of the local distribution of errors.

¹⁹⁰ If distance **h** is larger than variogram range then C(**h**) equals zero.

¹⁹¹ This implies that samples whose distances are greater than the range then its contribution becomes zero.

$$\sup{192} |x_0^* - x_0|$$

¹⁹³ Mean=53.3, EGM=49.9, variance=3562, EGV=3728, minimum=1, maximum=302.5, skewness=2.49, kurtosis=6.7, R index=1.4, centroid_x=367, centroid_y=216 and nearest_{mean}=21.

¹⁹⁴ 33% of the total variation is due to georeferenced data.

¹⁹⁵ Particularly if graphics and plots are viewed all together. Undoubtedly, dynamic brushing and linkage with cross-referred views to the central map is an imperative spatial analysis requirement nowadays.

¹⁹⁶ Particularly the Proc GIS and the Proc MIXED functions.

¹⁹⁷ Lower glacial for layers at <3.7m, intermediate for 3.7m to 6m and unoxidized between 6 and 20m.

¹⁹⁸ A DOS[®] waste management software by the University of California, Riverside.

¹⁹⁹ Find.neighbor, variogram, vario.fit, krige and predict.krige for OK and UK. They also wrote the programs for IDW-6 and IDW-12 approaches themselves.

²⁰⁰ Created by commercial GISs or manually.

²⁰¹ SPLANCS[®].

²⁰² The observed neighbor distribution within a certain distance is examined by the Kolmogorov-Smirnov test, respecting the theoretical Poisson random occurrences.

²⁰³ Spatial data are polluted, chaotic, voluminous, topological with geographical features and principally sensitive to scales with special time-consuming methods.

²⁰⁴ Where SAS[®], SPSS[®] and MatLab[®] are the elected ones.

²⁰⁵ The encompassing approach.

²⁰⁶ Unquestionably, there is a major shift in GIS from independent to COTS systems.

-
- ²⁰⁷ *On the other hand, this constitutes a disadvantage for independent software developers. According to the same questionnaire, 80% of users prefer to use software products of well-known companies than universities and research institutes.*
- ²⁰⁸ *A GIS that includes transit planning, routing, scheduling and facility location with the heuristic traveling salesman procedure.*
- ²⁰⁹ *Because TransCAD[®] has no programming language compiler, all run procedures should already be compiled in executable files and stored in procs subdirectory.*
- ²¹⁰ *AML[®] for ArcInfo[®], Avenue[®] for ArcView[®] and MapBasic[®] for MapInfo[®].*
- ²¹¹ *Unless an open system approach is adopted.*
- ²¹² *Bulgaria, Macedonia, Serbia, Romania and Greece.*
- ²¹³ *Remote Procedure Call.*
- ²¹⁴ *Univariate and bivariate CDFs for a specific defined region, a variogram cloud and lagged scatterplots.*
- ²¹⁵ *TIME and SPACE modules.*
- ²¹⁶ *Including features such as uphill and downhill.*
- ²¹⁷ *Correspondence courses, TV broadcasts, helpdesks, videoconferences, Web forums, virtual campus, whiteboard and virtual classes.*
- ²¹⁸ *Premiere[®], Javascript[®], Html, J++[®], Java[®] and Director[®].*
- ²¹⁹ *IRC within RCTS network, manual and CD-ROM.*
- ²²⁰ *Stressed by 60% of the total population.*
- ²²¹ *If technologies change, economic principles do not.*
- ²²² *Although this Appendix list is not a comprehensive, the following chapter completes it with a review of seventeen software solutions and a final Table comparing among them.*
- ²²³ *Angular counter clockwise degrees.*
- ²²⁴ *Both processes help inverse distance methods to avoid the issue of additional cluster samples.*
- ²²⁵ *$x_{\text{cartesian\_coord}} = \text{Lon}$ and $y_{\text{cartesian\_coord}} = \ln(\tan(\text{Lat}/2 + \pi/4))$, where Lon is longitude and Lat represents latitude coordinates.*
- ²²⁶ *Instead of taking the mean squared differences, one calculates the mean absolute difference, leading to a more robust measure of the correlation structure.*
- ²²⁷ *Covariance divided by the variance.*
- ²²⁸ *Based on the average squared difference of the standard weighted between observed and fitted model.*
- ²²⁹ *$\ln(x) = \exp(x - \text{transformed})$ and $\text{sqrt}(x) = x^2$.*
- ²³⁰ *Mean, median, deciles, minimum, maximum and mode, variance, standard deviation, interquartile range, skewness and variation coefficient.*
- ²³¹ *With ArcMap[®].*
- ²³² *<http://www.ems-i.com/gms>.*

