Masters Program in **Geospatial Technologies**

Spatio-temporal Modelling of Tornados with R-INLA, at the county-level in Texas and Oklahoma

Ângela Afonso Rodrigues

Dissertation submitted in partial fulfilment of the requirements for the Degree of *Master of Science in Geospatial Technologies*

NOVA IMS · UNIVERSITAT JAUME·I · Westfälische Wilhelms-Universität Münster

Spatio-temporal Modelling of Tornados with R-INLA, at the county-level in Texas and Oklahoma

by

Angela Afonso Rodrigues

Dissertation supervised by

Jorge Mateu Mahiques, Ph.D

Institute of New Imaging Technologies (INIT),

Universitat Jaume I, Castellón, Spain

Co-supervised by

Fernando Santa, M.Sc.

NOVA Information Management School (NOVA IMS),

Universidade Nova de Lisboa, Lisbon, Portugal

and

Edzer Pebesma, Ph.D

Institute for Geoinformatics,

University of Münster, Münster, Germany

February 2017

ACKNOWLEDGNEMENTS

I would like to express my uttermost and sincere gratitude my supervisor Prof. Dr. Jorge Mateu Mahiques, for all his support and guidance throughout the development of this thesis. It was a great honor to learn from such a great professional, and be advised by his positive and whole-hearted personality. I also express my gratitude to Dr. Pebesma and Fernado Santa, for accepting and co-supervising this thesis.

I am also indebted to Prof. Doctor Marco Painho, for his support and follow-up during the course of the whole program.

In addition, I would like to express my gratitude to Dr. Marta Blangiardo, Dr. Barry Rowlingson, and Dr. Peter Diggle for their prompt and effective advice.

I dedicate this thesis to my parents, for the comfort, support and spiritual shelter that only a family can provide. To my brother, for being the greatest human being alive, and for his limitless support and patience with me.

To my friends Pedro R., Chenah J., Dave M., Vânia O., and Carina D., who always cared for me with a truthfully and unbiased friendship. For all their support, advice and great moments shared over this period.

Last but not least, to Abuzar P., for the comfort and care; for the unlimited and unconditional support; for all the moments and places; for standing by, and holding on; for “this side of the ocean”.

Spatio-temporal Modelling of Tornados with R-INLA, at the county-level in Texas and Ocklahoma

ABSTRACT

The United States of America is the county in the world that is more prone to tornado occurrence. This fact led many researchers, for the past years, to study and formulate theories about tornado occurrence, and which factors promote tornadogenesis. The theories around tornados are always coupled with an attempt to predict their occurrence, for better disaster alertness, and response, in case they happen. At the country level, the tornado occurrence is highly studied and understood. But the same does not happen for the state level, or county level.

In this thesis, it is proposed a statistical model to characterize the occurrence of tornados in a state, given physical (terrain roughness and land-cover types)and demographic properties of its counties. This model also takes into consideration the spatial and temporal dimensions, as well as a space time interaction component. This model was applied for Oklahoma and Texas.

The model with the covariates fits Texas' tornado occurrence, but for Oklahoma, only the spatio-temporal formulation can be applied.

For Texas, the model explains the covariates as being congruent with the low-level inflow hypothesis, with tornados decreasing in zones where natural barriers for the flow can be constituted.

Under the Bayesian framework, maps of spatial risk and probability of tornado occurrence for Texas and Oklahoma were computed, that can be used to make predictions in the future.

KEYWORDS

Tornados

County-level tornado modelling

R-INLA

Arcpy

Python

Point-processes

Areal Modelling

Spatio-temporal analysis

Spatio-temporal modelling

Bayesian Statistics

ACRONYMS

LULC – Land-use Land-cover

FS – Fujita Scale

CWA – County Warning Areas

NOAA – National Oceanic Atmospheric Administration

SPC – Storm Prediction Center

USA – United States of America

WAIC – Wanabe-Akaike Information Criteria

INLA – Integrated Nested Laplace Approximation

TPI – Topographic Position Index

RI – Roughness Index

DEM – Digital Elevation Model

Kinhom – Inhomogeneous K-function

BYM – Besag-York-Mollie

PIT - Probability of Integral Transform

CPO - Conditional Predictive Ordinates

MCMC – Markov-Chain Monte Carlo

CrI – Credible Interval

CAR – Conditional autoregressive

RW – Random Walk

STI – Space-Time Interaction

STITI – Space-Time Interaction Type I

SD – Standard Deviation

ST – Space Time

SDTPI – Standard Deviation of Topographic Position Index

INDEX OF TEXT

ACKNOWLEDGNEMENTS ... iii
ABSTRACT ... iv
KEYWORDS ... v
ACRONYMS ... vi
INDEX OF FIGURES ... ix
INDEX OF TABLES ... xi
1. INTRODUCTION ... 1
1.1 Tornado Dataset: some statistics and remarks ... 2
1.2 The Tornado Alley ... 6
1.3 Why so many tornados? ... 7
1.4 Final Problem Statement ... 10
1.5 Objectives ... 10
1.6 Research Questions ... 11
1.7 Thesis Structure ... 11
2. THEORETICAL FRAMEWORK ... 12
2.1 Bayesian Framework ... 12
2.1.1 Bayesian Statistics ... 12
2.1.2 Bayesian Inference ... 12
2.1.3 R-INLA: The Integrated Nested Laplace Approximation ... 13
2.1.4 Spatio-temporal modelling under the Bayesian Framework ... 16
2.2 Modelling Tornado Occurrence ... 18
3 STUDY AREA ... 19
3.1 Texas ... 19
3.2 Oklahoma ... 22
4 RESOURCES USED ... 23
4.1 Data description ... 24
4.1.1 Tornado Occurrence in North-America ... 24
4.1.2 Digital Elevation Model ... 25
4.1.3 Population ... 26
4.1.4 Land Use/ Land Cover ... 28
4.2 Description of Software used ... 29
5 METHODOLOGY ... 30
5.1 Point Processes ... 30
5.1.1 Intensity ... 30
5.1.2 Intensity as a function of covariates ... 31
5.1.3 Correlation ... 32
5.1.4 Spatio-temporal Inhomogeneous K-function ... 33
5.2 Lattice Approach ... 34
5.2.1. Why INLA? ... 34
5.2.2. Data manipulation and database construction ... 34
5.2.3. Accessing model quality ... 35
5.2.3.3. Distribution of the random effects ... 36
5.2.4. Modelling Technique and formulation ... 36
5.2.5. About the model outputs ... 39
6. RESULTS AND DISCUSSION ... 40
6.1. Point Processes ... 41
6.2. Lattice Approach ... 46
7. CONCLUSIONS AND FUTURE WORK ... 63
8. BIBLIOGRAPHIC REFERENCES ... 65
9. ATTACHMENTS ... 71
A.1. Python script for DEM geoprocessing ... 71
A.2. Python script for Land-cover geoprocessing ... 72

A.3. Python Script to construct the buffer of 40Km outside and give the points of the polygon 75
A.4. Adjacency matrix for the counties 76
A.5. Visual representation of the point process pattern by FScale 77
A.6. R-Code for Point Processes.....80
A.7. R-code for Lattice approach.....88
A.8 Packages used in R.....102
A.9 Map of Texas with Counties Discrimination 104

INDEX OF FIGURES

Figure 1-1. Evolution of absolute tornado yearly counts in USA, during the period of 1950-2015; The linear smoother denotes the main tendency over the years. Data Source: SPC 2016a. 2
Figure 1-2. Absolute year counts of tornados per year, per Fujita scale. The trend line, computed by ´loess´ method, shows the general tendency for the mentioned period. Data Source: SPC 2016a. 4
Figure 1-3. Annual Totals of Property Losses, Fatalities and Injuries, subsequent of tornados, during the period of 1950-2015 (Data Source: SPC 2016a) 6
Figure 1-4. States that compose the Tornado Alley zone, or, in other words, states that have major incidence of tornado records in USA. 7
Figure 1-5. Number of absolute counts of tornados per year, for all states of the Tornado Alley. Data Source: SPC (2016a) 9
Figure 3-1. Location of Texas in USA, with representation of main cities, and surrounding states. 20
Figure 3-2. Number of tornado reports in Texas, over the period from 1970 to 2015, represented by FScale, with trend line computed by "loess" method. Source SPC 2016a. 21
Figure 3-3. Map of Texas with the number of tornados per county during the period 1970-2015. Please note that these values are the summation of tornado reports of all years. 22
Figure 3-4. Number of tornado reports in Oklahoma, over the period from 1970 to 2015, represented by FScale, with trend line computed by "loess" method. Source SPC 2016a. 23
Figure 3-5. Map of Oklahoma with the number of tornados per county during the period 1970-2015. Please note that these values are the summation of tornado reports of all years. 23
Figure 4-1. Topographic Position Index for Texas. Original DEM from USGS 2016. 26
Figure 4-2. Topographic Position Index for Texas. Original DEM from USGS 2016. 26
Figure 4-3. Left: Population Change given by percentage between the years of 2015-1970 for Texas; Right: 2015 Population density for Texas. 27
Figure 4-4. Left: Population Change given by percentage between the years of 2015-1970 for Oklahoma; Right: Population Density for the Oklahoma counties in 2015. 28
Figure 5-1. General output of the K-function, and its interpretation. The blue line indicates the expected random spatial pattern (Poisson). Red Line is the distribution of the sample under inspection. Envelopes represent the threshold for 95% confidence. 32
Figure 6-1. Graphic representation for the intensity function surface in 2-D (upper panels) and 3-D (lower panels), for different bandwidths. 41
Figure 6-2. Surfaces of the standard error for intensity. a) from intensity function computed with bandwidth 150 000; b) from intensity function computed with bandwidth 100 000; c) from intensity function computed with bandwidth of 50 000. 42
Figure 6-3. Intensity function p ˆ(z) against covariate values for elevation and TPI, together with 95% confidence bands assuming an inhomogeneous Poisson point process. 42
Figure 6-4. Estimated intensity function p ˆ(z) against covariate values for the logarithmic scale of the population, together with 95% confidence bands assuming an inhomogeneous Poisson point process. 44
Figure 6-5. Inhomogeneous K-function, Kinhom^(r), for tornado point processes, together with the theoretical K-function of the inhomogeneous Poisson process Kpois(r) = πr2. 44
Figure 6-6. Inhomogeneous spatio-temporal K-function ($Ku, v - 2\pi u2v$) for tornado occurrence in Texas. Top-left: for distance up to 70Km and time up to 15 days; Top right: time up to 30 days and distance up to 30Km; Bottom left: time up to 30 days and distance up to 100 Km; Bottom right: time up to 100 days and distance up to 30 Km. 45
Figure 6-7 Density plot for the spatial random effects distribution in the frailty model (spatially unstructured) 47
Figure 6-8 Map of the random effects for Texas, described by the frailty model. 48
Figure 6-9 Density plot of the distribution of: a) Random effects; b) Spatially Structured Effects; in the convolution model 48
Figure 6-10 Map of the random effects for the convolution model b) map of the spatial structured effects for the convolution model. 49
Figure 6-11 Spatial risk ζ (probability of tornado occurrence) in Texas, given by the convolution model . 50
Figure 6-12 Fitted Effects (θ) for the convolution model 50

FIGURE 6-13 DENSITY PLOTS FOR THE DISTRIBUTION OF THE RANDOM EFFECTS FOR THE CONVOLUTION MODEL PLUS AN UNSTRUCTURED TIME COMPONENT: A) SPATIALLY RANDOM EFFECTS; B) SPATIALLY STRUCTURED EFFECTS; C) TEMPORAL UNSTRUCTURED EFFECTS 51
FIGURE 6-14 DENSITY PLOTS FOR THE DISTRIBUTION OF THE (FROM LEFT TO RIGHT) SPATIALLY UNSTRUCTURED EFFECTS, SPATIALLY STRUCTURED RANDOM EFFECTS AND THE TEMPORAL STRUCTURED RANDOM EFFECTS FOR THE MODEL FORMULATED WITH BYM PLUS A STRUCTURED TIME COMPONENT. 51
FIGURE 6-15 FITTED EFFECTS (Θ) FOR THE SPATIO-TEMPORAL MODEL WITH STRUCTURED TIME EFFECTS 52
FIGURE 6-16 MARGINAL EFFECTS ζ(RISK) FOR THE SPATIO-TEMPORAL MODEL WITH STRUCTURED TIME EFFECTS 52
FIGURE 6-17 LEFT: OCCURRENCE RATE OF TORNADOS IN TEXAS (AVERAGE OVER YEARS; AS A PERCENTAGE OF DIFFERENCE FORM THE STATE AVERAGE); RIGHT: STANDARD ERROR OF OCCURRENCE RATE MAP 52
FIGURE 6-18 POSTERIOR MARGINALS OF THE MODEL DESCRIBED SPATIALLY BY THE BYM, PLUS TIME STRUCTURED AS RW1, AND THE COVARIATES. THE RED LINES ARE THE BENCHMARK FOR "NO CORRELATION". 54
FIGURE 6-19 LEFT: MAP OF THE OVERALL FITTED EFFECTS, AVERAGED ALONG THE YEARS. RIGHT: SPATIAL RISK FOR EACH AREA, COMPARED TO THE WHOLE STATE 58
FIGURE 6-20 BAYESIAN PROBABILITY OF TORNADO OCCURRENCE IN TEXAS. LEFT: FOR MORE THAN ONE TORNADO PER COUNTY; RIGHT: FOR MORE THAN TWO TORNADOS. 59
FIGURE 6-21 POSTERIOR MARGINALS OF THE TORNADO OCCURRENCE MODEL IN OKLAHOMA, DESCRIBED SPATIALLY BY THE BYM, PLUS TIME STRUCTURED AS RW1, AND THE COVARIATES. THE RED LINES ARE THE BENCHMARK FOR "NO CORRELATION". 60
FIGURE 6-22 SPATIAL RISK FOR THE TORNADO OCCURRENCE IN OKLAHOMA, GIVEN THE STITI MODEL FORMULATION, WITHOUT COVARIATES. 62
FIGURE 6-23 BAYESIAN PROBABILITY OF TORNADO OCCURRENCE IN OKLAHOMA. LEFT: FOR MORE THAN ONE TORNADO PER COUNTY 63

INDEX OF TABLES

TABLE 1-1. DETAILS ON FUJITA DAMAGE SCALE; DAMAGE DESCRIPTION FROM FUJITA (1971) AND SPC (2016B). 3
TABLE 1-2. CODES ATTRIBUTED AT THE TORNADO DATABASE (SPC 2016A) TO CHARACTERIZE CONSEQUENTIAL PROPERTY LOSS FROM EACH TORNADO 5
TABLE 4-1. CODES FOR EACH LAND COVER TYPE CLASSIFICATION, AND INTERDEPENDENCE BETWEEN CATEGORIES. CODE 1 CORRESPONDS TO THE CLASSIFICATION TYPE-KEYS USED FOR THE SCOPE OF THIS STUDY, AND IT IS A BROAD GENERALIZATION OF BOTH CODE 2 AND 3; CODE 2 CORRESPONDS TO THE CLASSIFICATION PRODUCED FOR 1992 (USGS, 2014A); CODE 3 WAS SHARED BY THE CLASSIFICATION PRODUCED FOR THE YEARS OF 2001 - USGS (2014B), 2006 - USGS (2014C) AND 2011, USGS (2014D). 29
TABLE 5-1 TEMPORAL GENERALIZATION FOR EACH LANDCOVER DATASET 35
TABLE 5-2 CORRESPONDENCE BETWEEN THE LATENT MODELS GIVEN BY THE PACKAGE R-INLA AND THE NAME OF THE MATHEMATICAL MODEL 36
TABLE 6-1. DIC AND WAIC VALUES FOR THE MODELS: NUMBER TORNADOS ~ YEAR WITH DIFFERENT FORMULATIONS: LINEAR TREND AND NON-LINEAR TREND FOR DIFFERENT YEAR STRUCTURE MODELS 46
TABLE 6-2 RESUME OF THE RESULTS FOR THE SPATIAL MODELS: WITH SPATIAL UNSTRUCTURED INTERACTION (FRAILTY) AND WITH SPATIAL STRUCTURE (CONVOLUTION) 48
TABLE 6-3 RESUME OF THE MODELS CREATED FOR THE ADDITION OF A COVARIATE TO THE SPATIO-TEMPORAL MODEL. DIC, WAIC AND BRIER – BRIER-SCORE – ARE MEASURES FOR QUALITY ASSESSMENT OF THE MODEL; LOG -LOG-SCORE ON THE CPO AND CVM P-VALUE ON PIT ARE MEASURES FOR THE PREDICTIVE QUALITY OF THE MODEL ASSESSMENT; FIXED EFFECTS ARE INHERENT VALUES TO THE MODELS 55
TABLE 6-4 RESUME OF THE FIXED EFFECTS FOR THE MODEL FORMULATED WITH SPATIAL BYM AND STRUCTURED TIME WITH RW1, AND ALL COVARIATES. 56
TABLE 6-5 RESUME OF THE MODEL QUALITY PARAMETERS ASSESSMENT FOR THE SPATIO-TEMPORAL MODELS WITH DIFFERENT FORMULATION, WITH COVARIATES 56
TABLE 6-6 MEAN FITTED EFFECTS FOR THE SPITI MODEL FORMULATION WITH COVARIATES, FOR TEXAS 57
TABLE 6-7 RESULTS FOR THE SPACE-TIME FORMULATIONS FOR OKLAHOMA WITH COVARIATES 60
TABLE 6-8 RESUME OF THE MEAN FIXED EFFECTS FOR THE STITI FOR OKLAHOMA WITH COVARIATES 61

1. INTRODUCTION

In a very general description and characterization, tornados are columns of air that touch the earth surface and hastily rotate over themselves in an axis that is defined at their center. These events are highly energetic, capable to inflict damage, and are connected or placed beneath a cumuliform, buoyant convective cloud. Their diameter can be somewhere between 10 m and 2 Km, but generally they are around values of 200 m (Bluestein 2013).

There are extensive studies on tornado occurrence, scientists that made their whole career in studying and predicting tornados (e.g. the work of Ted Fujita, mentor ofGregory Forbes, which continued his work) and even tornado chasers that lost their life in trying to spot and study these events (e.g. Tim Samaras and his son, Paul, both deceased during the El Reno tornado, Oklahoma).

The interest on tornados, and the attention they get amongst the scientific community is not recent, and is not confined to the scientific community; it involves political and social aspects of our society.

As explored further in this section, USA are highly prone to tornado occurrence, and, in this context, it was back in 1970 that President Nixon proposed the creation of NOAA, not exclusively, but also dedicated to tornado occurrence studies, "to serve a national need for better protection of life and property from natural hazards...for a better understanding of the total environment...[and] for exploration and development leading to the intelligent use of our marine resources..." (NOAA, 2016c).

But even before, in 1884, U.S. Army Signal Corps Sergeant John Finley, in charge of tornado investigation and development of forecasting methods, produced one of the first efforts to tornado forecast and study: he established 15 rules for early tornado detection, published after, in 1888, where he identified signs that a tornado is likely to occur.

The first official tornado report was executed by David Ludlam (1970), of a tornado that occurred in 1643, Massachusetts. The same author wrote in another piece, his review on local severe storms, that he made no consideration on tornado events for that publication, once they are "only a detail in the severe storms. However, its importance as a hazard and the interest of the problems which it poses make it desirable to indicate its probability place in the cumulonimbus[1] problem" (Ludlam 1963).

These events constituted the major advent for the upcoming massive tornado research, where NOAA plays a major role, but also some individual names such as James and Ian Elsner

[1]Cumulonimbus – Etymologically, from Latin "cumulus" means stockpile, amass, heap, and "nimbus" that means storm cloud. These are the most dangerous clouds on Earth, their horizontal and vertical dimensions are huge, as they can be seen as far as 400 Km, and are the principal first sign and cause of storms in general, and one of their outcomes are tornados.

(Florida University), Todd Moore (Towson University), Richard Dixon (Texas University), Thomas Grazulis (Oklahoma University) and Thomas Jagger (Florida University).

USA are prone to tornado occurrence, leading the list of highest annual tornado counts per country, with an outstanding average value of more than 1000 tornados recorded (after 1990) as denoted by Figure 1-1 (SPC 2016a). The following country in that very same list is Canada, with a much lesser annual tornado counts value: around 100 per year (NCEI 2016a).

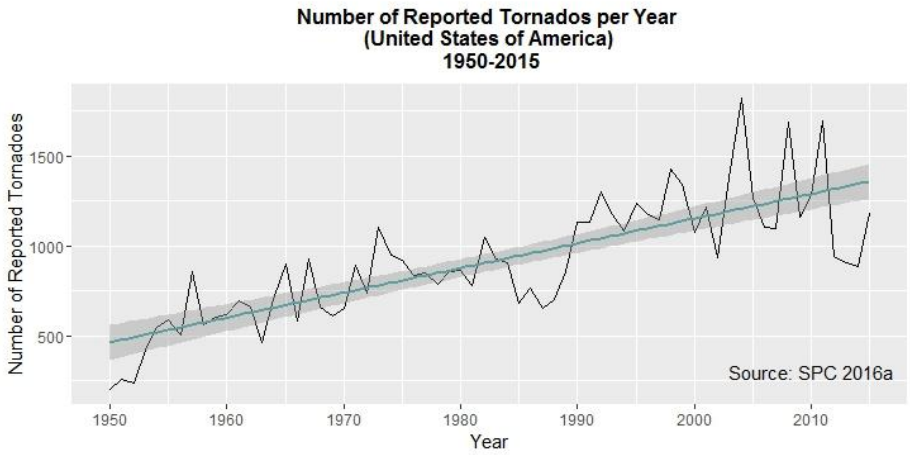


Figure 1-1. Evolution of absolute tornado yearly counts in USA, during the period of 1950-2015; The linear smoother denotes the main tendency over the years. Data Source: SPC 2016a.

After the heat waves in the last 10 years, tornados are the deadliest natural weather disasters in the United States (Romanic et al. 2016). And, even though each single tornado do not surpass the damage of a hurricane, the sum of losses in property and human lives that all the tornados can cause in a year in the country are very superior to the hurricane ones. Every year, in average, 60 people die from tornado occurrence, 1500 are injured and the losses sum up to 200 million dollars in damage (SPC 2016a).

1.1 Tornado Dataset: some statistics and remarks

Figure 1-2 shows the number of reported tornados, per Fujita Scale, during the time period comprehended between 1950-2015.

As a side note, Fujita Scale was developed by Theodoro Fujita (1971). Briefly, it organizes tornados by classes of damage, where the designation F0, F1, F2, F3, F4 and F5 are given to the classes “Light Damage”, “Moderate Damage”, “Considerable Damage”, “Severe Damage”, “Devastating Damage” and “Incredible Damage”, respectively. Table 1.1. displays some more detailed information about each damage class.

From simple observation of the plots on Figure 1-1 and 1-2, one can simply assume that the main trend in tornado occurrence in the United States of America is generally increasing with

time. More specifically, tornados from classes F0 and F1, after 1990, presented more registries.

Classes F2 and F3 are more irregular over time, with both having more incidence over the period 1950-1980, and remaining relatively constant until the present.

Table 1-1. Details on Fujita Damage Scale; Damage description from Fujita (1971) and SPC (2016b).

Scale	Wind Estimate (Km h-1)	Damage (Fujita 1971; SPC 2016b)
F0	<117	Some damage in chimneys, and antennas; some breaks in trees branches; shallow trees pushed over; sign boards damaged.
F1	117 – 180	Surfaces peeled off the roofs; windows broken; light trailer houses pushed; some trees uprooted; moving automobiles pushed off the road.
F2	182 – 253	Roofs torn off frame houses; weak buildings are demolished; trailer houses destroyed; large trees uprooted; light object missiles generated; cars lifted off ground.
F3	254 – 332	Roofs and some walls torn off from well-constructed houses; trains overturned; some rural buildings completely destroyed; most trees in a forest uprooted, snapped or leveled; heavy cars lifted off the ground.
F4	333 – 418	Well-constructed houses leveled, leaving piles of debris; structures with weaker foundation blown away some distance; cars and trains thrown and/or roll for considerable distances; large missiles generated.
F5	419 – 592	Strong frame houses leveled off foundations and swept away; steel-reinforced concrete structures badly damaged; automobile-sized missiles generated and fly through the air for more than a hundred meters; "Incredible phenomena will occur".

Class F4 is relatively constant through time, with mean values around 10 counts per year; and class F5 occurrence is reserved for only a few years, with a maximum of 6 tornados occurrences in 1974. There are some constrains in what concerns to this postulation, which will be discussed in more detail in Section 4.1.1., but as for now, the main idea is the fact that, as a major trend, tornado occurrence in United States is increasing over time. Even though this rate is more accentuated for the less harmful FS classes, it is worth to remark that even an F0-class tornado could already cause some damage (Table 1.1).

Figure 1-3 displays the annual total of losses in property, total fatalities, and injuries, per year in USA, consequent from tornado occurence.

It is important to point out that the methods to calculate property loss and input the values into the database were very much limited, prior to 1996; for each tornado, there was an attribution of property loss code, whose meanings are presented in Table 1.2. As observable, the thresholds are not, at all, proportional in their increment rate. Therefore, for representation on Figure 1-3, the mean value for each threshold was attributed to each year, to have an approximated mean of comparison with the values after 1996. Yet, this is only a very

unpolished approximation: e.g., if a code 4 is describing the property loss of a tornado, it could be any value between 5 000 and 50 000 dollars, but for the year summation computed for Figure 1-3, a value of 22 500 dollars was inputted. Moreover, analyzing the trends of property losses in dollars, without taking into consideration any inflation and wealth adjustments could lead to what Brooks and Doswell (2000) call as a "temporal myopia". In their study, they realized that the measurements of property loss in dollars subsequent from

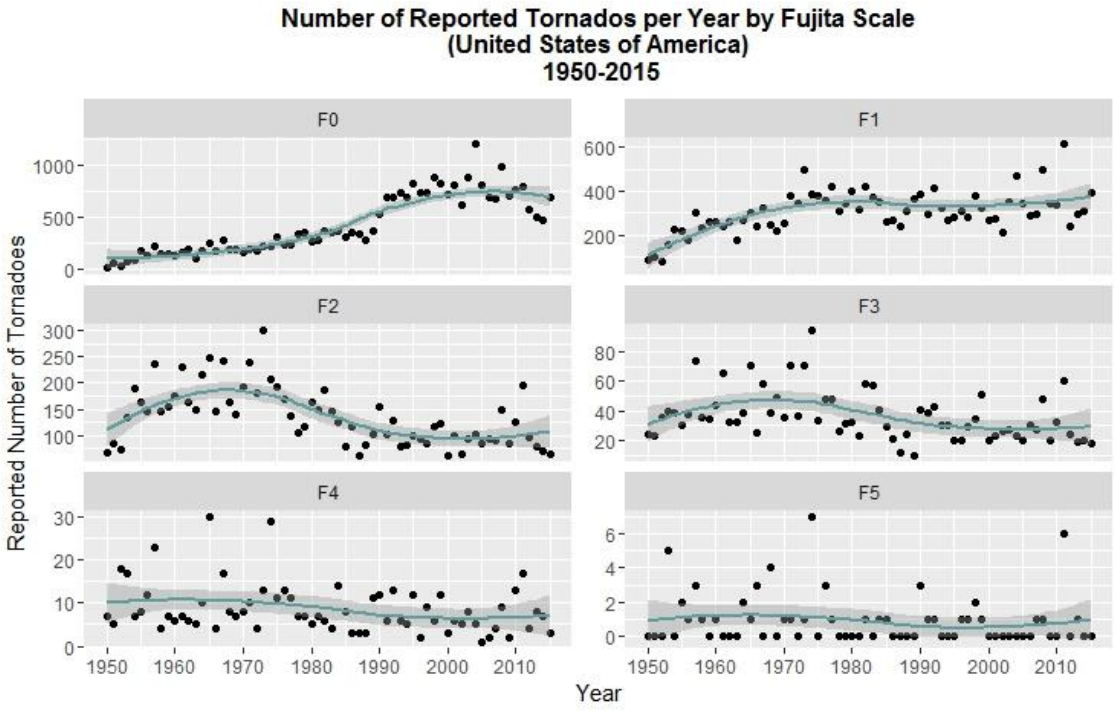


Figure 1-2. Absolute year counts of tornados per year, per Fujita scale. The trend line, computed by ´loess´ method, shows the general tendency for the mentioned period. Data Source: SPC 2016a.

tornados are not increasing over the years, but they reflect the changes in inflation and population wealth.

Nonetheless, even with this rough approximation, after the time of publication of the referred study, it can be assumed a general trend that reflects a yearly increase in property losses. This trend is also clearly accentuated by the 2011 Tornado Outbreak[2], with an astonishing value of almost 9 billion dollars in property losses, across the states of Mississippi, Alabama, Georgia, Tennessee, and Virginia (Knupp et al. 2014).

In fact, Figure 1-1 points displays at least 3 years after 2000 where records were broken in what concerns to absolute annual tornado counts (2004 – 1817 tornado counts; 2011 – 1691 tornado counts; 2008 – 1688 tornado counts).

[2]Out of curiosity, accordingly to NOAA (2016), 2011 was an unusual year, with some of the deadliest and destructive tornados ever registered (e.g. Joplin in Missouri (SPC 2016c)) and the second year with most registered tornados, with a total of 1691. Several records were broken, including the more number of tornados in a single month (758 in April) and the greatest daily total (200, on April 27th).

Regarding the evolution of fatalities over the time, there is a general decreasing trend over the period of 1950 to 1990, from 150 annual deaths to 50, a value that raises until the present day to an average to 150 deaths. It is worthy to point out that what seems to be an accentuated increase over the last 15 years, is highly dictated by the extreme values of 2011.So, having this fact into consideration, the apparent increase can be explained partially by the fact that the tornado occurrences are raising over time, and partially by the fact that, as the time increases, the accuracy of registries become more accurate, due to improvements in technology, measurement devices, etc.

The number of injuries seem to, after 1975, follow a general decrease over time, even though there are some outliers from the general tendency that express extreme values. These extreme values are scattered among time, and represent extremely high values.

Under these statements, two scenarios can be assumed to interpretation of Figure 1-3: either property losses, injuries and deaths subsequent from tornados are increasing over time, since 1950; or, in a more skeptical view, they have been remaining constant over the years (in what concerns to median values), and external factors to the database are creating an illusion of increment (including social and economic factors) as well as internal factors (tornado classification methods; property losses classification methods; machine-based spotting of events, such as radar, that could have misinterpretations).

Table 1-2. Codes attributed at the tornado database (SPC 2016a) to characterize consequential property loss from each tornado

Code	Threshold of Property Losses in Dollars
1	< 50
2	50 – 500
3	500 – 5 000
4	5 000 – 50 000
5	50 000 – 500 000
6	500 000 – 5 000 000
7	5 000 000 – 50 000 000
8	50 000 000 – 500 000 000
9	> 500 000 000

From 1950 to the present time, technology advanced and developed at an exponential rate, improving both forecasting methods and the disaster response strategy. In this sense, it would be expected that the social and economic consequences of tornados would represent a general decrease over time, which, in both scenarios, do not appear to be the most feasible option.

As a wrap-up from all this information, the conclusion to reach is that the changes on how tornados are reported make it difficult to formulize a general trend. But, as pinpointed by Brooks et al. (2014), one fact is for sure: the variability of occurrence has increased since the 1970s, given the standard deviations, and for the last years, several extreme events have been reported. Tippet (2014) endorses the theory of increasing variability by reporting an increase

in the volatility of annual tornado frequency, given by the standard deviation of the difference between annual tornado counts of consequent years.

On the other hand, in what concerns to the formulation of a generalized trend, the studies made so far reflect a vast heterogeneity. On one side, some studies have shown that the annual number of tornados in US have not increased over time, e.g. Brooks et al. (2014); Elsner et al. (2015); Tippett and Cohen (2016). On the other hand, there are studies that support the thesis that they are increasing (Gensini and Mote 2015; Seely and Romps 2015; Tippet et al. 2015).

From another perspective, Tippet and Cohen (2016) report a growing trend of the mean of number of tornados per outbreak per year.

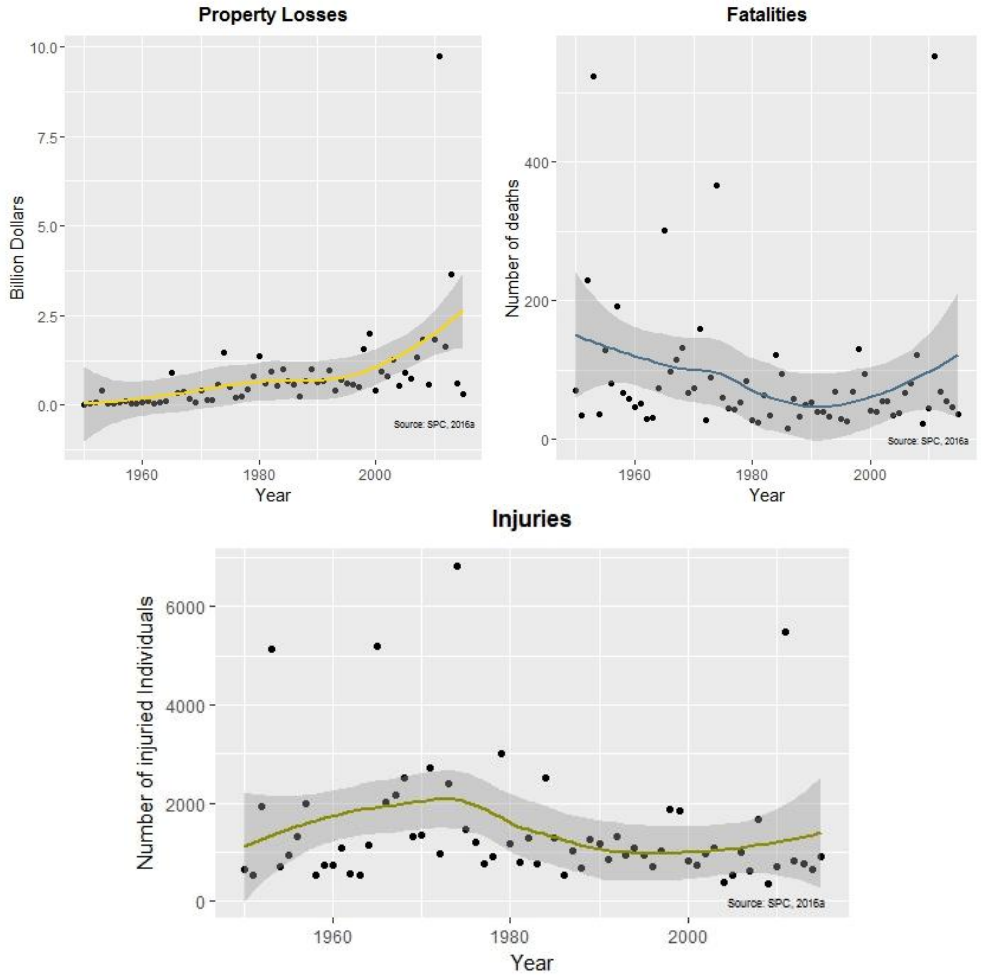


Figure 1-3. Annual Totals of Property Losses, Fatalities and Injuries, subsequent of tornados, during the period of 1950-2015 (Data Source: SPC 2016a)

1.2 The Tornado Alley

Under the thematic of USA tornado sensitivity, presented over the last paragraphs, there is a zone in the United States, nicknamed by the media as Tornado Alley zone (NSSL 2016). Bibliography varies in what concerns to which states belong to this area or not, e.g., in

Tornado (2017), Concannon et al. (2000) or Scolastic (2017); Coleman and Dixon (2014) present a great discussion about this matter. The difference of definition of this zone lies in the difference for quantification of tornados, as it could be expressed in many different ways: by all tornado counts, by tornado-county segments, or strong and violent tornados only (NSSL 2016). For the scope and purpose of this thesis, the following states were considered as inside the Tornado Alley Region (Figure 1-4.): Oklahoma, Kansas, Arkansas, Iowa, Missouri, Texas, Colorado, Louisiana, Minnesota, South Dakota, Mississippi, Illinois, Indiana, Nebraska, Tennessee, Kentucky and Wisconsin.

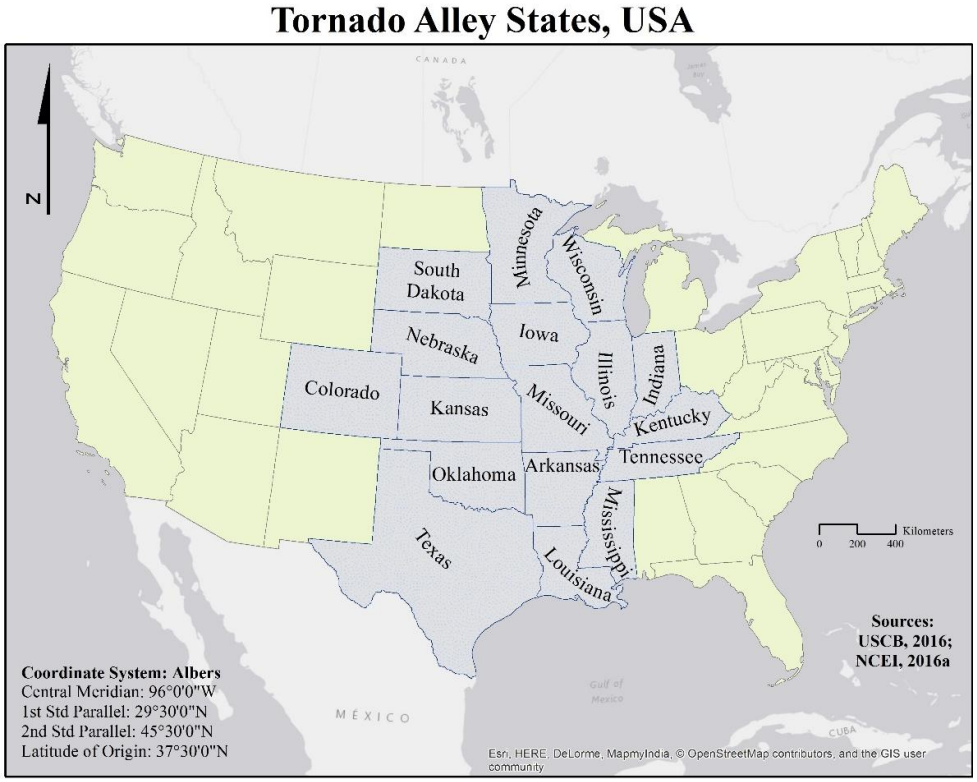


Figure 1-4. States that compose the Tornado Alley zone, or, in other words, states that have major incidence of tornado records in USA.

Figure 1-5 shows the yearly counts of tornados for each of the states of the referred zone. From all, Texas is the state that, by far, has more incidence of tornado occurrence, followed by Oklahoma and Kansas. All the other states follow, more or less the same trend, in tornado counts inside the threshold of 0-100 tornados year $^{-1}$.

1.3 Why so many tornados?

The physical and meteorological principles behind the great occurrence of tornados in the USA at the national level is well studied and understood (Jagger et al. 2015). As Brooks (2014) and Grazulis (2003) explain, the central US are the place where tornados are more likely to occur, due to the presence of the Rocky Mountains and the Gulf of Mexico. “The

surface winds from the south bring warm and moist air at low levels, whereas the winds from the west bring upward relatively cold and dry air, forming a north-south barrier.The temperature and moisture profile brings in the right conditions for thunderstorms and the change of the wind with height means storms will rotate" – Brooks (2014).

