NOVA SCHOOL OF SCIENCE & TECHNOLOGY

DEPARTMENT OF MATHEMATICS

VÍTOR ALEXANDRE CANHÃO AUGUSTO

Bachelor in Mathematics

ANALYSIS OF CHANGES IN ANNUAL PRECIPITATION PATTERNS IN ALENTEJO REGION USING LOG-LINEAR MODELS

MASTER IN MATHEMATICS AND APPLICATIONS

NOVA University Lisbon
March, 2022

NOVA SCHOOL OF SCIENCE & TECHNOLOGY

DEPARTMENT OF MATHEMATICS

ANALYSIS OF CHANGES IN ANNUAL PRECIPITATION PATTERNS IN ALENTEJO REGION USING LOG-LINEAR MODELS

VÍTOR ALEXANDRE CANHÃO AUGUSTO

Bachelor in Mathematics

Adviser: Elsa Estevão Fachadas Nunes Moreira
Assistant Professor, NOVA University Lisbon

Examination Committee

Chair: Isabel Cristina Maciel Natário
Associate Professor, NOVA School of Science and Technology

Rapporteur: Frederico Almeida Gião Gonçalves Caeiro
Associate Professor, NOVA School of Science and Technology

Adviser: Elsa Estevão Fachadas Nunes Moreira
Assistant Professor, NOVA School of Science and Technology

MASTER IN MATHEMATICS AND APPLICATIONS

NOVA University Lisbon
March, 2022

Analysis of changes in annual precipitation patterns in Alentejo region using log-linear models

This document was created with the (pdf/Xe/Lua)LaTeX processor and the NOVAthesis template (v6.9.10) Lourenço 2021.

To you and me.

Acknowledgements

First of all, I would like to express my sincere acknowledgment to my Adviser, Professor Elsa Moreira, for all the help, time and availability. Then, I have to thank to Professor Miguel Fonseca all of his kindness and the helpful tips.

After all this hard working years, I have to thank to NOVA School of Science and Technology and the Department of Mathematics for all the experiences I had!

I have also to thank to my friends in the coordination of NucM (Mathematics student group), from where I had the honour to be the president. It was quite a journey, and it would have been impossible without them!

And the last but not the least, I have to thank to my lovely family, specially my parents and my sister, my dearest girlfriend Ema and my closest friends Bruno, João and Tatiana. Without this people, I would probably be more crazy...

Abstract

Climatic changes are a topic of extreme importance nowadays, affecting all of the environment. With the aim of adapting viticulture in Alentejo to climate changes (regarding temperature and precipitation), in this thesis we tried to find statistically significant differences in the intra-annual cycles of precipitation in Alentejo, over the last 40 years. To do so, data from each location in Alentejo were divided into four decades and grouped into contingency tables in order to fit log-linear models with two categories: year and month. The analysis was performed using R software, obtaining ANOVA-type tables with residual deviations for the factors month, year and the interaction between them, which allowed us to know their statistical significance in the model. In addition to the previously exposed, the backward elimination method was also applied in an attempt to reduce the parameters of the models referring to the months that are not relevant to explain the variability of precipitation. In the end, we were able to conclude that in the oldest decade there was more intra-annual variability of precipitation, that is, there seems to be a trend towards smoothing out the differences in precipitation between the months of the year. Furthermore, with regard to the inter-annual variation, a cyclical behavior is observed when comparing the 4 decades, although there are more differences in terms of annual precipitation between the years of the most recent decade than between the years of the 3 previous decades.

Keywords: Climatic changes, Log-linear models, Contingency tables, Backward elimination method

Resumo

As alterações climáticas são um tema de extrema importância nos dias que correm, afetando todo o meio ambiente. Tendo como fim a adaptação da viticultura no Alentejo às mudanças do clima (nomeadamente de temperatura e precipitação), nesta dissertação procurou-se encontrar diferenças estatisticamente significativas nos ciclos intra-anuais de precipitação no Alentejo, ao longo dos últimos 40 anos. Para o fazer, os dados de cada localização no Alentejo foram divididos em quatro décadas e agrupados em tabelas de contingência de forma a serem a elas ajustados modelos log-lineares com duas categorias: o ano e o mês. A análise foi efetuada recorrendo ao software R obtendo-se com o output tabelas tipo ANOVA com os desvios residuais para os fatores mês, ano e a interação entre ambos, que nos permitiram saber a sua significância estatística no modelo. Além do exposto, aplicou-se também o método *Backward elimination* para tentar reduzir os parâmetros dos modelos referentes a meses menos relevantes para explicar a variabilidade da precipitação. No fim, pudemos concluir que na década mais antiga havia mais variabilidade intra-anual da precipitação, ou seja, parece estar-se a verificar uma tendência no sentido de se suavizar as diferenças de precipitação entre os meses do ano. Além disso, no que respeita à variação inter-anual, ela parece ter um comportamento cíclico quando comparamos as 4 décadas, embora haja mais diferenças de precipitação entre os anos da década mais recente do que entre os anos das 3 décadas anteriores.

Palavras-chave: Alterações climáticas, Modelos Log-lineares, Tabelas de Contingência, *Backward elimination*

Contents

List of Figures

List of Tables

1

Introduction

Climate changes are a problem we are facing nowadays. Those changes are influencing weather phenomena all around the globe, and in particular in Portugal. The subject of this thesis is inserted in a project that aims assessing the impact that climate change has had on the usual seasonal weather conditions in Portugal with focus on Alentejo, regarding climatic indicators like precipitation and temperature. This knowledge will be use to study vineyards in Alentejo wine region with the final aim of selecting which grapevine varieties can be better adapted to a changed climate. Up to this moment, there are several studies assessing the effects of climate change on viticulture in Portugal. For instance, Fraga et al. 2017 and Santos et al. 2020 show that climate change may cause significant impacts on Portuguese and Mediterranean viticulture, reinforcing the importance that changes in temperature, rainfall and other climatic indicators have in the wine earnings and its quality.

In this thesis will be treated the part regarding to understand if there have been statistically significant changes in the annual cycles of rainfall in Alentejo. To do that, we will select a set of locations representative of the region, with the amount of rainfall recorded monthly in the past forty years. Once this task is accomplished, contingency tables for each location are built with the number of mm^3 of precipitation occurred in each month during periods of 10 years. To these contingency tables, log-linear models with two categories, the year and the month will be fitted using software R. In the following, an analysis of variance (ANOVA) is performed to test the significance of the factor year, month and year-month interaction. Furthermore, in an attempt to reduce the number of the model parameters and test the significance of each month, we will use Backward elimination method.