-
- ²³³ SAKWeb[®] uses two executable ones, *gamv.exe*, a variogram calculation of scattered data, and *kb2d.exe*, a straightforward OK subroutine.
- ²³⁴ Commentaries lay down between */** and **/* characters.
- ²³⁵ <http://uncert.mines.edu>.
- ²³⁶ It compares quantiles between the predicted and the measured observations.
- ²³⁷ <http://www.ssg-surfer.com>.
- ²³⁸ <http://geoecosse.hypermart.net>.
- ²³⁹ Direction and major and minor range in conjunction with the anisotropy option activated.
- ²⁴⁰ Angle direction, angle tolerance, lags bandwidth, in conjunction with the existing search direction.
- ²⁴¹ Vicinity division (up to eight sectors) but highlighting points with a weight of more than 10%.
- ²⁴² Especially useful for the bivariate normality distribution.
- ²⁴³ The file size of *Kriging1_1.zip* is 10.845 KB.
- ²⁴⁴ Logarithm, exponential, standardization, inverse, absolute value, deviation from the mean and square root.
- ²⁴⁵ IDW power, contiguity and k-nearest neighbors.
- ²⁴⁶ $d_{ij}=3959 \times \arccos(\cos|Lon_i-Lon_j| \times \sin Lat_i \times \sin Lat_j + \cos Lat_i \times \cos Lat_j)$ where *Lat* and *Lon* correspond to the latitude and longitude coordinates.
- ²⁴⁷ Number of nonzero links, average weight and the average number of links.
- ²⁴⁸ Most and least connected, unconnected and link frequency distributions, average distance, minimum, maximum, first and third quartile distance.
- ²⁴⁹ A matrix that is constructed from $n^2 c_{ij}$ binary values, reflecting the existing multicollinearity index and the MLE computation of spatial autoregressive models.
- ²⁵⁰ Used in the LaGrange Multiplier assessment for the spatial autocorrelation test for regression models.
- ²⁵¹ According to Anselin [1992], the variance equals $\sigma^2 (X'X)^{-1} X' \Omega X (X'X)^{-1}$ where Ω is the error variance matrix scaled such that its trace equals *N* [Anselin, 1992], *X* represents the coefficients matrix of the dependent variables and σ^2 is the conventional variance.
- ²⁵² The file input option of SpaceStat[®] weights is able to read them.
- ²⁵³ 1 – Statistical Package; 2 - GIS Module; 3 - Independent Geosoftware.
- ²⁵⁴ It requires MS-Excel[®].
- ²⁵⁵ In order to reflect the estimated global variance for the variogram sill.
- ²⁵⁶ Licenciatura in computer science but including GIS subjects.
- ²⁵⁷ Licenciatura in statistics but including geostatistics subjects.
- ²⁵⁸ Licenciatura-raiz in MIS and statistics, ISEGI-4th year.
- ²⁵⁹ Licenciatura-terminal in MIS and statistics, ISEGI-2nd year.
- ²⁶⁰ Three from LabNT-ISEGI, one from IST and one from IGE.