This has more influence during the spring, across Oklahoma and Kansas (Schultz et al. 2014) and will spread towards north to the northern Plains and Midwest during summer, due to the migration of the jetstream northwards (Brooks and Doswell 2000; Jagger et al. 2015).

Nonetheless, the regional-scale dynamics is poorly understood (Jagger et al. 2015). This is due to several facts.

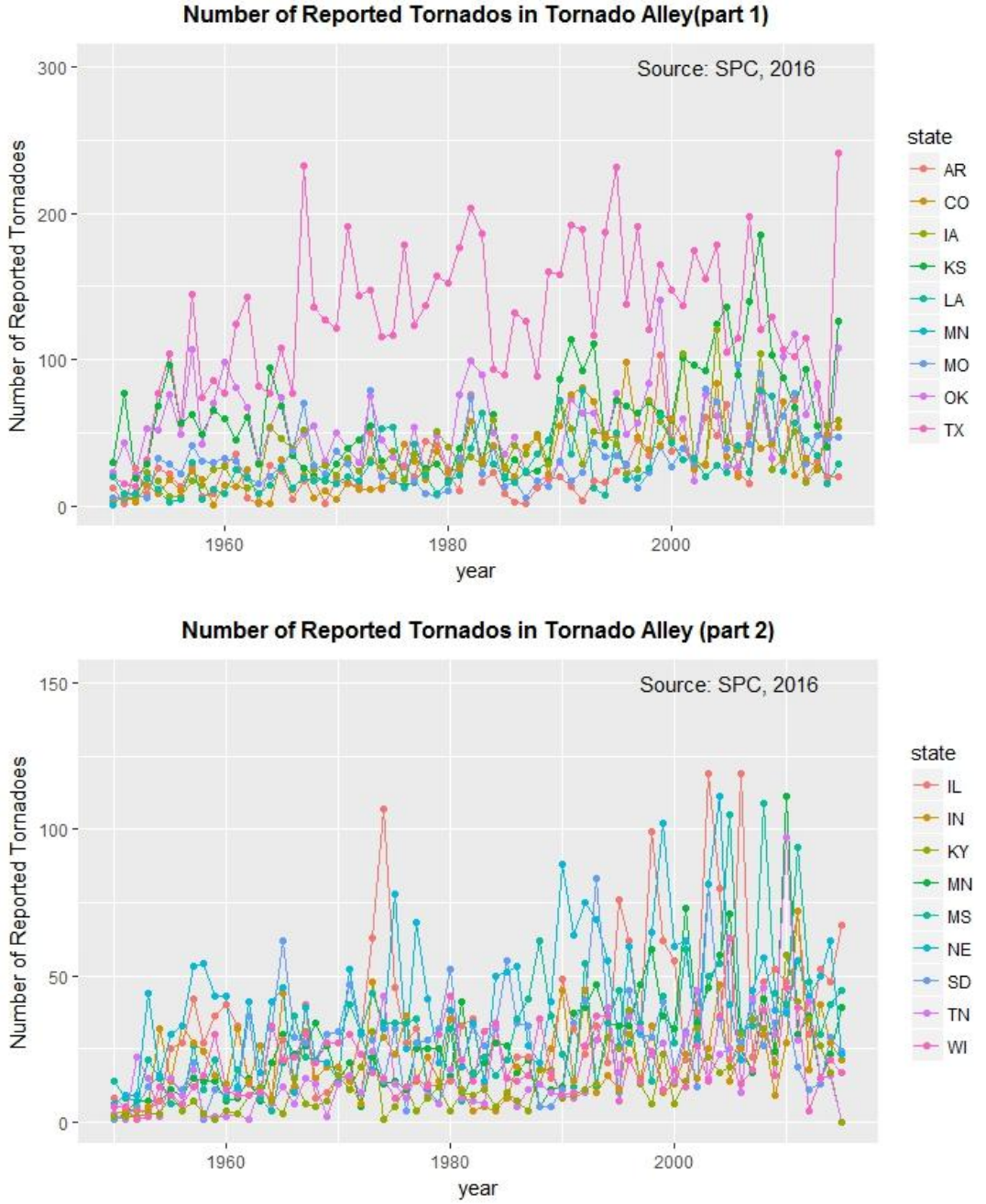


Figure 1-5. Number of absolute counts of tornados per year, for all states of the Tornado Alley[3]. Data Source: SPC (2016a)

Firstly, tornado events are locally rare, discrete and mostly clustered. Moreover, the quality for tornado records is uneven. The tornado national database for the United States, even though the biggest in the world, with records from 1950 to nowadays, it has many issues in what concerns to data interpretation for climatic studies, as referenced by countless studies,

[3]States Acronyms: "AR" – Arkansas; "CO" – Colorado; "IA" – Iowa; "KS" – Kansas; "LA" – Louisiana; "MN" – Minnesota; "MO" – Missouri; "OK" – Oklahoma; "TX" – Texas; "IL"- Illinois; "IN" – Indiana; "KY" – Kentucky; "MS" – Mississippi; "NE" – Nebraska; "SD" – South Dakota; "TN" – Tennessee; "WI" – Wisconsin.

e.g., Verbout et al. (2006), Doswell (2007), Coleman et al. (2011), Elsner et al. (2013), Kunkel et al. (2013), Widen et al. (2013), and Elsner et al. (2016)[4].

1.4 Final Problem Statement

USA are highly prone to tornados and this tendency seems to be increasing over time, as shown above. Not only the tendency of tornado occurrence seems to be increasing, but also, over the last decade, the extreme events, with tornado counts superior to 1500 tornados per year in the USA, values that were never reached before.

Tornados are a big source of property loss, injuries, and even fatalities; they can devastate huge zones, leaving nothing behind but destruction.

The physical climatology is well studied at the national level, but not so well for the regional scale (Jagger et al. 2015). In this sense, modelling tornado occurrence at the regional scale could help to better understand what are the mechanisms that favor tornadogenesis. This, in turn, could lead to serious improvements in what concerns to disaster management and response. Namely, a model that specifies the occurrence of tornado counts as a function of several covariates, characteristics of a given place, and that also has into consideration the spatial and temporal components (Moore 2017). Moreover, if defined at the county-level for a determined state, it could be of great interest to local authorities to improve both prevention and response to tornadoes. Furthermore, if it is clear what are the mechanisms that can control tornadogenesis for a county, the design of tornado-county alert, and risk zones alerts will be much more effective and precise. In this sense, it is also possible to understand on how a change in the territory could enhance or reduce the risk of tornados. Thus, such a model will not only help local policies in what concerns to disaster management and response, but also in what concerns to land management: e.g. in construction site definition or environmental management.

1.5 Objectives

Once Texas is the state in the Tornado Alley that has more occurrence of tornados, the main objective is to model spatio-temporal tornado occurrence at the county level for this state. The modelling strategy has into consideration three covariates: Elevation, Population and Land-Cover. Beyond the covariates, the model should have into consideration the spatial and temporal variability. The model is then tested against Oklahoma, the second state out of all 50 states that have more tornado occurrence.

[4] For more details on the database specifics please refer to section 4.1.1.

The study focus is mainly concentrated on understanding how tornados are distributed over space and time, at the county level for Texas, and the potential link between tornado occurrence and the above mentioned covariates, using two approaches: point processes and lattice. The first one will take into consideration each single tornado occurrence. The second one takes into consideration the amount of tornados at the county level, per year, from 1970 to 2015. For Oklahoma only the lattice approach will be performed to attest the quality of the state-based tornado model.

1.6 Research Questions

The main research questions for this study are:

- Would it be possible to come up with a reliable regional scale (for a state at the county level) model from a scattered and apparently non-reliable database?
- What is the spatial and spatio-temporal distribution of tornado occurrence in Texas?
- Are the events clustered somehow?
- Is there any link between the tornado occurrence and the terrain roughness?
- Is there any link between the tornado occurrence and population of a place?
- Is there any link between tornado occurrence and different kinds of land-cover classes?
- What is the statistical model equation that defines the spatial and temporal structure of tornado occurrence per state in Texas?
- What is the final statistical model equation that better describe tornado occurrence in Texas?
- Does this model fits another state?

1.7 Thesis Structure

In order to follow the reproducible research principle, all scripts (R and Arcpy) and data are given at https://github.com/AngRodrigues/Modelling-Tornado-Ocurrence-with-R-INLA.

For future reference, the coordinate reference system used for the whole geoprocessing and statistical analysis was the EPSG 102003, which is the Conic Contiguous Albers Equal Area projection for USA.

The second chapter deals with the theoretical framework, that gives an overview of the general principles applied in this thesis, from Bayesian statistics, to the R-INLA principles and spatio-temporal modelling within its framework. It also gives a «n overview of past research that had as an objective the tornado occurrence modelling.

Chapter three presents the study areas, Texas and Oklahoma, and gives a general overview of the distribution of occurrence of tornados in those states.
Chapter four refers to the resources used, from data (with a description of each dataset) to the software.
Chapter five presents all the steps used in the data analysis, and some complements on the theoretical framework, related to the model specifics.
Chapter six shows the results and discussion, and chapter seven highlights the main conclusions.
Bibliography is presented in chapter 8, and chapter 9 has all the attachments.

2. THEORETICAL FRAMEWORK

2.1 Bayesian Framework

During the last three decades, Bayesian methods have been suffering significant advances and are starting to be extensively recognized in many investigation areas (Blangiardo et al., 2012).

2.1.1 *Bayesian Statistics*

Bayesian statistics have a huge influence from conditional probability, as well documented and discussed by Hartmann and Sprenger (2010) and Hajek and Hartmann (2010), and even more detailed in Samandiego (2010) and Wakefield (2013). Bayes theorem (Bayes and Price 1763) is defined as:

$$\mathrm{P}\,(\mathrm{B}|\mathrm{A}) = \frac{\mathrm{P}(\mathrm{A}|\mathrm{B}) * \mathrm{P}\,(\mathrm{B})}{\mathrm{P}(\mathrm{A})}$$

In simpler words, the theorem follows the ideas developed by Bayes and Laplace regarding inverse probability: the probability of an event B, given that an event A occurs. So, it all occurs following the process: P(B) is computed before the event A is observed; then the P(A) is computed and used to access the P(B) and P(B|A) is then accessed.
In this sense, P(A) is the information on the event of interest available *a priori*, without carrying any experiment (also called prior information); this probability will affect the posterior probability of B. Thus, the P(B) will be dictated both by the prior information, as well as by the results of the experiment itself (Blangiardo and Cameletti 2015).

2.1.2 *Bayesian Inference*

Bayes theorem is well established in what concerns to observable events. But when it comes to Bayesian inference, i.e., general statistical analyses, where the parameters are less-known quantities and their prior distribution needs to be specified in order to reach the posterior distribution, its application becomes more controversial (Blangiardo and Cameletto 2015).

Let a random variable be Y, and the data available for its analysis be $y = (y1, \ldots, yn)'$. Its uncertainty is modeled using a probability function or a density function (for a discrete or continuous variable, respectively) which is always indexed by a parameter θ. The likelihood function $L(\theta) = p(Y = y|\theta)$, or, simpler, $p(y|\theta)$, specifies the distribution of the data y under the model defined by θ.

The variability on y depends on the sampling selection: it is assumed that the data are a random sample from the study population and uncertainty is generated by the fact that we only observe that sample instead of all the possible other ones (Blangiardo and Cameletti 2015).

The parameter θ is modeled through a suitable *prior* probability distribution $p(\theta)$, before any observation of a realization y. Given the two components, prior and likelihood, the inferential problem is solved by recurring to Bayes Theorem, to obtain the posterior distribution – $p(\theta|y)$ - which represents the uncertainty about the parameter θ after observing the data:

$$p(\theta|y) = \frac{p(y|\theta) \cdot p(\theta)}{p(y)}$$

The denominator $p(y)$ defines the marginal distribution of y, or the "prior predictive distribution of y" (Jeffreys 1961), is defined as:

$$p(y) = \int p(y|\theta)\, p(\theta)\, d(\theta)$$

and indicates what y should look like, given the model, before y has been observed (Statisticat LLC 2015); it is considered a normalization constant, so the Bayes theorem is also reported as:

$$p(\theta|y) \propto p(y|\theta) \times p(\theta)$$

In other words, the posterior distribution is proportional to the likelihood times the prior distribution, known as the "Bayesian Mantra" (Jackman 2009).

2.1.3 *R-INLA: The Integrated Nested Laplace Approximation*

2.1.3.1 The Algorithm

INLA algorithm was introduced by Rue et al. (2009). It is a deterministic algorithm for Bayesian inference. This is the key-point that makes it different from Monte Carlo and Markov Chain Monte-Carlo, because these are based in simulations. INLA is specially developed for latent Gaussian models and provides accurate results for an improved computing time, when compared to MCMC (Blangiardo and Cameletti, 2015). LGM's, or structure additive regression models, are a widely used class of models in statistical applications (Rue 2014). They include, amongst others, generalized linear models, smoothing

spline models, spatio and spatio-temporal models, log Gaussian Cox-processes and geostatistical and geo-additive models (Rue et al. 2009).

The very first step to define a latent Gaussian model within the Bayesian framework is to identify a distribution for the observed data $y = (y1, \dots, yn)$. As a general approach, it is specified a distribution for yi characterized by a parameter ϕi. This parameter is given by a function of a structured additive predictor ηi through a link function f(·), such that $f(\phi i) = \eta i$. In this sense, the additive linear predictor ηi is given by:

$$\eta_i = \beta_0 + \sum_{m=1}^{M} \beta_m x_{mi} + \sum_{I=1}^{L} f_l(z_{li})$$

Where:

- β_0 is a scalar representing the intercept;
- the coefficients $\boldsymbol{\beta}$= {β_1,… , β_m} quantify the (linear) effect of some covariates x=(x_1,…,x_m) on the response;
- f = {$f1(\cdot)$,… , $f_L(\cdot)$} is a collection of functions defined in terms of a set of covariates z=($z1$,… , zL).

The terms $f_l(\cdot)$ can assume different forms such as smooth and nonlinear effects of covariates, time trends, temporal or spatial random effects. For this reason, the latent Gaussian models can be used in a wide range of applications, from generalized and dynamic linear models, to spatial and spatio-temporal models (Blangiardo and Cameletti, 2015).

In this sense, all latent (non-observable) components of interest are collected, in a set of parameters designated by θ (θ = {β_0, β, f}. In addition, it is needed to specify a vector K of hyperparameters such as ψ= {ψ_1, … , ψ_K}.

By assuming conditional independence, the distribution of n observations is given by a likelihood:

$$p(y|\theta, \psi) = \prod_{i=1}^{n} p\,(y_i|\theta_i, \psi)$$

Where each data point y_i is connected to one element θ_i in the latent field θ.

On the logarithm, it is assumed a multivariate normal prior on θ, with mean 0, and precision matrix $Q(\boldsymbol{\psi})$, i.e., $\boldsymbol{\theta} \sim$ Normal($\mathbf{0}$, $Q^{-1}(\boldsymbol{\psi})$) with density function given by:

$$p(\theta|\psi) = (2\pi)^{\frac{-n}{2}} |Q(\psi)^{\frac{1}{2}} \exp\left(\frac{1}{2}\theta' Q\,(\psi)\theta\right)$$

This specification is known as Gaussian Markov random field, where the components of the latent Gaussian field $\boldsymbol{\theta}$are supposed to be conditionally independent with the consequence that $Q(\boldsymbol{\psi})$ is a sparse precision matrix.

The specification matrix is what improves computational times (Blangiardo and Cameletti, 2015). Here, the joint posterior distribution of $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ is given by:

$$p\,(\theta,\psi|y)\ \propto p(\psi)\cdot p(\theta|\psi)\cdot p(y|\theta,\psi)$$

$$\propto p\,(\psi)\cdot p(\theta|\psi)\cdot\prod_{i=1}^{n}p(y_i|\theta_i,\psi)$$

$$\propto p(\psi)\cdot|Q(\psi)|^{1/2}\exp\left(-\frac{1}{2}\theta'Q(\psi)\theta\right)\cdot\prod_{i=1}^{n}\exp\big(\log\big(p(y_i|\theta_i,\psi)\big)\big)$$

$$\propto p(\psi)\cdot|Q(\psi)|^{\frac{1}{2}}\exp\left(-\frac{1}{2}\theta'Q(\psi)\theta+\sum_{i=1}^{n}\log(p(y_i|\theta_i,\psi))\right)$$

Under the scope of R-INLA, the objectives of Bayesian inference are the marginal posterior distributions of each element of the parameter vector

$$p\,(\theta_i|y)=\int p(\theta_i,\,\psi|y)\cdot d\psi=\int p(\theta_i|\psi,y)\cdot p(\psi|y)\cdot d\psi$$

And also for each element of the hyperparameter vector,

$$p(\psi_k|y)=\int p(\psi\,|y)\,d\psi_{-k}$$

In this sense, two tasks are needed to achieve:

- the computation of $p(\boldsymbol{\psi}|\boldsymbol{y})$, from which also all the relevant marginals $p(\psi_k|\boldsymbol{y})$ can be obtained;
- the computation of $p(\theta i|\,\psi,\,\boldsymbol{y})$, which is needed to compute the parameter marginal posteriors $p(\theta i|\boldsymbol{y})$.

The idea to maintain here is that R-INLA perform these two tasks. The details on how they are processed and approached mathematically and computationally are given in Rue et al. (2009).

In this sense, "effectively only one form of uncertainty exists, which is described by suitable probability distributions" (Blangiardo et al. 2012). Consequently, there is no difference between observable data or unobservable parameters; these are also considered as random quantities. Under the Bayesian framework, the uncertainty about the realized value of the parameters given the current state of information (i.e. before observing any new data) is described by a prior distribution. Typically, but not constantly, the objective of the inference

is to deliver the posterior distribution; the inferential process, thus, combines the prior and the data model itself to derive the posterior distribution (Lindley, 2006; Blangiardo et al. 2012).

2.1.4 *Spatio-temporal modelling under the Bayesian Framework*

R-INLA has been used for an infinity of applications. Here are presented some recent studies that used INLA spatio-temporal capabilities:

- Laurini (2017) recurred to INLA to analyze the spatio-temporal gasoline price, arriving to a contiguous space model that allows the estimation of prices distribution throughout Brazil.
- Wang et al. (2017) estimated car crashes by crash types and crash severity, having into consideration factors such as crash type and severity counts on rural two lane highways.
- Braulio-Gonzalo et al. (2017) model the energy performance and indoor thermal comfort of residential stocks at a city-scale.
- Diaz-Avalos et al. (2016) modelled the wild fires in Castellón, Spain, for the period of 2001-06, using several spatial covariate information, in order to predict and, therefore, prevent wildfires in the zone.
- Tabb et al. (2016) modelled the relationship between alcohol outlets availability and violence, for the period of 2010-2013, in Seattle, USA.
- Breivik et al. (2017) used the INLA process to model historical bycatch in commercial fisheries, and project a prediction for the future.

The reason why many studies take advantage of the R-INLA, instead of other tools, is the computational time; MCMC takes days to compute, while INLA takes a few minutes. Moreover, there are advantages related to the modelling approach itself are quite something. For example, the specification of prior distributions allows the formal inclusion of information that can be obtained from previous studies or from expert opinion. In addition, the (posterior) probability that a parameter does/does not exceed a certain threshold is easily obtained from the posterior distribution, providing a more intuitive and interpretable quantity than a frequentist p-value (Blangiardo et al. 2012).

In addition, and perhaps the most important fact, is that, within the Bayesian approach, it possible to postulate a hierarchical structure on the data and/or parameters, which presents the added benefit of making prediction for new observations and missing data imputation relatively straightforward.

Theoretically, data that contains a spatial component are the so well-known spatial data, and are defined as a realization of a stochastic process indexed by space

$$Y(s) \equiv \{\ y(s), s \in D\}$$

where D is a fixed subset of $\mathbb{R}^d$ (Rue et al. 2009).

In what concerns to spatial data, Cressie (1993) and others, such as Gelfand et al. (2010), grounded and structured the distinction between three types of spatial data: Area (or lattice) data; point referenced (or geostatistical) data; and spatial point patterns.

Spatial data is considered by many as a "special kind of data", considering that the spatial trend has to be taken into consideration for its interpretation. The spatial component will provide such an additional information that, if neglected, could lead to serious biases in estimations. Under this scope, the Bayesian approach is particularly effective (Blangiardo et al. 2013; Gómez-Rubio et al. 2014; Bivand et al. 2015; Blangiardo and Cameletti 2015).

All above mentioned types of spatial data can be fitted into models, under the Bayesian framework, by extending the concept of hierarchical structure; such structure will acknowledge similarities that are based on the neighborhood or on the distance (Blangiardo and Cameletti 2015). For the purpose and scope of this thesis, area data will be explored under the Bayesian framework. For details on the other two types of spatial data, please see Blangiardo and Cameletti (2015).

At the area level of data, the neighbor structure dictates the spatial dependency. Given the area *i*, its neighbors N(*i*) are specified as the areas that share borders with it (first order neighbors), or that share border with its first order neighbors. Under the Markovian property that the parameter θ_i for the *i*th area is independent of all the other parameters, given the set of N(*i)* (local Markov property), then

$$\theta_i \perp\!\!\!\perp \theta_{-i} \mid \theta_{N\,(i)}$$

Where θ_{-i} indicates all elements in the parameter θ, except the *i*th. In this sense, the precision matrix **Q** of θ is sparse, a fact that enhances computational benefits. Thus, for any pair of elements (i, j) in θ

$$\theta_i \perp\!\!\!\perp \theta_j \mid \theta_{-ij} \leftrightarrow Q_{ij} = 0$$

Meaning that the precision matrix is now given by the neighbor structure of the event, by the pairwise Markov property. $Q_{ij} \neq 0$, only if $j \in \{i, N(i)\}$. The independence of θ_i from θ_j is not only conditioned by the hyperparameters, but also by the set of neighbors. The main formulation is specified in

The precision matrix can be now specified as a function of the structure matrix, as well documented by Rue and Held (2015):

$$Q = \tau R$$

Where

$$R_{ij} = \begin{cases} N(i), & if\ i = j \\ 1, & if\ i \sim j \\ 0, & otherwise \end{cases}$$

Where $i \sim j$ denotes that areas *i* and *j* are neighbors.

The spatial framework described does not allow to characterize the temporal variation of the data; it should be extended to the spatio-temporal case, by including a time dimension. Thus, both spatial and temporal models can be incorporated into a unique model to enhance the detection resolution. In a prior instance, the stochastic models were designed to, mainly, define a temporal model on a pixel, and, having it into consideration, subsequently describe the spatial model, as described by, for example, Ahmad and Collet (2016), or Friston et al. (1994).

Bayesian models are a reflection of this technique (Craigmile and Guttorp 2011).

In this sense, data should be defined by a process indexed by space and time:

$$Y(s,t) \equiv \{y(s,t), (s,t) \in D \subset \mathbb{R}^2 \cdot \mathbb{R}\}$$

That are observed at n spatial areas and at T time points.

More details in what concerns spatio-temporal model are given in section 5.2.4.

2.2 Modelling Tornado Occurrence

Literature in this matter is extensive. On one side, because of the above cited USA's tornado occurrence susceptibility. Secondly, because modelling techniques are different amongst the different studies and the study region scale largely varies from study to study.

The National Center for Environmental Information (NCEI) and Storm Prediction Center (SPC), included in the National Oceanic and Atmospheric Administration (NOAA), are the main suppliers for tornado information in USA (e.g. NCEI 2016a;NCEI 2016b; SPC 2016a). Data provided by this information center was the basis of inspiration for many studies that make an attempt to theoretically explain, model and forecast tornado occurrence in the USA (e.g. Boruff 2003; Ashley et al. 2008; Dixon et al. 2011; Dixon and Moore. 2012; Widen et al. 2013; Ashley et al. 2014; Coleman and Dixon 2014; Cusack 2014; Rosecrants and Ashley 2015; Elsner et al. 2016; Romanic et al. 2016; Moore 2017), to name a few.

Beyond these studies, there were over the last years, many authors that tried to compute statistical models to describe tornado occurrence. Here is a compilation of some of the methods and models proposed:

- Wikle and Anderson (2003) came up with a spatio temporal model constructed under the Bayesian framework, based on tornado count data for USA. In this study, it was used a zero-inflated Poisson likelihood to model the excess of zeros; the mean of this Poisson is then

modelled using different spatial and non-spatial random effects; a relationship with the El Niño/Southern Oscillation phenomenon was also taken into consideration.

- Elsner et al. (2013) built a model to predict violent tornados during springtime across US Central Plain, as a function of the nearest distance to a Poisson point process of non-violent tornados. It was also proposed a correction for the population bias (which they assume is expressed by less tornado reports in less populated areas), by including the distance to the nearest city.
- Karpman et al. (2013) suggested some spatio-temporal models, that were implemented by using an approach that is based on non-parametric kernel. The variables used in the model were tornado point patterns and topographic variation. The model was selected recurring to AIC[5].
- Elsner et al. (2014) proposed a model where tornado intensity is assumed to be distributed as a Weibull with the log-mean depending linearly on the path length and width which are strongly correlated to the Fujita categories.
- Akers et al. (2014) also explored the relationship between length and width of tornados and intensity, by recurring to a multinomial logistical model, without spatial random effects, to compute the probability of a particular tornado with a certain FS class occur.
- Gomez-Rubio et al. (2015) analyze the dataset by performing a model under INLA framework and estimated with stochastic Partial Differential Equations, to model the intensity of point patterns with marks.
- Jagger et al. (2015) proposed a regional scale model, that assumes a negative binomial distribution for tornado occurrence and is normalized with population, and has into account fluctuations in elevation and CWA.

3 STUDY AREA

3.1 Texas

Texas is the more southward state of USA, Figure 3-1. Toponymically, the word Texas was the Spanish pronunciation of *Tejas*. The latter was the Hasinai Indian word for *allies* or *friends*. Their state motto is, in fact, "friendship" (Dingus 1981).

Texas is the second most populous state in US, after California. This state has three of the top ten most populous cities in the USA, which are Houston, Dallas, and San Antonio, and 70% of its population lives within 200 miles of Austin, the capital of the state (USCB, 2016).

[5] Refer to section 5.2.

It covers 7.4% of the total USA's area, with around 80% of its own land being covered by farms, and 10% covered by forest, a class that includes four national and five state forests (USGS 2014d).

The highest point of Texas is the Guadalupe Peak, at 8749 feet and the lowest is the Gulf of Mexico (USGS 2016).

The deadliest natural disaster in this state, that there are records of, was Galveston Hurricane, in 1900, which killed around 8 000 to 12 000 people (Dar, 2008).

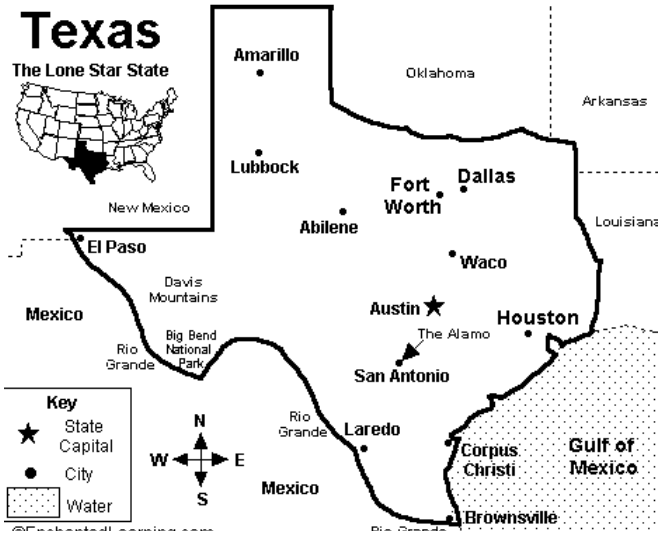


Figure 3-1. Location of Texas in USA, with representation of main cities, and surrounding states.

Out of all states that were considered as belonging to the well-known Tornado Alley, Texas is the one that has major tornado counts over the years (Figure 1-5.). In fact, it is the state in USA that has more tornado counts per year. Of course, this statement is largely biased by the fact that Texas covers, after Alaska, the greatest area, out of all the 49 states.

Texas is shared by two areas that are potentially prone to tornado occurrence: the already referred "Tornado Alley", reserved to the central plains of the US, and "Dixie Alley", which has its main concentration in states across the Gulf of Mexico. These zones are highly prone to tornado occurrence. This fact occurs because they constitute the meeting point of the humid air from the Gulf of Mexico, hot dry air from Arizona and New Mexico, and cool dry air from Canada. Specially in spring time, these masses of air work together and originate most of tornados.

Figure 3-2 shows the year tornado counts for Texas, per FS, over the period of 1970-2015[6] (Source SPC 2016a). It is clear that F0 tornados are the ones with more prevalence, with mean values generally above 75 counts per year; classes F1, F2 and F3 seem to follow a decreasing trend after the seventies and eighties, the very same period where F0 tornado counts seem to increase. This could be related to the aforementioned fragility of the database. What is worth to pinpoint is that weaker tornados occur more than stronger ones: F4 tornados fluctuate over time, but never get over the limit of four tornados per year, and, after 2010, there is one event every two years. F5 tornados are considerably rare in Texas.

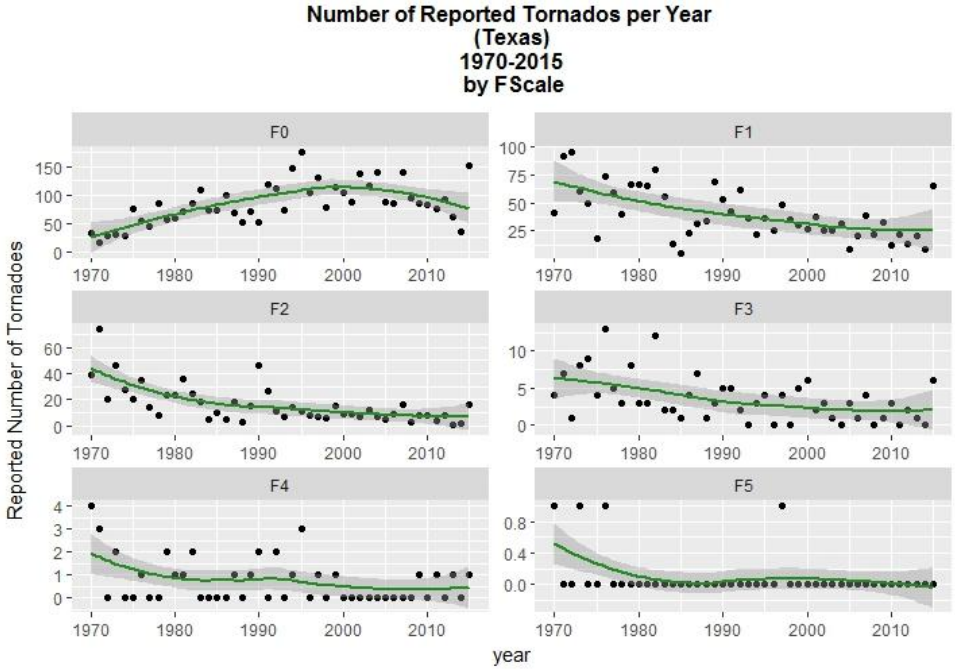


Figure 3-2. Number of tornado reports in Texas, over the period from 1970 to 2015, represented by FScale, with trend line computed by "loess" method. Source SPC 2016a.

Figure 3-3 displays the total number of tornado counts per county in Texas, over the period of 1970-2015. There are several counties with a minimum value of two tornado counts over the years. The county that has more total tornado counts over the years is Harris, with 189 tornados, followed by Hale and Galvinson, with 85 and 80, respectively. The mean value of total tornado counts is the summation of 26 tornados.

[6] Please refer to Section 4.1.1. for an explanation on why the time period was shorted to 46 years.

3.2 Oklahoma

Oklahoma is the state that is immediately northwards Texas. It covers around 180 000 Km^{-2} of American soil, being the twentieth biggest in area. With a population of 3 700 000 inhabitants, it is the second state that has more native American population.

In what regards to tornados, Figure 3-4. shows the tornado yearly tornado counts per FScale. As mentioned before, Oklahoma is the second state in the Tornado Alley with more occurrence of tornados per year. The distribution for each FS class follows, generically, the same as Texas, with an occurrence trend that decreases with increase in FS. Nonetheless, is worth to pin-point the great increase of F0 and F1 tornados after 1990. From the observation of Figure 3-5., the states that had a major occurrence of tornados during the period 1970 to 2015, have a total count of around 80, a much smaller value than the ones for Texas.

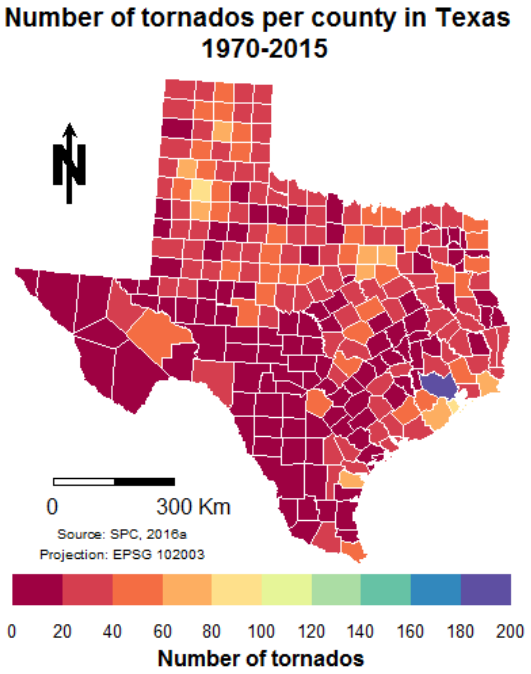


Figure 3-3. Map of Texas with the number of tornados per county during the period 1970-2015. Please note that these values are the summation of tornado reports of all years.

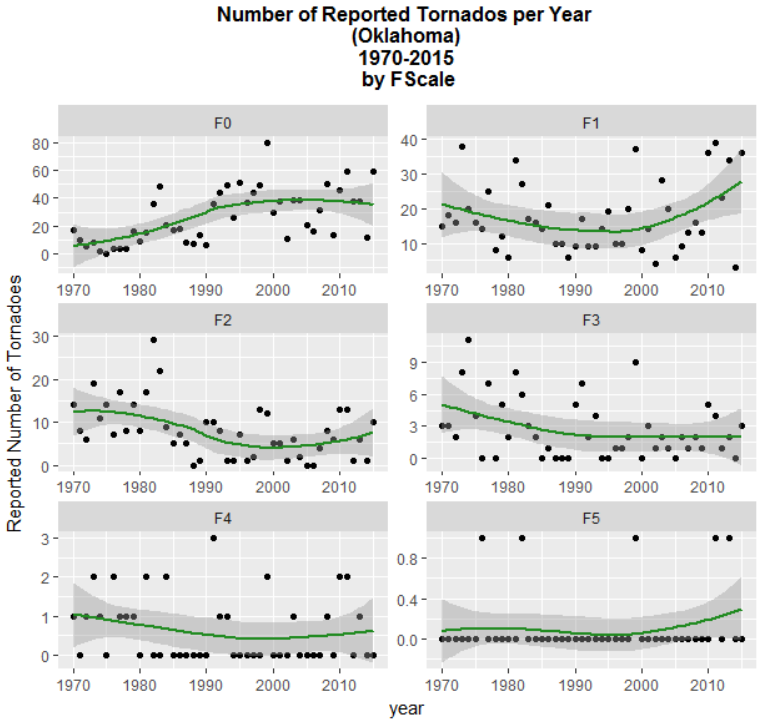


Figure 3-4. Number of tornado reports in Oklahoma, over the period from 1970 to 2015, represented by FScale, with trend line computed by "loess" method. Source SPC 2016a.

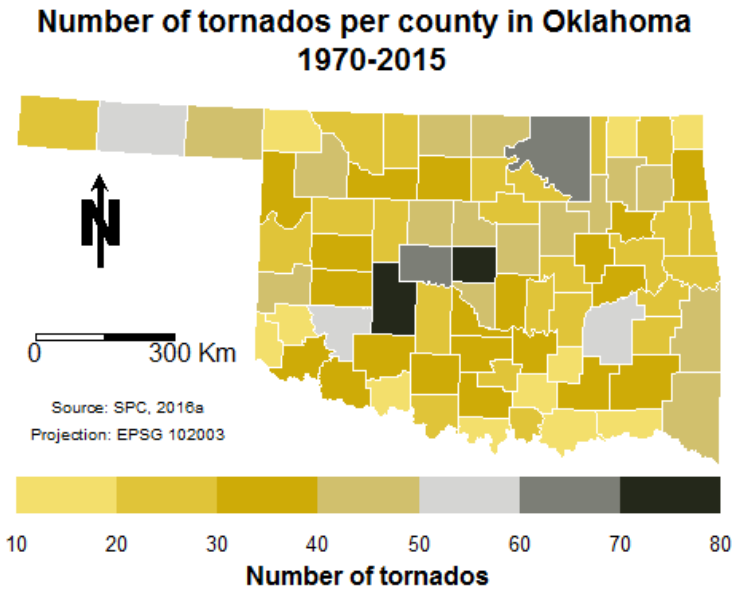


Figure 3-5. Map of Oklahoma with the number of tornados per county during the period 1970-2015. Please note that these values are the summation of tornado reports of all years.

4 RESOURCES USED

4.1 Data description

4.1.1 **Tornado Occurrence in North-America**

The tornado database used for this study is provided by the Storm Prediction Center, from the National Oceanic and Atmospheric Administration of USA (SPC 2016a). It was originally organized by the SPC, from newspaper accounts and reports (Corfidi 1999; Schaefer and Edwards 1999).

Generally, the database contains records from tornados that occurred in USA since 1950 until the present time. For each tornado are recorded several attributes, from time and date, to state, losses amount (for crops and property), magnitude, length, width, injuries, fatalities, among others. This specificity of data allows an infinity of different combination of space, time and attributes, originating a multitude of kind of analysis: from point processes to lattice data; at national, state, or county level; hourly, weekly, seasonal, monthly or yearly approaches; by attribute (FS, damage, etc.). Though the high specificity of the database, it has several limitations that many authors have been pointing out.

It all comes down to a simple limitation: the fact that the tornado records rely mostly on visual spotting and human annotation of occurrences. From this fact,

- Hart (1993) was one of the many authors that reported the fact that the number of reports are greater for places or regions with a greater population density. Therefore, factors such as highways distribution or distance to an official reporting station can influence the numbers is the database.
- The occurrence of F0 and F1 tornados shows a dramatic increase since 1980, while the stronger ones remain steady over time (Figure 1.2.). Otsby (1993) was one of the first authors to realize that reasons for this include an improvement of verification efforts by local offices but also a marked increase in tornado chasing.
- Doswell et al. (1999) and Verbout et al. (2006) advocate that due to technological developments and more tornado chasers, the probability of a tornado be reported will increase.
- Elsner et al. (2013) adds that the number of tornados reported in the database are smaller than the actual number of occurrences, but that this difference is shrinking over time.
- Widen et al. (2013) says that the database is imprecise and inhomogeneous, and that any resulting study from raw data will have a lower estimated risk of encountering a tornado.

In order to minimize some of this limitations, it was followed the advice from SPC not coun the tornados prior to 1970.

Moreover, several authors have been working intensively in this database, especially in what concerns to spatial and spatio-temporal analysis, as presented in section 2.2., and therefore, the database seems to be suitable for the scope of this study, with the temporal correction. In fact, this is one challenge to overcome: the difficulty of uniformizing the database.