A different kind of log-linear modeling was used in past years by the supervisor of this thesis to analyze climatic data, namely with the purpose of study and predict drought in Portugal (see for instance Moreira, Russo, and Trigo 2018, Moreira, Pires, and Pereira 2016, Moreira, Mexia, and Pereira 2013 or Moreira, Mexia, and Pereira 2012). Although log-linear models are a very known and used tool, as far as we known, the approach

herein presented is entirely new and never have been used for the purpose of this thesis.

The structure oh this thesis is described next. Chapter 2 contains the necessary concepts to address the problem in study. We will address theoretically the concepts of contingency tables and how to perform independence and homogeneity tests on these tables. Then, linear models like regressions and two-way ANOVA models without and with interaction between factors are also addressed. In the follow-up, we present non-linear models and, in particular, log-linear models for contingency tables with two dimensions. The maximum likelihood estimation method used for fitting log-linear models is presented, as well as the derivation of the likelihood equations for this type of models. Then, we address the model goodness of fit using chi-square test for log-linear models and the concept of model residual deviance, which is crucial for the analysis of the study results. To conclude chapter 2, a brief description of Backward elimination method is given. Chapter 3 is exclusively dedicated to the study of annual precipitation patterns evolution in Alentejo region, where, after a brief introduction, we describe the data and the methods used in their treatment, followed by the presentation of the results obtained and its discussion. The fourth and last chapter contains the main conclusions, and some topics to be developed as future work. Appendix A contains the demonstration of a theoretical result presented in chapter 2, and Appendix B and Annex I contain the graphs and tables supporting chapter 3.

2

Methodology review

2.1 Contingency Tables

Firstly named in Pearson 1904, contingency tables are commonly used to discover the relationship between two categorical variables, one displayed in the rows and the other one displayed in the columns (Two-dimension contingency tables). Contingency tables for 3 or more categorical variables can be built, however this work focus in two-dimension tables. As can be seen in Agresti 1990, a categorical variable has a measurement scale, and consists of a set of categories. For example, if we consider a variable that measures the choice of accommodation, we may use as categories "house", "condominium" or "apartment". Let us present some examples of contingency tables, all with fictional data:

a) Let us analyse the case where, in a supermarket, an employee takes notes on the number of wine bottles sold on one certain day, whether the wine was white, red or green, and the gender of the customer who bought the bottles: male or female. The employee noted that all of the 258 available bottles were sold, being 195 of it bought by women. Furthermore, we have that were sold 85 bottles of red wine, 101 bottles of white wine and 72 of green wine. Green wine bottles were divided equally by women and men, men took only 6 bottles of red wine, and women purchased 80 bottles of white wine. All of this data can be described with a contingency table, such as Table 2.1.

Table 2.1: Contingency table describing the sale of wine bottles

Gender	**White Wine**	**Red Wine**	**Green Wine**	**Total**
Male	21	6	36	63
Female	80	79	36	195
Total	101	85	72	258

b) Considering the data described in example a), we construct a similar contingency table, but this time with the probabilities. Let us consider $X \in \{\text{Male}, \text{Female}\}$, and $Y \in \{\text{White}, \text{Red}, \text{Green}\}$. This table will describe the probability of a bottle sold

being type Y and bought by a costumer of gender X. The contingency table for this case is presented by Table 2.2.

Table 2.2: Contingency table describing the the sale of wine bottles with probabilities

Gender	White Wine	Red Wine	Green Wine	Total
Male	$21/258 \approx 0.0814$	$6/258 \approx 0.0233$	$36/258 \approx 0.1395$	$63/258 \approx 0.2442$
Female	$80/258 \approx 0.3101$	$79/258 \approx 0.3062$	$36/258 \approx 0.1395$	$195/258 \approx 0.7558$
Total	$101/258 \approx 0.3915$	$85/258 \approx 0.3295$	$72/258 \approx 0.2790$	$258/258 = 1$

c) Let us consider the data described in example a), but this time without discriminating purchases in relation to the customer's gender. Let us also suppose that this data was collected in 2017. In this example we will consider the case where the supermarket manager studies the wine sales exactly on that day in the following years. An example of a contingency table for this case is presented in Table 2.3.

Table 2.3: Contingency table describing the the sale of wine bottles from 2017 to 2020

Year	White Wine	Red Wine	Green Wine
2017	101	85	72
2018	94	95	52
2019	145	58	4
2020	108	98	60

2.1.1 Definition

A contingency table is a matrix where each entry is a natural number, whose lines represent the categories of a random variable (r.v.) A, and whose columns represent the categories of random variable B. So if we have a matrix with dimension $r \times s$, we will have that the r.v. A has r categories and the r.v. B has s categories.

A contingency table can be used in two different situations. On the one hand, it may represent the data collected from a sample that is classified simultaneously in r different categories of a factor A, and s different categories of a factor B. On the other hand, it could represent the data collected from r samples that were classified in s different categories or measurements over time, for example. In the first situation, in general we want to test if both factors A and B are independent, while in the last case we want to test if each sample has equal probabilities when classified in different categories.

2.1.2 Hypothesis Test - Independence Test

Let us suppose that a sample of size n is classified according to two factors, A and B, having r and s categories respectively. A contingency table describing this variables can

be represented as Table 2.4, where $O_{i,j}$ represents the number of observations classified as being in j-th level of factor B $j = 1,\ldots,s$, and in the i-th level of factor A, $i = 1,\ldots,r$, and where $O_{\cdot,j} = \sum_{i=1}^{r} O_{i,j},\ j = 1,\ldots,s$ and $O_{i,\cdot} = \sum_{j=1}^{s} O_{i,j},\ i = 1,\ldots,r$. $O_{i,\cdot}$ and $O_{\cdot,j}$ are called marginal frequencies, where $O_{i,\cdot}$ is the i-th level of factor A total, and $O_{\cdot,j}$ is the j-th level of factor B total. So, we have that

$$\sum_{i=1}^{r}\sum_{j=1}^{s} O_{i,j} = n.$$

Each observation can only be classified in one level of each factor.