-
- ²⁶¹ *Three from IGE, five from LabNT-ISEGI, four from IST, all those from FCUL and one from CNIG.*
- ²⁶² *100% of these users recognize buffering, overlay and path finding operations.*
- ²⁶³ *Regardless the first conclusion, these conclusions were already confirmed previously by some interviews of the first questionnaire population.*
- ²⁶⁴ *Absolute and average amplitude, dip, azimuth, instantaneous phase and frequency and energy half-time.*
- ²⁶⁵ *Net pay, porosity, lithology, fluid contact and bed thickness.*
- ²⁶⁶ *Cartography, Desktop Mapping Systems, Image and Computer Aided Drawing and Database Management Systems.*
- ²⁶⁷ *Instead script ones.*
- ²⁶⁸ *Primarily, promoting software is advertising GUIs neglecting the business logic and the data server tiers.*
- ²⁶⁹ *VBasic[®], JavaScript[®], Avenue[®], VBScript[®], SAS[®] code, VC++[®], Perl[®], AML[®], Coldfusion[®], JSP[®], Tcl/Tk[®] or Excel[®] macros.*
- ²⁷⁰ *Pocket, palm talk and portable.*
- ²⁷¹ *But providing mechanisms to maintain communication continuity via Html hidden fields, cookies and finite time caching.*
- ²⁷² *According to Monteiro e Boavida [2001], Dense Wavelength Division Multiplexing WANs, using diverse wave lengths in an optical packets commutation within optical fiber cables, can reach 100Gbps exchange speed.*
- ²⁷³ *Using hubs, switchers, routers, repeaters, bridges, modems, satellite microwaves, multiplexers, twisted-optical-coaxial cables, transceivers and embeddable GIS components among computer languages.*
- ²⁷⁴ *Noticed that WML requires much less bandwidth and processing power than HTML and fits within tiny display screens. Furthermore, WAP supports most wireless network standards and operating systems for handheld computing devices such as PalmOS and Windows CE.*
- ²⁷⁵ *Another GIS crux.*
- ²⁷⁶ *The Knowledge Web concept.*
- ²⁷⁷ *Superb marketing skills and vision management towards what to achieve and how to achieve are decisive.*
- ²⁷⁸ *Latex[®], Nroff[®] and Troff[®].*
- ²⁷⁹ *Vines[®], XNS[®], Lan Manager[®], Apple Talk[®], DecNet[®], NETBeui[®], NetWare[®] or SNA DLC[®].*
- ²⁸⁰ *A server-side Java Applet[®].*
- ²⁸¹ *The three-tier model.*
- ²⁸² *Internet Server Application Programming Interface.*
- ²⁸³ *It is based on a set of tags, the Coldfusion Markup Language, instead of a scripting language.*
- ²⁸⁴ *File Transfer Protocol.*

- ²⁸⁵ *Inter Relay Chat.*
- ²⁸⁶ *HyperText Transfer Protocol.*
- ²⁸⁷ *Network News Transfer Protocol.*
- ²⁸⁸ *Simple Mail Transfer Protocol.*
- ²⁸⁹ *Web Application Manager.*
- ²⁹⁰ *Component Object Model.*
- ²⁹¹ *The ASP[®] parser.*
- ²⁹² *Domain Name Server.*
- ²⁹³ *Since the final host has not yet been setup, yet, it is not possible to indicate these parameters.*
- ²⁹⁴ *With the wish to create more than just a static display for the Web.*
- ²⁹⁵ *So that the Web server knows where to execute programs rather than just display it to the browser. If the Web server has an NCSA HTTPd server distribution version, for instance, then the /cgi-bin directory is the one chosen²⁹⁵.*
- ²⁹⁶ *Which is optimized for the thread management inside a process.*
- ²⁹⁷ *Server Side Includes.*
- ²⁹⁸ *ISAPI applications require all the programming and layout to be contained in a .dll file written in C++.*
- ²⁹⁹ *For an ASP[®] developer, a Website listens to the client browser with the Request object and then it answers with the Response one.*
- ³⁰⁰ *It creates an instance of an ASP[®] component.*
- ³⁰¹ *FileSystemObject, TextStream and Dictionary.*
- ³⁰² *It provides a virtual Web memory for an application to run correctly and retain no stateless.*
- ³⁰³ *It looks after the graphic side of Web applications by allowing any developer to code the invisible site back-end.*
- ³⁰⁴ *It encapsulates business logic rather than putting it in an ASP[®] page with the presentation code.*
- ³⁰⁵ *The developer may choose to use a component to perform queries, for example, and just call the component from the code.*
- ³⁰⁶ *A component that allows the user to upload files from one location to another using HTTP, for instance.*
- ³⁰⁷ *A collection of miscellaneous key/value pairs that the programmer may use within any page.*
- ³⁰⁸ *From the moment it is first run until the last end session is closed.*
- ³⁰⁹ *Sala.asp, grava_cookie_chat.asp, mensagem.asp and saida.asp.*
- ³¹⁰ *Created by the Actbar.cab and Activebar.lpk files.*
- ³¹¹ *Variogramfitness.asp, variogramfitness1.asp, variogramfitness2.asp, variogram.jpg and variogram.asp.*
- ³¹² *.swf file.*
- ³¹³ *Responsible for the variogram and OK routines.*
- ³¹⁴ *With the Salford[®] Software Fortran[®] 5.1 compiler.*