4.1.2 **Digital Elevation Model**

The DEM was retrieved from USGS (2016), with a pixel resolution of 1x1 Km. The same resolution was used throughout data analysis. The two states were clipped from the original dataset. It was needed a data simplification process, due to the high dimensions of each raster file. Therefore, the Topographic Position Index was calculated. It was developed by Weiss (2001) and compares the elevation of each cell in a DEM to the mean elevation of a specified neighborhood around that cell.

In this sense, focal statistics were applied for the 10 adjacent cells of each pixel, and the final equation to compute the TPI was:

$$TPI = \frac{mean_r - min_r}{max_r - min_r}$$

Where mean represents a smoothed DEM, computed by the mean of 10x10 adjacent cells of each pixel, and min and max are the DEM with minimum and maximum values of the 10x10 adjacent cells. The python script used to compute it is shown in Attachment A.1., for the case of Texas. The Oklahoma one followed the same procedure.

Figure 4-1 shows the TPI computed for Texas, and Figure 4-2. the TPI for Oklahoma. The index varies from 0 to 1, with values near 0 representing areas characterized by flat plains, and values close to 1 by peaks and ridges (Seif 2014). This generalization was imperative to simplify the inputted data on INLA, and worked as a simple but effective representation of elevation variability. Then, the standard deviation of this index was used to input in the modelling strategy, once it reflects itself a measure of Terrain Roughness (Riley et al. 1999; Ascione et al. (2008).

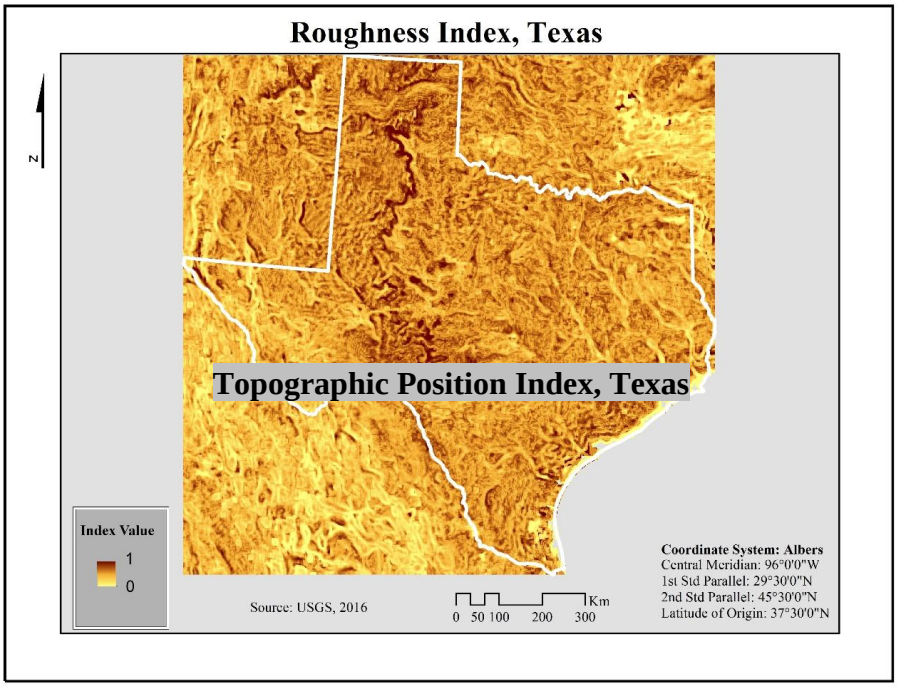


Figure 4-1. Topographic Position Index for Texas. Original DEM from USGS 2016.

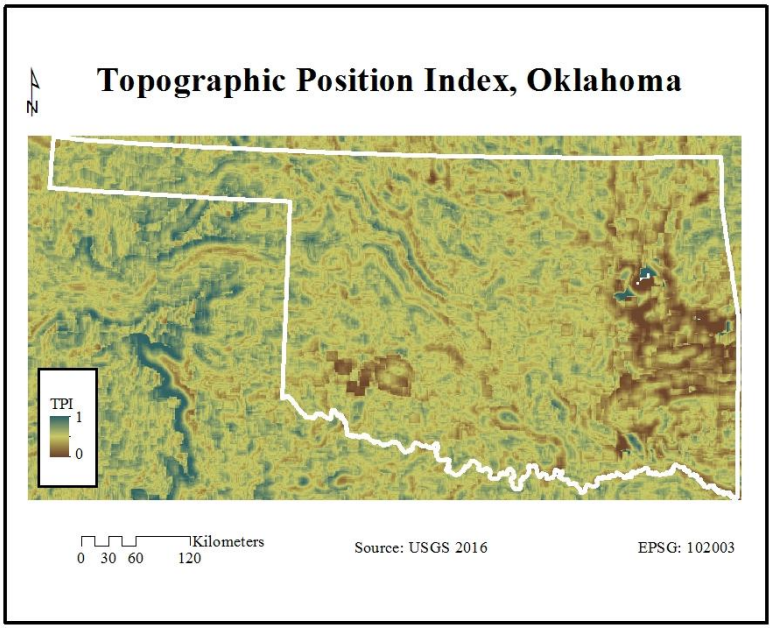


Figure 4-2. Topographic Position Index for Texas. Original DEM from USGS 2016.

Jagger et al. (2015) uses the same strategy: they represent the Roughness Index as a measure of the standard deviation of elevation.

The new approach in this study is to simplify the DEM with the TPI computation and use its standard deviation as a measure of Terrain Roughness.

4.1.3 **Population**

Historical Population estimations were retrieved from the NBER (2016) for both states. The data was organized as individual counts per state per year, with a temporal coverage for the

period 1970 to 2015. Figure 4.3. shows the population change in percentage over the study period, for Texas, as well as the most recent population density (2015) and Figure 4.4. shows the population change for Oklahoma, and respective population density, given by:

$$PC = \frac{(P_{2015} - P_{1970})}{P_{1970}} \cdot 100$$

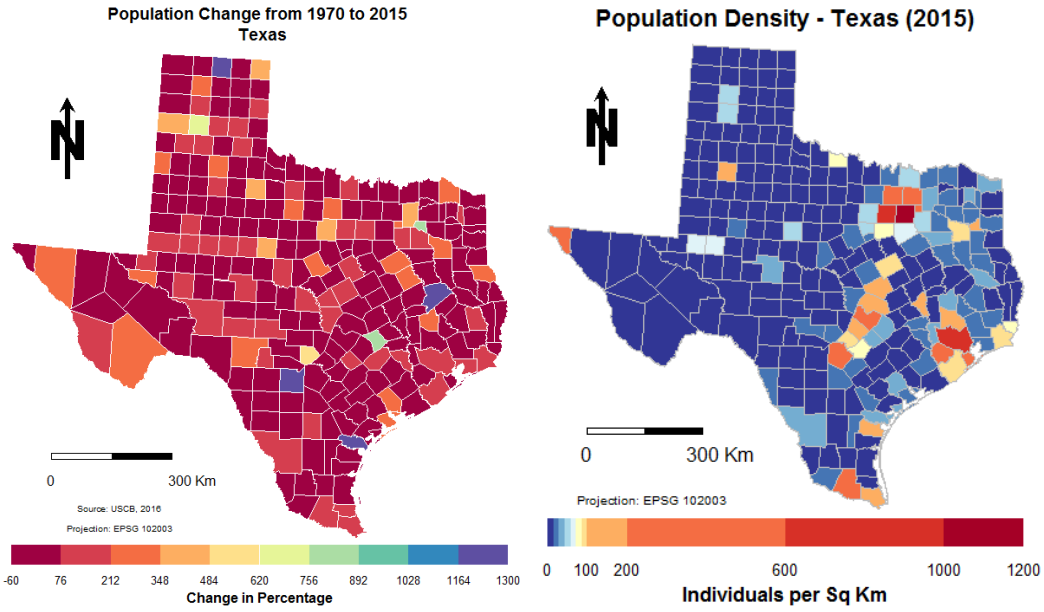


Figure 4-3. Left: Population Change given by percentage between the years of 2015-1970 for Texas; Right: 2015 Population density for Texas.

It is possible to see that from 1970 to the present day, in Texas, the population of more than half of the counties doubled, at least. The county of Ansford, Houston, Medina and San Patricio show an increase of population in the order of 1300% from the original values of 1970.

In order to account for the difference of area values between the counties, the unit used to input in the models was given by the population density (individuals per square kilometer).

Oklahoma shows a smaller percentage of change in the time period, which means that variability for this variable is lower spatially and temporally. The county with the highest percentage of change was Latimer, with 300% of positive change.

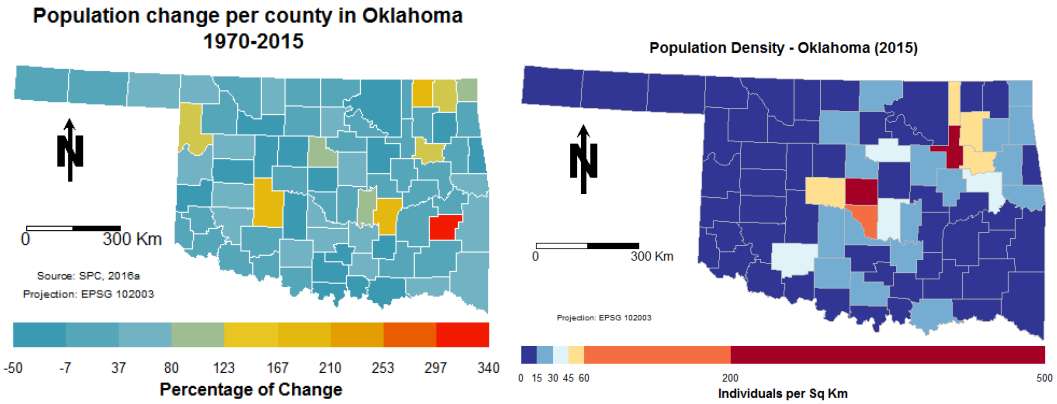


Figure 4-4. Left: Population Change given by percentage between the years of 2015-1970 for Oklahoma; Right: Population Density for the Oklahoma counties in 2015.

4.1.4 **Land Use/ Land Cover**

The land use/ land cover datasets were retrieved from USGS (2016 a, b, c and d), for the years of 1992, 2001, 2006, and 2011 respectively. Unfortunately, there was no other temporal resolution that could better represent the land cover changes for the period under study. To input land-cover data into the model two simplifications were needed. Attachment A.2. contains the python script that was used in ArcMap to process the datasets for Texas. The Oklahoma geoprocessing followed the same workflow.

First, it was needed some normalization of the land-cover to represent the differences of LC classes between the different counties.

First, the original datasets were characterized by a pixel resolution of 30x30m. Such a resolution created a very heavy file, hard to process in R. In this sense, each pixel was magnified 40 times, resulting in a resolution of 1.2 Km x 1.2 Km, given by the median values of the distribution of the original 40 pixels; the mean was not used, because it would not result in the desired values for land-cover.

Secondly, as shown in Table 4-1., the classes were not coincidental between the classifications for each year. Moreover, some classes had minimal differences amongst each other under the scope of this study, e.g., classes 41, 42, and 43. The interest is to understand if the occurrence of tornados is influenced by different kinds of land cover, not what different kind of plants would influence them. In this sense, the classes were generalized as shown in Table 4-1.

Table 4-1. Codes for each land cover type classification, and interdependence between categories. Code 1 corresponds to the classification type-keys used for the scope of this study, and it is a broad generalization of both code 2 and 3; Code 2 corresponds to the classification produced for 1992 (USGS, 2014a); Code 3 was shared by the classification produced for the years of 2001 - USGS (2014b), 2006 - USGS (2014c) and 2011, USGS (2014d).

Code 1	Classification	Code 2	Classification	Code 3	Classification
11	**Water**	11	Open Water	11	Open Water
		12	Perennial Ice	12	Perennial Ice
21	**Residential**	21	Low Intensity Residential	21	Developed, Open Space
				22	Low Intensity Residential
		22	High Intensity Residential	23	Medium Intensity Residential
		23	Commercial	24	High Intensity Residential
31	**Barren**	31	Bare Rock	31	Barren Land (Rock, Sand, Clay)
		32	Quarries/Strip Mines		
		33	Transitional		
41	**Forest**	41	Deciduous Forest	41	Deciduous Forest
		42	Evergreen Forest	42	Evergreen Forest
		43	Mixed Forest	43	Mixed Forest
51	**Low-grass**	51	Shrubland	51	Dwarf Scrub
				52	Shrub / Scrub
		61	Orchards/Vineyards		
		71	Grassland	71	Grassland/Herbaceous
				72	Sedge/Herbaceous
				73	Lichens
				74	Moss
		81	Pasture	81	Pasture
		82	Row crops	82	Cultivated Crops
		83	Small Grains		
		84	Fallow		
		85	Urban Grass		
91	**Wetland**	91	Woody wetlands	90	Woody Wetlands
		92	Herbaceous Wetlands	95	Herbaceous Wetlands

4.2 Description of Software used

The geoprocessing and statistical analysis had two main components: Data cleaning and uniformization and subsequent statistical analysis.

The first part was computed with ArcGIS Desktop 10.4, due to the fact that its toolboxes are quite unique and effective in what concerns to geoprocessing, and, especially in what concerns to raster files. The process was developed in python and presented in the attachments, so that the quality of generalizations, re-projections and geoprocessing can be tracked.

The second part, the core statistical analysis and graphical outputs was computed in R language, RStudio, R version 3.3.2. The code is shared in A.6 for point process analysis and A.7. for the lattice analysis. The packages used for both analyses are given in A.8.

5 METHODOLOGY

5.1 Point Processes

In the point process approach, the main objective was firstly to understand how the points are spatially distributed, and how this distribution can be spatially related to the value of the other covariates. The next step was to understand how they are distributed in time. These procedures represent an exploratory data analysis. Please refer to A.6. for the detailed code.

5.1.1 *Intensity*

The investigation of intensity of a point pattern is one of the first and most important steps in data analysis. In general, the intensity reflects the first moment (or expectation) analogous to the average of a population of numbers, and, compared to other properties of point processes, it requires few modelling assumptions (Baddeley et al. 2016).

The intensity is λ(u) for a spatial location u, where λ(u) is a function of location. In this sense, for any region B, one can imagine dividing B into pixels and calculate the expected number of points for each pixel, and adding up these expected numbers to obtain the expected total numbers of points in B. This summation is the integral of the intensity function:

$$\mathbb{E}[n(X \cap B)] = \int_B \lambda(u)du$$

For any region B, we can assume λ(u) as a surface, whose weights represent the intensity. From the equation, it is clear that the expected number of points falling in some region is equal to the volume under the surface. The values of intensity function are given in points per unit of area. This function can be estimated non-parametrically by kernel estimation.

The package in R used to perform the computation of intensity was spatstat (Baddeley et al. 2016). Its function 'density' computes a fixed-bandwidth kernel estimate (Diggle 1985) for the intensity function from a point pattern. By default, it computes the convolution of the isotropic Gaussian kernel (Normal distribution) of standard deviation of sigma, with point masses at each of the data points in x.

For this computation, it is needed a window of observation, which was defined with a buffer of 40 Km offshore, once tornados also happen in the sea, and they should be taken into

consideration. Attachment A.3. shows the python code used in ArcGIS Desktop 10.1 to compute the buffer and extract the points.

The intensity was computed adjusted with Diggle's improved edge correction (Diggle 2010), which have been proved to have better performance and accuracy (Jones 1993). For this case, the intensity of a point u is given by:

$$\hat{\lambda}(u) = \sum_i k(x_i - u) w_i e(x_i)$$

Where *k* is the Gaussian smoothing kernel, *e(u)* is an edge correction factor, and *w[i]* are the weights, which by default are 1.

Bandwidth selection controls the smoothing of the surface: a small value of sigma produces an irregular intensity surface, while a large value of sigma appears to oversmooth the intensity, increasing the bias and reducing the variance (Baddeley et al. 2016). In this sense, more estimations were made as an effort to improve bandwidth selection to compute the intensity function, selection of bandwidth was made using cross validation, firstly, using the method of Berman and Diggle (1989), which assume a Cox-process, and secondly using a likelihood cross-validation, which assumes a Poisson process[7].

5.1.2 *Intensity as a function of covariates*

For the next section, it is imperative to assume that the intensity of tornados is a function of a covariate Z. At any spatial location *u*, the intensity (λ(u)) of the point process is given by

$$\lambda(u) = \rho(Z(u))$$

Where Z(u) is the value of a covariate, and ρ is a function that reflects how the intensity of points depend on the value of the covariate. For a numerical covariate Z, the spatial distribution function is given by

$$G(z) = \frac{1}{|W|} \int_W 1\{Z(u) \leq z\}\, du$$

For a cumulative distribution function of the covariate Z(u) at a random point distributed in W (window of analysis) (Baddeley et al. 2012).

To quantify ρ having Z as elevation, the TPI dataset was inputted.

As for what concerns to the population, the time period was analyzed for 5 year intervals, and having Z as log10 of population.

To estimate the dependence of intensity as a function of land cover, the attributed codes presented in table 4.1. were used as a measure of the continuous covariate surface.

[7] For more information about the methods, please refer to the spatstat manual, available online at: https://cran.r-project.org/web/packages/spatstat/spatstat.pdf

5.1.3 *Correlation*

Correlation (or covariance) is a second moment quantity, and is usually explored to understand spatial dependence between points.

A very popular technique for analyzing spatial correlation is to use the K-function proposed by Ripley (1977), and is given by:

$$K(t) = A \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} I_{t(i,j)} / n^2$$

where n is the sample size, A is the area of the plot, wij corrects for edge effect (for more details, please refer to Badeley (2016).

If the resulting function has positive covariance, then it defines points that are clustered in space; a null value defines points that are completely randomly positioned, and negative covariance represent points that are equally sparse in space (Baddeley et al. 2016). For a better understanding on how the k-function works, it shown in Figure 5-1. Shows the traditional output of the spatial K-function.

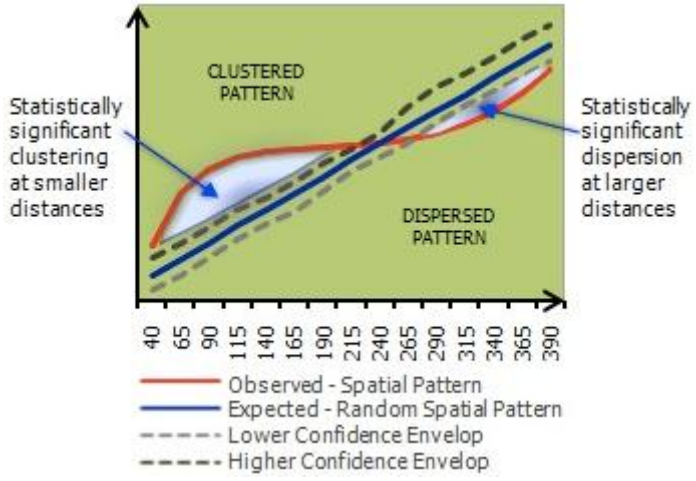


Figure 5-1. General output of the K-function, and its interpretation. The blue line indicates the expected random spatial pattern (Poisson). Red Line is the distribution of the sample under inspection. Envelopes represent the threshold for 95% confidence.

The observed point patterns are plotted against a theoretical distribution, under a selected model. The width of the envelopes is determined by finding the most extreme deviation from the theoretical *K*-function that is achieved by any of the simulated *K*-functions, at any distance *r* along the horizontal axis, to avoid "data snooping" (Baddeley et al. 2016).

In this sense, for each simulated dataset, the envelopes compute the maximum vertical deviation *D* between the graphs of graphs of *K* and *Kpois* over some range of distances. The envelopes are given by:

$$E\_(r) = K_{poisson}\ (r) - D_{max}$$

$$E_+(r) = K_{poisson}\,(r) + D_{max}$$

Whether or not the observed point patterns lie within the confidence threshold envelopes will demonstrate how (or not) the point processes are spatially clustered given the distance.

If there are suspicions that the point pattern under study is inhomogeneous, then the statistical analysis should take into consideration this lack of homogeneity (Diggle 2003). The method used to compute the Kinhom function for this study followed the generalization for the Kinhom function, to nonstationary point processes, proposed by Baddeley et al. (2000), and also implemented in 'spatstat'.

5.1.4 *Spatio-temporal Inhomogeneous K-function*

A separated spatial (or temporal) analysis itself might not evidence some relationships of interest. Therefore, there is the necessity of computing this function, once only the analysis of spatial-temporal patterns can realistically portray the spatial-temporal process under study; some relationships might occur too far apart in space (or in time) to be captured by the solely spatial or temporal analysis (Liang et al. 2010).

The spatio-temporal Inhomogeneous K-function is a second-order property that is used to analyse the spatio-temporal structure. It is useful to provide more insights about the behaviour of the point processes under study; these might present regularity, meaning that they scattered in space and time, or present clustering in both dimensions.

This method was proposed by Gabriel and Diggle (2009) and is given by:

$$K(u,v) = 2\pi \int_{-v}^{v}\int_{0}^{u} g(u^{'},v^{'})u^{'}du^{'}dv^{'}$$

Where $u = ||s - s'||$and $v = ||t - t'||$, and they denote spatial and temporal distances, respectively.

The equation was implemented by Gabriel (2014). She proposed an unbiased estimator for the spatio-temporal K-Inhomogeneous function, based on data with locations of events in $x_i : i = 1,\ldots n$ on a spatio-temporal region S × T, where S is an arbitrary polygon and T is a time interval:

$$\widehat{K}(u,v) = \sum_{i=1}^{n}\sum_{j\neq i}\frac{1}{w_{ij}}\frac{1}{\lambda(x_i)\,\lambda(x_j)}1_{\{\|s_i - s_j\|\leq u;\,|t_i - t_j|\leq v\}}$$

where $\lambda(x_i)$ is the intensity at $xi = (s_i, t_i)$ and w_{ij} is an edge correction factor to deal with spatial-temporal edge effects.

Theoretically, for an inhomogeneous spatio-temporal Poisson process, with intensity λ(x), its spatio-temporal Inhomogeneous K-function is $K(u,v) = 2\pi u^2 v$. In this sense, and accordingly to Gabriel (2014), $K(u,v) - 2\pi u^2 v$ can be used as a measure of the spatio-temporal aggregation or regularity, using an inhomogeneous Poisson process as benchmark. Thus, positive values of $K(u,v) - 2\pi u^2 v$ indicate clustering in space and time and negative values indicate regularity.

5.2 Lattice Approach

The R-code for this section is given in A.7.

5.2.1. Why INLA?

According to Simpson et al. (2011), Illian et al. (2012) and Krainski et al. (2016), there is a Latent Gaussian Random Field under the Log-Cox model assumption. In a log-Cox process models, the spatial variation is given by a random structure that presents continuity in space, and it is based on an underlying latent (or random field) ∧ (·) that describes the intensity of the point pattern, always assuming the independence amongst points in what concerns to this field (Illian et al. 2013), and the inference can be done with INLA (Krainski et al. 2016).

That being said, the point pattern can be described by the statistical model for complete spatial randomness, the Poisson process (Illian et al. 2008; Law et al. 2009).

In this sense, all inferences were based, for the i^{th} area, on:

$$y_i \sim Poisson\ (\lambda_i)$$

A common approach to fit the log-Cox process is to divide the study region into cells, that forms a lattice, and count the number of points into each one (Simpson et al. 2011). These counts are then modeled using the Poisson likelihood (Krainski et al. 2016).

For this case, the county level separation of Texas was used to represent the lattice proposed by Simpson et al. (2011).

5.2.2. Data manipulation and database construction

The data had to be uniformed for inputting into R-INLA framework. To accomplish it, a new database was build, to aggregate data into counties and order it spatially and temporally.

For each county (and FIP code) was created a spatial ID, that varies from 1 to 254, and is an unique identifier for each county, used throughout the whole analysis for simplification purposes.

Then, a neighboring matrix that expresses the adjacency between counties was determined by contiguity (Queen's rule), using functions from the spdep package. This matrix is shown in

Attachment A.4., and reflects all the spatial adjacencies between the counties, identified by the already attributed spatial ID. This matrix is the one that was used for all modelling purposes that had the spatial component.

The database was built having the variables grouped by year and county separately, sorted by date and then by spatial id. The variables are Population Density, Standard Deviation of TPI, and, for each land-cover class, the percentage of coverage for each county that varies temporally. The population has a single unique entry for each year and each county. the standard deviation of TPI is repeated over the years to all counties. The land-cover datasets had to be temporally generalized, according to table 5-1. This huge generalization had to be done, once there are no prior datasets that describe land-cover for Texas with some accuracy with a better temporal resolution. Even though it is a very rough generalization, it was expected that it could discriminate generally some differences in land-cover, that the model could recognize.

Table 5-1 Temporal Generalization for each landcover dataset

Dataset used	Time Period	Time span
1992	1970-1991	21
2001	1992-2000	8
2006	2001-2005	4
2011	2006-2015	9

5.2.3. Accessing model quality

5.2.3.1. DIC and WAIC

In order to access model quality for each stage of the modelling procedure, four model statistics were computed for each model.

The first one used was the Deviance Information Criterion (introduced by Spiegelhalter et al. 2002), defined by:

$$DIC = \overline{D} + p_D$$

Where D is the posterior mean of the deviance and Pd is the effective number of parameters in the model. A smaller value of DIC corresponds to a better model fit.

DIC is a really useful measure for comparison between models, but Gelman et al. (2013) suggest that comparison using Watanabe-Akaike information criterion (introduced by Watanabe 2010) is better, because it represents a more fully Bayesian approach for estimating the out-of-sample expectation. In this sense, the WAIC was also computed.

5.2.3.2. Brier score

Another measurement for model comparison is used in Jagger et al. (2015), and was adopted for this study. It reflects the mean squared difference between the predicted

probability and the actual count in each county for each year. The smaller the Brier score is, the better the model assessment.

5.2.3.3. Distribution of the random effects

The quality assessment based on the random effects was grounded on the premises that the distribution of the random effects (structured or unstructured in space or time) ensued from the model should be centred and symmetrical around zero, and normally distributed.

5.2.3.4. Assessment based on predictive scores

Predictive measures can be used to validate and compare models (Gelman et al. 2004). By using R-INLA, it is possible to compute the CPO and PIT. CPO are the Conditional Predictive Ordinates and are defined as (Martino and Rue 2009):

$$CPOi = \pi(yi \mid y_{-i})$$

Where the subscript -i indicates that the ith element of the vector is removed. Unusually small or large values of CPO indicate surprising observations.

PIT (Probability of Integral Transform) are the calibration for CPO's, given by:

$$PIT_i = Prob(y^{new}_i \leq y_i \mid y_{-i})$$

A small or large values indicate possible outliers.

5.2.3.4.1. Log Score on CPO

Jagger et al. (2015) suggest to use the cross-validated log-score for the values of CPO. A smaller value of this score indicates better model quality.

5.2.3.4.2. Test for uniformity on PIT values

According to Czado et al. (2007) the value of the PIT should follow a histogram that represents an uniform distribution. In this sense, the quality assessment was made by Cramer-Von-Mises Test of Goodness of fit for the uniform distribution, from package 'goftest'.

5.2.4. Modelling Technique and formulation

For the sake of simplicity, the nomenclature of each model in R-INLA and model correspondence are given in table 5-2. (Adapted from Bivand et al. 2015 and TRIP 2017).

Table 5-2 Correspondence between the latent models given by the package R-INLA and the name of the mathematical model.

Name in R-INLA :: f()	Model
Iid	Independent random variables
Besag	Intrinsic CAR (for spatial effects)
Besagproper	Proper CAR (for spatial effects)
Bym	Convolution Model (spatial effect plus random effect)

Table 5-2 (Cont)

rw1	Random-Walk order 1
rw2	Random Walk order 2
Mec	Classical measurement error model
Meb	Berkson measurement error model

For each model, the following were computed: Marginals for the linear predictor;Hyperparameters; Marginals for the latent field; DIC; CPO; PIT; Marginal Likelihood; Predictive ordinate andWAIC.

The modelling technique followed a workflow that started with the exploration of the general tendency of number of tornados as a function of year, and increases in complexity by first adding, first, the spatial component, and after the temporal one, with choice of covariates before the exploration of space-time interaction terms. This methodology is adopted by Blangiardo and Cameletti (2015), Blangiardo et al. (2012), and a similar approach is seen, e.g., in Jagger et al. (2015), Karpman et al. (2013), and DiMaggio (2015).

The first model formulation was number of yearly tornados as a function of year. This linear trend was computed just as an exploratory tool, for a better adjudication when choosing the temporal structure in the model.

After, models with spatial structure were computed. Firstly, it was computed what is known as **frailty** model. This model only has into consideration the random effect terms. It does not take covariates or time into account. Its known as frailty or susceptibility model, once it only reflects the characterization of an individual area, given some overall susceptibility (DiMaggio 2014), and is given by:

$$y_i = \beta_0 + u_i$$

where$\beta_o$is the intercept and $u_i$are the unstructured spatial random effects.

The second strategy was to extend to the spatial dependence. Here, the best model was BYM (Besag-York-Mollie formulation – Besag et al. 1991). This model is also known as a **convolution** model; it is a category of models that add a spatially-structured conditional autoregression term (v) to the a spatially unstructured heterogeneity random effect term (u) (DiMaggio, 2014). The linear model is then specified as:

$$y_i \sim Poisson\ (\lambda_i = e_i\theta_i)$$
$$\eta_i = \log(\theta_i) = \beta_0 + u_i + v_i$$
$$u \sim nl\ (0, \tau_v)$$
$$v \sim nl\ (\bar{v}_\delta, \tau_v/n_\delta$$

where:

a. The y_i counts in area i, have an independent and identical Poisson distribution and have an expectation in area i given by:

$$y_i, idd \sim Poisson\,(e_i\theta_i)$$

Where, e_i are the expected counts, and θ_i is the risk for the area i.

b. A logarithmic transformation (log (λ_i)) allows a linear, additive model of regression terms, along with
c. A spatially unstructured random effects component (u_i) that is iid, with a normal distribution and zero mean, and
d. A conditional autoregressive spatially structured component (v), where each neighborhood consists of adjacent spatial shapes that share a common border (DiMaggio, 2014).
e. β_0 is the intercept and it quantifies the average of tornados for the area (Blangiardo and Camelletti 2015; Blangiardo et al. 2013).

The u_i distributed or Gaussian random variation component (spatial unstructured random effects) is an effect characterized by a normally around the mean or the intercept. It represents mostly noise from the data that are not captured by them.

After this formulations, most of the authors generally proceed to the choice of covariates. Most of the times, they define and can explain a great part of the variability of the distribution of a variable.

The covariate STTPI is static in what concerns to time, but the same does not happen for the others – Population Density and the different kind of land cover classes, which vary both along the years. Consequently, the covariates were added after two models with spatio-temporal component were computed – a model with BYM formulation with an unstructured time component and a model with the BYM formulation with a structured time component, defined by the random walk type 1 model.

The covariates analyzed were – Population Density, Standard Deviation of the TPI, percentage of water as land-cover class, percentage of residential land-cover class, percentage of barren land-cover class, percentage of forest land-cover class, percentage of forest, percentage of low-grasslands and percentage of wetlands. The covariates selected in this step added some value to the overall tornado model.

After, the formulations used in the model were extended for a more broader spatio-temporal integration.

Firstly, the model was formulated with Bernardinelli et al. (1995) conception; this formulation expands the purely spatial nature of the BYM model to a spatio-temporal model. It is given by:

$$y_i = \beta_0 + u_i + v_i + (\beta + \delta_i) \cdot t$$

This formulation includes the same spatial structured and unstructured components as in the last model, a main linear trend β, which represents the global time effect, and a differential

trend δi, which identifies the interaction between time and space (Blangiardo et al. 2015). This specification assumes a linear effect on time for each area (δ_i).

Secondly, the Knorr-Held (Knorr-Held 2000) formulation was applied. This formulation releases the linearity in δ_i, and is given by:

$$y_i = \beta_0 + u_i + v_i + \gamma_t + \phi_t$$

Where β_0, u_i, and $v_i$have the same parameterization as in the Bernardinelli (2005) formulation, the term $\gamma_t$represents the temporally structured effect, modeled dynamically. $\phi_t$is specified by means of a Gaussian exchangeable prior: $\phi t \sim$ Normal(0, $1/\tau\phi$) (Blangiardo et al. 2013).

Also, for the purpose of this study, it was interesting to investigate the space-time interactions, which would explain differences of tornado occurrence along the space and time interaction trend for different areas. In this sense, the model presented for Knorr-Held can be extended for an interaction between space and time, such as:

$$y_i = \beta_0 + u_i + v_i + \gamma_t + \phi_t + \delta_{it}$$

where δ_{it} represents the interaction between space and time.

The interaction term can be defined in a number of different ways. Here, it is assumed that the effects v_i and ϕ_i interact. The precision matrix of the parameter δ_i defines the neighboring structure and is given by $\tau\delta \boldsymbol{R}\delta$, where the first is an unknown scalar and the second the precision matrix. The latter can be factorized as the Kronecker product of the structure matrix (Clayton 1996). There are four ways to define the structure matrix, as presented in Knorr-Held (2000), but, for the scope of this study, type I was computed.

It assumes that the two unstructured effects v and ϕ interact, and is given by:

$$R_\delta = R_v \otimes R_\phi = I \otimes I = I$$

Consequently, no spatial and/or temporal structure are assumed on the interaction either and, therefore, $\delta_{it} \sim$ Normal(0, $1/\tau\delta$).

For all these formulations, the basal expected count of tornados (e_i) was provided, for a better model performance. For the spatial model, the expected number of tornados was given by the mean for sum of tornados that occurred in the 46 years, which was an average of 26 tornados for each county, for Texas. The basal number of tornados, for year per county was given by 0.56 tornados per state, per year, and this was the parameter used for the spatio-temporal models. In the case of Oklahoma, the average of tornado occurrence is 32 per state, which divided by the 46 years gives a basal number of 0.71 per year per state.

5.2.5. About the model outputs

There are many outputs for R-INLA models. Here will be described the ones that were used to support each decision made along the workflow.

The first one are the random effects. These should be random in space, and its distribution should be symmetrical around zero.

The marginals of the fixed effects were also used to access the relationship between the covariates and the tornado occurrence. If these are centred around zero, then there is no relationship.

Then, three parameters were used for the model visualization: fitted values, spatial risk and spatial exceedance.

The fitted values are given by

$$\theta = \alpha + u + v$$

These represent the expected number of tornados, given the model under study. Under the R-INLA these are accessed in the model output in R-INLA at `summary.fitted.values`. For more information on the code, please, refer to the attachment A.7. A map (or maps per time) of the fitted values, simply shows how the model fits the data, or how much of the data can be explained by the model itself.

The spatial risk can be seen of how much of the tornado occurrence at the state-level is explained by the spatial disposition of its counties, or the number of tornados expected in a county, given its relation to its neighbours. It is given as:

$$\zeta = u + v$$

At the INLA output, these values are accessed by applying the `inla.emarginal`function over the marginal of the structured effects. This R-INLA function computes the overall expected values of the model for each area.

Lastly, the probability exceedance. This is useful for predictive measures. It returns the probability of the tornado occurrence being higher than a given value (in this case, x). It is given by:

$$\Pr(\zeta_i > x$$

To access these values, it was used `inla.pmarginal` function of R-INLA. This function computes the distribution function of the marginal of the model, given a value (or, more precisely, a threshold). If it is considered `1-inla.pmarginal` of this threshold then what is being calculated is it complementary occurrence – the distribution function of the occurrence being superior of this value.

6. RESULTS AND DISCUSSION

6.1. Point Processes

Attachment A.5. shows the point processes analyzed in this section for the period of 1970-2015.

Figure 6-1. shows the values of intensity computed for different values of bandwidth (σ). The Bandwidth of 150 000 was selected by default with Diggle's improved edge correction, bw.diggle and bw.ppl are selected by cross validation, and 100 000 was selected as a medium point between the default and the cross validation methods. As expected and shown on the surfaces, a bigger bandwidth over-smooths the density function surface and vice-versa.

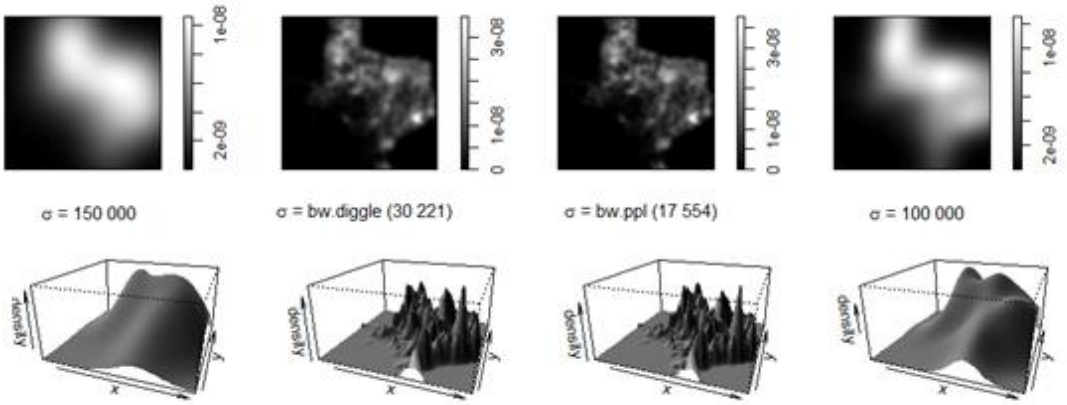


Figure 6-1. Graphic representation for the intensity function surface in 2-D (upper panels) and 3-D (lower panels), for different bandwidths.

The values of intensity are given, for this case, as the number of tornados per square meter (tornados · m^{-2}). A value of, e.g., $1x10^{-8}$ indicates a value of 0.01 Tornados per square Km [8]. Even for different bandwidths the results show the same: the density is not spatially homogeneous: in all intensity plots, the values vary in a great scale.

For the smaller bandwidths, it is possible to see with more detail where are the hotspots of intensity, and for the bigger values of bandwidth, these hotspots are more dissolved, giving a broader and generalized idea for the intensity.

Figure 6-2. shows the comparison of the standard deviation of intensities, computed for bandwidth 150 000 (a), 100 000 (b), and 50 000 (c) with Diggle's improved edge correction.

The standard error shows that, for smaller bandwidths, the error seems to be higher; in this sense, the choice of it should be a balance between the amount of smoothing desired and the error associated. The error increases as we spatially move towards the hotspots identified in (ultima)

Figure 6-3. shows the estimated function $\rho\hat{}(z)$ against covariate values *z*, in this case, for elevation and TPI, together with 95% confidence bands assuming an inhomogeneous Poisson

[8] $\frac{10^{-8}}{m^2} = \frac{10^{-2}}{Km^2} = 0.01 \cdot Km^{-2}$

point process. The plots indicate that the tornados are more likely occur at elevations until 1000 meters, and in more rough terrains than would be expected if the intensity was constant.