Table 2.4: Generic contingency table

	Factor B			
Factor A	**1**	$\cdots$	**s**	**Row sum**
1	$O_{1,1}$	$\cdots$	$O_{1,s}$	$O_{1,\cdot}$
$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
r	$O_{r,1}$	$\cdots$	$O_{r,s}$	$O_{r,\cdot}$
Column sum	$O_{\cdot,1}$	$\cdots$	$O_{\cdot,s}$	n

Table 2.5: Generic probabilities contingency table

	Factor B			
Factor A	**1**	$\cdots$	**s**	**Row sum**
1	$p_{1,1}$	$\cdots$	$p_{1,s}$	$p_{1,\cdot}$
$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
r	$p_{r,1}$	$\cdots$	$p_{r,s}$	$p_{r,\cdot}$
Column sum	$p_{\cdot,1}$	$\cdots$	$p_{\cdot,s}$	$p_{\cdot,\cdot} = 1$

Let $p_{i,j}$ be the probability of an observation belongs to i-th level of factor A and j-th level of factor B, with $i = 1,\ldots,r$ and $j = 1,\ldots,s$. The relations between all the $p_{i,j}$ is presented in Table 2.5. The hypothesis of statistical independence are:

$$\begin{aligned} & H_0: \quad p_{i,j} = p_{i,\cdot} \times p_{\cdot,j},\ \forall i \in \{1,\ldots,r\}\ \forall j \in \{1,\ldots,s\} \\ vs. & \\ & H_1: \quad \exists i \in \{1,\ldots,r\}\ \exists j \in \{1,\ldots,s\}:\ p_{i,j} \neq p_{i,\cdot} \times p_{\cdot,j} \end{aligned} \tag{2.1}$$

Let us consider the random vector $(O_{1,1}, O_{1,2},\ldots, O_{r,s})$, with components representing the counts of observations classified in each of the rs combinations of factor A and B categories. Under the null hypothesis and since we consider that the sample size is fixed, equal to n, basing in Agresti 1990 and Everitt 1977, we can assume that the previous vector has a Multinomial distribution, with parameters $p_{1,1}, p_{1,2},\ldots, p_{r,s}$, where the parameters $p_{i,j}$ are unknown ($i = 1,\ldots,r,\ j = 1,\ldots,s$). So, under the null hypothesis, we have $p_{i,j} = p_{i,\cdot} \times p_{\cdot,j}$, and as we can estimate the probabilities $p_{i,\cdot}$ and $p_{\cdot,j}$ as $\hat{p}_{i,\cdot} = \dfrac{O_{i,\cdot}}{n}$ and

$\hat{p}_{\cdot,j} = \dfrac{O_{\cdot,j}}{n}$, respectively, with $i = 1,\dots,r-1$ and $j = 1,\dots,s-1$. The expected frequency in each category is $E_{i,j} = np_{i,j} = np_{i,\cdot}p_{\cdot,j}$, and can be estimated by

$$\hat{E}_{i,j} = n\hat{p}_{i,\cdot}\hat{p}_{\cdot,j} = n\frac{O_{i,\cdot}}{n}\frac{O_{\cdot,j}}{n} = \frac{O_{i,\cdot}O_{\cdot,j}}{n}. \tag{2.2}$$

With this process we can estimate $(r-1)+(s-1)$ parameters, since we have that $\hat{p}_{r,\cdot} = 1 - \sum_{k=1}^{r-1}\hat{p}_{k,\cdot}$ and $\hat{p}_{\cdot,s} = 1 - \sum_{k=1}^{s-1}\hat{p}_{\cdot,k}$.

To study these differences we will use the Chi-Squared test, which is based in the Theorem 2.1.1.

Theorem 2.1.1. *Let $X = (X_1,\dots,X_k)$ be a random variable with a Multinomial distribution, with parameters $p_1,\dots,p_k$. So, the distribution function of the random variable*

$$\chi^2 = \sum_{i=1}^{k}\frac{(O_{i,\cdot} - np_i)^2}{np_i}$$

converges to the Chi-Squared distribution with $(k-1)$ degrees of freedom, when $n \to \infty$.

As a result, the larger the sample size n, the better the approximation given by the theorem.

The statistic used to perform the independence test is

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{\left(O_{i,j} - \hat{E}_{i,j}\right)^2}{\hat{E}_{i,j}}. \tag{2.3}$$

Taking into account that row and column marginal totals are fixed, when the null hypothesis of independence is true we have that the degrees of freedom of the statistic χ^2, which is approximated by the chi-squared distribution, is equal to the number of independent terms in (2.3). Despite the total number of terms in the previous expression being $r \times s$ (the number of cells in the table), there are some terms that are determined knowing the row and column totals. For example, in the i-th row we know that the row total is $O_{i,\cdot}$, which means that we can calculate $O_{i,s}$ by doing $O_{i,s} = O_{i,\cdot} - \sum_{j=1}^{s-1} O_{i,j}$, having then only $s-1$ independent terms. This fact is verified for all $i \in \{1,\dots,r\}$. An equivalent fact can be verified if we proceed with the same logic but applied to each one of the s columns, where we will have $r-1$ independent terms. So, the number of degrees of freedom (df) of the independence statistic is

$$\text{df} = (r-1)(s-1).$$

Being $\chi^2_{(r-1)(s-1)}(1-\alpha)$ the $(1-\alpha)$ quantile of the Chi-square distribution with $(r-1)(s-1)$ degrees of freedom and a chosen level of significance α, we should reject the null hypothesis, with a significance level of α, if

$$\chi^2 > \chi^2_{(r-1)(s-1)}(1-\alpha),$$

that is, the probabilities of belonging to i-th level of factor A and j-th level of factor B are significantly different of the product between the probability of belonging to i-th level of factor A and the probability of belonging to j-th level of factor B, which occurs when there are large differences between the observed and expected frequencies. Consequently, we can conclude that the factors are not independent.

The value of statistic in (2.3) depends on the values of $(O_{i,j} - E_{i,j})$, since the value of these differences will be lower if the variables are independent. By consequence, χ^2 will be smaller when H_0 is true than when it is false.

Conditions to apply the Independence Test

It is advisable to perform the independence test only if no more than 20% of the observed frequencies have a value lesser than 5, and there should be no observed frequencies equal to 0. If these conditions cannot be verified, some categories must be merged in order to obtain higher expected frequencies.

We have to be careful taking the decision of merging some categories in a contingency table. In some cases it can cause the loss of the meaning of the categories. In this case, this test should not be applied.

2.1.3 Hypothesis Test - Homogeneity Test

Let us suppose now that we have r independent random samples, with dimensions $O_{i,.}$, $i = 1,...,r$, obtained from r populations, where each sample is classified in one of the s different categories. The contingency table for this case will be similar to that in Table 2.4, but instead of factor A with r categories, we have the i-th population with $i = 1,...,r$ and instead of factor B we have the s categories.