-
- ³¹⁵ In order to include these ActiveX[®] files into Html files.
- ³¹⁶ Formerly known as OLE Controls or OCXs.
- ³¹⁷ For copyright reasons, the license package file must also be included using the License Package Authoring Tool.
- ³¹⁸ Java DataBase Connectivity.
- ³¹⁹ The **W** covariogram contiguity approach for an inferior distance of the variogram range generates standardized weights between zero and one.
- ³²⁰ Appendix K proves this by assigning several weights to the corresponding neighbors.
- ³²¹ Half of the countries have military governments.
- ³²² sill=2170, nugget-effect=1 and range= 26.5 ($R^2=0.39$).
- ³²³ Covariogram_Neighbor_Population_Density=0.036×Population_Density+35.301 ($R^2=0.007$).
- ³²⁴ The Moran I reveals no spatial autocorrelation for this wider range, either.
- ³²⁵ 0.45 and 0.48, respectively.
- ³²⁶ 0.03 and 0.78, respectively.
- ³²⁷ One tenth of the range between the mean nearest distance and the variogram range.
- ³²⁸ This approach requires bounded variograms.
- ³²⁹ The mean, standard deviation, skewness, kurtosis and standard error deviation of the mean are 0.382, 0.206, 1.31, 2.74 and 0.019, respectively, while $\gamma(h)=0.0521 \times \left(\frac{3}{2} \frac{|h|}{68.2} - \frac{1}{2} \left(\frac{|h|^3}{68.2^3} \right) \right)$.
- ³³⁰ Covariogram_Vicinity=0.975×Pb+ 0.0085, $R^2=0.439$.
- ³³¹ Nine are omitted for the first one.
- ³³² These samples always layout near the x-axis or y-axis of the Moran scatterplot.
- ³³³ Plotted with the strong brown dots in the previous figure.
- ³³⁴ Note that SAKWeb[®] discriminates both transition samples according to both uncertainties.
- ³³⁵ According to tests, on average, the range of observations position due to **W** misspecification was 15% of the population average (see section 2.7.1 for further discussion).
- ³³⁶ Domain=mean_± range=0.3820_± 0.02 (7.5%×0.3820).
- ³³⁷ In the end, the where issue of section 2.1 must be taken into consideration.
- ³³⁸ Uncertainty region due to cutoff mean (average housing costs where neighborhood costs are high.).
- ³³⁹ Low housing costs against average neighbor housing costs.
- ³⁴⁰ According to the conventional Moran scatterplot, its plot is always the same.
- ³⁴¹ According to the Bartlet and Levene indices.
- ³⁴² The z-axis is necessarily low.
- ³⁴³ Note that the cutoff value can be different for each axis.
- ³⁴⁴ The central location is lower than the mean.
- ³⁴⁵ SAKWeb[®] uses different colours to differentiate groups but for reasons of legibility, lines were drawn.

³⁴⁶ $\gamma(h)=8130\times(1-\exp(-|h|/45))$ fills **W** specification.

³⁴⁷ According to their number order.

³⁴⁸ Note that the conventional variance is lower then the EGV.

³⁴⁹ Concerning the nearest neighborhood analysis, the R index equals 1 (a low moderate dispersed sampling), the weighted centroid and its centroid are located at (75,101) and (75,100), respectively, for a point density of 0.002 and a mean nearest distance of 12.

³⁵⁰ Looking more closely to its scattergram (Appendix N), major outliers are placed along the perimeter region.

³⁵¹ If the number of local samples is zero or one for a particular estimate cell, SAKWeb[®] considers the local variance equals global one.

³⁵² The non-convexity property. According to Isaaks and Srivastava [1989], if OK produce estimates that are negative, one is perfectly justified in setting such estimates to zero.

³⁵³ The prediction standard error is zero.

³⁵⁴ Only with OK, SK and UK is measurement error possible [ESRI, 2001].

³⁵⁵ Analogous to a wider range.

³⁵⁶ By increasing the range and the number of points, it will not improve Kriging variances.

³⁵⁷ A micro-scale component plus a measurement error variance due to a noise process.

³⁵⁸ A parameter of the nearest neighborhood analysis.

³⁵⁹ Note that the choice of the micro-scale model is the same as the long-range struture.

³⁶⁰ The smooth micro-scale variation and measurement error.

³⁶¹ $C_0+\sigma_{\gamma1}+\sigma_{\gamma2}$.

³⁶² With field studies of electromagnetic radiometry, according to Atkinson [1997], it is easy to attain it.

³⁶³ Especially when the site to be estimated is the sample location (instead of the zero assignment).