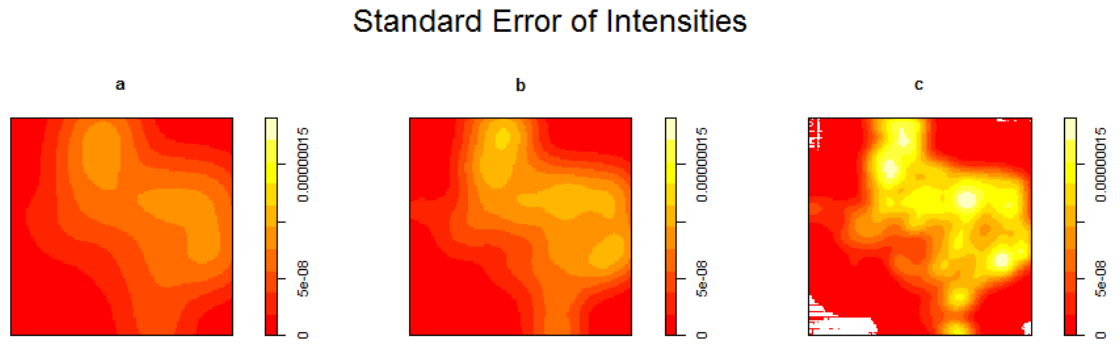


Figure 6-2. Surfaces of the standard error for intensity. a) from intensity function computed with bandwidth 150 000; b) from intensity function computed with bandwidth 100 000; c) from intensity function computed with bandwidth of 50 000.

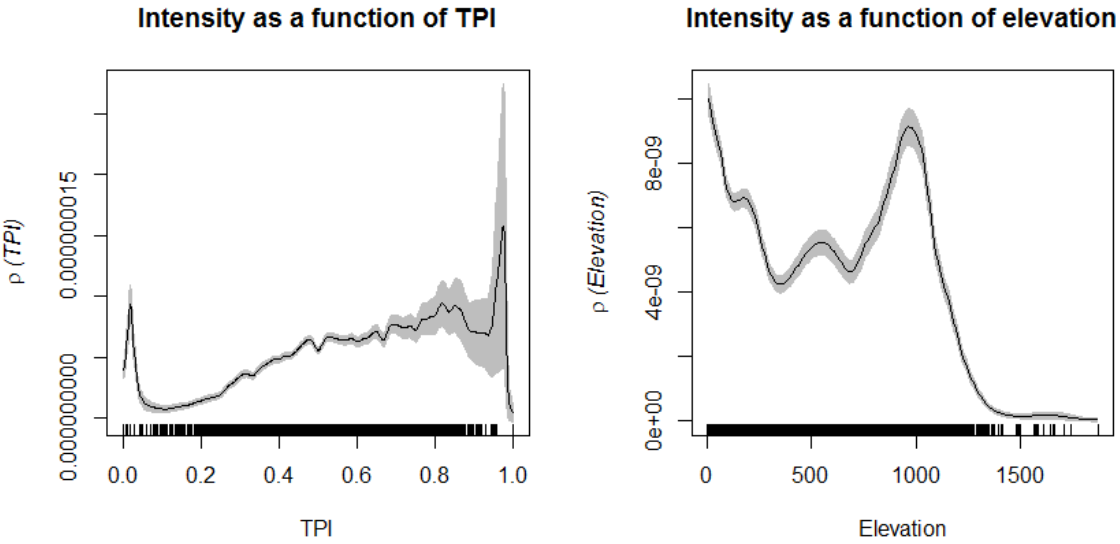


Figure 6-3. Intensity function ρ ˆ(z) against covariate values for elevation and TPI, together with 95% confidence bands assuming an inhomogeneous Poisson point process.

Figure 6-4. shows the estimated function ρ ˆ(z) against covariate values *z*, in this case, for the logarithmic scale of the population, together with 95% confidence bands assuming an inhomogeneous Poisson point process. If the number of tornados reports was dictated by the population amount, meaning that the reports increase as population increases, then a clear trend would be here expected: high values of ρ ˆ(z) for higher values of population.

This trend is visible on the charts for the periods of 1995-2010, but the same does not happen for the other ones. For example, for the period between 2010-2015, it seems that tornados are more likely to occur at places where the population is less. The same happens between 1970-1995.

Figure 6-5. shows the spatial inhomogeneous K-function [Kinhom^(r)] for the tornado point processes (black line), together with the theoretical *K*-function of the inhomogeneous Poisson process $Kpois(r) = \pi r2$, which serves as the benchmark of 'no correlation' (red-dotted line). Please, note that this visualization is the same for all different intensities computed above, which shows a very consistent behavior of the point process under study, and, for that reason, only one representation of this function is shown. From its visualization, it is possible to understand that the points are spatially correlated for values of distance between 50 Km and 200 Km in space, compared to the inhomogeneous Poisson case. The conclusion to understand from here is that the process of tornadoes exhibits spatial structure in form of interaction among the events. In this sense,and recalling figure 5-1., these values for $Kinhom^r$, comparing with the theoretical for the inhomogeneous Poisson process, indicate that the events exhibit a clustered pattern for distances between 50 and 200 Km. In simpler words, tornado events in Texas exhibit spatial clustering, characterized by a radius given by the referred distances.

As referred in section 5.1.4., the spatial inhomogeneous K-function is a measure of correlation that analyzes all events (all years) in space, but do not separate them in what concerns to the time dimension. Thus, a spatio-temporal technique is needed, once some relationships might end up not being found, if time as a dimension is not considered. In this sense, the spatio-temporal inhomogeneous K-function was computed (Figure 6-6), for different time and space thresholds.

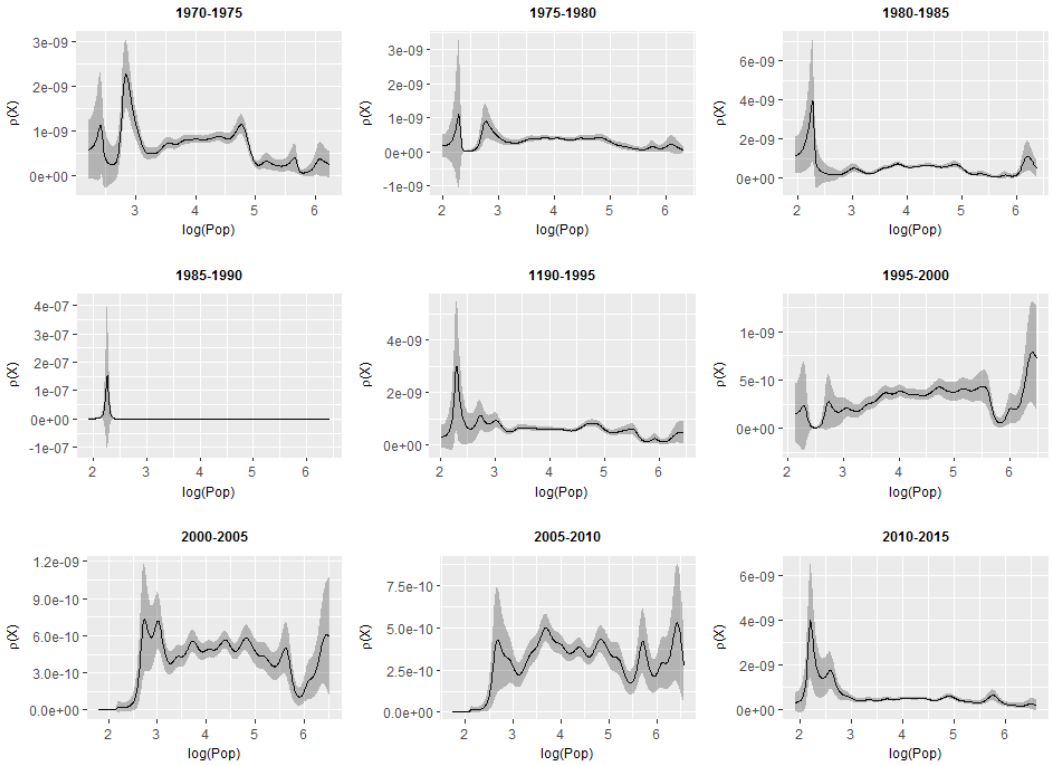


Figure 6-4. Estimated intensity function ρ ˆ(z) against covariate values for the logarithmic scale of the population, together with 95% confidence bands assuming an inhomogeneous Poisson point process.

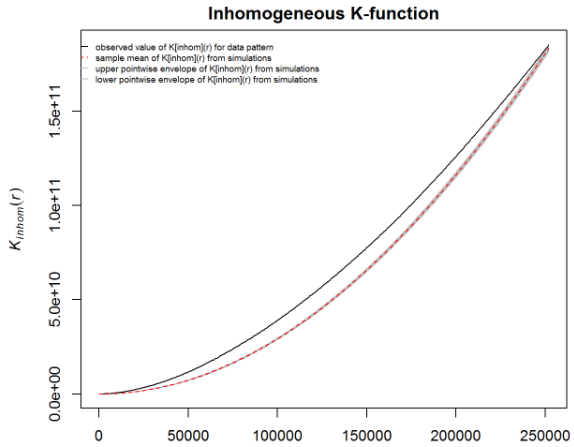


Figure 6-5. Inhomogeneous K-function, Kinhom^(r), for tornado point processes, together with the theoretical K-function of the inhomogeneous Poisson process Kpois(r) = πr2.

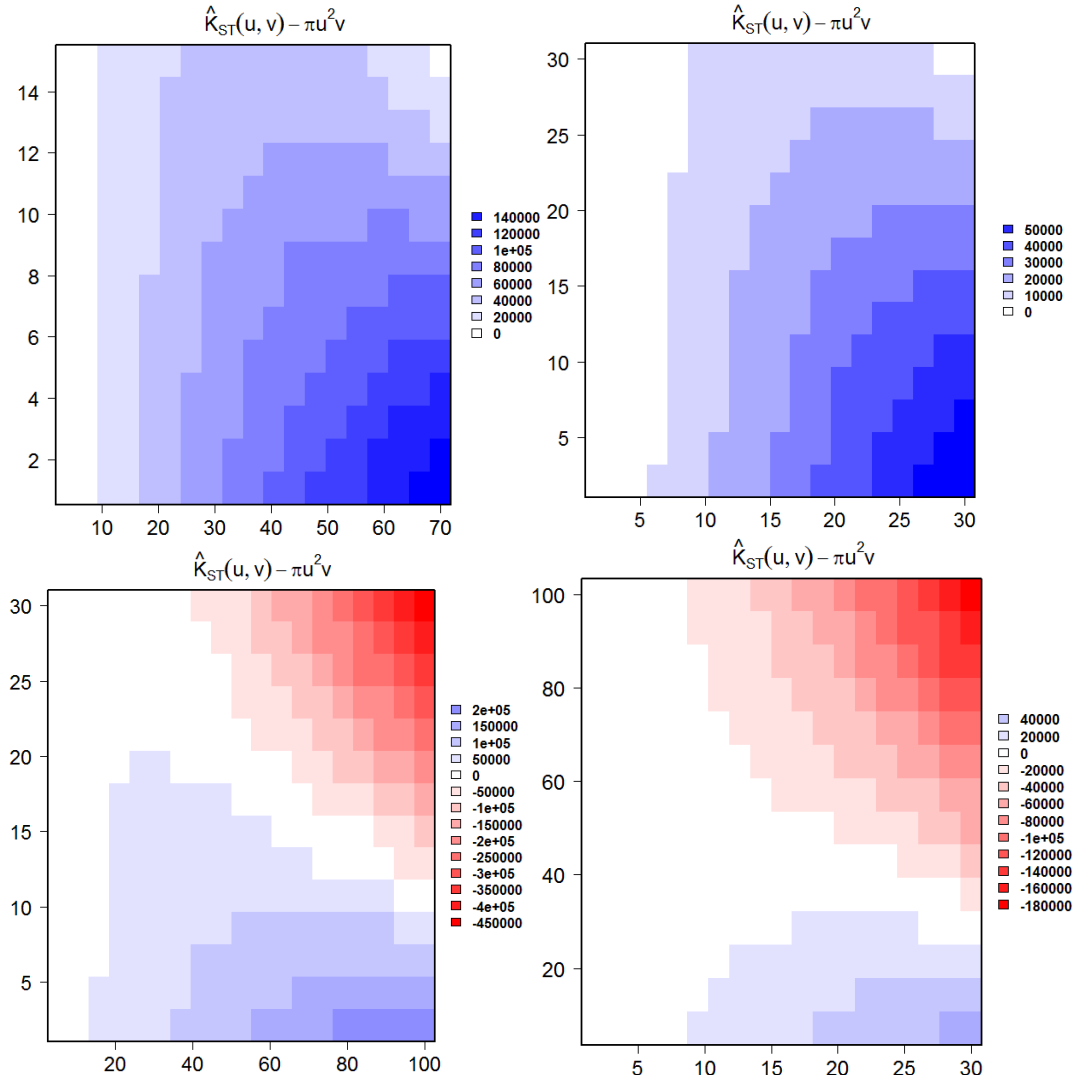


Figure 6-6. Inhomogeneous spatio-temporal K-function ($K(u,v) - 2\pi u^2 v$) for tornado occurrence in Texas. Top-left: for distance up to 70Km and time up to 15 days; Top right: time up to 30 days and distance up to 30Km; Bottom left: time up to 30 days and distance up to 100 Km; Bottom right: time up to 100 days and distance up to 30 Km.

The value of $K(u,v) - 2\pi u^2 v$ indicates that for small thresholds of space and time, the occurence of tornados is clustered. For example, from figure 6.6, upper images, it is evident that up to 70 Km, the events are clustered in space; for 30 Km, and time up to a month, the events are also clustered.

When the space is extended to values of 100Km, the events show clustering up to 10-20 days, for distances of 100 Km and 30 Km, respectively. If the time and space are increased more than these values (for example, more than 40 days (Figure 6.6, bottom right), then the tornados show regularity in space and time.

This analysis gives more insight than the spatial Inhomogeneous K-function (Figure 6.5.), where there is an apparent clustering from 50 Km to 200 Km, a value that can occur, but for a

very limited time threshold (for example, for 100Km, the clustering occur up to 10 days), as shown for the inhomogeneous spatio-temporal K-function.

6.2. Lattice Approach

From the results of the previous section, it is possible to assume that one possible model to describe the occurrence of tornados is the Cox processes. They are frequently applied for aggregated spatial point patterns where the aggregation is due to a stochastic environmental heterogeneity (Cressie 1993, Møller et al. 2017). In this sense, it is possible to use the Bayesian framework to modelling procedures[9].

Table 6-1. shows the DIC and WAIC values for the simple approach of modelling the total number of tornados as a function of time. From the values of DIC and WAIC, it is possible to understand that the occurrence of tornados is structured (related) in time. Indications for this temporal structural component were also given by the spatio-temporal Kinhom function presented in the last section. The models that best described this structure are iid (uncorrelated time), rw1 and meb (correlated time)[10].

Table 6-1. DIC and WAIC values for the models: number Tornados ~ year with different formulations: linear trend and non-linear trend for different year structure models

	Model	DIC	WAIC
Linear Trend	-	832	858
Non-Linear Trend	iid	401	392
	rw2	403	402
	rw1	402	397
	crw2	405	406
	mec	407	413
	meb	401	392

The models with the spatial component were subsequently computed. The resume of the results is given in Table 6-2.

The frailty model, is a simple random effects model; it was computed in order to access the individual area level of susceptibility, and to have a baseline to compare with other models. The DIC on the model is 1723, and WAIC is 1681. The log score is 4.8, brier score is 988.31, and the CVM test (on the PIT values)[11] p-value is $3e^{-10}$. The brier score is indicative of a poor model. The CVM also indicates a bad model, once the p-value should be superior to 0.10 at least, to indicate an uniform distribution.

The fixed effect, the intercept, has a mean of -0.2316 with standard deviation of 0.046 (95% CrI -0.30, -0.15). These values are translated to values of mean 0.79, with standard deviation

[9] Please refer to section 5.2.1.

[10] Please refer to section 5.2.4., table 5.2.

[11] In this context, the Cramer-Von-Mises performs a test where the null hypothesis is: "the PIT distribution being uniform".

1.047 (95% CrI 0.73, 0.86), when exponentiated [12]. This would, in turn, reveal that the tornado occurrence, overall for each state, in the period under study, increased 21%, at a rate of 0.45% per year.

The density of the distribution for random effects for this model, the term that is of interest, is given in figure 6-7,and shows that these are not really normally distributed, and, even though approximated, they are not symmetric around zero, which could attest that this model is not a good approximation to the reality. Figure 6-8 shows the map of these effects, for the same model. They indicate the overall risk for each county.The county more prone to tornado occurrence is Harris (marked in red), followed by Galveston, Hale, Brazoria and Tarrant[13].

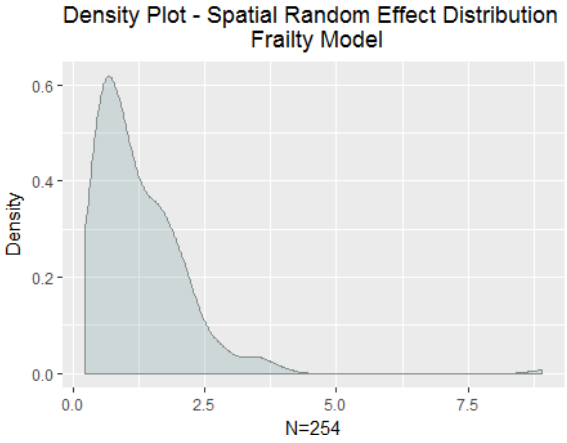
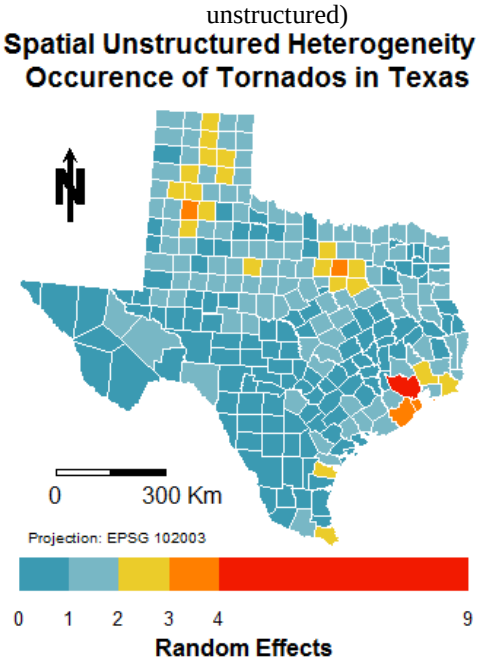


Figure 6-7Density plot for the spatial random effects distribution in the frailty model (spatially unstructured)



[12] This is done because the outputs of the models come in the log-linear form. In this sense, and for the sake of simplicity, the results shown from now on reflect always the exponentiated values from the outputs.

[13] Please, refer to A.9 to a map of Texas counties with respective FIP code.

Figure 6-8 Map of the random effects for Texas, described by the frailty model.

The next model to be computed was the convolution model. This model adds a structured spatial component to the frailty model. The results for this model include a DIC of 1705, and a WAIC of 1666, and the brier score slightly increases. The decrease on these values (except in the Brier score), compared to the frailty model, can attest that the insertion of the spatial structure in the modelling technique for tornados in Texas has a positive effect in the model quality. Nonetheless, the values of the log score (4.33) and CVM p-value still reflect a poor predictive quality. Figures 6-9a and 6-9b show the density plot for the random effects distribution of the model, and the density plot for the spatially structured effects distribution, respectively. Both should be symmetrical around zero and normally distributed, which happens until a certain extent – both are approximately centred around one. This indicates that there is other heterogeneity in the model that has to be taken into account.

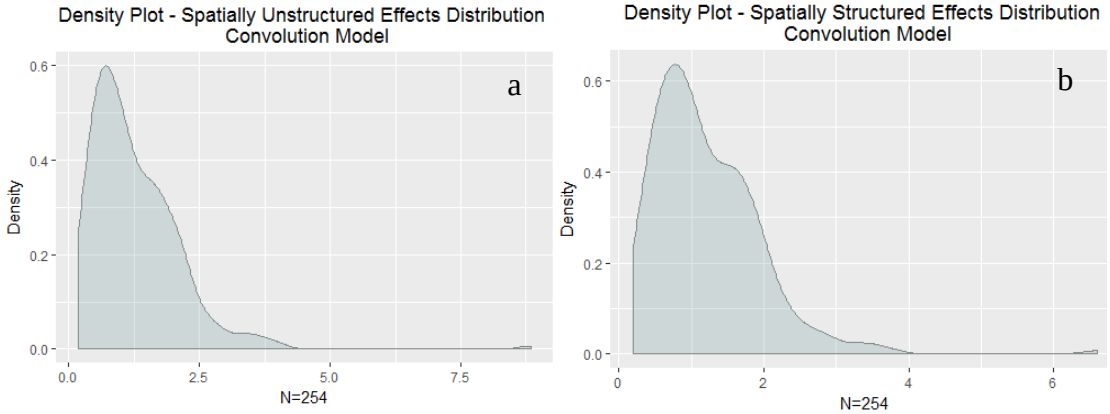


Figure 6-9 Density plot of the distribution of: a) Random effects; b) Spatially Structured Effects; in the convolution model

Table 6-2 Resume of the results for the spatial models: with spatial unstructured interaction (frailty) and with spatial structure (convolution)

Model	**DIC**	**WAIC**	**Log Score**	**Brier Score**	**CVM**	**Fixed Effects**			
						Intercept (Mean)	Standard Deviation	Quantiles **5%**	**95%**
FRAILTY	1723	1681	4.88	988.31	$3e^{-10}$	0.79	1.05	0.72	0.86
CONVOLUTION	1705	1666	4.33	988.48	$4e^{-12}$	0.79	1.02	0.76	0.81

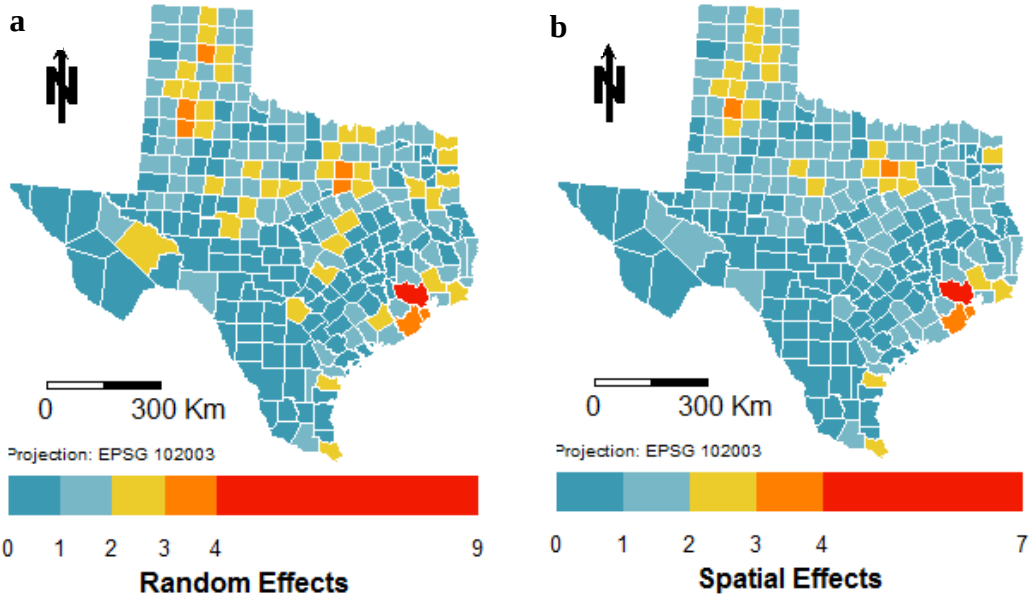


Figure 6-10 Map of the random effects for the convolution model b) map of the spatial structured effects for the convolution model.

Figure 6-10a and 6-10b show the maps for the random and spatial effects for this model, respectively. They are shown for a better understanding on how the modelling technique affects the way the occurrence is given for each county. In figure 6-10a, there is no spatial structure in how tornado occurrence is modelled, but, in contrast, the spatial neighbouring structure is taken into account in figure 6-10b, where the risk is different. In fact, the spatial structure seems to smooth the occurrence, even though the more prone counties are still highlighted. These are shown separately, but they can be merged into a single prediction model that reflects the overall tornado occurrence, given in Figure 6-11. The map of this value corresponds to θ, given by $\beta_0 + u_i + v_i$ and it represents the overall data that is now tailored by the model. A map of spatial risk (ζ) given by $\zeta = u_i + v_i$, can be also computed, from the values of the model marginals (Figure 6-12), and it displays the counties that have higher probability of tornado occurrence, under the scope of this model. With the spatial effects taken into account, we can now observe that the counties with more probability of having tornado occurrence are Harris, and Hale, followed by Galveston, Tarrant, Brazoria, Lubbock, Johnson Carson and Dallas.

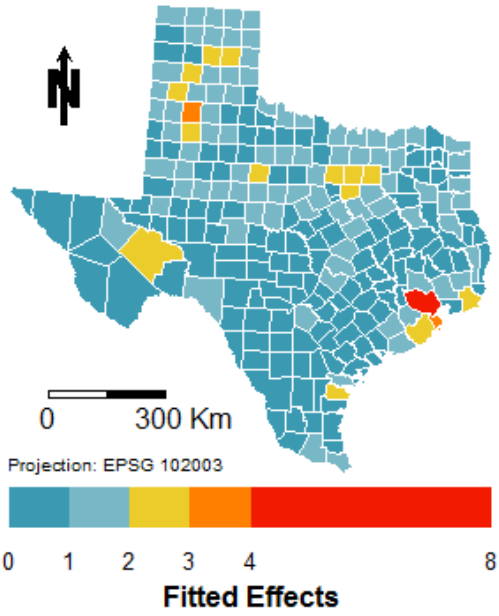


Figure 6-12 Fitted Effects (θ) for the convolution model

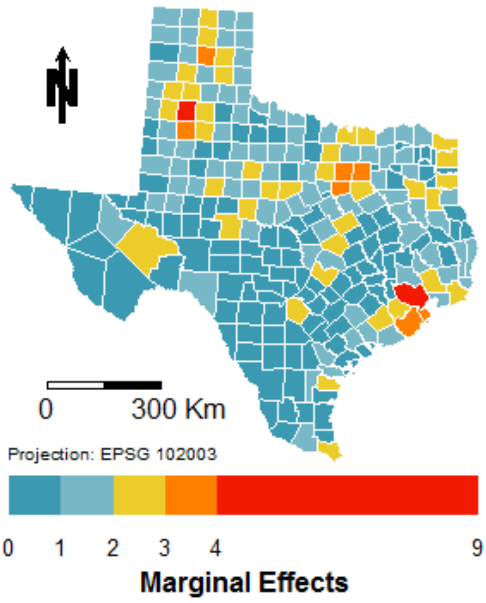


Figure 6-11Spatial risk ζ (probability of tornado occurrence) in Texas, given by the convolution model

An unstructured time component was given to the BYM formulation, and the subsequent model was computed. The DIC for this model is 23859, and the WAIC is 24075, with a log-score of 1.04. Even though the values of the DIC and WAIC are much higher than the latter models without the temporal component, the log score shows a better model fit. The brier score is 1.45 and the CVM p-value is 0.16. These last values suggest that the predictive quality of the model is better than the ones without the temporal component, once now the brier score is much less, and the PIT seem to be in an uniform distribution. The intercept is now 0.76 (95% CrI 0.70, 0.82).

Figure 6-13a, 6-13b, and 6-13c show the density plots for the distribution of the spatial random effects, spatial structured effects and the temporal unstructured effects for this model. All distributions show now a major concentration around zero. This is a sign of model improvement; even though they are not symmetrical: They exhibit a bump on the right tail that still reflects some geographical and temporal heterogeneity that is not being taken into account.

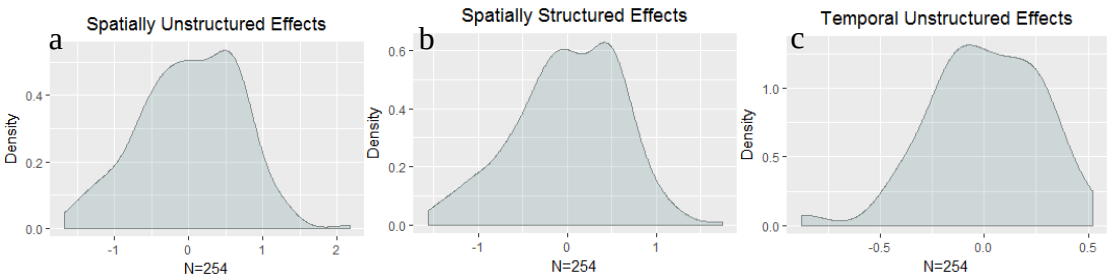


Figure 6-13 Density plots for the distribution of the random effects for the convolution model plus an unstructured time component: a) spatially random effects; b) spatially structured effects; c) temporal unstructured effects

The model formulated with the BYM and a structured temporal component, defined by the rw1 (the best model to describe the temporal simple model – Table 6-1) was subsequently computed. The DIC and WAIC are now 23861 and 24074, respectively, with a log score of 1.03. Even though the value for the DIC is greater, the value of WAIC is slightly smaller than the last model, where the time was unstructured. In this sense, and assuming that WAIC should be preferred over DIC, it is assumed that the structured time is the best temporal description for the tornado occurrence in Texas. The brier score is 1.45 and the intercept is now given by a mean of 0.76 (95% CrI 0.73, 0.79). Figure 6-14 shows the random effects for the different spatio-temporal components of the model.

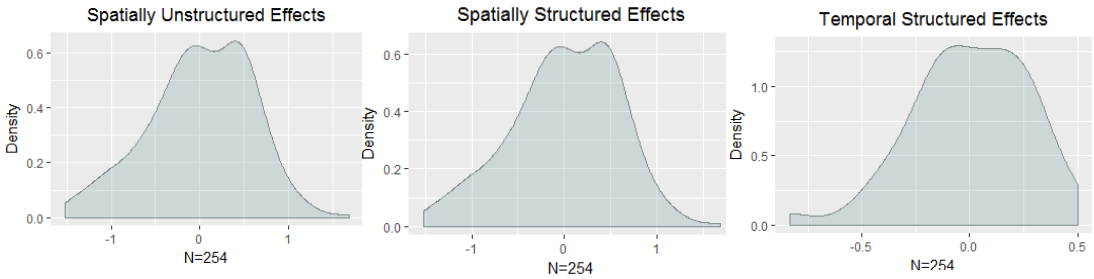


Figure 6-14 Density plots for the distribution of the (from left to right) spatially unstructured effects, spatially structured random effects and the temporal structured random effects for the model formulated with BYM plus a structured time component.

As for the last model, there is a component in the spatial variability that the model can not explain, given by the bimodal appearance of the spatial random effect. Nonetheless, it seems that it explains the temporal variability, given the almost perfect symmetry of the temporal random effects.

Figure 6-15 shows the averaged fitted effects for all years for this model. The insertion of the temporal structured component dissolved some of the fitted values. Some of the counties that had more tornado occurrence decreased in the fitted values class. This is due to the fact that now some of the occurrences can be explained by the temporal trend. The same happens for

the marginal terms, given in Figure 6-16. The maximum probability of tornado occurrence in a county was 9, in an upcoming year. But with the temporal component insertion, this maximum value decreased to 8, which, again, is justified by the fact that the temporal distribution affects the modelling process of occurrence of tornados.

These results can also be seen from another perspective. Figure 6.17 shows the map for the occurrence rate in Texas, and its random error. The areas (counties) with more concentration of occurrence are the ones that represent more variability. This is due to the simple fact that, the more samples are in a given state, the variability will, imperatively, increase.

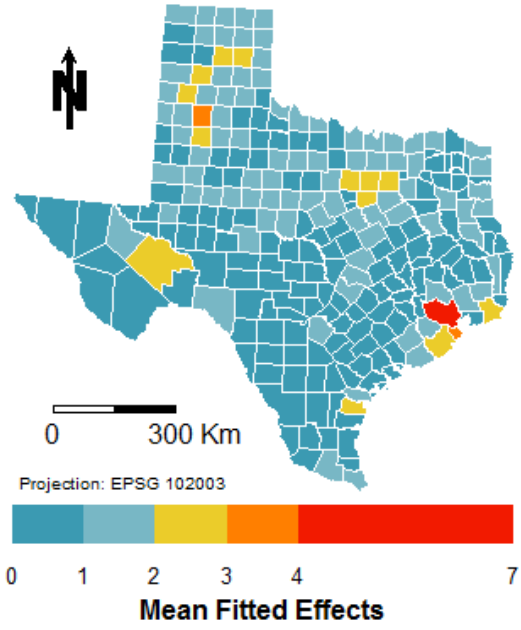


Figure 6-15Fitted Effects (θ) for the spatio-temporal model with structured time effects

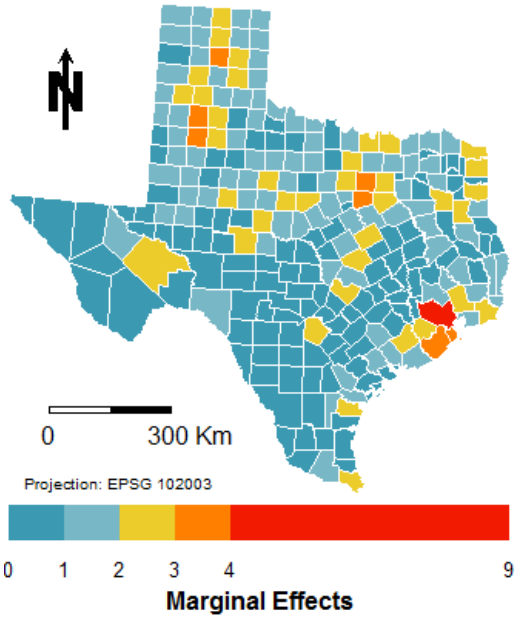


Figure 6-16Marginal effects ζ (Risk) for the spatio-temporal model with structured time effects

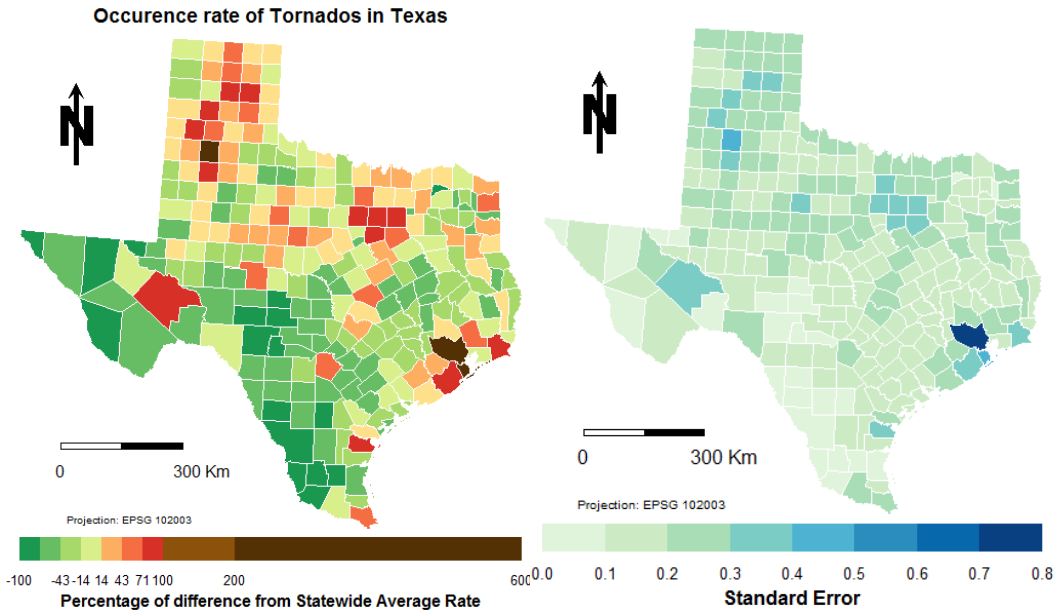


Figure 6-17 Left: Occurrence Rate of tornados in Texas (average over years; as a percentage of difference form the state average); Right: Standard Error of occurrence rate map

The covariates explain a great part of the variability of a spatio-temporal model, and in this sense, they were added in this step of the modelling process, before analysing the temporal component and adding more noise into the model. The resume of these results are given in table 6-3.

First, the population density term seems to decrease the model quality, given the values of DIC and WAIC, compared to the model without this covariate. This gives a contrary perspective on the studies that advocate that population density has an effect on the tornado occurrence, with an underlying idea that the more population a county has, the higher is the probability of a tornado being spotted and reported.

The measure of terrain roughness gives an improvement to the model, given the values of WAIC and DIC.

The percentage of land-cover type is more dubious to interpret. Given the values of WAIC, it seems that only the percentage of forest has a positive influence in the model quality. But, for the values of DIC, barren, forest and low-grass land cover types seem to make an improvement on the model, while water, wetlands and residential seem to not improve the models for the best.

In this sense, further indications of model quality with the insertion of covariates need to be taken into consideration. Thus, a model with spatial BYM formulation, plus a structured temporal component, plus the covariates altogether, was computed, and the distribution of the posterior marginals was accessed.

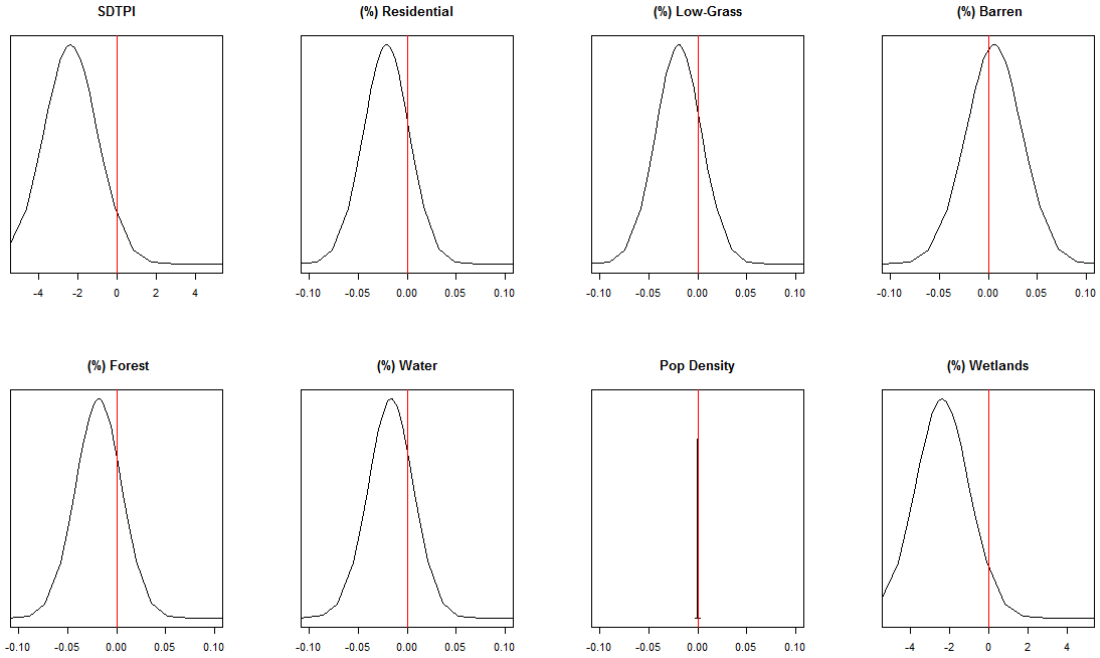


Figure 6-18 Posterior marginals of the model described spatially by the BYM, plus time structured as rw1, and the covariates. The red lines are the benchmark for "no correlation".