We want to test the hypothesis:

$$\begin{aligned} & H_0: \quad p_{1,j} = \ldots = p_{r,j} = p_j, \ \forall j = 1,\ldots,s \\ vs. & \\ & H_1: \quad \exists i,k \in \{1,\ldots,r\}: \ i \neq k \ \wedge p_{i,j} \neq p_{k,j} \end{aligned} \tag{2.4}$$

In other words, we want to test whether the probability of an observation belonging to a certain category is the same, regardless of the population from which it was obtained.

Let us consider the random vector $(O_{i,1}, O_{i,2}, \ldots, O_{i,s})$, that is, the vector with the observations from the sample i, belonging to the categories $j = 1,\ldots,s$. Under the null hypothesis, and, again, according to Agresti 1990 and Everitt 1977, we also assume that the previous vector has a Multinomial distribution, with parameters $(O_{i,.}, p_1, p_2, \ldots, p_s)$, where $p_j = p_{1,j} = \ldots = p_{r,j}$, $\forall j = 1,\ldots,s$. So, the expected frequency in each category is $E_{i,j} = O_{i,.}p_j$. If the null hypothesis is in fact true, we will have a low value for $O_{i,j} - O_{i,.}p_j$. To study these differences we will use the Chi-Squared test as in subsection 2.1.2, being the test statistic obtained as follows.

Let us consider the random variables

$$\chi_i^2 = \sum_{j=1}^{s} \frac{\left(O_{i,j} - E_{i,j}\right)^2}{E_{i,j}},\ i = 1,\ldots,r, \tag{2.5}$$

having each variable a Chi-Square asymptotic distribution, with $(s-1)$ degrees of freedom, basing in Theorem 2.1.1. Since our samples are independent, we have that the variables $\chi_i^2,\ i = 1,\ldots,r$, are also independent, and if we consider the variable

$$\chi^2 = \sum_{i=1}^{r} \chi_i^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(O_{i,j} - E_{i,j}\right)^2}{E_{i,j}}, \tag{2.6}$$

we will have that this new variable χ^2 has also a Chi-Square asymptotic distribution, now with $r \times (s-1)$ degrees of freedom.

Denoting the estimator of p_j by $\hat{p}_j = \frac{O_{.,j}}{n}$, we have $\hat{E}_{i,j} = O_{i,.}\hat{p}_j = O_{i,.}\frac{O_{.,j}}{n}$ where $\hat{p}_s = 1 - \prod_{j=1}^{s-1} \hat{p}_j$, so it is necessary to estimate only $(s-1)$ parameters $\hat{p}_j,\ j = 1,...,s-1$.

Replacing the estimators in expression (2.6), we will have that

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(O_{i,j} - \hat{E}_{i,j}\right)^2}{\hat{E}_{i,j}}, \tag{2.7}$$

has also a Chi-Square asymptotic distribution, but this time with $r \times (s-1) - (s-1) = (r-1)(s-1)$ degrees of freedom.

Conditions to apply the Homogeneity Test

To perform the Homogeneity Test, exactly the same conditions as the Independence Test must be verified.

2.2 Regression Models

Named by Francis Galton in 1885, a regression is a statistical technique that studies the prediction of the values taken by a group of variables as a function of the values taken by another group of variables (in particular cases, these groups may be constituted just by one variable). In typical models, there is only one variable to be explained, which we will call response variable, and the variables that will explain the values of this previous variable are called explanatory variables.

Obviously, if we are trying to explain one variable with the values of a group of different values, there are some errors that will be committed. Usually, we admit that those errors are not correlated, and that the expected value of those error terms are equal to 0 and, according to the errors structure, that will indicate which model should be used, between linear, non-linear and generalized linear regression models. In this paper we will work with log-linear models, which are a particular case of the non-linear ones. However, as an introduction, we will also introduce linear regression models and non-linear regression models.

2.2.1 Multiple Linear Regression Models

Linear regression is called linear because the response variable is considered to be a linear function of their parameters. In a linear model it is usual to assume a Normal distribution for the error term with an expected value equal to zero, and a unknown variance σ^2. It is also assumed that the error committed in any two predictions of the response variable are independents.

Describing a model with mathematical symbols, let Y be the response variable of our model, and let us consider we have p explanatory variables $\mathrm{x}_1,\ldots,\mathrm{x}_p$, and we have the vectors $\boldsymbol{x}_i = (x_{i1},\ldots,x_{in})'$, $i = 1,\ldots,p$, which represent, respectively, a n sized sample of the variables x_i. Also, let us consider $\beta_i \in \mathbb{R}$, $i = 0,1,\ldots,p$ unknown values that can be estimated, called regression coefficients or parameters. So, our linear regression model can be represented by

$$Y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} + \varepsilon_i, \quad i = 1,\ldots,n, \tag{2.8}$$

with

$$\varepsilon_i \sim N(0,\sigma^2) \quad \text{and} \quad Cov(\varepsilon_i,\varepsilon_j) = 0, \forall i,j \in \{1,\ldots,n\}, i \neq j, \tag{2.9}$$

where Y_i represent the n values taken by the response variable, and ε_i represent the error committed in each prediction. We can also consider the representation with the vectors, having

$$\boldsymbol{Y} = \beta_0 \mathbf{1}_n + \beta_1 \boldsymbol{x}_1 + \ldots + \beta_p \boldsymbol{x}_p + \boldsymbol{\varepsilon}, \quad \text{with} \quad \boldsymbol{\varepsilon} \sim N_n(\mathbf{0}_n, \sigma^2 I_n), \tag{2.10}$$

where $\boldsymbol{Y} = (Y_1,\ldots,Y_n)'$ and $\boldsymbol{\varepsilon} = (\varepsilon_1,\ldots,\varepsilon_n)'$, where $\mathbf{0}_n$ and $\mathbf{1}_n$ represent a n sized vector with each position equal to 0 and 1, respectively, and where $N_n(\mathbf{0}_n,\sigma^2 I_n)$ represents the multivariate Normal distribution, with mean vector equal to $\mathbf{0}_n$, and covariance matrix equal to $\sigma^2 I_n$, denoting the identity matrix of order n by I_n.

Assumptions to establish a Linear Regression Model

As can be seen in Dunn and Smyth 2018, there are a set of assumptions there are necessary for establish a linear regression model like the one exposed in (2.8). And those are:

- Suitability: The regression model is apropriate for all the observations.
- Homoscedasticity: The variance (σ_i^2, $i = 1,\ldots,p$) of the errors ε_i, $i = 1,\ldots,p$ is constant, that is, we have $\sigma_1^2 = \ldots = \sigma_p^2 = \sigma^2$.
- Independence: The responses Y are independent of each other.

This last condition leads us to the fact that the errors are independent and identically distributed (i.i.d.) variables.