³⁶⁴ SAKWeb[®] allows the user to reach 30% of the total measurement error.

³⁶⁵ This means closer to the overall mean.

³⁶⁶ The small spikes of the non-exact OK with measurement error relate to the existing samples.

³⁶⁷ The smoothing effect.

³⁶⁸ The first three cases are exact while the following ones are non-exact interpolators.

³⁶⁹ Counties 2, 10, 16, 22, 30, 33, 40, 45, and 48 were taken previously for validation purposes.

³⁷⁰ If this last option has not been defined, then a warning message appears to do so.

³⁷¹ This choice regards future developments although SK is already installed.

³⁷² The IDW map presented in this figure is based on the GS+[®] Pb dataset.

³⁷³ The same strategy is used by varmap.exe and gamv.exe GSLib[®] routines.

³⁷⁴ `response.write("sakwebsessionid").`

³⁷⁵ `$_cookie["sakwebsessionid"].`

³⁷⁶ `response.cookies("sakwebsessionid").`

³⁷⁷ It is assumed that the estimation range is centered on the standard deviation Gaussian curve.

³⁷⁸ *If cleaning is the objective, false positives will lead to clean areas that are not really contaminated.*

³⁷⁹ *False negatives will lead to retain contaminated sites that have not been cleaned, thus carrying environmental and health costs.*

³⁸⁰ *As always, a pre-loader with a progressive status bar was setup.*

³⁸¹ *Keep It Simple Stupid.*

³⁸² *The authors of these .pdf files are Georg Kosakowski of Paul Scherrer Institute and Luc Anselin of University of Illinois, Urbana-Champaign.*

³⁸³ *The search form uses the Indexing Service to search the text index. If IIS[®] is not running, it uses the WAIS (Wide Area Information Server) search engine.*

³⁸⁴ *In Microsoft[®] Access 2000 format.*

³⁸⁵ *With the exception of the name of the site.*

³⁸⁶ *Because the mapping must be done within ASP code, it is still difficult to improve the processing time. For instance, with the 470 samples of Srivastava and Isaaks dataset, it takes 2 seconds to assess the C code but 45 seconds to layout the final maps.*

³⁸⁷ *The sum of the weights constraint does not exist but the overall mean is considered known.*

³⁸⁸ *Taking into consideration the experts' main opinions.*

³⁸⁹ *This feature was, thus, installed after this survey.*

³⁹⁰ *Among the three methods, IDW was chosen to be added in version 2.0.*

³⁹¹ *Note that the OK variance for 0% measurement error equals zero.*

³⁹² *Because SAKWeb[®] does not compute IK, it was not possible to estimate the probability of exceeding this U.S. Department of Agriculture limit or to compare the areas of infestation exceeding 9 grasshoppers per m² by OK and IK.*

³⁹³ *Currently, the rangeland grasshopper densities must exceed 8 grasshoppers per square meter before federal funding is available for control measures [Davis et al., 1992].*

³⁹⁴ *Most of the land (70%) is privately owned. The remainder is under the jurisdiction of the states (22%) and the Federal government (7%).*

³⁹⁵ *If the all area taken into consideration has a similar spatial distribution.*

³⁹⁶ *This decision could also be supported by the coefficient of variation since the four datasets presents values greater than one. Another possibility to decrease the coefficient of variation close to one is to perform outlier sensitivity with the bivariate scatterplot, for instance.*

³⁹⁷ *To avoid this issue, according to Clark and Hosking [1986], summary measures can be based on transformations of the straight arithmetic scale such as $\phi = -\log_2 b$, where b is the grasshopper density. According to these authors, by using base 2 logarithms, ϕ units change one unit for every doubling in size of the grasshopper density and, by setting the transformation negatively, the ϕ increase as grasshopper density gets smaller.*

³⁹⁸ *Recognizing samples that are outside the general cluster is possible and particularly useful when the coefficient of variation is greater than one. Although it may be a controversial decision,*

dropping samples whose values are more than five standard deviations above the mean will reduced the coefficient of variation.