Table 6-3 Resume of the models created for the addition of a covariate to the spatio-temporal model. DIC, WAIC and BRIER – Brier-score – are measures for quality assessment of the model; LOG -log-score on the CPO and CVM p-value on PIT are measures for the predictive quality of the model assessment; Fixed Effects are inherent values to the models

Covariate added to the spatio-temporal model		DIC	WAIC	LOG	BRIER	CVM	Fixed Effects				
									Quantiles		
							Effect	Mean	ST	5%	95%
Population Density		23863	24077	1.03	1.45	0.16	Intercept	0.76	1.01	0.74	0.78
							Population Density	1	1	0.99	1
Standard deviation of the TPI		23860	24074	1.03	1.45	0.16	Intercept	0.95	1.18	0.72	1.26
							SDTPI	0.17	3.90	0.018	0.16
Land-cover class percentage	**Water**	23862	24078	1.03	1.45	0.16	Intercept	0,75	1,02	0,7282	0,789
							Water	1,003	1,004	0,995	1,01
	Residential	23862	24079	1.03	1.45	0.16	Intercept	0,76	1,02	0,73	0,795
							Residential	0,999	1	0,996	1,002
	Barren	23859	24076	1.03	1.45	0.16	Intercept	0,7561	1,02	0,73	0,78
							Barren	1,027	1,01	1,001	1,05
	Forest	23856	24073	1,03	1,45	0,16	Intercept	0,745	1,027	0,712	0,778
							Forest	1,001	1	1	1,002
	Low-Grass	23858	24076	1,03	1,45	0,16	Intercept	0,827	1,04	0,76	0,895
							Low-grass	0,9989	1	0,99	0,999
	Wetland	23863	24079	1,03	1,45	0,16	Intercept	0,7613	1,02	0,73	0,79
							Wetland	0,999	1	0,99	1

It is possible to see that the only factor that indeed does not affect, or has no relationship with the tornado occurrence is the population density. The terrain roughness and the percentage of wetland have the major contribution for the tornado occurrence. The rest of the land-cover classes seem to have some influence, as well, in the distribution of tornado occurrence.

Table 6-4 shows the summary for the fixed effects for this model.

Table 6-4 Resume of the fixed effects for the model formulated with spatial BYM and structured time with rw1, and all covariates.

	Mean	**Standard deviation**	**Quantiles**	
			5%	95%
Intercept	7.352	10.38	0.158	353.67
(%) Residential	0.979	1.02	0.941	1.01
(%) Barren	1.005	1.02	0.959	1.05
(%) Forest	0.982	1.02	0.94	1.01
(%) Water	0.983	1.02	0.94	1.02
(%) Low-grass	0.980	1.02	0.94	1.01
(%) Wetland	0.978	1.02	0.94	1.01
Terrain Roughness	0.094	3.934	0.009	0.89
Population Density	0.999	1	0.999	1

The mean values of the fixed effects indicate the tornado occurrence varies inversely to the increase of all covariates[14] except percentage of barren, which seems to facilitate the tornado occurrence, and except for population density, which seems to have no relationship with the tornado occurrence distribution (it would influence the tornado occurrence) only by 0.01%. In this sense, the population density will be the only covariate that will not be taken into account for further model formulations.

Now that the covariates are selected, it is time to explore the spatio-temporal trends for the tornado occurrence in Texas.

Table 6-5 shows the main model quality assessment parameters for the spatio-temporal models formulated by Bernardinelli, Knorr-Held and SPITI.

Table 6-5 Resume of the model quality parameters assessment for the spatio-temporal models with different formulation, with covariates

Model	**DIC**	**WAIC**	**Log Score**	**Brier Score**	**CVM**
Bernardinelli	23860	24081	1.03	1.45	0.16
Knorr-Held	23860	24081	1.03	1.45	0.16
SPITI	20693	20534	2.59	0.57	0.107

[14] e.g. for every percentpoint increase in the wetland land-cover type, there is a decrease of 2.2% in the tornado occurrence risk for that county.

The log-score and the CVM p-value indicate a worst model, but the brier score and the DIC and WAIC indicate that the STITI is the best fit. The huge difference from the other formulations in the latter values make this model acceptable, compared to Bernardinelli and Knorr-Held formulations.

In this sense, the SPITI with the covariates is the model that better fits the tornado occurrence in Texas, and is given by:

$$y_i \sim Poisson\,(\lambda_i = e_i\theta_i)$$

$$y_i = \beta_0 + u_i + v_i + \gamma_t + \phi_t + \delta_{it} + \beta_n x_n$$

Where β_0 is the intercept, u_i is the spatially unstructured random effects component with a normal distribution and zero mean, and v_i is a conditional autoregressive spatially structured component; term $\gamma_t$represents the temporally structured effect, modeled dynamically. $\phi_t$is specified by means of a Gaussian exchangeable prior: $\phi t \sim$ Normal(0, $1/\tau\phi$), and δ_{it} represents the interaction between space and time, which is unstructured, and $\beta_n x_n$ are the effects of the terrain roughness, and the percentage of the different land-cover types.

Table 6-6 shows the fixed effects for this model.

Table 6-6 Mean fitted effects for the SPITI model formulation with covariates, for Texas

	Mean	**Standard deviation**	**Quantiles**	
			0.25%	97.5%
Intercept	1.58	23.5	1.73	809.882
(%) Residential	0.993	1.03	0.93	1.05
(%) Barren	0.980	1.03	0.92	1.04
(%) Forest	1.01	1.03	0.94	1.09
(%) Water	0.985	1.03	0.92	1.04
(%) Low-grass	0.983	1.03	0.92	1.04
(%) Wetland	0.978	1.03	0.91	1.04
Terrain Roughness	0.264	3.91	0.01	1.93

These values indicate that all of the covariates have negative influence in the tornado occurrence, except for Forest, that for each percent of increase in land-cover, there is a decrease of 1% in the risk of tornado occurrence for that state.

The terrain roughness measures how steep the zone is. Under this scope, there is a theory, denominated the low-level flow hypotheses that advocates that a tornado is more likely to occur in zones where the low-level inflow is unimpeded. Some studies show that the terrain roughness indeed affects negatively the tornado occurrence, e.g., Leslie (1978). In this case, it is observable that for an unit increase in the SD of the TPI, the occurrence of tornados will decrease 26%. This is in accordance with Jagger et al. (2015), that uses another measure for the terrain roughness and arrived to a value of 18% tornado reduction for an unit increase in the terrain roughness for Kansas.

Moreover, all land-cover types that reduce the tornado occurrence (all except forest) are the ones that allow the low-level inflow to circulate. Forest works as a barrier for the tangencial velocity of tornados and specially for the low-level flow of the tornados. In this sense, the results make sense.

Figure 6-19 shows the mean spatio-temporal fitted effects, given by θ. These are the spatio-temporal mean of tornado occurrences that the model can explain for each county. In addition, in figure 6-19 is also given a spatial risk map, given by ζ. This is now interpreted as the residual relative risk for each area (compared to the whole state).

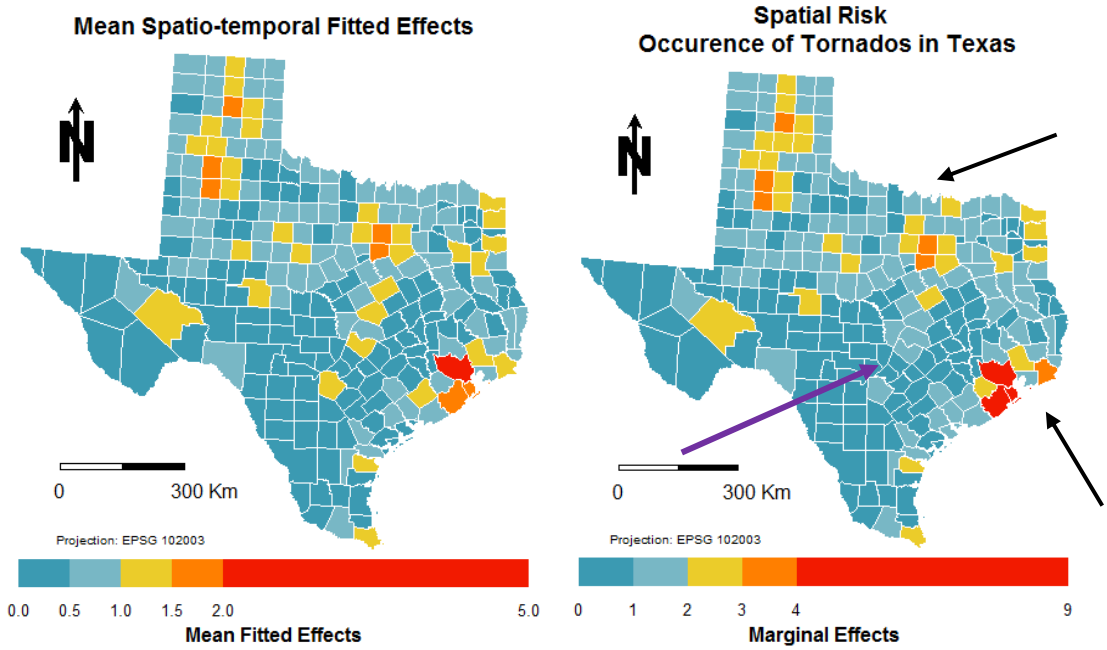


Figure 6-19 Left: Map of the overall fitted effects, averaged along the years. Right: Spatial risk for each area, compared to the whole state

These maps are shown, to attest that, by comparison with figure 6-15 and 6-16, the space and time interactions and the covariates have, indeed, effect on the overall tornado occurrence. It not only increased the risk for some counties, e.g., black arrow, but also smoothed other areas, e.g. purple arrow.

As an additional information, the probability of exceedance was computed for 1 and 3 tornados per state and the results are given in Figure 6-20. It is possible to observe that the north eastern part of Texas is the most problematic one, where, with a 100% certainty, there will be at least one tornado per county.

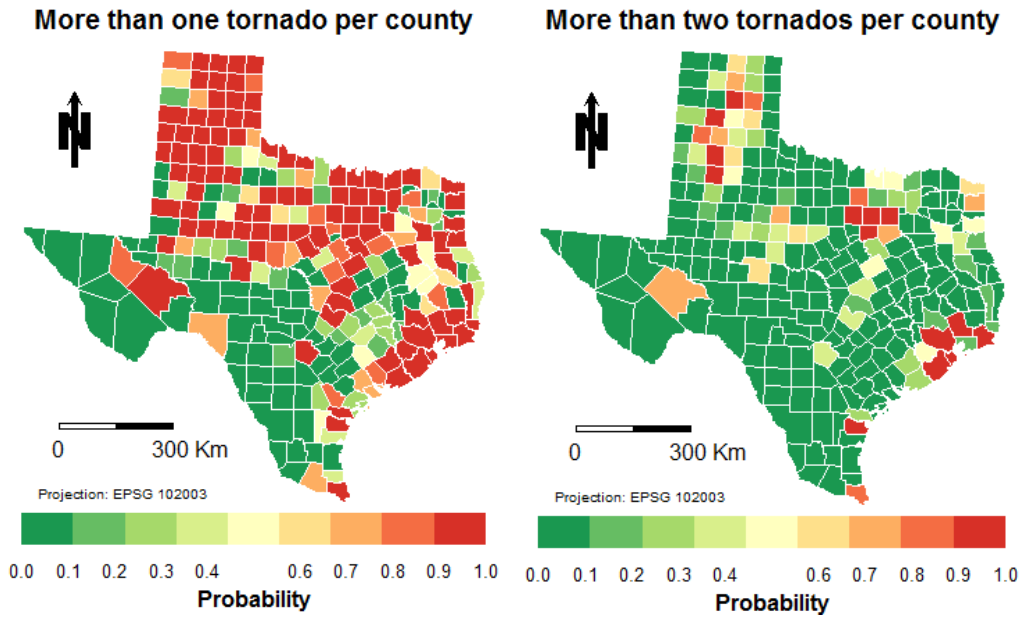


Figure 6-20 Bayesian Probability of tornado occurrence in Texas. Left: for more than one tornado per county; Right: for more than two tornados.

As now is established that the tornado occurrence can be given by the spatial and temporal component of its distribution, the exploration of results in Oklahoma started with a model formulated with spatial BYM plus time structured with random walk 1.

This gave a DIC of 8460, WAIC of 8562, log-score of 1.2, brier score of 1.59, and a CVM test p-value of 0.01. To this model was added the covariates of land-cover percentage and terrain roughness. The new DIC is now 8457, and 8566, with the log-score, brier score and CVM test p-value equal to the baseline model. These values indicate that the insertion of covariates affects the distribution of tornado occurrence in a small way. This postulation is also shown by the distribution of the posterior marginals, Figure 6-21, which shows that all covariates have little relationship with y_i (given the proximity of the peak of the distributions to zero). In addition, the covariate that most influenced the occurrence of tornados in Texas seems now to be the one that has less influence, together with the percentage of residential and wetland land-cover percentage. The Bernardinelli, Knorr-Held and SPITI with the covariates were also computed and the resume of the quality assessment model parameters are given in table 6-7.

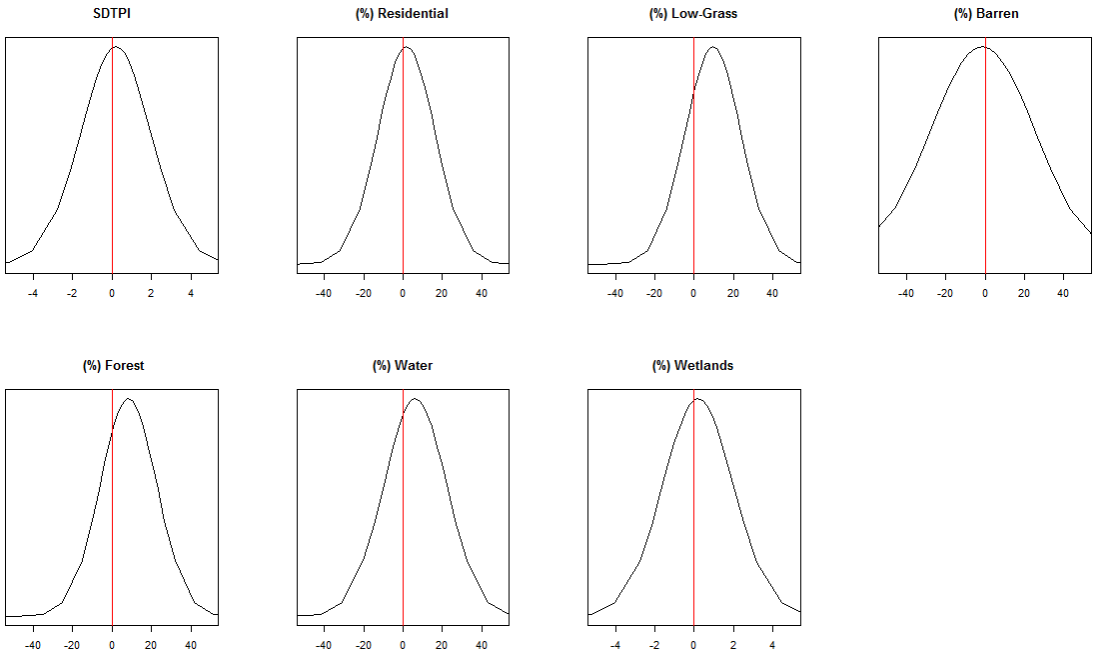


Figure 6-21 Posterior marginals of the tornado occurrence model in Oklahoma, described spatially by the BYM, plus time structured as rw1, and the covariates. The red lines are the benchmark for "no correlation".

Table 6-7 Results for the space-time formulations for Oklahoma with covariates

Model	DIC	WAIC	Log Score	Brier Score	CVM
Bernardinelli	8455	8564	1.2	1.59	0.01

Knorr-Held	8455	8564	1.2	1.59	0.016
SPITI	7339	7269	3.09	0.72	0.05

These values show that the SPITI is the best approximation to the tornado occurrence in Oklahoma, given the values of DIC, WAIC, brier score and CVM p-value. Regarding the latter, even though the p-value indicates that is unlikely that the distribution of the PIT values is uniform, it is still the better p-value for all models. Table 6-8 shows the mean of the fixed effects for this model.

Table 6-8 Resume of the mean fixed effects for the STITI for Oklahoma with covariates

	Mean
Intercept	1.1e-1
(%) Water	1.74e3
(%) Residential	3.56e2
(%) Barren	3.25e-7
(%) Forest	5.43e4
(%) Low-grass	2.62e5
(%) Wetland	3.83e-10
Terrain Roughness	2.35

The SPITI was also computed without the covariates, to access whether these have influence or not on the results. The results show a DIC of 7340, and a WAIC of 7270, which indicates that the covariates have little or no effect in the tornado occurrence for Oklahoma. A model with the STITI with the terrain roughness was also ran, to understand if this covariate has effect on the overall model, but the DIC and WAIC showed there is also no correlation (7341 and 7274, respectively).

These results show that the distribution of tornados in Oklahoma are not correlated with the percentage of the different types of land-cover, neither with the terrain roughness. Nonetheless, they are correlated in space and time,given by:

$$y_i \sim Poisson\,(\lambda_i = e_i\theta_i)$$

$$y_i = \beta_0 + u_i + v_i + \gamma_t + \phi_t + \delta_{it}$$

Where β_0 is the intercept, u_i is the spatially unstructured random effects component with a normal distribution and zero mean, and v_i is a conditional autoregressive spatially structured component; term $\gamma_t$represents the temporally structured effect, modeled dynamically. $\phi_t$is specified by means of a Gaussian exchangeable prior: $\phi t \sim$ Normal(0, $1/\tau\phi$), and δ_{it} represents the interaction between space and time, which is unstructured.

The only fixed term is the intercept, which is 0.39, with a standard deviation of 1.09 (97.5% CrI 0.33, 0.47). This value indicates the average risk across all counties, which correspond to one tornado per county per year. The spatial risk for this county is shown in figure 6-22. It is possible to see that this model shows that the risk increases as a gradient, from the outer parts to the inner parts of the county.

The probability exceedance of 1, is given in figure 6-23, that shows some spatial heterogeneity in what concerns to this probability.

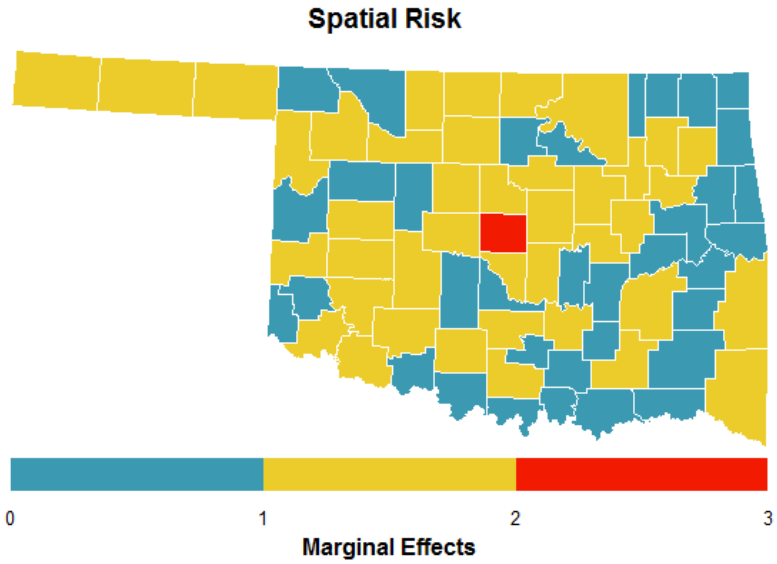


Figure 6-22 Spatial Risk for the tornado occurrence in Oklahoma, given the STITI model formulation, without covariates.

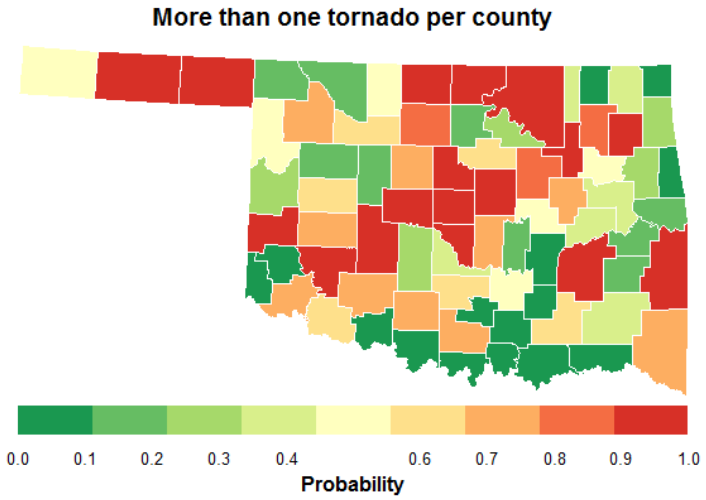


Figure 6-23 Bayesian Probability of tornado occurrence in Oklahoma. Left: for more than one tornado per county

7. CONCLUSIONS AND FUTURE WORK

Modelling tornado occurrence at the county level is not an easy and direct task, and the database heterogeneity is one of its biggest limitations. The registers depend, most of the times, on human eye and their quantification is based on a scale that has into consideration the damage produced, and approximated wind speed values (FS).

Moreover, there is a big loss of information during the process. For example, in this thesis, the tornado occurrence was averaged by year. This will make a loss of information in what concerns, for example, to seasonal variation, which could be a factor to take into consideration. Moreover, the occurrence was modelled at the area level, or more specifically, at the state level. Thus, there is also a loss of information in what concerns to the specific geographical information of each tornado.

On the other hand, the generalizations done on covariates – such as the computing of the standard deviation of the TPI as a measure of terrain roughness; or the generalization done on the datasets for land cover to enhance the pixel size – also contribute for loss of information that could be vital for the models. However, these generalizations are needed.

Despite the loss of information, and the fragility of the database that serves as a base of the modelling procedure, there were some few good outcomes from modelling tornados.

Firstly, it was shown that population density has no effect on the distribution of tornados, at the county level. This is contrary to some of the studies made before. But, specifically in the case of Texas, it is shown that, even for the point processes, population density has little or no effect on the distribution of tornados.

In addition, it was possible to understand that the tornado occurrence in Texas is best described by a spatio-temporal model that has an unstructured spatial component, a structured spatial component, an unstructured temporal component, a time structured component and an unstructured space-time interaction component, a formulation that had not been used so far in tornado occurrence modelling application, and that produced the best results. The same formulation when applied to another state produced the best results, as well. More studies should be done, in order to extend to another states, but this is a good start to define the spatio-temporal structure of the occurrence of tornados in USA.

For Texas, it was possible to define a model that takes all the above mentioned parameters, plus the land-cover percentage of coverage, plus the terrain roughness. The results can be explained by the low-level inflow hypothesis, and advocate that all land-cover classes, except forest, can enhance the occurrence of tornados in a given state. The surface roughness also attests the same hypothesis, showing that states with less plain land will have less occurrence of tornados.

For the case of Oklahoma, the covariates had no effect on the distribution of tornado occurrence. This could be explained by the big differences of area from one state to another. Oklahoma is much smaller than Texas, and this "zoom in" in the area level could be dissolving the effects of land-cover in the tornado occurrence in Oklahoma. Moreover, the covariate variability can be much smaller than Texas, which, in turn, will make the model less susceptible to them.

Even though it was not possible to extend the model that has the terrain characteristics into account for Oklahoma, it was possible to extend the spatio-temporal model, which is already a great adaptation.

Moreover, this modelling technique under the Bayesian framework is especially useful when thinking about risk maps and probability of occurrence maps to help decision makers in what concerns to disaster management and response. It is even more important in the case of Texas, where these maps have into consideration the land-cover types.

Thus, it is possible to say that, from a scattered and heterogeneous database, it is possible to create spatio-temporal models that can be applied not only at the county level, but as well as the state level. The application of these spatio-temporal models should be in the future extended to more states, and compared. Moreover, the USA covers a great area, and probably, as one gets away from the tornado alley zone, the spatio-temporal structure will be different.

This is one of the biggest problems of tornado modelling - tornados occur scattered in space and time. And the time and space span that is in between can be huge – from kilometres to months or even years. So the statistical models used should be prepared for this kind of heterogeneity. In this context, the Bayesian probability is a good approach, because it does not rely on the frequentist approach, but yes on the inverse probability, which allows the

"inference on unknown quantities, adapt the models, make predictions and learn from data" (Ghahramani 2012).

A great future direction on this study would be to model the occurrence of tornados having into consideration atmospheric variables. It would be interesting to derive measurements from, for example, MODIS, and by performing inference from certain measurements, reach to an approximation to an atmospheric pressure spatio-temporal dataset, in the form of, for example, an index. This covariate could be inserted in the model, in conjugation with temperature products, or radiance.

Another measure that could be interesting to input in the modelling technique would be the concentrations of certain particulate matters. More interesting would the modelling of a complete monthly spatio-temporal dataset, for a time period of, let us say 10 years.

Another covariate to relate the occurrence of tornados would be indexes from remote sensed data, such as the Normalized Difference Vegetation Index, or the Normalized Building Vegetation Index, to better understand how land-cover can be related to the occurrence of tornados.

Another interesting way to view the problem would be to insert a new dimension. In this study, the dimensions are: Cartesian coordinates (x and y) plus time (t). What about if altitude (z) is inputted in the model as a fourth dimension? This would mean to have into consideration the column of mass of air that composes the Troposphere or the Stratosphere, and its composition. In fact, is easier to find column data (for an area w, we find the values for x, y and z).

8. BIBLIOGRAPHIC REFERENCES

Akers, C., Smith, N., Shifa, N. (2014) Multinomial Logistic Regression Model for Predicting Tornado Intensity Based on Path Length and Width. *Scientific and Academic Publishing*, 4, 2, 61-66.

Ascione, A., Cinque, A., Miccadei, E., Villani, F., Berti, C. (2008) The Plio-Quaternary uplift of the Apennine chain: New data from the analysis of topography and river valleys in Central Italy. *Geomorphology*, 102, 105-118.

Bayes, T., Price, R. (1763) "*An Essay Towards Solving a Problem in the Doctrine of Chances. By the late Rev. Mr. Bayes, communicated by Mr. Price, in a letter to John Canton, MA. and F.R.S*." Philosophical Transactions of the Royal Society of London, 53, 370–418.

Baddeley, A., Moller, J., Waagepetersen, R. (2000) Non- and semiparametric estimation of interaction in inhomogeneous point patterns. *Statistica Neerlandica* 54, 329–350.

Baddeley, A., Chang, Y.-M., Song, Y. and Turner, R. (2012) Nonparametric estimation of the dependence of a point process on spatial covariates. *Statistics and Its Interface* 5 (2), 221–236.

Baddeley, A., Rubak, E. Turner, R. (2015) *Spatial Point Patterns: Methodology and Applications with R*. Chapman and Hall/CRC Press.

Bernardinelli, L., Clayton, D., Pascutto, C., Montomoli, C., Ghislandi, M., Songini, M. (1995) Bayesian analysis of space-time variation in disease risk. *Statistics in Medicine*, 14(21–22), 2433–2443.

Blangiardo, M., Cameletti, M., Baio, G., Rue, H. (2013) Spatial and spatio-temporal models with R-INLA. *Spatial and Spatio-temporal Epidemiology*, 4, 33–49.

Blangiardo, M., Cameletti, M. (2015) *Spatial and Spatio-temporal Bayesian Models with R-INLA*. John Wiley & Sons.

Bluestein, H. B. (2013) *Severe Convective Storms and Tornadoes: Observations and Dynamics*. Springer-Verlag Berlin Heidelberg

Bivand, R., Gómez-Rubio, V., Rue, H. (2015) Spatial data analysis with R-INLA with some extensions. *Journal of Statistical Software*, 63(20), 1–31.

Boruff, B. J., Easoz, J. A., Jones, S. D., Landry, H. R., Mitchem, J. D., Cutter, S. L. (2003) Tornado hazards in the United States. *Climate Research*, *24*(2), 103–117.

Braulio-Gonzalo, M., Bovea, M. D., Ruá, M. J., Juan, P. (2016) A methodology for predicting the energy performance and indoor thermal comfort of residential stocks on the neighbourhood and city scales. A case study in Spain. *Journal of Cleaner Production*, 139, 646-665.

Berman, M., Diggle, P. (1989) Estimating weighted integrals of the second-order intensity of a spatial point process. *Journal of the Royal Statistical Society, series B*, 51, 81–92.

Besag J., York J., Mollie, A. (1991) Bayesian Image Restoration, with Two Applications in Spatial Statistics.*Annals of the Institute of Statistical Mathematics*, 43(1), 1–59.

Breivik, O. N., Storvik, G., Nedreaas, K. (2017) Latent Gaussian models to predict historical bycatch in commercial fishery. *Fisheries Research*, 185, 62-72.

Brooks, H., Doswell, C. (2000) Normalized Damage from Major Tornadoes in The United States: 1890-1999 [Online]. *NOAA: Weather and Forecasting*. Available at: http://www.nssl.noaa.gov/users/brooks/public_html/damage/tdam1.html

Brooks, H., Carbin, G., Marsh, P. (2014) Increased variability of tornado occurrence in the United States. *Science*, 346, 6207, 349-352.

Brooks, H. (2014) The U.S. Gets More Tornadoes Than Anywhere Else In The World. But Why?. Interview conducted by Ria Misra [Online]. Available at: http://io9.gizmodo.com/the-u-s-gets-more-tornadoes-than-anywhere-else-in-the-1561468276 [Accessed January 2017]

Clayton, D. (1996). Generalised linear mixed models. In Gilks, W., Richardson, S., Spiegelhalter, D. (1996*) Markov Chain Monte Carlo in Practice*. Chapman & Hall.

Concannon, P., Brooks, H., Doswell, C. (2000) Climatological Risk of Strong and Violent Tornadoes in the United States [Online]. Available at: http://www.nssl.noaa.gov/users/brooks/public_html/concannon/ [Accessed January 2017].

Coleman, T., Dixon, P. G. (2014). An objective analysis of tornado risk in the United States. *Weather and Forecasting*, *29*(2), 366–376.

Coleman, T., Knupp, K., Spann, J., Elliot, J., Peters, B. (2011) The history (and future) of tornado warning dissemination in the United States. *Bulletin of American Meteorological Society*, 92, 567 – 582.

Corfidi, F. (1999) The Birth and Early Years of the Storm Prediction Center. *Weather and Forecasting*, *14*(4), 507–525.

Craigmile, P. F., Guttorp, P. (2011) Space-time modelling of trends in temperature series. *Journal of Time Series Analysis*, *32*(4), 378–395.

Cressie, N. (1993) *Statistics for Spatial Data*. New York: John Wiley & Sons.

Cusack, S. (2014) Increased tornado hazard in large metropolitan areas. *Atmospheric Research*, *149*, 255–262.

Czado, C., Gneiting, T., Held, L. (2007) Predictive model assessment for count data. *Technical Report 518, University of Washington, Dep. of Statistics*.

Dar, A. (2008). *Insight Guide Texas*. New York: Langenscheidt Publishers.

Díaz-Avalos, C., Juan, P., Serra-Saurina, L. (2016) Modeling fire size of wildfires in Castellon (Spain), using spatiotemporal marked point processes. *Forest Ecology and Management, 381*, 360–369.

Diggle, P. (1985) A kernel method for smoothing point process data. Applied Statistics *(Journal of the Royal Statistical Society, Series C)*, 34, 38–147.

Diggle, P. (2003) *Statistical Analysis of Spatial Point Patterns*. London: Hodder Arnold.

Diggle, P. (2010) Nonparametric methods. From: Gelfand, A., Diggle, P., Fuentes, M., Guttorp, P. (eds.) *Handbook of Spatial Statistics*. Florida: CRC Press.

DiMaggio, C. (2014) Notes and Codes for small area NYC pedestrian injury spatio-temporal analyses with INLA [Online] Available at: http://www.injuryepi.org/resources/spatial/inlaEpidemCode.pdf [Accessed January 2017].

DiMaggio, C. (2015) Small-area spatiotemporal analysis of pedestrian and bicyclist injuries in New York City.*Epidemology*, 26 (2), 247-254.

Dingus, A. (1981). *The book of Texas Lists*. Austin: Texas Monthly Press.

Dixon, P. (2002) *Ripley's K-function*. In: El-Shaawari, A., Piergorsch, W. (2002) Encyclopedia of Environmetrics. Chichester: John Wiley & Sons, Ltd.

Dixon, R. W., Moore, T. W. (2012) Tornado Vulnerability in Texas. *Weather, Climate, and Society*, *4*(1), 59–68.

Dixon, P.G., Mercer, A. E., Choi, J., Allen, J. S. (2011) Tornado Risk Analysis: Is Dixie Alley an Extension of Tornado Alley? *American Meteorological Society*.

Doswell, C. (2007) Small sample size and data quality issues illustrated using tornado occurrence data. *Electronic Journal of Severe Storms Meteorology*, 116(2), 1–10.

Elsner, J., Michaels L., Scheitlin, K., Elsner, I. (2014) The decreasing population bias in tornado reports. *Weather, Climate, and Society*, 5, 221–232.

Elsner, J., Jagger, T., Elsner, I., 2014. Tornado intensity estimated from damage path dimensions. *PLoS One* 9 (9).

Elsner, J., Murnane, R., Jagger, T., Widen, H. (2013) A spatial point process model for violent tornado occurrence in the US Great Plains. *Mathematical Geoscience*, 45 (6), 667–679.

Elsner, J. B., Fricker, T., Jagger, T. H., Mesev, V. (2016) Statistical Models for Predicting tornado rates: Case studies from Oklahoma and the mid-south USA. *Internal Journal of Safety and Security Engineering*, 6 (1), 1-9.

Elsner, J., Michaels, L., Scheitlin, K., Elsner, I. (2013) The Decreasing Population Bias in Tornado Reports across the Central Plains. *Weather Climate Society*, 5, 221–232.

Fujita, T. (1971). Proposed characterization of tornadoes and hurricanes by area and intensity. *Satellite and Mesometeorology Research Project Research Paper*. Univeristy of Chicago.

Friston, K. J., Jezzard, P., Turner, R. (1994) Analysis of functional MRI time-series. *Human Brain Mapping, 1*(2), 153–171.

Gabriel, E., Diggle P. (2009) Second-order analysis of inhomogeneous spatio-temporal point process data. *Statistica Neerlandica*, 63, 43-51.

Gabriel, E. (2014) Estimating second order characteristics of inhomogeneous spatio-temporal point processes: influence of edge correction method and intensity estimates. *Methodology and computing in Applied Probability*, 16, 1.

Gelfand, A., Diggle, P., Fuentes, M., Guttorp, P. (2010) *Handbook of Spatial Statistics*. Chapman & Hall.

Gelman, A., Carlin, J., Stern, H., Rubin, D. (2004) *Bayesian Data Analysis*. Texts in Statistical Science Series. Florida: Chapman & Hall/CRC.

Gelman, A., Hwang, J., Vehtari, A. (2014) Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, *24*(6), 997–1016.

Gensini, V., Mote, T. (2015) Downscaled estimated of late convective weather in the U.S. using dynamic downscaling. *Electronic Journal of Severe Storms Meteorology*, 6, 1-40.

Gómez-Rubio, V., Bivand, R., Rue, H. (2014) Spatial models using Laplace Approximation methods. In: Fisher, M., Nijkamp, P. (2013*) Handbook of Regional Science*. Berlin: Springer.

Gómez-Rubio, V., Cameletti, M., Finazzi, F. (2015) Analysis of massive marked point patterns with stochastic partial differential equations. *Spatial Statistics*, *14*, 179–196.

Grazulis, T. P. (2003). *The Tornado Nature's Ultimate Windstorm*. University of Oklahoma Press.

Hajek, A., Hartmann, S. (2010) Bayesian Epistemology. In: Dancy, J., Sosa, E., Steup, M. (eds). *A companion to Epistemology*. Oxford: Wiley-Blackwell.

Hart, A. (1993) SVRPLOT: A New Method of Accessing and Manipulating the NSSFC Severe Weather Data Base. Preprints of the 17th Conf. On Severe Local Storms, St. Louis. *American Meteorological Society*, 40-41.

Hartmann, S., Sprenger, J. (2010) Bayesian Epistemology. In: Bernecker, S., Pritchard, D. (eds) *The Routledge Companion to Epistemology*. London: Routledge.

Illian, J., Martino, S., Sørbye, S., Gallego-Fernandéz, J., Zunzunegui, M., Esquivias, M., Travis, J. (2013) Fitting complex ecological point process models with integrated nested Laplace approximation. *Methods in Ecology and Evolution*, 4, 4, 305-315.

Illian, J., Penttinen, A., Stoyan, D., Stoyan, H. (2008) *Analysis and Modelling of Spatial Point Patterns*. Chichester: Wiley.

Illian, J., Sørbye, H., Rue, H. (2012). A toolbox for fitting complex spatial point process models using Integrated Nested Laplace Approximation (INLA). *Annuals of Applied Statistics*, 6(4),1499–1530.

Jackman, S. (2009) *Bayesian Analysis for the Social Science*s. West Sussex: John Wiley & Sons.

Jagger, T. H., Elsner, J. B., Widen, H. M. (2015) A Statistical Model for Regional Tornado Climate Studies. *PLoS ONE*, 10(8).

Jeffreys, H. (1961) *Theory of Probability*. Oxford: Oxford University Press.Longley, P., Goodchild, M., Maguire, D., Rhind, D. (2011) Geographical Information Systems and Science. West Sussex: John Wiley & Sons.

Jones, M. (1993) Simple boundary corrections for kernel density estimation. *Statistics and Computing*, 3, 135–146.

Laurini, M. P. (2017) The spatio-temporal dynamics of ethanol/gasoline price ratio in Brazil. *Renewable and Sustainable Energy Reviews*, 70, 1-12.

Law, R., Illian, J., Burslem, D., Gratzer, G., Gunatilleke, C., Gunatilleke, I. (2009) Ecological information from spatial patterns of plants: insights from point process theory. *Journal of Ecology*, 97, 616–628.

Liang, L., Xu, B., Chen, Y., Liu, Y., Chao, W., Fang, L., Feng, L., Goodchild, M., Gong, P. (2010) Combining Spatial-Temporal and Phylogenetic Analysis Approaches for Improved Understanding on Global H5N1 Transmission. PLOS One [Online]. Available at: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0013575 [Accessed January 2017].

Lindley, D. (2006) *Understanding Uncertainty*. Wiley-Blackwell.

Longley, P. A., Goodchild, M. F., Maguire, D. J., Rhind, D. W. (2011) *Geographical Information Systems and Science*. *City* (Vol. 83).

Ludlam, D. M. (1963) Severe Local Storms: A Review. In: Atlas, D. (Ed.) *Severe Local Storms*. American Meteorological Society Monographs, 27, 5.

Ludlam, D. M. (1970) *Early American Tornadoes 1586-1870*. American Meteorological Society.

Karpman, D., Ferreira, M., Wikle, C., (2013) A point process model for tornado report climatology. *Statistica*, 2, 1, 1–8.

Knorr-Held, L. (2000) Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in Medicine*, 19 (17-18), 2555–2567.

Knupp, K., Murphy, T., Coleman, T., Wade, R., Mullins, S., Schultz, C., Schultz, E., Carey, L., Sherrer, A., McCaul, E., Carcione, B., Latimer, S., Kula, A., Laws, K., Marsh, P., Klockow K. (2014) Meteorological overview of the devastating 27 April 2011 tornado outbreak. *Bulletin of the American Meteorological Society*, 95, 1041–1062.

Kraisnki, E., Lindgren, F., Simpson, D., Rue, H. (2016) The R-INLA Tutorial on SPDE methods [Online]. Available at: https://www.math.ntnu.no/inla/r-inla.org/tutorials/spde/spde-tutorial.pdf [Accessed January 2017).