2.2.2 Non-Linear Regression Models

While we have a linear regression model when the response variable is explained through a linear combination of the explanatory variables, as the name suggests we have a non-linear regression model when the response variable is explained through a non-linear combination of the explanatory variables.

We can divide the non-linear models in two types: those who are linearizables, and those were not. We will call a non-linear model as *non-linear but linearizable* to each model that can be written as

$$Y_i = h(\beta_0, \beta_1 x_{1i}, \ldots, \beta_p x_{pi}, \varepsilon_i), \tag{2.11}$$

where h is non-linear function, and for that exist functions f, g_0, g_1, ..., g_p such that

$$f(Y_i) = g_0(\beta_0) + \beta_1 g_1(x_{1i}) + \ldots + \beta_p g_p(x_{pi}) + \varepsilon_i. \tag{2.12}$$

On the other hand, we will say that a non-linear model is not linearizable if it can be written as the form exposed in (2.11), but there are no functions f, g_0, g_1, ..., g_p that allow to rewrite the model in the form exposed in (2.12).

As an example, we can consider the non-linear regression model given by

$$Y_i = \beta_0 x_{1i}^{\beta_1} x_{2i}^{\beta_2} \ldots x_{pi}^{\beta_p} e^{\varepsilon_i}. \tag{2.13}$$

Despite the fact that this model is non-linear, if we consider apply the logarithm function to the model, and assume that $\forall i \in \{1, \ldots, n\}$, $\varepsilon_i \sim N(0, \sigma^2)$, we will obtain a linear model, since we have

$$\begin{aligned} \log(Y_i) &= \log\left(\beta_0 x_{1i}^{\beta_1} x_{2i}^{\beta_2} \ldots x_{pi}^{\beta_p} e^{\varepsilon_i}\right) \\ &= \log(\beta_0) + \beta_1 \log(x_{1i}) + \beta_2 \log(x_{2i}) + \ldots + \beta_p \log(x_{pi}) + \varepsilon_i,\ i = 1, \ldots, n. \end{aligned} \tag{2.14}$$

A model to which the logarithm function is applied as presented in (2.14) is named as a log-linear regression model.

An example of a non-linear model that is not linearizable is

$$Y_i = \beta_0 x_{1i}^{\beta_1} x_{2i}^{\beta_2} \ldots x_{pi}^{\beta_p} + \varepsilon_i,\ i = 1, \ldots, n. \tag{2.15}$$

Here, it is not possible to find a function $h(\cdot)$ that verifies the condition presented in condition (2.12).

Log-linear models are a particular case of the regression models. However, as they are the models which will be used in the study, they will treated in a separate section.

2.3 Analysis of Variance

2.3.1 Introduction

Analysis of Variance (commonly named as ANOVA) is a technique used to test if there are significant differences between several levels of the same factor. For example, if a

set of patients with a certain disease are being treated with three different techniques, ANOVA can test if there are significant differences between the results of each treatment. In this case, we are in the presence of an one-way ANOVA, which compares the levels of just one factor (Montegomery 2012).

Samples often come from the observation of experimental units with different treatments applied to them. The sample size for each treatment group can be the same or differ from sample to sample. The treatments usually are fixed, which means that in the case of repetition of the experience, they will be considered the same, otherwise we say they are random.

To perform an ANOVA, it is assumed that the samples are all independent of each other, identically distributed with normal distribution and the same variance σ^2. However, as we will explain in section 3.3, we can still apply ANOVA even if data does not verify these assumptions.

When there is a set of two or more levels of a random variable, ANOVA is a better technique to process the data than the usual t-tests, since ANOVA only needs one test to be performed. Without ANOVA it will be necessary to perform several t-tests to compare the levels. This reduction in the number of tests performed is an advantage, since we are reducing the probability of reject the null hypothesis for the absence of significant differences between each group means, being this hypothesis true (type I error).

ANOVA with one factor is the simplest linear model in analysis of variance, but our interest for this thesis is focused on the two factor models, with and without interaction between the two factors (Montegomery 2012), described next.

2.3.2 Two-Factor Model With Interaction

Let us consider a model with two factors: factor A with r levels, and factor B with s levels. Then, let us consider that there are n replicates of the study. Using the notation used in Montegomery 2012, we can write an ANOVA linear model with fixed effects and interaction as

$$Y_{i,j,k} = \mu + \tau_i + \beta_j + (\tau\beta)_{i,j} + \varepsilon_{i,j,k}, \tag{2.16}$$

for $i = 1,\ldots,r$, $j = 1,\ldots,s$ and $k = 1,\ldots,n$, where μ is the overall mean effect, τ_i is the effect of the i-th level of the row factor A, with $\sum_{i=1}^{r} \tau_i = 0$, β_j is the effect of the j-th level of column factor B, with $\sum_{j=1}^{s} \beta_i = 0$, $(\tau\beta)_{i,j}$ is the two-factor interaction effect, with $\sum_{i=1}^{r}(\tau\beta)_{i,j} = 0$ and $\sum_{j=1}^{s}(\tau\beta)_{i,j} = 0$, $\varepsilon_{i,j,k} \sim N(0,\sigma^2)$ i.i.d. are random error components, and $Y_{i,j,k}$ is the k-th observed response when factor A is at the i-th level and factor B is at the j-th level, according to the observation matrix represented in Table 2.6. Each cell in Table 2.6 is considered a treatment, that is, a combination of factor levels. For each treatment in a two-way ANOVA with interaction must exist several observations (replicates).

Table 2.6: Generic model table for a model with two factors, n replicates of the study with interaction and fixed effects

	Factor B		
Factor A	**1**	$\cdots$	**s**
1	$Y_{1,1,1}, Y_{1,1,2} \ldots, Y_{1,1,n}$	$\cdots$	$Y_{1,s,1}, Y_{1,s,2} \ldots, Y_{1,s,n}$
$\vdots$	$\vdots$	$\cdots$	$\vdots$
r	$Y_{r,1,1}, Y_{r,1,2} \ldots, Y_{r,1,n}$	$\cdots$	$Y_{r,s,1}, Y_{r,s,2} \ldots, Y_{r,s,n}$

In a one-way ANOVA, the model only include the overall mean, the single factor effect with r levels and random error components for the r levels and n replicates.