³⁹⁹ *These results were obtained by SPSS[®].*

⁴⁰⁰ *Confirmed by the Moran location scatterplot.*

⁴⁰¹ *The mean center can be described as the centre of gravity of a spatial distribution of points.*

⁴⁰² *The weighted mean center includes the grasshopper density distribution.*

⁴⁰³ *The standard deviational ellipse takes into account the direction bias of observations.*

⁴⁰⁴ *The range as a length measure of the heterogeneity correlation (see Section 3.2 for a further review).*

⁴⁰⁵ *According to Appendix E, the error sum of the validation dataset ranked the second worst among the four years. Certainly, the smoothing effect of the wide variogram range helps to avoid the worst interpolation.*

⁴⁰⁶ *In fact, the division of the political boundaries does not fit the grasshopper outbreaks boundaries.*

⁴⁰⁷ *This conclusion can be reached by Pythagoras formula where triangle sides are half of the length of the standard deviation ellipse.*

⁴⁰⁸ *Notice that the 1995 and 1997 range should be around 100 km (if isotropic) since it is the distance at which the Moran I correlogram equals zero. Nevertheless, this will lead to a nonsense OK result in the Rocky Mountains in Colorado on account of the lack of observations.*

⁴⁰⁹ *These variograms were based on the following experimental parameters: The tolerance angle of 23°, the lag distance and the horizontal bandwidth is 25 km, the number of lags equals 11 km, the tolerance lag is 14 km with a -1° deviation of principal axis.*

⁴¹⁰ *As expected, the variogram values for the y-axis are stronger for 1994 than 1993.*

⁴¹¹ *In addition, a major variogram limitation is that only one dominant correlation scale can be assumed.*

⁴¹² *The lower the threshold values are, the greater the number of outliers (yellow), high variability region (black) and medium central-neighbors (green). Naturally, the number of observations classified as a stable regions (blues dots) becomes smaller too.*

⁴¹³ *According to the U.S. Department of Agriculture, the previous two areas fulfill the requirements for state funding aid.*

⁴¹⁴ *Which ranges between +1 and -1 (according to several neighborhood ranges) regardless of the data values.*

⁴¹⁵ *In terms of estimation, a negative spatial autocorrelation is preferable to then a zero one because there is a major probability of interpolating nearby locations (exactly the inverse value of the sample ones). With a null situation, the vicinity could be equal to or completely different than the available observation site.*

⁴¹⁶ *Further research is necessary to prove this conclusion.*

⁴¹⁷ *And if so, exactly how?*

- ⁴¹⁸ *In order to take into account the preferential sampling issue.*
- ⁴¹⁹ *Note that Kriging estimates do not change with this rescaling operation.*
- ⁴²⁰ *Because of the numerous graphics presented in this section, only OK variance with measurement error is presented.*
- ⁴²¹ *Or the differences between the approaches, presented in Appendix E, are near zero.*
- ⁴²² *Small Sample Interval.*
- ⁴²³ *As it should be.*
- ⁴²⁴ *Note that the non-sense negative estimates correspond to areas with no samples at all (the Rocky Mountains of Colorado). Moreover, the interpolation of this region leads to a minimal difference among OK approaches.*
- ⁴²⁵ *It assumes a positive Moran I in the variogram range. Model 3 is also implicit in these comparison statistics although models 1 and 2 follow the same pattern.*
- ⁴²⁶ *In addition, the average Kriging standard error is even more closer to the root-mean-square prediction error.*
- ⁴²⁷ *It assumes a micro and a macro variogram structure.*
- ⁴²⁸ *It assumes $\gamma(0)=0$ but discontinuity as $h \rightarrow 0$.*
- ⁴²⁹ *It assumes $\gamma(0)=0$.*
- ⁴³⁰ *A situation not considered here.*
- ⁴³¹ *Geography is not a homogeneous discipline but a complex system since the study object can be the same for different sciences.*
- ⁴³² *The view of Ratzel determinism.*
- ⁴³³ *Nomothetic geography versus idiographic geography.*
- ⁴³⁴ *Like spatial interpolation.*
- ⁴³⁵ *Such as spatial autocorrelation.*
- ⁴³⁶ *In order to reflect the estimated global variance for the variogram sill.*
- ⁴³⁷ *Curiously, major virtual courses concern computer matters.*
- ⁴³⁸ *However, system interoperability is a major concern for this population.*
- ⁴³⁹ *Confirmed by the first survey.*
- ⁴⁴⁰ *Leading, quite often, to the need to program in several environments.*
- ⁴⁴¹ *Accomplished with Flash[®] with SAKWeb[®].*
- ⁴⁴² *Overlooked by classical statistical estimators.*
- ⁴⁴³ *With a confidence level of 95%.*
- ⁴⁴⁴ *“Do not reinvent the wheel but improve the wheel, invent a new wheel or use the wheel for a new car.”*