Kunkel, K., Karl, T., Brooks, H., Kossin, J., Lawrimore, J., Arndt, D., Bosart, L., Changnon, D., Cutter, S., Doesken, N., Emanuel, K., Groisman, P., Katz, R., Knutson, T.,Brien, J., Paciorek, C., Peterson, T., Robinson, T., Trapp, J., Vose, R., Weaver, S., Weahner, M., Wolter, K., Wuebbles, D. (2013) Monitoring and understanding the trends in extreme storms: State of knowledge. *Bulletin of American Meteorological Society*, 94, 405-407.

Martino, S., Rue, H. (2010) Case studies in Bayesian computation using INLA. In *Complex Data Modeling and Computationally Intensive Statistical Methods*. 99–114.

Martino, S., Rue, H. (2009) Implementing Approximate Bayesian Inference using Integrated Nested Laplace Approximation: a manual for the inla program [Online]. Available at: https://www.math.ntnu.no/inla/r-inla.org/doc/inla-manual/inla-manual.pdf [Accessed January 2017].

Møller, J., Syversveen, A., Waagepetersen, R. (2017) Log Gaussian Cox Processes [Online] Available at: https://pdfs.semanticscholar.org/0796/0627be6d23f3fad40e7d153936b2ca96a704.pdf [Accessed January 2017].

Moore, T. W. (2017) On the temporal and spatial characteristics of tornado days in the United States. *Atmospheric Research*, 184, 56-65.

National Oceanic and Atmospheric Administration (2016) 2011 Tornado Information [Online]. Available at: http://www.noaanews.noaa.gov/2011_tornado_information.html [Accessed January 2017].

National Center Environmental Information (2016a) *Historical Records and Trends* [Online]. Available at: https://www.ncdc.noaa.gov/climate-information/extreme-events/us-tornado-climatology/trends [Accessed December 2016]

National Center Environmental Information (2016b) *U.S. Tornado Climatology* [Online]. Available at: https://www.ncdc.noaa.gov/climate-information/extreme-events/us-tornado-climatology [Accessed December 2016]

National Center Environmental Information (2016c) NOAA historical Background [Online] Available at :http://www.publicaffairs.noaa.gov/grounders/noaahistory.html [Accessed Januar 2017].

National Severe Storms Laboratory (2016). Severe Weather – Tornado Basics [Online]. Available at: http://www.nssl.noaa.gov/education/svrwx101/tornadoes/ [Accessed December 2017].

Ostby, F. (1993) The Changing Nature of Tornado Climatology. *17th Conference On Severe Local Storms, St. Louis (Boston)*, 1-5.

Riley, J., DeGloria, S., Elliot, R (1999) A terrain ruggedness index that quantifies topographic heterogeneity. *Intermountain Journal of Sciences* 5(1-4), 23-27.

Ripley, D. (1977) Modelling spatial patterns (with discussion). *Journal of the Royal Statistical Society, Series B*, 39, 172–212.

Romanic, D., Refan, M., Wu, C. H., Michel, G. (2016) Oklahoma tornado risk and variability: A statistical model. *International Journal of Disaster Risk Reduction*, *16*, 19–32.

Rosencrants, T. D., Ashley, W. S. (2015) Spatiotemporal analysis of tornado exposure in five US metropolitan areas. *Natural Hazards*, *78*(1), 121–140.

Rue, H. (2014) *Tutorial: Bayesian computing with INLA: An introduction* [Online]. Available at:

https://www.r-project.org/nosvn/conferences/useR-2013/Tutorials/Rue.html [Accessed December 2016]

Rue, H., Held, L. (2005) *Gaussian Markov Random Fields: Theory and Applications*. Monographs on Statistics and Probability, 104. London: Chapman & Hall.

Rue, H., Martino, S., Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *71*(2), 319–392.

Seely, J., Romps, D. (2015) The Effect of Global Warming on Severe thunderstorms in the United States. *Journal of Climate*, 28, 2443-2458.

Samandiego, F. (2010) *A comparison of the Bayesian and Frequentist Approaches to Estimation*. London: Springer.

Schaefer, J., R. Edwards, 1999: The SPC tornado/severe thunderstorm database. *11th Conference on Applied Climatology, American Meteorological Society*, 215–220.

Schultz, D., Richardson, Y., Markowski, P., Doswell, C. (2014) Tornados in the central United States and the "Clash of Air Masses". *Bulletins for the American Meteorological Society*, 95, 1704-1712.

Scolastic (2017) Map Tornado States [Online] Available at: http://teacher.scholastic.com/activities/scholasticnews/articles/SN3_051513_Map_TornadoStates.html [Accessed January 2017].

Seif, A. (2014) Using Topography Position Index for Landform Classification (Case study: Grain Mountain). *Bulletin of Environment, Pharmacology and Life Sciences*, 3(11), 33-39.

Simpson, D., Illian, B., Lindren, F., Sørbye, S., Rue, H. (2011) Going off grid: Computationally efficient inference for log-gaussian cox processes. *submitted*.
Online Submitted Paper Version available at: https://arxiv.org/pdf/1111.0641v3.pdf [Accessed January 2017].

Spiegelhalter, D., Best, G., Carlin, P., Linde, A. (2002) Bayesian measures of model complexity and fit. *Journal Royal Statistical Society series B*, 64(4), 583–639.

Storm Prediction Center (2016a) *Severe Weather Database Files (1950–2015)* [Online]. Available at: http://www.spc.noaa.gov/gis/svrgis/zipped/tornado.zip[Accessed December 2016]

Storm Prediction Center (2016b) Fujita Tornado Damage Scale [Online]. Available at: http://www.spc.noaa.gov/faq/tornado/f-scale.html [Accessed December 2016].

Storm Prediction Center (2016c). The 25 Deadliest Tornados [Online]. http://www.spc.noaa.gov/faq/tornado/killers.html [Accessed January 2017].

Statisticat LLC (2015) *Bayesian Inference*. Vignette for LaplacesDemon: Complete Environment for Bayesian Inference. R package version 15.03.19, URL http://www.bayesian-inference.com/software.

Tabb, L. P., Ballester, L., Grubesic, T. H. (2016) The spatio-temporal relationship between alcohol outlets and violence before and after privatization: A natural experiment, Seattle, Wa 2010–2013. *Spatial and Spatio-temporal Epidemiology*, 19, 115–124.

The R-INLA project (2017). Latent Models [Online] Available at: http://www.r-inla.org/models/latent-models [Accessed January 2017]

Tippet, M. (2014) Changing volatility of U.S. annual tornado report. *Geophysical Research Letters*, 41, 6956-6961.

Tippet, M., Cohen, J. (2016) Tornado outbreak variability follows Taylor's power law of fluctuation scaling and increases dramatically with severity. *Nature Communications*, 7, 10668.

Tippet, M., Allen, J., Gensini, V., Brooks, H. (2015) Climate and Hazardous convective weather. *Current Climate Change Reports*. 1, 2, 60-73.

Tornado, T. (2017) Where is the Tornado Alley [Online]. Available at: http://www.tornadochaser.net/tornalley.html [Accessed January 2017].

National Bureau for Economic Research (2016). Census U.S. Intercensal County Population Data, 1970-2014 [Online] Available at: http://www.nber.org/data/census-intercensal-county-population.html [Accessed December 2016].

United States Census Bureau (2016) *Cartographic Boundary Shapefiles* [Online]. Available at: https://www.census.gov/geo/maps-data/data/tiger-cart-boundary.html [Accessed December 2016]

United States Geological Survey (2014a) NLCD 1992 Land Cover (2011 Edition, amended 2014). *National Geospatial Data Asset (NGDA) Land Use Land Cover*. Sioux Falls: U.S. Geological Survey.

United States Geological Survey (2014b) NLCD 2001 Land Cover (2011 Edition, amended 2014). *National Geospatial Data Asset (NGDA) Land Use Land Cover*. Sioux Falls: U.S. Geological Survey.

United States Geological Survey (2014c) NLCD 2006 Land Cover (2011 Edition, amended 2014). *National Geospatial Data Asset (NGDA) Land Use Land Cover*. Sioux Falls: U.S. Geological Survey.

United States Geological Survey (2014d) NLCD 2011 Land Cover (2011 Edition, amended 2014). *National Geospatial Data Asset (NGDA) Land Use Land Cover*. Sioux Falls: U.S. Geological Survey.

United States Geological Survey (2016) Global 30 Arc-Second Elevation (GTOPO30) [Online]. Available at: https://lta.cr.usgs.gov/GTOPO30 [Accessed December 2016]

Verbout, M., Brooks, H., Leslie, M., Schultz, D. (2006) Evolution of the U.S. tornado database: 1954–2003. *Weather Forecasting*, 21, 86-93.

Wakefield, J. (2013) *Bayesian and Frequentist Regression Methods*. Washington: Springer.

Watanabe, S. (2010). Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory. *Journal of Machine Learning Research*, 11, 3571-3594.

Wang, K., Ivan, J. N., Ravishanker, N., Jackson, E. (2017) Multivariate poisson lognormal modeling of crashes by type and severity on rural two lane highways. *Accident Analysis and Prevention*, *99*, 6–19.

Weiss, A. (2001). "Topographic Positions and Landforms Analysis" (Conference Poster). *ESRI International User Conference*. San Diego, CA, 9-13.

Widen, H., Elsner, J., Cruz, R., Xing, G., Fraza, E., Migliorelli, L., Strazzo, S., Amrine, C., Mulholland, B., Patterson, M., Michaels, L. (2013) Adjusted Tornado Probabilities. *E-Journal Of Severe Storms Meteorology*, 8, 7.

Wikle, C., Anderson, C. (2003) Climatological Analysis of tornado report counts using a hierarchical Bayesian spatio temporal model. *Journal of Geophysical Research: Atmospheres*, 128, D24.

9. ATTACHMENTS

A.1. Python script for DEM geoprocessing

```
import arcpy

arcpy.env.workspace =r"D:\TEXAS"

arcpy.MakeRasterLayer_management("D:\TEXAS\gt30w100n40_dem(1)\gt30w140n40.dem",
"dem1")
arcpy.MakeRasterLayer_management("D:\TEXAS\gt30w100n40_dem(1)\gt30w100n40.dem",
"dem2")
arcpy.MakeFeatureLayer_management("texas.shp", "texas")

#PROJECT RASTER TO EPSG 102003
# Process: Project Raster
arcpy.ProjectRaster_management("dem1", "D:\\TEXAS\\DEM\\dem1",
"PROJCS['USA_Contiguous_Albers_Equal_Area_Conic',GEOGCS['GCS_North_American_1983',D
ATUM['D_North_American_1983',SPHEROID['GRS_1980',6378137.0,298.257222101]],PRIMEM['
Greenwich',0.0],UNIT['Degree',0.0174532925199433]],PROJECTION['Albers'],PARAMETER['
False_Easting',0.0],PARAMETER['False_Northing',0.0],PARAMETER['Central_Meridian',-
96.0],PARAMETER['Standard_Parallel_1',29.5],PARAMETER['Standard_Parallel_2',45.5],P
```

```
ARAMETER['Latitude_Of_Origin',37.5],UNIT['Meter',1.0]]", "NEAREST",
"1075,53344334831 1075,53344334831", "WGS_1984_(ITRF00)_To_NAD_1983", "",
"GEOGCS['GCS_WGS_1984',DATUM['D_WGS_1984',SPHEROID['WGS_1984',6378137.0,298.2572235
63]],PRIMEM['Greenwich',0.0],UNIT['Degree',0.0174532925199433]]")

# Process: Project Raster (2)
arcpy.ProjectRaster_management("dem2", "D:\\TEXAS\\DEM\\dem2",
"PROJCS['USA_Contiguous_Albers_Equal_Area_Conic',GEOGCS['GCS_North_American_1983',D
ATUM['D_North_American_1983',SPHEROID['GRS_1980',6378137.0,298.257222101]],PRIMEM['
Greenwich',0.0],UNIT['Degree',0.0174532925199433]],PROJECTION['Albers'],PARAMETER['
False_Easting',0.0],PARAMETER['False_Northing',0.0],PARAMETER['Central_Meridian',-
96.0],PARAMETER['Standard_Parallel_1',29.5],PARAMETER['Standard_Parallel_2',45.5],P
ARAMETER['Latitude_Of_Origin',37.5],UNIT['Meter',1.0]]", "NEAREST",
"1040,08018938805 1040,08018938805", "WGS_1984_(ITRF00)_To_NAD_1983", "",
"GEOGCS['GCS_WGS_1984',DATUM['D_WGS_1984',SPHEROID['WGS_1984',6378137.0,298.2572235
63]],PRIMEM['Greenwich',0.0],UNIT['Degree',0.0174532925199433]]")

#merge rasters
# Process: Mosaic To New Raster
arcpy.MosaicToNewRaster_management("dem1;dem2", "D:\\TEXAS\\DEM", "merged.tif",
"PROJCS['USA_Contiguous_Albers_Equal_Area_Conic',GEOGCS['GCS_North_American_1983',D
ATUM['D_North_American_1983',SPHEROID['GRS_1980',6378137.0,298.257222101]],PRIMEM['
Greenwich',0.0],UNIT['Degree',0.0174532925199433]],PROJECTION['Albers'],PARAMETER['
False_Easting',0.0],PARAMETER['False_Northing',0.0],PARAMETER['Central_Meridian',-
96.0],PARAMETER['Standard_Parallel_1',29.5],PARAMETER['Standard_Parallel_2',45.5],P
ARAMETER['Latitude_Of_Origin',37.5],UNIT['Meter',1.0]]", "16_BIT_SIGNED", "", "1",
"BLEND", "MATCH")

# Process: Clip
arcpy.Clip_management("merged.tif", "-999736,164899999 -1295597,9295
235280,481100001 -88854,1662999997", "D:\\TEXAS\\DEM\\clipped", "texas", "-9",
"NONE", "NO_MAINTAIN_EXTENT")

# Process: Focal Statistics
arcpy.gp.FocalStatistics_sa("clipped", "D:\\TEXAS\\DEM\\minDem", "Rectangle 10 10
CELL", "MINIMUM", "DATA")

# Process: Focal Statistics (2)
arcpy.gp.FocalStatistics_sa("clipped", "D:\\TEXAS\\DEM\\maxDem", "Rectangle 10 10
CELL", "MAXIMUM", "DATA")

# Process: Focal Statistics (3)
arcpy.gp.FocalStatistics_sa("clipped", "D:\\TEXAS\\DEM\\10x10", "Rectangle 10 10
CELL", "MEDIAN", "DATA")

# Process: Raster Calculator
arcpy.gp.RasterCalculator_sa("Float(\"%10x10%\" - \"%minDem%\") /
Float(\"%maxDem%\" - \"%minDem%\")", "D:\\TEXAS\\DEM\\R_index")
```

A.2. Python script for Land-cover geoprocessing

```
import arcpy
import os
import urllib
import zipfile

arcpy.env.workspace =r"C:\Angela\OK"

#extract filed from URL
urllib.urlretrieve
("http://www.landfire.gov/bulk/downloadfile.php?TYPE=nlcd2011&FNAME=nlcd_2011_landc
over_2011_edition_2014_10_10.zip",
"nlcd_2011_landcover_2011_edition_2014_10_10.gz")
```

```
urllib.urlretreive
("http://www.landfire.gov/bulk/downloadfile.php?TYPE=nlcd2006&FNAME=nlcd_2006_landc
over_2011_edition_2014_10_10.zip",
"nlcd_2006_landcover_2011_edition_2014_10_10.gz")
urllib.urlretrieve
("http://www.landfire.gov/bulk/downloadfile.php?TYPE=nlcd2001v2&FNAME=nlcd_2001_lan
dcover_2011_edition_2014_10_10.zip",
"nlcd2001v2&FNAME=nlcd_2001_landcover_2011_edition_2014_10_10.gz")
urllib.urlretreive
("http://www.landfire.gov/bulk/downloadfile.php?TYPE=nlcd92&FNAME=nlcd_1992_30meter
_whole.zip", "nlcd92&FNAME=nlcd_1992_30meter_whole (1).gz")


#UNZIP
directory ="C:\\Angela\\OK"
zip=".zip"

for RASTER in os.listdir(directory):
if RASTER.endswith(zip):
        rastername = os.path.abspath(RASTER)
        zip_ref = zipfile.ZipFile(rastername)
        zip_ref.extractall(directory)
        zip_ref.close()
        os.remove(rastername)

#START GEOPROCESSING!
n1992_USA = ["C:\\Angela\\OK\\nlcd_1992_30meter_whole (1)",
"nlcd_1992_30meter_whole.img"]
n2001_USA =
["C:\\Angela\\OK\\nlcd_2001_landcover_2011_edition_2014_10_10\\nlcd_2001_landcover_
2011_edition_2014_10_10", "nlcd_2001_landcover_2011_edition_2014_10_10.img"]
n2006_USA =
["C:\\Angela\\OK\\nlcd_2006_landcover_2011_edition_2014_10_10\\nlcd_2006_landcover_
2011_edition_2014_10_10", "nlcd_2006_landcover_2011_edition_2014_10_10.img"]
n2011_USA =
["C:\\Angela\\OK\\nlcd_2011_landcover_2011_edition_2014_10_10\\nlcd_2011_landcover_
2011_edition_2014_10_10", "nlcd_2011_landcover_2011_edition_2014_10_10.img"]


#import admin boundaries
admin_boundaries_usa2 ="C:\\Angela\\tmp\\cb_2013_us_county_5m.shp"
arcpy.MakeFeatureLayer_management(admin_boundaries_usa2, "admin_boundaries_usa2",
"","", "FID FID VISIBLE NONE;Shape Shape VISIBLE NONE;STATEFP STATEFP VISIBLE
NONE;COUNTYFP COUNTYFP VISIBLE NONE;COUNTYNS COUNTYNS VISIBLE NONE;AFFGEOID
AFFGEOID VISIBLE NONE;GEOID GEOID VISIBLE NONE;NAME NAME VISIBLE NONE;LSAD LSAD
VISIBLE NONE;ALAND ALAND VISIBLE NONE;AWATER AWATER VISIBLE NONE")

#crop the latter to texas
# Process: Select Layer By Attribute
arcpy.SelectLayerByAttribute_management("admin_boundaries_usa2", "NEW_SELECTION",
"\"STATEFP\" = '48'")

# Process: Copy Features
arcpy.CopyFeatures_management("admin_boundaries_usa2", "tx_shp2", "", "0", "0",
"0")
arcpy.SelectLayerByAttribute_management("admin_boundaries_usa2", "CLEAR_SELECTION")

#dissolve to have the whole state
arcpy.Dissolve_management("tx_shp2", "tx_shp_Dissolve2", "STATEFP", "",
"MULTI_PART", "DISSOLVE_LINES")

#loop to crop
inputfolder ="C:\\Angela\\LC"
outputfolder ="C:\\Angela\\LC"

n1992_USA ="nlcd_1992_30meter_whole.img"
```

```
n2001_USA ="nlcd_2001_landcover_2011_edition_2014_10_10.img"
n2006_USA ="nlcd_2006_landcover_2011_edition_2014_10_10.img"
n2011_USA ="nlcd_2011_landcover_2011_edition_2014_10_10.img"

rasterlist=[n1992_USA, n2001_USA, n2006_USA, n2011_USA]

for inraster in rasterlist:
print inraster
    inpath = os.path.join(inputfolder, inraster)
    tx_name = os.path.join(outputfolder, "2"+inraster)
    arcpy.ProjectRaster_management(inraster, tx_name,
"PROJCS['USA_Contiguous_Albers_Equal_Area_Conic',GEOGCS['GCS_North_American_1983',D
ATUM['D_North_American_1983',SPHEROID['GRS_1980',6378137.0,298.257222101]],PRIMEM['
Greenwich',0.0],UNIT['Degree',0.0174532925199433]],PROJECTION['Albers'],PARAMETER['
False_Easting',0.0],PARAMETER['False_Northing',0.0],PARAMETER['Central_Meridian',-
96.0],PARAMETER['Standard_Parallel_1',29.5],PARAMETER['Standard_Parallel_2',45.5],P
ARAMETER['Latitude_Of_Origin',37.5],UNIT['Meter',1.0]]", "NEAREST", "30 30", "",
"",
"PROJCS['Albers_Conical_Equal_Area',GEOGCS['GCS_North_American_1983',DATUM['D_North
_American_1983',SPHEROID['GRS_1980',6378137.0,298.257222101]],PRIMEM['Greenwich',0.
0],UNIT['Degree',0.0174532925199433]],PROJECTION['Albers'],PARAMETER['false_easting
',0.0],PARAMETER['false_northing',0.0],PARAMETER['central_meridian',-
96.0],PARAMETER['standard_parallel_1',29.5],PARAMETER['standard_parallel_2',45.5],P
ARAMETER['latitude_of_origin',23.0],UNIT['Meter',1.0]]")

n1992_USA ="2nlcd_1992_30meter_whole.img"
n2001_USA ="2nlcd_2001_landcover_2011_edition_2014_10_10.img"
n2006_USA ="2nlcd_2006_landcover_2011_edition_2014_10_10.img"
n2011_USA ="2nlcd_2011_landcover_2011_edition_2014_10_10.img"
arcpy.MakeRasterLayer_management(n1992_USA, "n1992")
arcpy.MakeRasterLayer_management(n2001_USA, "n2001")
arcpy.MakeRasterLayer_management(n2006_USA, "n2006")
arcpy.MakeRasterLayer_management(n2011_USA, "n2011")

rasterlist=[n1992, n2001, n2006, n2011]

for inraster in rasterlist:
print inraster
    outpathname = os.path.join(outputfolder, "3"+ inraster)
    arcpy.gp.Reclassify_sa(inraster, "Value",
"0 0;11 11;12 11;21 21;22 21;23 21;24 21; 31 31;32 31;33 31;41 41;42 41;43 41;51
51;52 51; 61 51;71 51;81 51;82 51;83 51;84 51;85 51;90 91; 91 91;92 91; 95 91",
                            outpathname, "DATA")

n1992_USA ="3nlcd_1992_30meter_whole.img"
n2001_USA ="3nlcd_2001_landcover_2011_edition_2014_10_10.img"
n2006_USA ="3nlcd_2006_landcover_2011_edition_2014_10_10.img"
n2011_USA ="3nlcd_2011_landcover_2011_edition_2014_10_10.img"

rasterlist=[n1992_USA, n2001_USA, n2006_USA, n2011_USA]

for inraster in rasterlist:
print inraster
    outpathname = os.path.join(outputfolder, "4"+ inraster)
    arcpy.gp.Aggregate_sa(inraster, outpathname, "400", "MEDIAN", "EXPAND", "DATA")


n1992_USA ="43nlcd_1992_30meter_whole.img"
n2001_USA ="43nlcd_2001_landcover_2011_edition_2014_10_10.img"
n2006_USA ="43nlcd_2006_landcover_2011_edition_2014_10_10.img"
n2011_USA ="43nlcd_2011_landcover_2011_edition_2014_10_10.img"

rasterlist=[n1992_USA, n2001_USA, n2006_USA, n2011_USA]
tx_shp2="C:\\Users\\Administrator\\Documents\\ArcGIS\\Default.gdb\\tx_shp2"
for inraster in rasterlist:
print inraster
```

```
    outpathname = os.path.join(outputfolder, "4"+ inraster)
    arcpy.gp.TabulateArea_sa(tx_shp2, "GEOID", inraster, "Value", outpathname,
"30")

n1992_USA ="443nlcd_1992_30meter_whole.img"
n2001_USA ="443nlcd_2001_landcover_2011_edition_2014_10_10.img"
n2006_USA ="443nlcd_2006_landcover_2011_edition_2014_10_10.img"
n2011_USA ="443nlcd_2011_landcover_2011_edition_2014_10_10.img"

rasterlist=[n1992_USA, n2001_USA, n2006_USA, n2011_USA]
os.makedirs("C:\\Angela\\Area_Per_County_sqM"
for inraster in rasterlist:
    outpath="C:\\Angela\\Area_Per_County_sqM"
    outname = inraster +"final"
    TableToTable_conversion (inraster, out_path, out_name)
```

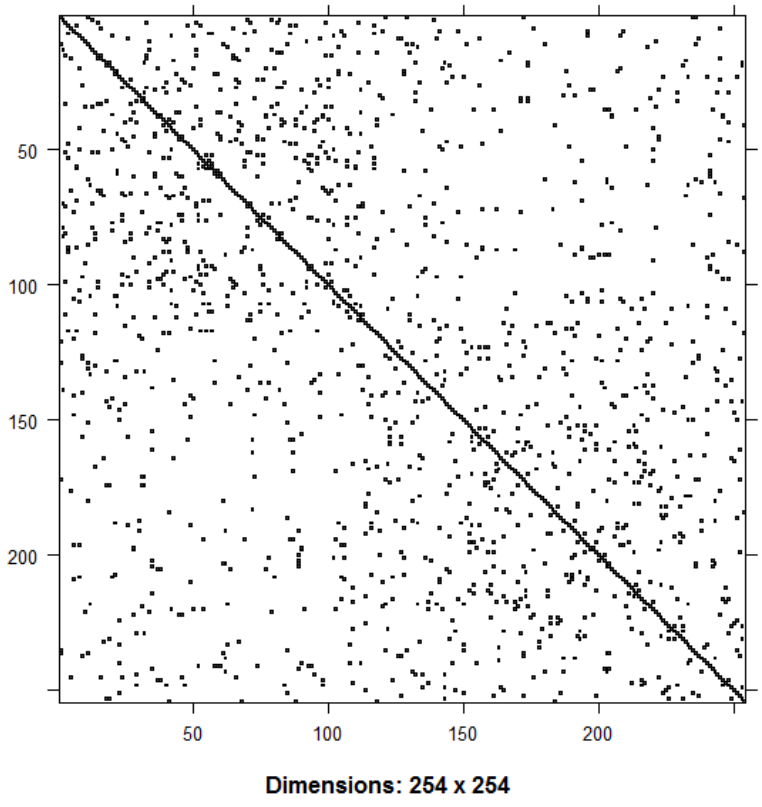
A.3. Python Script to construct the buffer of 40Km outside and give the points of the polygon

```
import arcpy

arcpy.env.workspace =r"D:\Tornado_Thesis\cluster\arcgis"
arcpy.MakeFeatureLayer_management("STATES.shp", "ALL_STATES")


#zone in the shapefile that has texas:
zonamento=2

# Process: Select Layer By Attribute
arcpy.SelectLayerByAttribute_management("ALL_STATES", "NEW_SELECTION", '"Zone" =
%i'% zonamento )

#SAVE THE SELECTION
 arcpy.CopyFeatures_management ("ALL_STATES", 'stateszone%i'%zonamento)

#dissolve the zone field
arcpy.Dissolve_management ('stateszone%i'%zonamento, 'dissolved_zone%i'%zonamento,
"zone")

#CLEAR SELECTION
arcpy.SelectLayerByAttribute_management("ALL_STATES", "CLEAR_SELECTION")

#apply the buffer
arcpy.Buffer_analysis('dissolved_zone%i'%zonamento, 'buffer_zone%i'%zonamento, "50
Kilometers", "OUTSIDE_ONLY", "ROUND", "NONE", "", "PLANAR")

#merge the buffer with the original polygon
arcpy.Merge_management(['buffer_zone%i'%zonamento,'dissolved_zone%i'%zonamento],
'merged_zone%i'%zonamento,"Zone \"Zone\" true true false 2 Short 0 0
,First,#,buffer_zone1,Zone,-1,-1,zone1_dissolved,Zone,-1,-1;Shape_Length
\"Shape_Length\" false true true 8 Double 0 0 ,First,#,buffer_zone1,Shape_Length,-
1,-1,zone1_dissolved,Shape_Length,-1,-1;Shape_Area \"Shape_Area\" false true true 8
Double 0 0 ,First,#,buffer_zone1,Shape_Area,-1,-1,zone1_dissolved,Shape_Area,-1,-
1")

#prepare for selection
arcpy.MakeFeatureLayer_management ('merged_zone%i'%zonamento,
'merging%i'%zonamento)

#select both polygons to make the eliminate
```

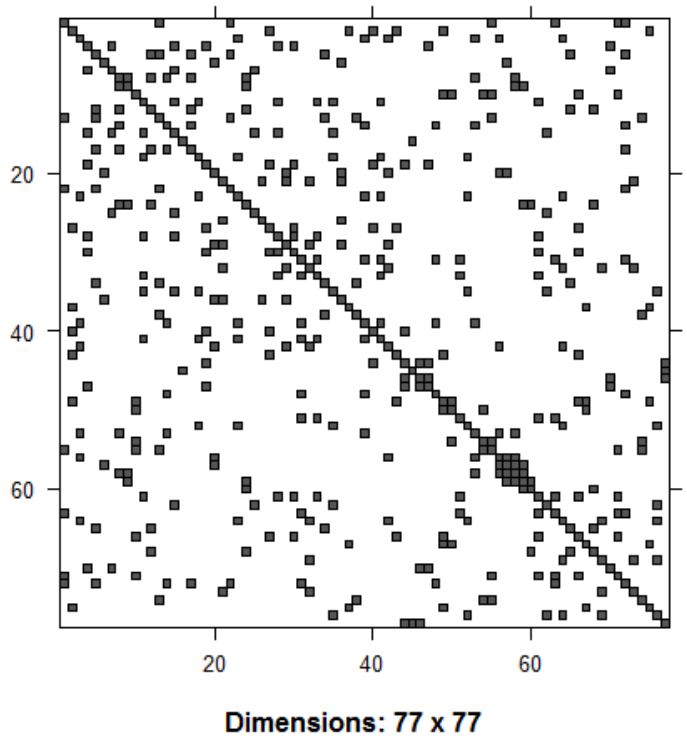
```
arcpy.SelectLayerByAttribute_management('merging%i'%zonamento, "NEW_SELECTION",
'"Zone" = %i'% zonamento)

#eliminate process
arcpy.Eliminate_management('merging%i'%zonamento, 'eli%i'%zonamento, "LENGTH", "",
"")

#clear selection
arcpy.SelectLayerByAttribute_management('merging%i'%zonamento, "CLEAR_SELECTION")

#from feature to point
arcpy.FeatureVerticesToPoints_management('eli%i'%zonamento, 'points%i'%zonamento ,
"ALL")

#add x and y geometry fields to the new shp.
arcpy.AddGeometryAttributes_management('points%i'%zonamento, "POINT_X_Y_Z_M",
"METERS", "",
"PROJCS['USA_Contiguous_Albers_Equal_Area_Conic',GEOGCS['GCS_North_American_1983',D
ATUM['D_North_American_1983',SPHEROID['GRS_1980',6378137.0,298.257222101]],PRIMEM['
Greenwich',0.0],UNIT['Degree',0.0174532925199433]],PROJECTION['Albers'],PARAMETER['
False_Easting',0.0],PARAMETER['False_Northing',0.0],PARAMETER['Central_Meridian',-
96.0],PARAMETER['Standard_Parallel_1',29.5],PARAMETER['Standard_Parallel_2',45.5],P
ARAMETER['Latitude_Of_Origin',37.5],UNIT['Meter',1.0]]")

#save the file
zone_2_coords ="D:\\Tornado_Thesis\\cluster\\zone_2_coords"

arcpy.ExportXYv_stats('points%i'%zonamento, "POINT_X;POINT_Y", "SEMI-COLON",
'D:\\Tornado_Thesis\\cluster\\coords_zone%i'%zonamento, "ADD_FIELD_NAMES")
```

A.4. Adjacency matrix for the counties

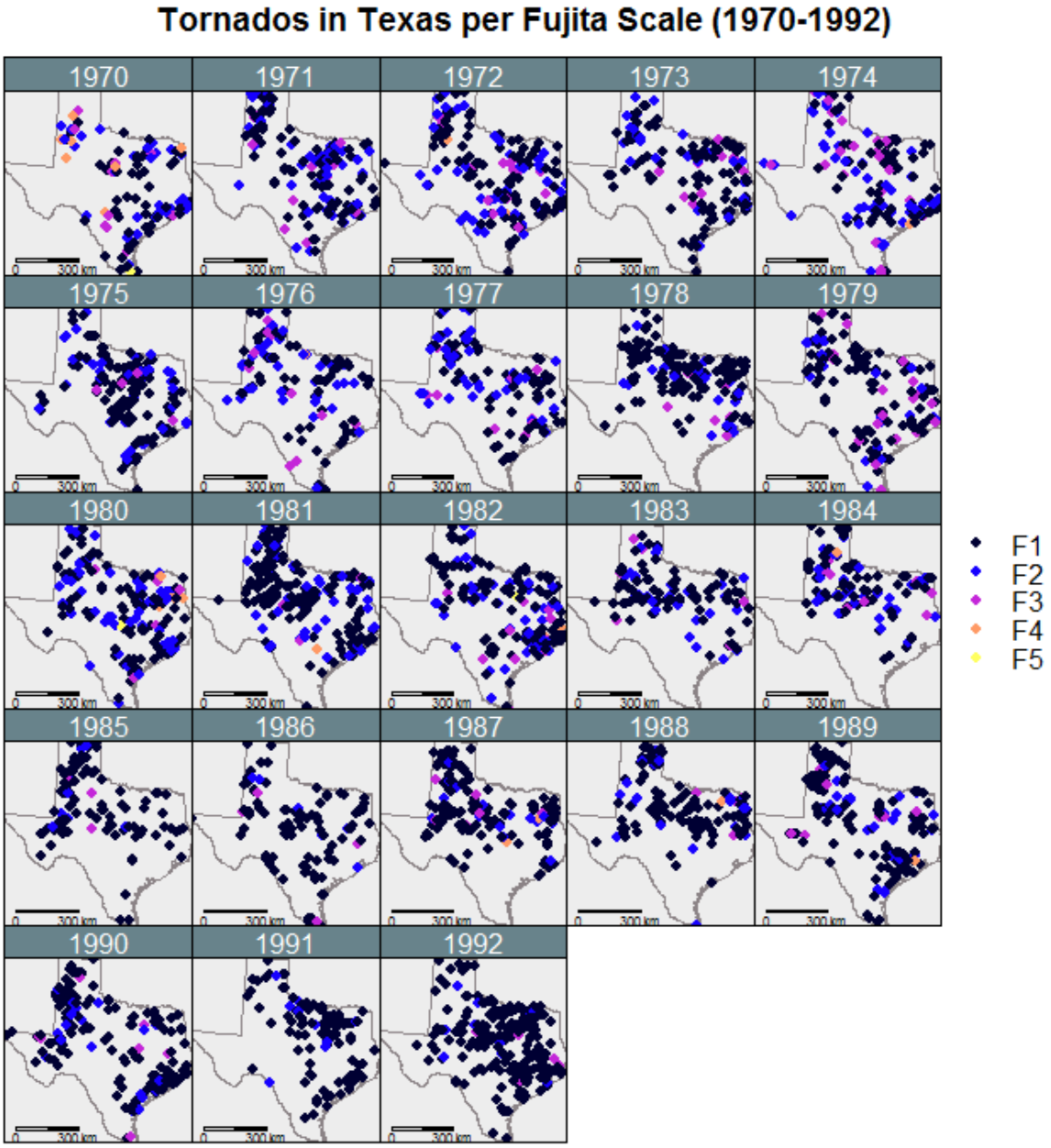
Texas



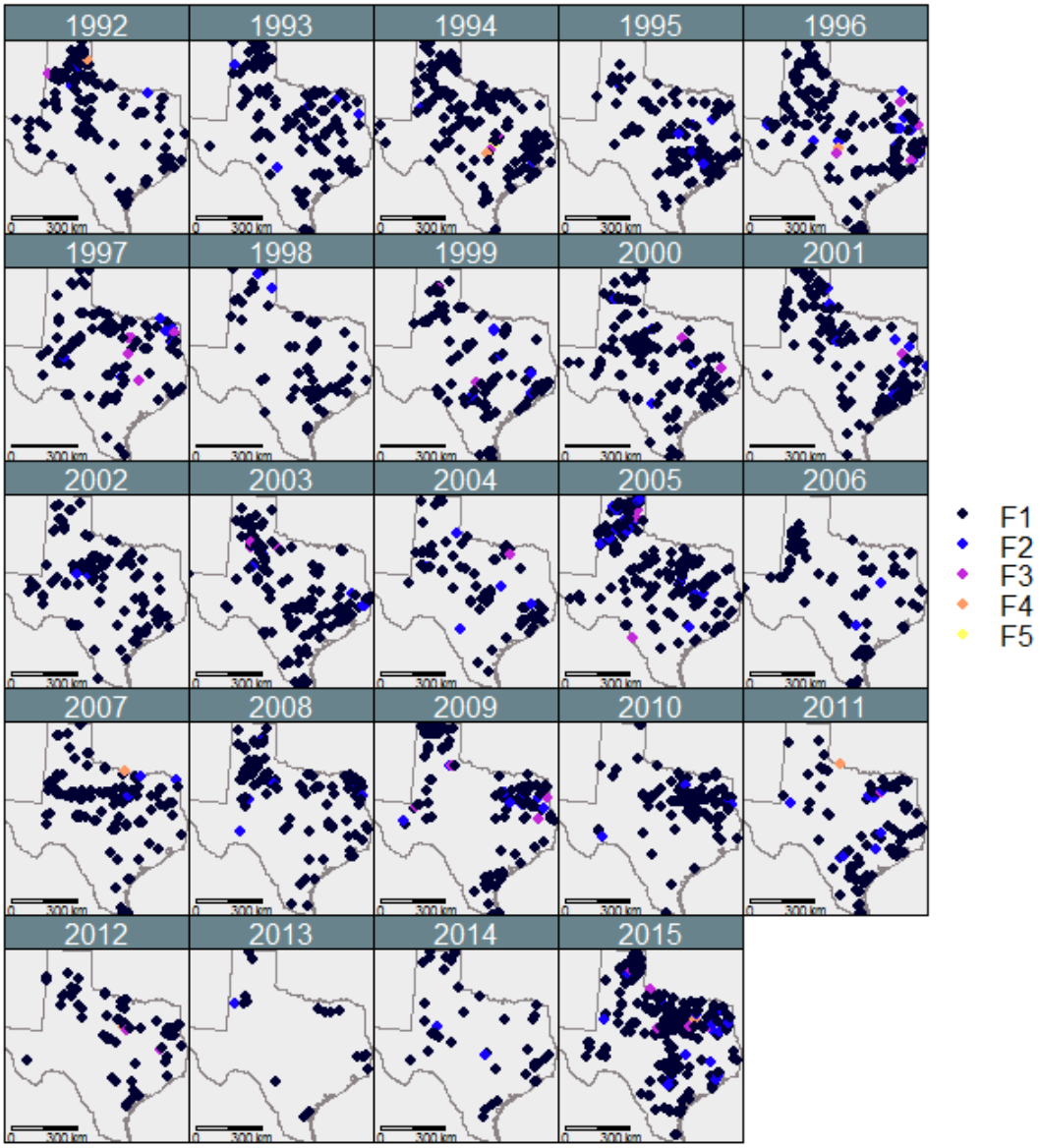
Ocklahoma



A.5. Visual representation of the point process pattern by FScale


Tornados in Texas per Fujita Scale (1970-1992)
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
0 300 km
F1
F2
F3
F4
F5

Tornados in Texas per Fujita Scale (1992-2015)