For this kind of models, we may be interested in testing hypotheses about the equality of factor A levels, that is

$$\begin{aligned} & H_0: \quad \tau_1 = \cdots = \tau_r = 0 \\ vs. & \\ & H_1: \quad \exists i \in \{1, \ldots, r\} : \tau_i \neq 0 \end{aligned} \tag{2.17}$$

and about the equality of factor B levels, that is

$$\begin{aligned} & H_0: \quad \beta_1 = \cdots = \beta_s = 0 \\ vs. & \\ & H_1: \quad \exists j \in \{1, \ldots, s\} : \beta_j \neq 0 \end{aligned} \tag{2.18}$$

It will also be of interest to determine if factor A interacts with factor B, that is,

$$\begin{aligned} & H_0: \quad \forall i \in \{1, \ldots, r\} \; \forall j \in \{1, \ldots, s\}, \; (\tau\beta)_{i,j} = 0 \\ vs. & \\ & H_1: \quad \exists i \in \{1, \ldots, r\} \; \exists j \in \{1, \ldots, s\} : (\tau\beta)_{i,j} \neq 0 \end{aligned} \tag{2.19}$$

To test the hypothesis above, we have to present some definitions and notations, as in Montegomery 2012.

Let $y_{i..}$ be the total of all observations under the i-th level of factor A, $y_{.j.}$ be the total of all observations under the j-th level of factor B, $y_{i,j,.}$ be the total of all observations in the i, j-th cell and $y_{...}$ be the total of all the observations. We will also define the row, column, cell and global averages of Table 2.6 as $\overline{y}_{i..}$, $\overline{y}_{.j.}$, $\overline{y}_{i,j}$ and $\overline{y}_{...}$, respectively. Expressing

mathematically these means, we have:

$$\begin{aligned}
y_{i\cdot\cdot} = \sum_{j=1}^{s}\sum_{k=1}^{n} y_{i,j,k} &\mapsto \overline{y}_{i\cdot\cdot} = \frac{y_{i\cdot\cdot}}{sn}, \qquad i = 1,\ldots,r \\
y_{\cdot j\cdot} = \sum_{i=1}^{r}\sum_{k=1}^{n} y_{i,j,k} &\mapsto \overline{y}_{\cdot j\cdot} = \frac{y_{\cdot j\cdot}}{rn}, \qquad j = 1,\ldots,s \\
y_{i,j,\cdot} = \sum_{k=1}^{n} y_{i,j,k} &\mapsto \overline{y}_{i\cdot\cdot} = \frac{y_{i,j,\cdot}}{n}, \qquad i = 1,\ldots,r,\ j = 1,\ldots,s \\
y_{\cdots} = \sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{n} y_{i,j,k} &\mapsto \overline{y}_{\cdots} = \frac{y_{\cdots}}{rsn}
\end{aligned}$$

So, the total corrected sum of squares (SS_T) may be written as

$$\begin{aligned}
\sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{n}(y_{i,j,k} - \overline{y}_{\cdots})^2 &= sn\sum_{i=1}^{r}(\overline{y}_{i\cdot\cdot} - \overline{y}_{\cdots})^2 + rn\sum_{j=1}^{s}(\overline{y}_{\cdot j\cdot} - \overline{y}_{\cdots})^2 + \\
&\quad + n\sum_{i=1}^{r}\sum_{j=1}^{s}(\overline{y}_{i,j,\cdot} - \overline{y}_{i\cdot\cdot} - \overline{y}_{\cdot j\cdot} + \overline{y}_{\cdots})^2 + \\
&\quad + \sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{n}(y_{i,j,k} - \overline{y}_{i,j,\cdot})^2 \qquad (2.20)
\end{aligned}$$

As can be seen in expression (2.20), the total sum of squares can be divided in four summands, where the first is a sum of squares due to factor A (SS_A), the second is a sum of squares due to factor B (SS_B), the third is a sum of squares due to interaction between factors A and B (SS_{AB}) and the last is a sum of squares due to error (SS_E). So, we may rewrite expression (2.20) as

$$SS_T = SS_A + SS_B + SS_{AB} + SS_E. \qquad (2.21)$$

The independence between the three sums of squares SS_A, SS_B, SS_{AB} and the SS_E is established by the Cochran's theorem presented in Theorem 2.3.1, since they have chi-square distribution, when divided by σ^2 with the correspondent degrees of freedom as presented in Table 2.7. As was exposed previously, factors A and B have, respectively, r and s levels, which lead us to the fact that effects A and B have, respectively, $r-1$ and $s-1$ degrees of freedom. For the interaction, the number of degrees of freedom is the difference between the degrees of freedom of the number of cells ($rs-1$) and the degrees of freedom of effects A and B, that is,

$$rs - 1 - (r-1) - (s-1) = rs - r - s + 1 = r(s-1) - (s-1) = (r-1)(s-1).$$

As we have n replicates of all the rs cells, each cell will have $n-1$ degrees of freedom. As we have rs cells, the error will have $rs(n-1)$ degrees of freedom.

Table 2.7: Degrees of freedom associated to each sum of squares

Effect	Degrees of Freedom
A	$r-1$
B	$s-1$
$A-B$ interaction	$(r-1)(s-1)$
Error	$rs(n-1)$
Total	$rsn-1$

Theorem 2.3.1 (**Cochran's Theorem**). *Let $X_1, X_2, \ldots, X_v$ be independent random variables with standard normal distribution $N(0,1)$, $\sum_{i=1}^{k} X_i^2 = Q_1 + Q_2 + \ldots + Q_s$, $s \leq k$ and Q_i has v_i degrees of freedom, then the random variables $Q_1, Q_2, \ldots, Q_k$ are independent and have distributions $\chi^2_{v_1}, \chi^2_{v_2}, \ldots, \chi^2_{v_s}$ if and only if $v = v_1 + v_2 + \ldots + v_s$.*

So, we can write:

$$\begin{aligned}
\frac{SS_A}{\sigma^2} &\sim \chi^2_{r-1} \\
\frac{SS_B}{\sigma^2} &\sim \chi^2_{s-1} \\
\frac{SS_{AB}}{\sigma^2} &\sim \chi^2_{(r-1)(s-1)} \\
\frac{SS_E}{\sigma^2} &\sim \chi^2_{rs(n-1)}
\end{aligned}$$

As was exposed previously, factors A and B have, respectively, r and s levels, which lead us to the fact that effects A and B have, respectively, $r-1$ and $s-1$ degrees of freedom. For the interaction, the number of degrees of freedom is the difference between the degrees of freedom of the number of cells $(rs-1)$ and the degrees of freedom of effects A and B, that is,

$$rs - 1 - (r-1) - (s-1) = rs - r - s + 1 = r(s-1) - (s-1) = (r-1)(s-1).$$

As we have n replicates of all the rs cells, each cell will have $n-1$ degrees of freedom. As we have rs cells, the error will have $rs(n-1)$ degrees of freedom.

Then we consider the mean of the sum of squares, dividing it by its degrees of freedom and call it a mean squares. It's proven that the expected values of those mean squares (see Montegomery 2012) are given by

$$\begin{aligned}
E(MS_A) &= E\left(\frac{SS_A}{r-1}\right) = \sigma^2 + \frac{sn\sum_{i=1}^{r}\tau_i^2}{r-1} \\
E(MS_B) &= E\left(\frac{SS_B}{s-1}\right) = \sigma^2 + \frac{rn\sum_{j=1}^{s}\beta_j^2}{s-1} \\
E(MS_{AB}) &= E\left(\frac{SS_{AB}}{(r-1)(s-1)}\right) = \sigma^2 + \frac{n\sum_{i=1}^{r}\sum_{j=1}^{s}(\tau\beta)_{i,j}^2}{(r-1)(s-1)} \\
E(MS_E) &= E\left(\frac{SS_E}{rs(n-1)}\right) = \sigma^2
\end{aligned}$$

and under the null hypotheses (2.17), (2.18), (2.19), they are unbiased estimators of σ^2. Thus, assuming that the model exposed in (2.4.2) fits in the data set, and that $\varepsilon_{i,j,k} \sim N(0,\sigma^2)$ are i.i.d., where σ^2 is a constant, the quotient of two chi-square random variables divided by the respective degrees of freedom has F distribution, and comes

$$\begin{aligned} \frac{MS_A}{MS_E} &\sim F_{r-1,\ rs(n-1)} \\ \frac{MS_B}{MS_E} &\sim F_{s-1,\ rs(n-1)} \\ \frac{MS_{AB}}{MS_E} &\sim F_{(r-1)(s-1),\ rs(n-1)} \end{aligned}$$

ANOVA table is presented in Table 2.8.

Table 2.8: ANOVA table for a two-factor fixed effects model with interaction

Source of Variation	**Sum of Squares**	**Degrees of Freedom**	**Mean Square**	**F**
A	SS_A	$r-1$	$MS_A = \frac{SS_A}{r-1}$	$F_0 = \frac{MS_A}{MS_E}$
B	SS_B	$s-1$	$MS_B = \frac{SS_B}{s-1}$	$F_0 = \frac{MS_B}{MS_E}$
Interaction	SS_{AB}	$(r-1)(s-1)$	$MS_{AB} = \frac{SS_{AB}}{(r-1)(s-1)}$	$F_0 = \frac{MS_{AB}}{MS_E}$
$Error$	SS_E	$rs(n-1)$	$MS_E = \frac{SS_E}{rs(n-1)}$	
Total	SS_T	$rsn-1$		

To test if the effects are relevant, we compare the values of F_0 with the $1-\alpha$ quantile of the F distribution. We should conclude that an effect is statistically significant with significance level alpha if $F_0 > F_{a,rs(n-1)}(1-\alpha)$, with a being the number of degrees of freedom belonging to the effect considered, since the corresponding hypotheses is rejected.

2.3.3 Two-Factor Model Without Interaction

A general model of ANOVA with two factors, fixed effects and without interaction is, according to Montegomery 2012, given by

$$Y_{i,j} = \mu + \tau_i + \beta_j + \varepsilon_{i,j}, \tag{2.22}$$

for $i = 1,\ldots,r$ and $j = 1,\ldots,s$, where, as for the model with interaction, μ represents the overall mean effect, τ_i the effect of the i-th level of the row factor A, with $\sum_{i=1}^{r} \tau_i = 0$, β_j the effect of the j-th level of column factor B, with $\sum_{j=1}^{s} \beta_i = 0$, $\varepsilon_{i,j,k} \sim N(0,\sigma^2)$ i.i.d. the random error components, and $Y_{i,j}$ the observed response for the i-th level of factor A and j-th level of factor B, according to the observation matrix 2.9.
This model is a particular case of the models exposed in section 2.3.2, considering $n = 1$. ANOVA table for these models is presented in Table 2.10, where the sum of squares for the residuals SS_R is the difference between the total sum of squares and the summation

Table 2.9: Generic model matrix for a model with two factors without interaction and fixed effects

	Factor B		
Factor A	**1**	$\cdots$	**s**
1	$Y_{1,1}$	$\cdots$	$Y_{1,s}$
$\vdots$	$\vdots$	$\cdots$	$\vdots$
r	$Y_{r,1}$	$\cdots$	$Y_{r,s}$

Table 2.10: ANOVA table for a two-factor fixed effects model without interaction

Source of Variation	**Sum of Squares**	**Degrees of Freedom**	**Mean Square**	**Expected Mean Square**	**F**
A	$\sum_{i=1}^{r} \frac{y_{i,\cdot}^2}{s} - \frac{y_{\cdot\cdot}^2}{rs}$	$r-1$	$MS_A = \frac{SS_A}{r-1}$	$\sigma^2 + \frac{s\sum_{i=1}^{r}\tau_i^2}{r-1}$	$F_0 = \frac{MS_A}{MS_R}$
B	$\sum_{j=1}^{s} \frac{y_{\cdot,j}^2}{r} - \frac{y_{\cdot\cdot}^2}{rs}$	$s-1$	$MS_B = \frac{SS_B}{s-1}$	$\sigma^2 + \frac{r\sum_{j=1}^{s}\beta_j^2}{s-1}$	$F_0 = \frac{MS_B}{MS_R}$
Residual	SS_R	$(r-1)(s-1)$	$MS_R = \frac{SS_R}{(r-1)(s-1)}$	σ^2	
Total	$\sum_{i=1}^{r}\sum_{j=1}^{s} y_{ij}^2 - \frac{y_{\cdot\cdot}^2}{rs}$	$rs-1$			

of the sum of squares for factors A and B. Assuming that model presented in (2.22) is appropriate for the data, we will have that MS_R present in Table 2.10 is an unbiased estimator of σ^2, we can test the hypothesis (2.18) and (2.19) for the main effects of factor A and B as described in last section.

2.4 Log-linear Models for Contingency Tables

2.4.1 Introduction

Log-linear models are a particular case of non-linear regression models where the function $h(\cdot)$ in section 2.2.2 is the logarithm. This thesis will focus on these models and their inference.

Log-linear model for contingency tables is inspired by the model used in analysis of variance for two factors with or without interaction (Montegomery 2012), in order to fit the data represented in contingency tables. The main objective of the log-linear models is the adjustment of models representing the observed data, and by consequence the estimation of the parameters of those models.

In a log-linear model it is considered that the logarithm of the expected frequencies are given by a linear combination of a certain number of parameters, which may be estimated by well known methods, as least squares or maximum likelihood.