A.6. R-Code for Point Processes

```
#download and save the tornado file
download.file ("http://www.spc.noaa.gov/gis/svrgis/zipped/tornado.zip",
"tornado.zip", mode ="wb")
unzip ("tornado.zip",exdir="./tmp")

#read data into R; create reference number to identify single events
Tornados <-read.dbf("./tmp/torn/torn.dbf")
Tornados$ref <- ""
Tornados$ref <- seq(1,60114, by=1)

#make a table to input as point patterns into ARCgis; EPSG: 102003, GET COORDINATES
tornado_events <- as.data.frame(Tornados)
coords <- cbind(tornado_events$slon, tornado_events$slat)
tornados <- SpatialPointsDataFrame(coords, tornado_events, proj4string
=CRS("+init=epsg:4326"))
CRS.new <-CRS("+proj=aea +lat_1=29.5 +lat_2=45.5 +lat_0=37.5 +lon_0=-96 +x_0=0
+y_0=0+datum=NAD83 +units=m +no_defs +ellps=GRS80 +towgs84=0,0,0")  #EPSG:102003
tornados <-spTransform(tornados, CRS.new)
tornado_events<-as.data.frame(tornados@coords)
Tornados$x <- ""
Tornados$x <- tornado_events$coords.x1
Tornados$y <- tornado_events$coords.x2

#Plot Total annual counts w/ trend line
begin <-Tornados$yr[1]
end <-as.numeric(Tornados$yr[length(Tornados$yr)])
Count <-as.integer(table(Tornados$yr))
AnnualCountALL.df <-data.frame(Year = (begin:end), Count = Count)

ggplot(AnnualCountALL.df, aes(x = Year, y = Count)) +geom_line()+
geom_smooth(method ="gam", color="cadetblue") +ylab("Number of Reported Tornadoes") +
theme_gray()+
ggtitle("Number of Reported Tornados per Year \n (United States of America) \n 1950-2015")+
theme(plot.title =element_text(size=12, face="bold",
margin =margin(10, 0, 10, 0), hjust=0.5))+
annotate(geom="text", x=2010, y=300, label="Source: SPC 2016a", size=4
  )+
scale_x_continuous(breaks=c(1950,1960, 1970, 1980, 1990, 2000, 2010))

#Plot per Fscale
TornTable <-as.data.frame(table(Tornados$yr, Tornados[,11]))
TornTable$year <-as.numeric(levels(TornTable$Var1))
TornTable$Fscale <-paste("F", TornTable$Var2, sep ="")

ggplot(TornTable[TornTable$Var2 !=-9, ], aes(x = year, y = Freq)) +
geom_point() +geom_smooth(span =0.9, color="cadetblue") +
facet_wrap(~Fscale, ncol =2, scales="free_y") +
theme_gray()+
ggtitle("Number of Reported Tornados per Year by Fujita Scale\n (United States of America) \n
1950-2015")+
theme(plot.title =element_text(size=12, face="bold",
margin =margin(10, 0, 10, 0), hjust=0.5))+
xlab("Year")+ylab("Reported Number of Tornadoes")+
scale_x_continuous(breaks=c(1950,1960, 1970, 1980, 1990, 2000, 2010))+
scale_y_continuous(breaks = scales::pretty_breaks(n=4))

## `geom_smooth()` using method = 'loess'

#plot per capital losses
#before 1996, losses are from 0-9; after 1996, they classify it for millions of dollars.
#give mean values for before 1996
Tornados$tloss<-""
Tornados$tloss<-Tornados$loss*1000000
Tornados$tloss[Tornados$yr %in%c(1950:1995) &Tornados$loss=="0"]<-"0"
Tornados$tloss[Tornados$yr %in%c(1950:1995) &Tornados$loss=="1"]<-"25"
Tornados$tloss[Tornados$yr %in%c(1950:1995) &Tornados$loss=="2"]<-"225"
Tornados$tloss[Tornados$yr %in%c(1950:1995) &Tornados$loss=="3"]<-"2250"
Tornados$tloss[Tornados$yr %in%c(1950:1995) &Tornados$loss=="4"]<-"22500"
Tornados$tloss[Tornados$yr %in%c(1950:1995) &Tornados$loss=="5"]<-"225000"
Tornados$tloss[Tornados$yr %in%c(1950:1995) &Tornados$loss=="6"]<-"2250000"
Tornados$tloss[Tornados$yr %in%c(1950:1995) &Tornados$loss=="7"]<-"22500000"
```

```
Tornados$tloss[Tornados$yr %in%c(1950:1995) &Tornados$loss=="8"]<-"225000000"
Tornados$tloss[Tornados$yr %in%c(1950:1995) &Tornados$loss=="9"]<-"5000000000"
Tornados$tloss<-as.numeric(Tornados$tloss)

losses<-aggregate(Tornados$tloss, by=list(Tornados$yr), "sum")
AnnualLossALL.df <-data.frame(Year = (begin:end), Count = losses$x/1000000000)

q1<-ggplot(AnnualLossALL.df, aes(x = Year, y = Count)) +
geom_point()+geom_smooth(method ="loess", color="gold")+
ylab("Billion Dollars") +theme_gray()+
ggtitle("Property Losses")+
theme(plot.title =element_text(size=12, face="bold",
margin =margin(10, 0, 10, 0), hjust=0.5))+
annotate(geom="text", x=2010, y=-0.5, label="Source: SPC, 2016a", size=2)

###plot number of fatalities
fatal<-aggregate(Tornados$fat, by=list(Tornados$yr), "sum")
AnnualfatalALL.df <-data.frame(Year = (begin:end), Count = fatal$x)

q2<-ggplot(AnnualfatalALL.df, aes(x = Year, y = Count)) +geom_point()+
geom_smooth(method ="loess", col="skyblue4") +ylab("Number of deaths") +theme_gray()+
ggtitle("Fatalities")+
theme(plot.title =element_text(size=12, face="bold",
margin =margin(10, 0, 10, 0), hjust=0.5))+
annotate(geom="text", x=2010, y=-0.5, label="Source: SPC, 2016a", size=2)

inj<-aggregate(Tornados$inj, by=list(Tornados$yr), "sum")
AnnualinjALL.df =data.frame(Year = (begin:end), Count = inj$x)

q3<-ggplot(AnnualinjALL.df, aes(x = Year, y = Count)) +
geom_point()+geom_smooth(method ="loess", col="yellow4")+
ylab("Number of injuried Individuals") +theme_gray()+
ggtitle("Injuries")+
theme(plot.title =element_text(size=12, face="bold",
margin =margin(10, 0, 10, 0), hjust=0.5))+
annotate(geom="text", x=2010, y=-0.5, label="Source: SPC, 2016a", size=2)

multiplot(q1, q2, cols=2)

q3

#plot texas and the other Tornado Alley states
#Estados de Oklahoma, Kansas, Arkansas, Iowa e Missouri
#Texas, Colorado, Luisiana, Minnesota e Dacota do Sul,
#Mississippi, Illinois, Indiana, Nebraska, Tennessee, Kentucky Wisconsin.


torn_alley<-subset(Tornados, Tornados$st=="TX" |Tornados$st=="OK" |Tornados$st=="KS"
|Tornados$st=="AR"
                   |Tornados$st=="IA" |Tornados$st=="MO"|Tornados$st=="CO"|Tornados$st=="LA"
                   |Tornados$st=="MN"|Tornados$st=="SD"|Tornados$st=="MS"|Tornados$st=="IL"
                   |Tornados$st=="IN"|Tornados$st=="NE"|Tornados$st=="TN"|Tornados$st=="KY"
                   |Tornados$st=="WI")
year<-rep(begin:end, 17)
alleycounts<-as.data.frame(year)
alleycounts$state <-rep(c("TX", "OK", "KS", "AR", "IA", "MO", "CO", "LA",
"MN","SD", "MS", "IL", "IN", "NE", "TN", "KY",
"WI"), times=1, each=66)
tx <-subset(Tornados, Tornados$st=="TX")
tx <-as.integer(table(tx$yr))
ok <-subset(Tornados, Tornados$st=="OK")
ok <-as.integer(table(ok$yr))
ks <-subset(Tornados, Tornados$st=="KS")
ks <-as.integer(table(ks$yr))
ar <-subset(Tornados, Tornados$st=="AR")
ar <-as.integer(table(ar$yr))
ia <-subset(Tornados, Tornados$st=="IA")
ia <-as.integer(table(ia$yr))
mo <-subset(Tornados, Tornados$st=="MO")
mo <-as.integer(table(mo$yr))
co <-subset(Tornados, Tornados$st=="CO")
co <-as.integer(table(co$yr))
la <-subset(Tornados, Tornados$st=="LA")
la <-as.integer(table(la$yr))
```

```
mn <-subset(Tornados, Tornados$st=="MN")
mn <-as.integer(table(mn$yr))
sd <-subset(Tornados, Tornados$st=="SD")
sd <-as.integer(table(sd$yr))
ms <-subset(Tornados, Tornados$st=="MS")
ms <-as.integer(table(ms$yr))
il <-subset(Tornados, Tornados$st=="IL")
il <-as.integer(table(il$yr))
ind <-subset(Tornados, Tornados$st=="IN")
ind <-as.integer(table(ind$yr))
ne <-subset(Tornados, Tornados$st=="NE")
ne <-as.integer(table(ne$yr))
tn <-subset(Tornados, Tornados$st=="TN")
tn <-as.integer(table(tn$yr))
tn<-c(tn, 0)
ky <-subset(Tornados, Tornados$st=="KY")
ky <-as.integer(table(ky$yr))
ky<-c(ky, 0)
wi <-subset(Tornados, Tornados$st=="WI")
wi <-as.integer(table(wi$yr))
all<-c(tx, ok, ks, ar, ia, mo, co, la, mn, sd, ms, il, ind, ne, tn, ky, wi )
alleycounts$counts<-all

alleycounts1<-alleycounts[1:529,]
p =ggplot(alleycounts1, aes(x = year, y = counts, group=state)) +
geom_line(aes(color=state))+
geom_point(aes(color=state))
p +ylab("Number of Reported Tornadoes") +
labs(title="Number of Reported Tornados in Tornado Alley(part 1)")+
theme(plot.title =element_text(size=12, face="bold",
margin =margin(10, 0, 10, 0), hjust=0.5))+
annotate(geom="text", x=2010, y=300, label="Source: NOAA, SPC, 2016", size=3)

alleycounts2<-alleycounts[529:1122,]
q =ggplot(alleycounts2, aes(x = year, y = counts, group=state)) +
geom_line(aes(color=state))+
geom_point(aes(color=state))
q +ylab("Number of Reported Tornadoes") +
labs(title="Number of Reported Tornados in Tornado Alley (part 2)")+
theme(plot.title =element_text(size=12, face="bold",
margin =margin(10, 0, 10, 0), hjust=0.5))+
annotate(geom="text", x=2010, y=150, label="Source: NOAA, SPC, 2016", size=3)

#SUBSET TEXAS and year >1970
torn_texas<-subset(Tornados, Tornados$st=="TX"&yr>=1970 )
torn_texas$date<-as.Date(torn_texas$date, format="%Y-%m-%d")
Tornados$yr<-as.numeric(Tornados$yr)

##plot per Fscale
TornT <-as.data.frame(table(torn_texas$yr, torn_texas[,11]))
TornT$year <-as.numeric(levels(TornT$Var1))
TornT$Fscale <-paste("F", TornT$Var2, sep ="")
ggplot(TornT[TornT$Var2 !=-9, ], aes(x = year, y = Freq)) +
geom_point()+geom_smooth(span =0.9, color="forestgreen") +
facet_wrap(~Fscale, ncol =2, scales ="free") +
theme_gray()+
ggtitle("Number of Reported Tornados per Year \n (Texas) \n 1970-2015 \n by FScale")+
theme(plot.title =element_text(size=12, face="bold",
margin =margin(10, 0, 10, 0), hjust=0.5))+
ylab("Reported Number of Tornadoes")

## `geom_smooth()` using method = 'loess'

#ousiders removal
#zone texas - refs 2949, 5732, 9752, 10216, 10641, 57954 are outside
torn_texas<-subset(torn_texas,
                   torn_texas$ref !=2949&
torn_texas$ref !=5732&
torn_texas$ref !=9752&
torn_texas$ref !=10216&
torn_texas$ref !=10641&
torn_texas$ref !=57954)
```

```
#1. Plotting Point Processes
TornALL <-readOGR(dsn ="./tmp/torn", layer ="torn", stringsAsFactors =FALSE)

CRS.new <-CRS("+proj=aea +lat_1=29.5 +lat_2=45.5 +lat_0=37.5 +lon_0=-96 +x_0=0
+y_0=0+datum=NAD83 +units=m +no_defs +ellps=GRS80 +towgs84=0,0,0")  #EPSG:102003
TornALL <-spTransform(TornALL, CRS.new)
TornALL$Ref<-seq(1, 60114)
TornTexas<-subset(TornALL, TornALL$st=="TX"&
TornALL$Ref !=2949&
TornALL$Ref !=5732&
TornALL$Ref !=9752&
TornALL$Ref !=10216&
TornALL$Ref !=10641&
TornALL$Ref !=57954&yr>=1970)

x<-TornTexas$slon
y<-TornTexas$slat
coords<-cbind(x,y)
we<-SpatialPoints(coords, CRS("+init=epsg:4326"))
we<-spTransform(we, CRS.new)
coords<-as.matrix(we@coords)
num<-as.numeric(torn_texas$mag)
data1<-as.data.frame(num)
sp1<-SpatialPoints(coords=coords, proj4string=CRS.new)
endTime<-as.POSIXct(31-12-2015, format="%d-%m-%Y", origin ="01-01-2016")
TornTexas$date<-as.Date(TornTexas$date, format="%Y-%m-%d")
asd<-STIDF(sp=sp1, time=TornTexas$date, data=data1)
US.sp <-readOGR(dsn ="./tmp", layer ="cb_2013_us_county_5m",
stringsAsFactors =FALSE)

TX.sp <-US.sp[US.sp$STATEFP ==48, ]
county <-paste(TX.sp$STATEFP, TX.sp$COUNTYFP, sep ="")
county2 <-geometry(spChFIDs(TX.sp, county))
counties <-spTransform(county2, CRS.new)
a<-RColorBrewer::brewer.pal(5, "Set3")
wcounty<-unionSpatialPolygons(counties, ID =rep("1", length(row.names(counties))))
years<-c(1970:1992)
w<-list("sp.lines", wcounty, col="lavenderblush4", cex=1.5)
scale <-list("SpatialPolygonsRescale", layout.scale.bar(),
scale =400000,
fill =c("transparent", "black"), offset =c(-850000, -1200000))
text1 <-list("sp.text", c(-850000, -1240000), "0", cex=0.5, col="black")
text2 <-list("sp.text", c(-500000, -1240000), "300 km", cex=0.5, col="black")
years<-c(1970:1991)
stplot(asd[1:3763,], names.attr=years, number=22, cuts=5,
sp.layout=list(w, scale,text1, text2),
main="Tornados in Texas per Fujita Scale (1992-2015)", cex=0.7,
legendEntries =c("F1", "F2", "F3", "F4", "F5"), key.space="right",
par.settings =list(panel.background=list(col="gray93"),
strip.background=list(col="lightblue4"),
add.text=list(col="white")))

years2<-c(1992:2015)
stplot(asd[3764:6678,], names.attr=years2, number=24, cuts=5,
sp.layout=list(w, scale,text1, text2),
main="Tornados in Texas per Fujita Scale (1992-2015)", cex=0.5,
legendEntries =c("1", "2", "3", "4", "5"), key.space="right",
par.settings =list(panel.background=list(col="gray93"),
strip.background=list(col="lightblue4"),
add.text=list(col="white")))
```

```
###create owin, ppp objects
torn_texas<-torn_texas[!duplicated(torn_texas$x),]
a<-min(torn_texas$x)
b<-max(torn_texas$x)
c<-min(torn_texas$y)
d<-max(torn_texas$y)

window_texas<-owin(c(a,b), c(c,d))

x <-torn_texas$x
y <-torn_texas$y
```

```
#arcpy script in attachment
poly<-read.csv("points_out_zone2.csv", header=T, sep=";", dec =",")
poly[,1]<-rev(poly[,1])
poly[,2]<-rev(poly[,2])
poly<-as.matrix(poly)
par(mfrow=c(1,2))
plot(poly[,1], poly[,2])
plot(x, y)

options(scipen=4)
tornado_texas_ppp<-ppp(x, y, window=window_texas, check=T)

#1.1. density calculated w/ standard; Diggle and ppl
a <-density.ppp(tornado_texas_ppp, diggle = T)
b <-density.ppp(tornado_texas_ppp, sigma=bw.diggle, adjust=2)
c <-density.ppp(tornado_texas_ppp, sigma=bw.ppl,adjust=2)
d <-density.ppp(tornado_texas_ppp, sigma =100000)

par(mar=c(2, 2, 2, 2), mfrow=c(2,4), cex=0.8, oma=c(0, 0, 3, 0))
my_palette <-colorRampPalette(c("black", "white"))(n =299)
plot(a, main=expression(paste(sigma, " = 150 000")), col=my_palette)
plot(b, main=expression(paste(sigma, " = bw.diggle")), col=my_palette)
plot(c, main=expression(paste(sigma, " = bw.ppl")), col=my_palette)
plot(d, main=expression(paste(sigma, " = 100 000")), col=my_palette)

persp(a, theta=20, phi=20, zlab="density", border=NA, col="grey", shade=0.75,
main=expression(paste(sigma, " = 150 000")))
persp(b, theta=20, phi=20, zlab="density", border=NA, col="grey", shade=0.75,
main=expression(paste(sigma, " = bw.diggle (30 221)")))
persp(c, theta=20, phi=20, zlab="density", border=NA, col="grey", shade=0.75,
main=expression(paste(sigma, " = bw.ppl (17 554)")))
persp(d, theta=20, phi=20, zlab="density", border=NA, col="grey", shade=0.75,
main=expression(paste(sigma, " = 100 000")))
mtext("Spatial Density Study for tornado occurence in Texas", outer = T, cex=1.5)

#1.1.2. Standard deviation of intensity
#Estimate of standard error for the kernel estimate of intensity
#Uniform edge correction, bandwidth 1 metre

f <-density.ppp(tornado_texas_ppp, sigma =150000, se=T, diggle=T)$SE
q <-density.ppp(tornado_texas_ppp, sigma =100000, se=T, diggle=T)$SE
g <-density.ppp(tornado_texas_ppp, sigma =50000, se=T, diggle = T)$SE

## Warning in sqrt(structure(c(-4.99272059035671e-31, 0,
## -2.18431525828106e-31, : NaNs produced

my_palette<-heat.colors(10)
Zlist <-list(a=f, b=q, c=g)
Zrange <-range(unlist(
lapply(Zlist, function(x){summary(x)$range})))
plot(as.listof(Zlist), zlim=Zrange, ncols=3,
main="Standard Error of Intensities",
sub="(without covariation with RI)", col=my_palette)

#Density w/ covariate Elevation&TPI

RI <-raster("Index_Value1.tif")
RI<-as.im.RasterLayer(RI)

w1<-rhohat(tornado_texas_ppp, RI)

RI2 <-raster("final.tif")
RI2 <-as.im.RasterLayer(RI2)

w2<-rhohat(tornado_texas_ppp, RI2)

par(mfrow=c(1,2))
plot(w1, legend=F, ylab=expression(paste(rho, " (TPI)")),
xlab="TPI", main="Intensity as a function of TPI")
plot(w2, legend=F, ylab=expression(paste(rho, " (Elevation)")),
xlab="Elevation", main="Intensity as a function of elevation")
```

```
#tests for the others covariates
population<-read.csv("Population_final2.csv", header=T, sep=";", dec=".")
TX.sp<-TX.sp[,5]
colnames(population)[1] <- "GEOID"
sp_pop<-merge(TX.sp, population, by ="GEOID")
sp_pop$log1970<-log10(population$pop1970)
sp_pop$log1975<-log10(population$pop1975)
sp_pop$log1980<-log10(population$pop1980)
sp_pop$log1985<-log10(population$pop1985)
sp_pop$log1990<-log10(population$pop1990)
sp_pop$log1995<-log10(population$pop1995)
sp_pop$log2000<-log10(population$pop2000)
sp_pop$log2005<-log10(population$pop2005)
sp_pop$log2010<-log10(population$pop2010)
sp_pop$log2015<-log10(population$pop2015)

sp_pop<-spTransform(sp_pop, CRS.new)

#shp2raster function from
#https://github.com/brry/misc/blob/master/shp2raster.R
popr1970 <-shp2raster(shp=sp_pop, column="log1970", ascname ="l1970", overwrite=TRUE)
popr1975 <-shp2raster(shp=sp_pop, column="log1975", ascname ="l1975", overwrite=TRUE)
popr1980 <-shp2raster(shp=sp_pop, column="log1980", ascname ="l1980", overwrite=TRUE)
popr1985 <-shp2raster(shp=sp_pop, column="log1985", ascname ="l1985", overwrite=TRUE)
popr1990 <-shp2raster(shp=sp_pop, column="log1990", ascname ="l1990", overwrite=TRUE)
popr1995 <-shp2raster(shp=sp_pop, column="log1995", ascname ="l1995", overwrite=TRUE)
popr2000 <-shp2raster(shp=sp_pop, column="log2000", ascname ="l2000", overwrite=TRUE)
popr2005 <-shp2raster(shp=sp_pop, column="log2005", ascname ="l2005", overwrite=TRUE)
popr2010 <-shp2raster(shp=sp_pop, column="log2010", ascname ="l2010", overwrite=TRUE)

#1970
t1970<-subset(torn_texas, torn_texas$yr>=1970&torn_texas$yr<1976)
t1970<-ppp(t1970$x, t1970$y, window=window_texas, check=T)
den1<-rhohat(t1970, as.im.RasterLayer(popr1970))

#1975
t1975<-subset(torn_texas, torn_texas$yr>=1976&torn_texas$yr<1980)
t1975<-ppp(t1975$x, t1975$y, window=window_texas, check=T)
den2<-rhohat(t1975, as.im.RasterLayer(popr1975))

#1980
t1980<-subset(torn_texas, torn_texas$yr>=1980&torn_texas$yr<1986)
t1980<-ppp(t1980$x, t1980$y, window=window_texas, check=T)
den3<-rhohat(t1980, as.im.RasterLayer(popr1980))

#1985
t1985<-subset(torn_texas, torn_texas$yr>=1986&torn_texas$yr<1990)
t1985<-ppp(t1985$x, t1985$y, window=window_texas, check=T)
den4<-rhohat(t1985, as.im.RasterLayer(popr1985))

#1990
t1990<-subset(torn_texas, torn_texas$yr>=1990&torn_texas$yr<1996)
t1990<-ppp(t1990$x, t1990$y, window=window_texas, check=T)
den5<-rhohat(t1990, as.im.RasterLayer(popr1990))

#1995
t1995<-subset(torn_texas, torn_texas$yr>=1996&torn_texas$yr<2000)
t1995<-ppp(t1995$x, t1995$y, window=window_texas, check=T)
den6<-rhohat(t1995, as.im.RasterLayer(popr1995))

#2000
t2000<-subset(torn_texas, torn_texas$yr>=2000&torn_texas$yr<2006)
t2000<-ppp(t2000$x, t2000$y, window=window_texas, check=T)
den7<-rhohat(t2000, as.im.RasterLayer(popr2000))

#2000
t2005<-subset(torn_texas, torn_texas$yr>=2006&torn_texas$yr<2010)
t2005<-ppp(t2005$x, t2005$y, window=window_texas, check=T)
den8<-rhohat(t2005, as.im.RasterLayer(popr2005))

#2010
t2010<-subset(torn_texas, torn_texas$yr>=2010&torn_texas$yr<2016)
t2010<-ppp(t2010$x, t2010$y, window=window_texas, check=T)
den9<-rhohat(t2010, as.im.RasterLayer(popr2010))
```

```
q1<-(ggplot(den1, aes(x=X, y=rho))+geom_line(aes(X, rho))+
geom_ribbon(data=den1,aes(ymin=hi,ymax=lo),alpha=0.3))+
ylab(expression(paste(rho, "(X)"))) +xlab("log(Pop)") +
labs(title="1970-1975")+theme(axis.title.y =element_text(size =9),
axis.title.x =element_text(size =9),
plot.title =element_text(size=9, face="bold",
margin =margin(10, 0, 10, 0), hjust=0.5))
q2<-(ggplot(den2, aes(x=X, y=rho))+geom_line(aes(X, rho))+
geom_ribbon(data=den2,aes(ymin=hi,ymax=lo),alpha=0.3))+
ylab(expression(paste(rho, "(X)"))) +xlab("log(Pop)") +
labs(title="1975-1980")+
theme(axis.title.y =element_text(size =9),
axis.title.x =element_text(size =9),plot.title =element_text(size=9, face="bold",
margin =margin(10, 0, 10, 0), hjust=0.5))
q3<-(ggplot(den3, aes(x=X, y=rho))+geom_line(aes(X, rho))+
geom_ribbon(data=den3,aes(ymin=hi,ymax=lo),alpha=0.3))+
ylab(expression(paste(rho, "(X)"))) +xlab("log(Pop)") +
labs(title="1980-1985")+
theme(axis.title.y =element_text(size =9),
axis.title.x =element_text(size =9),
plot.title =element_text(size=9, face="bold",
margin =margin(10, 0, 10, 0), hjust=0.5))
q4<-(ggplot(den4, aes(x=X, y=rho))+geom_line(aes(X, rho))+
geom_ribbon(data=den4,aes(ymin=hi,ymax=lo),alpha=0.3))+ylab(expression(paste(rho, "(X)")))
+xlab("log(Pop)") +
labs(title="1985-1990")+
theme(axis.title.y =element_text(size =9),
axis.title.x =element_text(size =9),
plot.title =element_text(size=9, face="bold",
margin =margin(10, 0, 10, 0), hjust=0.5))

q5<-(ggplot(den5, aes(x=X, y=rho))+geom_line(aes(X, rho))+
geom_ribbon(data=den5,aes(ymin=hi,ymax=lo),alpha=0.3))+
ylab(expression(paste(rho, "(X)"))) +xlab("log(Pop)") +
labs(title="1190-1995")+
theme(axis.title.y =element_text(size =9),
axis.title.x =element_text(size =9),
plot.title =element_text(size=9, face="bold",
margin =margin(10, 0, 10, 0), hjust=0.5))
q6<-(ggplot(den6, aes(x=X, y=rho))+geom_line(aes(X, rho))+
geom_ribbon(data=den6,aes(ymin=hi,ymax=lo),alpha=0.3))+
ylab(expression(paste(rho, "(X)"))) +xlab("log(Pop)") +
labs(title="1995-2000")+
theme(axis.title.y =element_text(size =9),
axis.title.x =element_text(size =9),
plot.title =element_text(size=9, face="bold",
margin =margin(10, 0, 10, 0), hjust=0.5))
q7<-(ggplot(den7, aes(x=X, y=rho))+geom_line(aes(X, rho))+
geom_ribbon(data=den7,aes(ymin=hi,ymax=lo),alpha=0.3))+
ylab(expression(paste(rho, "(X)"))) +xlab("log(Pop)") +
labs(title="2000-2005")+
theme(axis.title.y =element_text(size =9),
axis.title.x =element_text(size =9),
plot.title =element_text(size=9, face="bold",
margin =margin(10, 0, 10, 0), hjust=0.5))
q8<-(ggplot(den8, aes(x=X, y=rho))+geom_line(aes(X, rho))+
geom_ribbon(data=den8,aes(ymin=hi,ymax=lo),alpha=0.3))+
ylab(expression(paste(rho, "(X)"))) +xlab("log(Pop)") +
labs(title="2005-2010")+
theme(axis.title.y =element_text(size =9),
axis.title.x =element_text(size =9),
plot.title =element_text(size=9, face="bold",
margin =margin(10, 0, 10, 0), hjust=0.5))
q9<-(ggplot(den9, aes(x=X, y=rho))+geom_line(aes(X, rho))+
geom_ribbon(data=den9,aes(ymin=hi,ymax=lo),alpha=0.3))+
ylab(expression(paste(rho, "(X)"))) +xlab("log(Pop)") +
labs(title="2010-2015")+
theme(axis.title.y =element_text(size =9),
axis.title.x =element_text(size =9),
plot.title =element_text(size=9, face="bold",
margin =margin(10, 0, 10, 0), hjust=0.5))
```

```r
multiplot(q1,q4 ,q7 , q2, q5,q8, q3,q6,q9, cols=3 )

#INHOMOGENEOUS k-Function
kinhomtexas<-envelope(tornado_texas_ppp, Kinhom, nsim=99,
simulate =expression(rpoispp(a)))

kinhomtexas2<-envelope(tornado_texas_ppp, Kinhom, nsim=99,
simulate =expression(rpoispp(d)))

par(mar=c(2,4,2,4))
v<-plot(kinhomtexas, legend=FALSE, main="Inhomogeneous K-function")
legend(-20000, 190022224257, legend=v$meaning, lty=v$lty, col=v$col, cex=0.6, bty="n")

####space-time
## Space-time inhomogeneous K-function


data<-torn_texas[ ,c(24, 25, 5)]
TX <-as.3dpoints(data[,1]/1000, data[,2]/1000, data[,3])
Poly <-poly/1000

# Estimation of the temporal intensity
Mt <-density(TX[ ,3], n =1000)
mut <-Mt$y[findInterval(TX[ ,3], Mt$x)] *dim(TX)[1]

# Estimation of the spatial intensity
# Finding the optimal bandwidth for kernel smoothing
h <-mse2d(as.points(TX[,1:2]), Poly, nsmse =100, range =4)
h <-h$h[which.min(h$mse)]
Ms <-kernel2d(as.points(TX[ ,1:2]), Poly, h = h, nx =500, ny =500)
atx <-findInterval(x = TX[ ,1], vec = Ms$x)
aty <-findInterval(x = TX[ ,2], vec = Ms$y)
mhat <-NULL
for(i in 1:length(atx)) mhat <-c(mhat, Ms$z[atx[i],aty[i]])

# Estimation of the STIK function
#for dx=400km, dt=100days
u <-seq(0,400, leng =20)
v <-seq(0,100, leng=20)
stik <-STIKhat(xyt = TX, s.region = Poly, t.region =c(1,24056),
lambda = mhat*mut/7355, dist = u, times = v, infectious = T)

plotK(stik, L=FALSE,type="persp", theta =30, phi =20, legend=TRUE)
plotK(stik, L=TRUE, type="persp",theta=30, phi =30,legend=T)

#for dx=400 km, and dt=4years
u <-seq(0,400, leng =20)
v <-seq(0,1460, leng=20)
stik2 <-STIKhat(xyt = TX, s.region = Poly, t.region =c(1,24056),
lambda = mhat*mut/7355, dist = u, times = v, infectious = T)

plotK(stik2, L=FALSE,type="persp", theta =30, phi =20, legend=TRUE)
plotK(stik2, L=TRUE, type="persp",theta=30, phi =30,legend=T)

#for dx=100 km, and dt=4years
u <-seq(0,100, leng =20)
v <-seq(0,1460, leng=20)
stik3 <-STIKhat(xyt = TX, s.region = Poly, t.region =c(1,24056),
lambda = mhat*mut/7355, dist = u, times = v, infectious = T)
plotK(stik3, L=FALSE,type="persp", theta =30, phi =20, legend=TRUE)
plotK(stik3, L=TRUE, type="persp",theta=30, phi =30,legend=T)

#for dx=100 km, and dt=4years
u <-seq(0,100, leng =20)
v <-seq(0,1460, leng=20)
stik3 <-STIKhat(xyt = TX, s.region = Poly, t.region =c(1,24056),
lambda = mhat*mut/7355, dist = u, times = v, infectious = T)

plotK(stik4, L=FALSE,type="persp", theta =20, phi =20, legend=TRUE)
plotK(stik4, L=TRUE, type="persp",theta=30, phi =30,legend=T)
```

A.7. R-code for Lattice approach

```
setwd("D:/TEXAS")

# Load required libraries in R
packages<-function(){
library(sp)
library(ggplot2)
library(foreign)
library(stpp)
library(dplyr)
library(spatstat)
library(rgdal)
library(maptools)
library(raster)
library(data.table)
library(plyr)
library(gridExtra)
library(spdep)
library(INLA)
library(goftest)
library(spacetime)
}
packages()

##Load all tornados
TornALL <-readOGR(dsn ="./tornado/torn", layer ="torn", stringsAsFactors =FALSE)
Data_correction<-function(i=TornALL){
  i$yr <-as.integer(i$yr)
  i$mo <-as.integer(i$mo)
  i$EF <-as.integer(i$mag)
  i$Date <-as.Date(i$date, format="%Y-%m-%d")
  i$Length <-as.numeric(i$len) *1609.34
  i$Width <-as.numeric(i$wid) *0.9144
  i$fat <-as.integer(i$fat)
  i$slon <-as.numeric(i$slon)
  i$slat <-as.numeric(i$slat)
  i$elon <-as.numeric(i$elon)
  i$elat <-as.numeric(i$elat)
  i$inj <-as.numeric(i$inj)
  i$Ref <-1:nrow(i)
return(i)
}

TornALL <-Data_correction(TornALL)
CRS.new <-CRS("+proj=aea +lat_1=29.5 +lat_2=45.5 +lat_0=37.5 +lon_0=-96 +x_0=0
+y_0=0+datum=NAD83 +units=m +no_defs +ellps=GRS80 +towgs84=0,0,0")  #EPSG:102003
TornALL <-spTransform(TornALL, CRS.new)

#Load Boundaries
US.sp <-readOGR(dsn ="./tmp", layer ="cb_2013_us_county_5m",
stringsAsFactors =FALSE)
TX.sp <-US.sp[US.sp$STATEFP ==48, ]
county <-TX.sp$GEOID
county2 <-geometry(spChFIDs(TX.sp, county))
counties <-spTransform(county2, CRS.new)
county<-as.numeric(county)

#Load pop
Pop <-read.csv("Population_final2.csv", header=T, sep=";", dec=".")
Pop <-Pop[,-2]
Pop<-Pop[,-48]
Pop.df =melt(Pop, id.vars ="FIP")
Pop.df$Year =as.numeric(substring(Pop.df$variable, first =4, last =7))
names(Pop.df)[3:4] =c("pop", "YearPop")
Pop.df$lpop =log10(Pop.df$pop)
Pop.df$ID <- ""
Pop.df$ID<-match(Pop.df$FIP, county)



###POP Change
##pop changes by county #http://pages.uoregon.edu/rgp/PPPM613/class8a.htm
```

```
PC <-Pop.df %>%group_by(ID) %>%
summarize(Change = (pop[YearPop ==max(YearPop)] -pop[YearPop ==min(YearPop)])/pop[YearPop
==min(YearPop)] *100)
PC.df =as.data.frame(PC)
row.names(PC.df) =county


####Preparatrion for inla
#1. Subset Texas from big shape
TornTexas<-subset(TornALL, TornALL$st=="TX")
TornTexas<-subset(TornTexas, TornTexas$Ref !=2949&
TornTexas$Ref !=5732&
TornTexas$Ref !=9752&
TornTexas$Ref !=10216&
TornTexas$Ref !=10641&
TornTexas$Ref !=57954&yr>=1970)
#2. Return Number of tornados, first by state, per year, starting in 1970
ct =over(counties, TornTexas, returnList =TRUE)
names(ct) =county
TornAll <-ldply(ct, data.frame)
nTornados <-TornAll %>%filter(!duplicated(Ref)) %>%
group_by(yr, .id) %>%
dplyr::summarize(numberTorn =n())

#3. Creation of Dataframe Number counts/county/year
years<-c(1970:2015)
yearTorn<-rep(years, each=length(county))

countyTorn<-rep(sort(county), times=length(years))

Random.df<-data.frame(County=countyTorn, Year=yearTorn)
colnames(nTornados) [1:2] <-c("Year", "County")
TornInla <-merge(Random.df,nTornados,by=c("Year","County"), all=TRUE)
TornInla[is.na(TornInla)] <-0

#4. Prep. of INLA graph
spdf =SpatialPolygonsDataFrame(counties, PC.df)
spdf$ID<-seq(1:254)
View(spdf)
spdf$area =round((rgeos::gArea(counties, byid =TRUE)/10^6), 5)
spdf$Name =TX.sp$NAME
spdf$FIP =county
a <-aggregate(TornInla$numberTorn, by=list(TornInla$County), FUN=sum)
colnames(a)<-c("FIP", "nT")
spdf<-merge(spdf, a, by="FIP")
View(spdf)

nb =poly2nb(spdf)
nb2INLA("tornb.inla", nb)
tornb.inla =inla.read.graph("tornb.inla")
image(inla.graph2matrix(tornb.inla),xlab="",ylab="")

##Generate spatial ID in INLAtORN. DF
TornInla$ID <- ""
spdf234<-as.data.frame(spdf)
spdf234<-arrange(spdf234, FIP)
ID<-rep(spdf234$ID, 46)
TornInla$ID=ID

##Now ID in INLAtable and ID in INLA graph match!!!
TornInla$ID2<-TornInla$ID
TornInla$Year2<-TornInla$Year

##several controls:
control <-list(
predictor =list(compute =TRUE),
results =list(return.marginals.random =TRUE, return.marginals.predictor=TRUE),
compute =list(hyperpar=TRUE, return.marginals=TRUE, dic=TRUE, mlik =TRUE, cpo =TRUE,
po =TRUE, waic=TRUE, graph=TRUE, gdensity=TRUE, openmp.strategy="huge"),
group =list(model="rw2"))

#1 Model linear trend (or non-linear) only of tornados/year
b<-aggregate(TornInla$numberTorn, by=list(TornInla$Year), FUN=sum)
colnames(b)<-c("Year", "nT")
```

```
b<-b[-47,]
spdf<-merge(spdf, a, by="FIP")

formula<-nT ~Year
modelt0 =inla(formula = formula,
family ="poisson",
quantiles =c(.05, .5, .95),
data = b,
control.compute = control$compute,
control.predictor = control$predictor)
summary(modelt0)

b$Year2<-b$Year

#non-linear trend
formula<-nT ~Year +f(Year2, model="iid")
modelt1 =inla(formula = formula,
family ="poisson",
quantiles =c(.05, .5, .95),
data = b,
control.compute = control$compute,
control.predictor = control$predictor)
summary(modelt1)

###non -linear trend better fit!! But which model? - by DIC, CRW2
formula<-nT ~Year +f(Year2, model="rw2")
modelt2 =inla(formula = formula,
family ="poisson",
quantiles =c(.05, .5, .95),
data = b,
control.compute = control$compute,
control.predictor = control$predictor)
summary(modelt2)

formula<-nT ~Year +f(Year2, model="rw1")
modelt3 =inla(formula = formula,
family ="poisson",
quantiles =c(.05, .5, .95),
data = b,
control.compute = control$compute,
control.predictor = control$predictor)
summary(modelt3)

formula<-nT ~Year +f(Year2, model="crw2")
modelt4 =inla(formula = formula,
family ="poisson",
quantiles =c(.05, .5, .95),
data = b,
control.compute = control$compute,
control.predictor = control$predictor)
summary(modelt4)

formula<-nT ~Year +f(Year2, model="mec")
modelt5 =inla(formula = formula,
family ="poisson",
quantiles =c(.05, .5, .95),
data = b,
control.compute = control$compute,
control.predictor = control$predictor)
summary(modelt5)

formula<-nT ~Year +f(Year2, model="meb")
modelt6 =inla(formula = formula,
family ="poisson",
quantiles =c(.05, .5, .95),
data = b,
control.compute = control$compute,
control.predictor = control$predictor)
summary(modelt6)

#Prepare data for spatial component
nTornadospatial <-TornInla %>%
group_by(County) %>%
dplyr::summarize(numberTorn =sum(numberTorn))
```

```
nTornadospatial$ID<-spdf234$ID

############################Frailty model
#simple random effect (spatially unstructured heterogeneity model)

frailtyformula<-numberTorn~f(ID, model="iid")
E <-mean(nTornadospatial$numberTorn)
frailtymodel=inla(formula=frailtyformula, family ="poisson",
data=nTornadospatial, E=E,
quantiles =c(.05, .5, .95),
control.compute = control$compute,
control.results = control$results,
control.predictor = control$predictor)
summary(frailtymodel)

brier.score <-function(x, m){
with(m, {mean(x^2) -2 *mean(x *mean) +mean(mean^2 +sd^2)})
}
-mean(log(frailtymodel$cpo$cpo))
brier.score(nTornadospatial[["numberTorn"]], frailtymodel[["summary.fitted.values"]])
goftest::cvm.test(frailtymodel$cpo$pit, null="punif")

exp(frailtymodel$summary.fixed)

####analyse random effects
refm<-exp(frailtymodel$summary.random$ID[,2])
a<-data.frame(refm)
a <-ggplot(a, aes(refm))
a+geom_density(fill="cadetblue4", alpha=0.2, colour="azure4")+
ggtitle("Density Plot - Spatial Random Effect Distribution \n Frailty Model")+
theme(plot.title =element_text(hjust =0.5))+
xlab("N=254")+ylab("Density")
a
a<-data.frame(refm)
#plot random effects
#1.merge dataframes
spdf_img<-spdf[,3]
spdf_img$re1<-a[,1]
#3. spplot
range(spdf_img$re1)
rng =c(seq(0, 4, length=5), 9)
rnq =c("#3B9AB2", "#78B7C5", "#EBCC2A", "darkorange1", "#F21A00")
scale =list("SpatialPolygonsRescale", layout.scale.bar(),
offset =c(-900000,-1100000),
scale =300000, fill=c("transparent","black"))
text1 =list("sp.text", c(-900000,-1150000), "0")
text2 =list("sp.text", c(-550000,-1150000), "300 Km")
text4<-list("sp.text", c( -730000, -1270000), cex=0.6, "Projection: EPSG 102003")
arrow =list("SpatialPolygonsRescale", layout.north.arrow(),
offset =c(-900000, -400000), scale =200000)
spplot(spdf_img, "re1", col ="white", at = rng,
col.regions = rnq,
colorkey =list(
space ="bottom", labels=list(
at=round(rng, 1))),
sp.layout=list(scale, text1, text2, text4, arrow),
par.settings =list(axis.line =list(col =NA)),
sub ="Random Effects",
main="Spatial Unstructured Heterogeneity \n Occurence of Tornados in Texas")


#####Convolution model
convolformula <-numberTorn ~f(ID, model ="bym", graph=tornb.inla)
convolmodel <-inla(formula = convolformula, family ="poisson",
quantiles =c(.05, .5, .95),
data = nTornadospatial, E=E,
control.compute = control$compute,
control.predictor = control$predictor)


summary(convolmodel)

exp(convolmodel$summary.fixed)
-mean(log(convolmodel$cpo$cpo))
```

```
brier.score(nTornadospatial[["numberTorn"]], convolmodel[["summary.fitted.values"]])
goftest::cvm.test(convolmodel$cpo$pit, null="punif")


####analyse random effects
refm1<-exp(convolmodel$summary.random$ID[1:254,2])
plot(density(refm))
a<-data.frame(refm)
a <-ggplot(a, aes(refm))
a+geom_density(fill="cadetblue4", alpha=0.2, colour="azure4")+
ggtitle("Density Plot - Spatially Unstructured Effects Distribution \n Convolution Model")+
theme(plot.title =element_text(hjust =0.5))+
xlab("N=254")+ylab("Density")
a
refm<-exp(convolmodel$summary.random$ID[255:508,2])
plot(density(refm))
a<-data.frame(refm)
a <-ggplot(a, aes(refm))
a+geom_density(fill="cadetblue4", alpha=0.2, colour="azure4")+
ggtitle("Density Plot - Spatially Structured Effects Distribution \n Convolution Model")+
theme(plot.title =element_text(hjust =0.5))+
xlab("N=254")+ylab("Density")
a

#plot random effects
#1.merge dataframes
re2<-refm1
spdf_img$re2<-re2
#3. spplot
range(spdf_img$re2)
rng =c(seq(0, 4, length=5), 9)
rnq =c("#3B9AB2", "#78B7C5", "#EBCC2A", "darkorange1", "#F21A00")
scale =list("SpatialPolygonsRescale", layout.scale.bar(),
offset =c(-900000,-1100000),
scale =300000, fill=c("transparent","black"))
text1 =list("sp.text", c(-900000,-1150000), "0")
text2 =list("sp.text", c(-550000,-1150000), "300 Km")
text4<-list("sp.text", c( -730000, -1270000), cex=0.6, "Projection: EPSG 102003")
arrow =list("SpatialPolygonsRescale", layout.north.arrow(),
offset =c(-900000, -400000), scale =200000)
spplot(spdf_img, "re2", col ="white", at = rng,
col.regions = rnq,
colorkey =list(
space ="bottom", labels=list(
at=round(rng, 1))),
sp.layout=list(scale, text1, text2, text4, arrow),
par.settings =list(axis.line =list(col =NA)),
sub ="Random Effects",
main="Spatial Unstructured Heterogeneity \n Occurence of Tornados in Texas \n Convolution
Model")

re3<-refm
spdf_img$re3<-re3
#3. spplot
range(spdf_img$re3)
rng =c(seq(0, 4, length=5), 7)
rnq =c("#3B9AB2", "#78B7C5", "#EBCC2A", "darkorange1", "#F21A00")
scale =list("SpatialPolygonsRescale", layout.scale.bar(),
offset =c(-900000,-1100000),
scale =300000, fill=c("transparent","black"))
text1 =list("sp.text", c(-900000,-1150000), "0")
text2 =list("sp.text", c(-550000,-1150000), "300 Km")
text4<-list("sp.text", c( -730000, -1270000), cex=0.6, "Projection: EPSG 102003")
arrow =list("SpatialPolygonsRescale", layout.north.arrow(),
offset =c(-900000, -400000), scale =200000)
spplot(spdf_img, "re3", col ="white", at = rng,
col.regions = rnq,
colorkey =list(
space ="bottom", labels=list(
at=round(rng, 1))),
sp.layout=list(scale, text1, text2, text4, arrow),
par.settings =list(axis.line =list(col =NA)),
sub ="Spatial Effects",
main="Spatial Structured Heterogeneity \n Occurence of Tornados in Texas \n Convolution
```

```
Model")


#fitted
CARfit<-NULL
CARfit<-convolmodel$summary.fitted.values[,1]
CARfit<-data.frame(CARfit=CARfit, ID=nTornadospatial$ID)
View(spdf_img)
spdf_img<-spdf[,3]

spdf_img<-merge(spdf_img, CARfit, by="ID")
range(spdf_img$CARfit)
rng =c(seq(0, 4, length=5), 8)
rnq =c("#3B9AB2", "#78B7C5", "#EBCC2A", "darkorange1", "#F21A00")
scale =list("SpatialPolygonsRescale", layout.scale.bar(),
offset =c(-900000,-1100000),
scale =300000, fill=c("transparent","black"))
text1 =list("sp.text", c(-900000,-1150000), "0")
text2 =list("sp.text", c(-550000,-1150000), "300 Km")
text4<-list("sp.text", c( -730000, -1270000), cex=0.6, "Projection: EPSG 102003")
arrow =list("SpatialPolygonsRescale", layout.north.arrow(),
offset =c(-900000, -400000), scale =200000)
spplot(spdf_img, "CARfit", col ="white", at = rng,
col.regions = rnq,
colorkey =list(
space ="bottom", labels=list(
at=round(rng, 1))),
sp.layout=list(scale, text1, text2, text4, arrow),
par.settings =list(axis.line =list(col =NA)),
sub ="Fitted Effects",
main="Fitted Effects \n Occurence of Tornados in Texas \n Convolution Model")



#spatial risk
CARmarginals<-convolmodel$marginals.random$ID[1:254]
CARzeta<-lapply(CARmarginals, function (x) inla.emarginal(exp, x))
risk<-data.frame(CARzeta=unlist(CARzeta), ID=seq(1, 254, 1))
risk<-merge(nTornadospatial, risk, by="ID")
spdf_img<-spdf[,3]
spdf_img<-merge(spdf_img, risk, by="ID")
View(spdf_img)
rng =c(seq(0, 4, length=5), 9)
rnq =c("#3B9AB2", "#78B7C5", "#EBCC2A", "darkorange1", "#F21A00")
spplot(spdf_img, "CARzeta", col ="white", at = rng,
col.regions = rnq,
colorkey =list(
space ="bottom", labels=list(
at=round(rng, 1))),
sp.layout=list(scale, text1, text2, text4, arrow),
par.settings =list(axis.line =list(col =NA)),
sub ="Marginal Effects",
main="Spatial Risk \n Occurence of Tornados in Texas \n Convolution Model")


#######spatio-temporal
#SP+TIME uncorrelated
E=mean(TornInla$numberTorn)
convolformulat <-numberTorn ~f(ID, model ="bym", graph=tornb.inla)+f(Year, model="iid")
convolmodelt <-inla(formula = convolformulat, family ="poisson",
quantiles =c(.05, .5, .95),
data = TornInla, E=E,
control.compute = control$compute,
control.predictor = control$predictor)
summary(convolmodelt)
exp(convolmodelt$summary.fixed)
-mean(log(convolmodelt$cpo$cpo))
brier.score(TornInla[["numberTorn"]], convolmodelt[["summary.fitted.values"]])
goftest::cvm.test(convolmodelt$cpo$pit, null="punif")

re<-convolmodelt$summary.random$ID[1:254,2]
plot(density(re))
a<-data.frame(re)
a <-ggplot(a, aes(re))
```

```
a+geom_density(fill="cadetblue4", alpha=0.2, colour="azure4")+
ggtitle("Spatially Unstructured Effects")+
theme(plot.title =element_text(hjust =0.5))+
xlab("N=254")+ylab("Density")

care<-convolmodelt$summary.random$ID[255:508, 2]
a<-data.frame(care)
a <-ggplot(a, aes(care))
a+geom_density(fill="cadetblue4", alpha=0.2, colour="azure4")+
ggtitle("Spatially Structured Effects")+
theme(plot.title =element_text(hjust =0.5))+
xlab("N=254")+ylab("Density")

tre<-convolmodelt$summary.random$Year[,2]
a<-data.frame(tre)
a <-ggplot(a, aes(tre))
a+geom_density(fill="cadetblue4", alpha=0.2, colour="azure4")+
ggtitle("Temporal Unstructured Effects")+
theme(plot.title =element_text(hjust =0.5))+
xlab("N=254")+ylab("Density")


##sp-time correlated
convolformulat2<-numberTorn ~f (ID, model="bym", graph = tornb.inla)+
f(Year, model="rw1")
convolmodelt2<-inla(formula = convolformulat2, family ="poisson",
quantiles =c(.05, .5, .95),
data = TornInla, E=E,
control.compute = control$compute,
control.predictor = control$predictor)

summary(convolmodelt2)
exp(convolmodelt2$summary.fixed)
-mean(log(convolmodelt2$cpo$cpo))
brier.score(TornInla[["numberTorn"]], convolmodelt2[["summary.fitted.values"]])
goftest::cvm.test(convolmodelt2$cpo$pit, null="punif")


re<-convolmodelt2$summary.random$ID[1:254,2]
a<-data.frame(re)
a <-ggplot(a, aes(re))
a+geom_density(fill="cadetblue4", alpha=0.2, colour="azure4")+
ggtitle("Spatially Unstructured Effects")+
theme(plot.title =element_text(hjust =0.5))+
xlab("N=254")+ylab("Density")

care<-convolmodelt2$summary.random$ID[255:508, 2]
a<-data.frame(care)
a <-ggplot(a, aes(care))
a+geom_density(fill="cadetblue4", alpha=0.2, colour="azure4")+
ggtitle("Spatially Structured Effects")+
theme(plot.title =element_text(hjust =0.5))+
xlab("N=254")+ylab("Density")

tre<-convolmodelt2$summary.random$Year[,2]
a<-data.frame(tre)
a <-ggplot(a, aes(tre))
a+geom_density(fill="cadetblue4", alpha=0.2, colour="azure4")+
ggtitle("Temporal Structured Effects")+
theme(plot.title =element_text(hjust =0.5))+
xlab("N=254")+ylab("Density")


#map fitted effects
carfit<-data.frame(FIT=convolmodelt2$summary.fitted.values[,1], ID=TornInla$ID,
Year=TornInla$Year)
meanfit<-carfit %>%
group_by(ID) %>%
dplyr::summarize(meannumberTorn =mean(FIT))
spdf_img<-merge(spdf_img, meanfit, by="ID")

rng =c(seq(0, 4, length=5), 7)
rnq =c("#3B9AB2", "#78B7C5", "#EBCC2A", "darkorange1", "#F21A00")
spplot(spdf_img, "meannumberTorn", col ="white", at = rng,
```

```
col.regions = rnq,
colorkey =list(
space ="bottom", labels=list(
at=round(rng, 1))),
sp.layout=list(scale, text1, text2, text4, arrow),
par.settings =list(axis.line =list(col =NA)),
sub ="Mean Fitted Effects",
main="Mean Spatio-temporal Fitted Effects")


#Marginal Effects
CARmarginals<-convolmodelt2$marginals.random$ID[1:254]
CARzeta<-lapply(CARmarginals, function (x) inla.emarginal(exp, x))
risk<-data.frame(CARzeta=unlist(CARzeta), ID=seq(1, 254, 1))
risk<-merge(nTornadospatial, risk, by="ID")
spdf_img<-spdf[,3]
spdf_img<-merge(spdf_img, risk, by="ID")

rng =c(seq(0, 4, length=5), 9)
rnq =c("#3B9AB2", "#78B7C5", "#EBCC2A", "darkorange1", "#F21A00")
spplot(spdf_img, "CARzeta", col ="white", at = rng,
col.regions = rnq,
colorkey =list(
space ="bottom", labels=list(
at=round(rng, 1))),
sp.layout=list(scale, text1, text2, text4, arrow),
par.settings =list(axis.line =list(col =NA)),
sub ="Marginal Effects",
main="Spatial Risk \n Occurence of Tornados in Texas")


df =convolmodelt2$summary.fitted.values
names(df) =c("mean", "sd", "QL", "QM", "QH", "mode")
df$ID<-TornInla$ID
df$Year<-TornInla$Year
df =df %>%group_by(ID) %>%
dplyr::summarize(mean=mean(mean), sd=mean(sd),
QL=mean(QL), QM=mean(QM),
QH=mean(QH))

df<-df%>%mutate(QL = QL -1,
QH = QH -1,
Sig =sign(QL) ==sign(QH),
sd = sd,
ctyPerState = (mean -1)*100)

sum(df$Sig)

spdfR =spdf
spdfR@data =df

range(spdfR$ctyPerState)
rng =c(seq(-100, 100, length=8), 200, 600)
rnq =c(rev(RColorBrewer::brewer.pal(8, "RdYlGn")), "#8c510a", "#543005")
spplot(spdfR, "ctyPerState", col ="white", at = rng,
col.regions = rnq,
colorkey =list(
space ="bottom", labels=list(
at=round(rng))),
sp.layout=list(scale, text1, text2, text4, arrow),
par.settings =list(axis.line =list(col =NA)),
sub ="Percentage of difference from Statewide Average Rate",
main="Occurence rate of Tornados in Texas")

range(spdfR$sd)
rng<-seq(0, 0.80, length=9)
rnq<-c("#e0f3db", "#ccebc5", "#a8ddb5", "#7bccc4", "#4eb3d3", "#2b8cbe", "#0868ac", "#084081")
spplot(spdfR, "sd", col ="white", at = rng,
col.regions = rnq,
colorkey =list(
space ="bottom", labels=list(
at=round(rng, 2))),
sp.layout=list(scale, text1, text2, text4, arrow),
par.settings =list(axis.line =list(col =NA)),
```

```
sub ="Standard Error")


###################################ADD COVARIATES
TornInla$Pop<-Pop.df$pop
TornInla$Lpop<-Pop.df$lpop
a<-data.frame(area=round(spdf$area, 5), County=spdf$FIP)

TornInla<-merge(TornInla, a, by="County")
TornInla <-dplyr::arrange(TornInla, Year)
TornInla$DPop<-""
TornInla$DPop<-TornInla$Pop/TornInla$area

convolformulatc2<-numberTorn ~f (ID, model="bym", graph = tornb.inla)+
f(Year, model="rw1") +DPop
convolmodeltc2<-inla(formula = convolformulatc2, family ="poisson",
quantiles =c(.05, .5, .95),
data = TornInla, E=E,
control.compute = control$compute,
control.predictor = control$predictor)

summary(convolmodeltc2)
exp(convolmodeltc2$summary.fixed)
-mean(log(convolmodeltc2$cpo$cpo))
brier.score(TornInla[["numberTorn"]], convolmodeltc2[["summary.fitted.values"]])
goftest::cvm.test(convolmodeltc2$cpo$pit, null="punif")

###Insert Roughness Index
wcounty<-unionSpatialPolygons(counties, ID =rep("1", length(row.names(counties))))
Ind<-raster("Index_Value1.tif")
Ind<-as(crop(Ind, extent(wcounty)), "SpatialGridDataFrame")
proj4string(Ind) =proj4string(wcounty) #same projection, different datum & ellipsoid
RI.data<-over(counties, Ind, returnList =TRUE)

Elev.df =data.frame(county =rep(county, sapply(RI.data, nrow)),
Elev =unlist(RI.data),
ID =rep(spdf$ID, sapply(RI.data, nrow)),
stringsAsFactors =FALSE)

CE.df =Elev.df %>%group_by(ID) %>%
dplyr::summarize(elev =mean(Elev, na.rm =TRUE),
elevS =sd(Elev, na.rm =TRUE),
elevCV = elevS/elev)
all(spdf$ID ==CE.df$ID)
TornInla =merge(TornInla, CE.df, by ="ID")


####Models with st of RI
formulamodeltc3 <-numberTorn ~f(ID, model ="bym", graph=tornb.inla) +f(Year, model="rw1")
+elevS
modeltc3 <-inla(formula = formulamodeltc3, family ="poisson",
quantiles =c(.05, .5, .95),
data = TornInla, E=E,
control.compute = control$compute,
control.predictor = control$predictor,
control.results=control$results)

summary(modeltc3)
exp(modeltc3$summary.fixed)
-mean(log(modeltc3$cpo$cpo))
brier.score(TornInla[["numberTorn"]], modeltc3[["summary.fitted.values"]])
goftest::cvm.test(modeltc3$cpo$pit, null="punif")


####INSERT LAND-COVER
###Adding Land-Cover
lc1992<-read.dbf("./Area_Per_County_sqM/1992.dbf")
lc2001<-read.dbf("./Area_Per_County_sqM/2001.dbf")
lc2006<-read.dbf("./Area_Per_County_sqM/2006.dbf")
lc2011<-read.dbf("./Area_Per_County_sqM/2011.dbf")

# percentages:
tarea<-data.frame(Area= spdf$area*10^6, GEOID=spdf$FIP)
lc1992<-merge(lc1992, tarea, by="GEOID")
```

```
lc1992$perc11<-lc1992$VALUE_11/lc1992$Area*100
lc1992$perc21<-lc1992$VALUE_21/lc1992$Area*100
lc1992$perc31<-lc1992$VALUE_31/lc1992$Area*100
lc1992$perc41<-lc1992$VALUE_41/lc1992$Area*100
lc1992$perc51<-lc1992$VALUE_51/lc1992$Area*100
lc1992$perc91<-lc1992$VALUE_91/lc1992$Area*100


lc2001<-merge(lc2001, tarea, by="GEOID")
lc2001$perc11<-lc2001$VALUE_11/lc2001$Area*100
lc2001$perc21<-lc2001$VALUE_21/lc2001$Area*100
lc2001$perc31<-lc2001$VALUE_31/lc2001$Area*100
lc2001$perc41<-lc2001$VALUE_41/lc2001$Area*100
lc2001$perc51<-lc2001$VALUE_51/lc2001$Area*100
lc2001$perc91<-lc2001$VALUE_91/lc2001$Area*100

lc2006<-merge(lc2006, tarea, by="GEOID")
lc2006$perc11<-lc2006$VALUE_11/lc2006$Area*100
lc2006$perc21<-lc2006$VALUE_21/lc2006$Area*100
lc2006$perc31<-lc2006$VALUE_31/lc2006$Area*100
lc2006$perc41<-lc2006$VALUE_41/lc2006$Area*100
lc2006$perc51<-lc2006$VALUE_51/lc2006$Area*100
lc2006$perc91<-lc2006$VALUE_91/lc2006$Area*100

lc2011<-merge(lc2011, tarea, by="GEOID")
lc2011$perc11<-lc2011$VALUE_11/lc2011$Area*100
lc2011$perc21<-lc2011$VALUE_21/lc2011$Area*100
lc2011$perc31<-lc2011$VALUE_31/lc2011$Area*100
lc2011$perc41<-lc2011$VALUE_41/lc2011$Area*100
lc2011$perc51<-lc2011$VALUE_51/lc2011$Area*100
lc2011$perc91<-lc2011$VALUE_91/lc2011$Area*100

a<-rep(lc1992$perc11, times=23)
b<-rep(lc2001$perc11, times=10)
c<-rep(lc2006$perc11, times=5)
d<-rep(lc2011$perc11, times=8)

asd<-c(a,b,c,d)

TornInla$perc11<-asd

a<-rep(lc1992$perc21, times=23)
b<-rep(lc2001$perc21, times=10)
c<-rep(lc2006$perc21, times=5)
d<-rep(lc2011$perc21, times=8)

asd<-c(a,b,c,d)

TornInla$perc21<-asd

a<-rep(lc1992$perc41, times=23)
b<-rep(lc2001$perc41, times=10)
c<-rep(lc2006$perc41, times=5)
d<-rep(lc2011$perc41, times=8)

asd<-c(a,b,c,d)
TornInla$perc41<-asd

a<-rep(lc1992$perc51, times=23)
b<-rep(lc2001$perc51, times=10)
c<-rep(lc2006$perc51, times=5)
d<-rep(lc2011$perc51, times=8)

asd<-c(a,b,c,d)

TornInla$perc51<-asd

a<-rep(lc1992$perc31, times=23)
b<-rep(lc2001$perc31, times=10)
c<-rep(lc2006$perc31, times=5)
d<-rep(lc2011$perc31, times=8)

asd<-c(a,b,c,d)
```

```
TornInla$perc31<-asd

a<-rep(lc1992$perc91, times=23)
b<-rep(lc2001$perc91, times=10)
c<-rep(lc2006$perc91, times=5)
d<-rep(lc2011$perc91, times=8)

asd<-c(a,b,c,d)

TornInla$perc91<-asd


formulamodeltc4 <-numberTorn ~f(ID, model ="bym", graph=tornb.inla) +f(Year, model="rw1")
+perc11
modeltc4 <-inla(formula = formulamodeltc4, family ="poisson",
quantiles =c(.05, .5, .95),
data = TornInla, E=E,
control.compute = control$compute,
control.predictor = control$predictor,
control.results=control$results)

summary(modeltc4)
exp(modeltc4$summary.fixed)
-mean(log(modeltc4$cpo$cpo))
brier.score(TornInla[["numberTorn"]], modeltc4[["summary.fitted.values"]])
goftest::cvm.test(modeltc4$cpo$pit, null="punif")


formulamodeltc5 <-numberTorn ~f(ID, model ="bym", graph=tornb.inla) +f(Year, model="rw1")
+perc21
modeltc5 <-inla(formula = formulamodeltc5, family ="poisson",
quantiles =c(.05, .5, .95),
data = TornInla, E=E,
control.compute = control$compute,
control.predictor = control$predictor,
control.results=control$results)

summary(modeltc5)
exp(modeltc5$summary.fixed)
-mean(log(modeltc5$cpo$cpo))
brier.score(TornInla[["numberTorn"]], modeltc5[["summary.fitted.values"]])
goftest::cvm.test(modeltc5$cpo$pit, null="punif")


formulamodeltc6 <-numberTorn ~f(ID, model ="bym", graph=tornb.inla) +
f(Year, model="rw1") +perc31
modeltc6 <-inla(formula = formulamodeltc6, family ="poisson",
quantiles =c(.05, .5, .95),
data = TornInla, E=E,
control.compute = control$compute,
control.predictor = control$predictor,
control.results=control$results)

summary(modeltc6)
exp(modeltc6$summary.fixed)
-mean(log(modeltc6$cpo$cpo))
brier.score(TornInla[["numberTorn"]], modeltc6[["summary.fitted.values"]])
goftest::cvm.test(modeltc6$cpo$pit, null="punif")


formulamodeltc7 <-numberTorn ~f(ID, model ="bym", graph=tornb.inla) +
f(Year, model="rw1") +perc41
modeltc7 <-inla(formula = formulamodeltc7, family ="poisson",
quantiles =c(.05, .5, .95),
data = TornInla, E=E,
control.compute = control$compute,
control.predictor = control$predictor,
control.results=control$results)

summary(modeltc7)
exp(modeltc7$summary.fixed)
-mean(log(modeltc7$cpo$cpo))
brier.score(TornInla[["numberTorn"]], modeltc7[["summary.fitted.values"]])
goftest::cvm.test(modeltc7$cpo$pit, null="punif")
```

```
formulamodeltc8 <-numberTorn ~f(ID, model ="bym", graph=tornb.inla) +
f(Year, model="rw1") +perc51
modeltc8 <-inla(formula = formulamodeltc8, family ="poisson",
quantiles =c(.05, .5, .95),
data = TornInla, E=E,
control.compute = control$compute,
control.predictor = control$predictor,
control.results=control$results)
summary(modeltc8)
exp(modeltc8$summary.fixed)
-mean(log(modeltc8$cpo$cpo))
brier.score(TornInla[["numberTorn"]], modeltc8[["summary.fitted.values"]])
goftest::cvm.test(modeltc8$cpo$pit, null="punif")


formulamodeltc9 <-numberTorn ~f(ID, model ="bym", graph=tornb.inla) +
f(Year, model="rw1") +perc91
modeltc9 <-inla(formula = formulamodeltc9, family ="poisson",
quantiles =c(.05, .5, .95),
data = TornInla, E=E,
control.compute = control$compute,
control.predictor = control$predictor,
control.results=control$results)

summary(modeltc9)
exp(modeltc9$summary.fixed)
-mean(log(modeltc9$cpo$cpo))
brier.score(TornInla[["numberTorn"]], modeltc9[["summary.fitted.values"]])
goftest::cvm.test(modeltc9$cpo$pit, null="punif")


##ALL covariates
formulamodeltc5 <-numberTorn ~f(ID, model ="bym", graph=tornb.inla) +
f(Year, model="rw1") +perc21+perc31+perc41+perc11+perc51+perc91+elevS+DPop
modeltc5 <-inla(formula = formulamodeltc5, family ="poisson",
quantiles =c(.05, .5, .95),
data = TornInla, E=E,
control.compute = control$compute,
control.predictor = control$predictor,
control.results=control$results)

MOD=modeltc5
par(mfrow=c(2,4))

plot(MOD$marginals.fixed$elevS, main="SDTPI", xlab="", ylab="", xlim=c(-5,5),  yaxt='n',
type="l")
abline(v=0, col="red")
plot(MOD$marginals.fixed$perc21, main="(%) Residential", xlim=c(-0.1,0.1),  xlab="", ylab="",
yaxt='n', type="l")
abline(v=0, col="red")
plot(MOD$marginals.fixed$perc51, main="(%) Low-Grass", xlim=c(-0.1,0.1), xlab="", ylab="",
yaxt='n', type="l")
abline(v=0, col="red")
plot(MOD$marginals.fixed$perc31, main=" (%) Barren", xlim=c(-0.1,0.1),  xlab="", ylab="",
yaxt='n', type="l")
abline(v=0, col="red")
plot(MOD$marginals.fixed$perc41, main=" (%) Forest",xlim=c(-0.1,0.1),  xlab="", ylab="",
yaxt='n', type="l")
abline(v=0, col="red")
plot(MOD$marginals.fixed$perc11, main="(%) Water", xlim=c(-0.1,0.1),  xlab="", ylab="",
yaxt='n', type="l")
abline(v=0, col="red")
plot(MOD$marginals.fixed$DPop, main="Pop Density", xlim=c(-0.1,0.1),  xlab="", ylab="",
yaxt='n', type="l")
abline(v=0, col="red")
plot(MOD$marginals.fixed$elevS, main="(%) Wetlands", xlim=c(-5,5),  xlab="", ylab="",
yaxt='n', type="l")
abline(v=0, col="red")
dev.off()

exp(MOD$summary.fixed)
#####Bernardinelli
```

```
formulamodeltc11 <-numberTorn ~f(ID, model ="bym", graph=tornb.inla) +
f(Year, model="iid") +Year2+
elevS +perc31 +perc41 +perc11 +perc51 +perc91 +perc21

modeltc11 <-inla(formula = formulamodeltc11, family ="poisson",
quantiles =c(.05, .5, .95),
data = TornInla, E=E,
control.compute = control$compute,
control.results=control$results)

summary(modeltc11)
exp(modeltc11$summary.fixed)
-mean(log(modeltc11$cpo$cpo))
brier.score(TornInla[["numberTorn"]], modeltc11[["summary.fitted.values"]])
goftest::cvm.test(modeltc11$cpo$pit, null="punif")


##Knorr held
formulamodeltc12 <-numberTorn ~f(ID, model ="bym", graph=tornb.inla) +
f(Year, model="rw1")+f(Year2, model ="iid") +
elevS +perc11+perc21+perc31 +perc41 +perc51+perc91
modeltc12 <-inla(formula = formulamodeltc12, family ="poisson",
quantiles =c(.05, .5, .95),
data = TornInla, E=E,
control.compute = control$compute,
control.results=control$results)

summary(modeltc12)
exp(modeltc12$summary.fixed)
-mean(log(modeltc11$cpo$cpo))
brier.score(TornInla[["numberTorn"]], modeltc11[["summary.fitted.values"]])
goftest::cvm.test(modeltc11$cpo$pit, null="punif")


#type I interaction
TornInla$area.year <-seq(1,length(countyTorn))

formTypeI <-numberTorn~+f(ID, model="bym", graph=tornb.inla)+
f(Year, model="rw1") +f(Year2, model="iid")+
f(area.year, model="iid")+
perc11+perc21+perc31+perc41+perc51+perc91+elevS

mod.intI <-inla(formTypeI,family="poisson",data=TornInla,
control.predictor=control$predictor,
control.compute=control$compute)

summary(mod.intI)
exp(mod.intI$summary.fixed)
-mean(log(mod.intI$cpo$cpo))
brier.score(TornInla[["numberTorn"]], mod.intI[["summary.fitted.values"]])
goftest::cvm.test(mod.intI$cpo$pit, null="punif")

save.image("hopefully.RData")

#exploring final model:
MOD<-mod.intI
exp(MOD$summary.fixed)

summary(MOD)
#map fitted effects
carfit<-data.frame(FIT=MOD$summary.fitted.values[,1], ID=TornInla$ID, Year=TornInla$Year)
meanfit<-carfit %>%
group_by(ID) %>%
dplyr::summarize(meannumberTorn =mean(FIT))
spdf_img<-spdf[,3]
spdf_img<-merge(spdf_img, meanfit, by="ID")

range(spdf_img$meannumberTorn)
rng =c(seq(0, 2, length=5), 5)
rnq =c("#3B9AB2", "#78B7C5", "#EBCC2A", "darkorange1", "#F21A00")
spplot(spdf_img, "meannumberTorn", col ="white", at = rng,
col.regions = rnq,
colorkey =list(
space ="bottom", labels=list(
```

```
at=round(rng, 1))),
sp.layout=list(scale, text1, text2, text4, arrow),
par.settings =list(axis.line =list(col =NA)),
sub ="Mean Fitted Effects",
main="Mean Spatio-temporal Fitted Effects")


#spatial risk
#Marginal Effects
CARmarginals<-MOD$marginals.random$ID[1:254]
CARzeta<-lapply(CARmarginals, function (x) inla.emarginal(exp, x))
risk<-data.frame(CARzeta=unlist(CARzeta), ID=seq(1, 254, 1))
risk<-merge(nTornadospatial, risk, by="ID")
spdf_img<-spdf[,3]
spdf_img<-merge(spdf_img, risk, by="ID")

rng =c(seq(0, 4, length=5), 9)
rnq =c("#3B9AB2", "#78B7C5", "#EBCC2A", "darkorange1", "#F21A00")
spplot(spdf_img, "CARzeta", col ="white", at = rng,
col.regions = rnq,
colorkey =list(
space ="bottom", labels=list(
at=round(rng, 1))),
sp.layout=list(scale, text1, text2, text4, arrow),
par.settings =list(axis.line =list(col =NA)),
sub ="Marginal Effects",
main="Spatial Risk \n Occurence of Tornados in Texas")


#exceedence Probability
a=log(1)
stexceed<-lapply(MOD$marginals.random$ID[1:254],
                 function (X) {
1-inla.pmarginal(a, X)
                 })
stexceed<-unlist(stexceed)
risk<-data.frame(stexceed=stexceed, ID=seq(1, 254, 1))
spdf_img<-spdf[,3]
spdf_img<-merge(spdf_img, risk, by="ID")
rng =seq(0, 1, length=10)
rnq=rev(RColorBrewer::brewer.pal(9, "RdYlGn"))
spplot(spdf_img, "stexceed", col ="white", at = rng,
col.regions = rnq,
colorkey =list(
space ="bottom", labels=list(
at=round(rng, 1))),
sp.layout=list(scale, text1, text2, text4, arrow),
par.settings =list(axis.line =list(col =NA)),
main ="More than one tornado per county",
sub="Probability")


a=log(2)
stexceed<-lapply(MOD$marginals.random$ID[1:254],
                 function (X) {
1-inla.pmarginal(a, X)
                 })
stexceed<-unlist(stexceed)
risk<-data.frame(stexceed=stexceed, ID=seq(1, 254, 1))
spdf_img<-spdf[,3]
spdf_img<-merge(spdf_img, risk, by="ID")
rng =seq(0, 1, length=10)
rnq=rev(RColorBrewer::brewer.pal(9, "RdYlGn"))
spplot(spdf_img, "stexceed", col ="white", at = rng,
col.regions = rnq,
colorkey =list(
space ="bottom", labels=list(
at=round(rng, 1))),
sp.layout=list(scale, text1, text2, text4, arrow),
par.settings =list(axis.line =list(col =NA)),
main ="More than two tornados per county",
sub="Probability")
```

```
pop density map

dens<-subset(TornInla, TornInla$Year==2015 )
dens1<-merge(spdf, dens, by="ID")
rng =c(seq(0, 100, length=8),200, 600, 1000, 1200)
rnq=rev(RColorBrewer::brewer.pal(11, "RdYlBu"))
spplot(dens1, "DPop", col ="grey", at = rng,
col.regions = rnq,
colorkey =list(
space ="bottom", labels=list(
at=c(0,100,200,600,1000,1200))),
sp.layout=list(scale, text1, text2, text4, arrow),
par.settings =list(axis.line =list(col =NA)),
main ="Population Density - Texas (2015)",
sub="Individuals per Sq Km")
```

A.8 Packages used in R

#sp

Pebesma, E. J., Bivand, R. S. (2005) Classes and methods for spatial data in R. *R News*, 5, 2. Available at: https://cran.r-project.org/doc/Rnews/

Bivand, R., S., Pebesma, E., Gomez-Rubio, V. (2013) *Applied spatial data analysis with R*. Use R! Series Springer.

#ggplot2

Wickham, H. (2009) *ggplot2: Elegant Graphics for Data Analysis*. New-York: Springer-Verlag.

#foreign

R Core Team (2016) foreign: Read Data Stored by Minitab, S, SAS, SPSS, Stata, Systat, Weka, dBase, ... *R package version 0.8-67*. Available at: https://CRAN.R-project.org/package=foreign

#stpp

Gabriel, E., Diggle P., J., Rowlingson, B. (2014) stpp: Space-Time Point Pattern simulation, visualization and analysis. *R package version 1.0-5*. Available at: https://CRAN.R-project.org/package=stpp

#dplyr

Wickham, H., Francois, R. (2016) dplyr: A Grammar of Data Manipulation. *R package version 0.5.0*. Available at: https://CRAN.R-project.org/package=dplyr

#spatstat

Baddeley, A., Rubak, E., Turner, R. (2015) Spatial Point Patterns: Methodology and Applications with R. London: Chapman and Hall/CRC Press, 2015.

#rgdal

Bivand, R., Keitt, T., Rowlingson, B. (2016) rgdal: Bindings for the Geospatial Data Abstraction Library. *R package version 1.2-5*. Available at: https://CRAN.R-project.org/package=rgdal

#maptools

Bivand, R., Lewin-Koh, N. (2017) maptools: Tools for Reading and Handling Spatial Objects. *R package version 0.8-41*. Available at: https://CRAN.R-project.org/package=maptools

#raster

Hijmans, R., J. (2016) raster: Geographic Data Analysis and Modeling. *R package version 2.5-8*. Available at: https://CRAN.R-project.org/package=raster

#data.table
Dowle, M., Srinivasan, A. (2016) data.table: Extension of `data.frame`. *R package version 1.10.0*. Available at: https://CRAN.R-project.org/package=data.table

#plyr
Wickham, H. (2011) The Split-Apply-Combine Strategy for Data Analysis. *Journal of Statistical Software*, 40(1), 1-29.

#spded
Bivand, R., Piras, G. (2015) Comparing Implementations of Estimation Methods for Spatial Econometrics. *Journal of Statistical Software*, 63(18), 1-36.

Bivand, R. S., Hauke, J., and Kossowski, T. (2013) Computing the Jacobian in Gaussian spatial autoregressive models: An illustrated comparison of available methods. *Geographical Analysis*, 45(2), 150-179.

#INLA
Rue, H., Martino, S., Chopin, N. (2009) Approximate Bayesian Inference for Latent Gaussian Models Using Integrated Nested Laplace Approximations (with discussion). *Journal of the Royal Statistical Society B*, 71, 319-392.

Martins, T., G., Simpson, D., Lindgren, F., Rue, H. (2013) Bayesian computing with INLA: New features. *Computational Statistics and Data Analysis*, 63, 68-83.

Lindgren, F., Rue, H., Lindstrom, J. (2011) An Explicit Link Between Gaussian Fields and Gaussian Markov Random Fields: The Stochastic Partial Differential Equation Approach (with discussion). *Journal of the Royal Statistical Society B*, 73(4), 423-498.

Lindgren, F., Havard R. (2015) Bayesian Spatial Modelling with R-INLA. *Journal of Statistical Software*, 63(19), 1-25. Available at: http://www.jstatsoft.org/v63/i19/

#spacetime
Pebesma, E. (2012) spacetime: Spatio-Temporal Data in R. *Journal of Statistical Software*, 51(7), 1-30. Available at: http://www.jstatsoft.org/v51/i07/.

Bivand, R., Pebesma, E., Gomez-Rubio, V. (2013) Applied spatial data analysis with R. New York: Springer.

#gridExtra
Baptiste, A. (2016) gridExtra: Miscellaneous Functions for "Grid" Graphics. *R package version 2.2.1*. Available at: https://CRAN.R-project.org/package=gridExtra

#goftest
Faraway, J., Marsaglia, G., Marsaglia, J., Baddeley, A. (2015) goftest: Classical goodness of fit Test for Univariate Distributions. R package version 1.0-3. Available at: https://CRAN.R-project.org/package=goftest

#wesanderson
Karthik, R., Wickham, H. (2015) wesanderson: A Wes Anderson Palette Generator. *R package version 0.3.2*. https://CRAN.R-project.org/package=wesanderson

A.9 Map of Texas with Counties Discrimination