2.4.2 The Model For Two Dimensions

Let us go back to the section 2.1.2, where was exposed how to perform an independence test for a two-dimension contingency table. From now on, in order to simplify the formula presentation, we change the notation of $O_{i,j}$ to $n_{i,j}$ and $E_{i,j}$ to $m_{i,j}$. There, we considered the following hypothesis for statistical independence for two factors:

$$\begin{aligned} & H_0: \quad p_{i,j} = p_{i,\cdot} \times p_{\cdot,j}, \ \forall i \in \{1,\dots,r\}, \ \forall j \in \{1,\dots,s\} \\ vs. & \\ & H_1: \quad \exists i \in \{1,\dots,r\} \, \exists j \in \{1,\dots,s\} : \ p_{i,j} \neq p_{i,\cdot} \times p_{\cdot,j} \end{aligned} \tag{2.23}$$

This relation specifies a model for the data obtained from the studied population, where the probability of an observation coming from the i-th category from factor A and j-th category from factor B, reduces itself to the product between the probability of an observation coming from the i-th sample, and the probability of an observation belonging to the j-th category, separately.

If we apply the logarithm function to both sides of the null hypothesis, we have that $p_{i,j}$ can be expressed as a sum between $p_{i,\cdot}$ and $p_{\cdot,j}$, that is,

$$\log(p_{i,j}) = \log(p_{i,\cdot}) + \log(p_{\cdot,j}). \tag{2.24}$$

Relation (2.24) can be expressed in terms of expected frequencies, using the expression previously introduced $m_{i,j} = np_{i,j}$. So, we have

$$\log(m_{i,j}) = \log\left(\frac{n^2 p_{i,j}}{n}\right) = \log\left(\frac{(np_{i,\cdot})(np_{\cdot,j})}{n}\right) = \log(np_{i,\cdot}) + \log(np_{\cdot,j}) - \log(n), \tag{2.25}$$

or, using the expected frequency notation, we have

$$\log(m_{i,j}) = \log(m_{i,\cdot}) + \log(m_{\cdot,j}) - \log(n), \tag{2.26}$$

If we sum the expression (2.26) over i, we will have

$$\sum_{i=1}^{r} \log(m_{i,j}) = \sum_{i=1}^{r} \log(m_{i,\cdot}) + r\log(m_{\cdot,j}) - r\log(n), \tag{2.27}$$

and summing the same expression over j we will have

$$\sum_{j=1}^{s} \log(m_{i,j}) = s\log(m_{i,\cdot}) + \sum_{j=1}^{s} \log(m_{\cdot,j}) - s\log(n). \tag{2.28}$$

So, summing the expression over i and j, we will have

$$\sum_{i=1}^{r}\sum_{j=1}^{s} \log(m_{i,j}) = s\sum_{i=1}^{r} \log(m_{i,\cdot}) + r\sum_{j=1}^{s} \log(m_{\cdot,j}) - rs\log(n). \tag{2.29}$$

Using the expressions (2.27) to (2.29), we can rewrite the condition (2.26) as

$$\log(m_{i,j}) = u + u_i^L + u_j^C, \tag{2.30}$$

with

$$\begin{aligned} u &= \frac{\sum_{i=1}^{r}\sum_{j=1}^{s}\log(m_{i,j})}{rs}, \\ u_i^L &= \frac{\sum_{j=1}^{s}\log(m_{i,j})}{s} - u, \\ u_j^C &= \frac{\sum_{i=1}^{r}\log(m_{i,j})}{r} - u. \end{aligned}$$

The proof of this fact is written in Appendix A.

Considering ANOVA model with two factors, fixed effects and without interaction, as is presented in expression (2.22), we can calculate the expected value of $Y_{i,j}$, having

$$E(Y_{i,j}) = E(\mu + \tau_i + \beta_j + \varepsilon_{i,j}) = \mu + \tau_i + \beta_j + \underbrace{E(\varepsilon_{i,j})}_{=0} = \mu + \tau_i + \beta_j, \tag{2.31}$$

where $i = 1,\ldots,r$ and $j = 1,\ldots,s$. So, we have that expression (2.26) can be rewritten as a model similar to ANOVA model in expression (2.31), as can be seen in expression (2.30). So, we obtained a linear model to the logarithm of the expected frequencies (that is, a log-linear model), where u represent the global mean, u_i^L represents the main effect of the i-th sample and u_j^C represents the main effect of the j-th category.

In section 2.1.2, for the purposes of testing independence, we considered that the counts in the contingency table have multinomial distribution, since the total number of observations n was fixed. However, many times this is not correct, because the count data do not result from a fixed number of observations. For instance, if we decide to count the number of patients that arrive to an hospital with k different pathologies, the

total number of observations will be unknown. As a result, the Poisson distribution is the appropriated one for these cases (see Agresti 1990). So, from now one we will assume that the observed frequencies are i.i.d. random variables with Poisson distribution.

To estimate the parameters in this models it is used the maximum likelihood method. How to obtain the maximum likelihood estimators (MLE) assuming Poisson distribution will be addressed in the section 2.4.3.

Just as in ANOVA models, principal effects are represented as deviations between the mean of each factor and the global mean, all in terms of the logarithm of the expected frequencies. Furthermore, considering the definitions of u_i^L and u_j^C, we can see that these parameters are expressed as deviations of row or column means of the logarithm of the expected frequencies from the overall mean, and consequently we can conclude that

$$\sum_{i=1}^{r} u_i^L = 0 \qquad \text{and} \qquad \sum_{j=1}^{s} u_j^C = 0,$$

and by consequence we will have

$$u_r^L = -\left(\sum_{i=1}^{r-1} u_i^L\right) \qquad \text{and} \qquad u_s^C = -\left(\sum_{j=1}^{s-1} u_j^C\right).$$

Remembering ANOVA model with two factors and without interaction, we can use the following hypothesis regarding the main effects:

$$\begin{aligned} H_{0L}:\ u_1^L = u_2^L = \cdots = u_r^L = 0 \quad vs. \quad H_{1L}:\ \exists i \in \{1\ldots,r\}:\ u_i^L \neq 0 \\ H_{0C}:\ u_1^C = u_2^C = \cdots = u_s^C = 0 \quad vs. \quad H_{1C}:\ \exists j \in \{1\ldots,s\}:\ u_j^C \neq 0. \end{aligned}$$

When concluded that the independence model (2.30) fits the data, it will be necessary estimate the expected frequencies the parameters of the model using the maximum likelihood estimation (MLE). However, in what concerns to contingency tables, it is more relevant the case where the two factors are not independent, that is, when exists an association between them, which in analysis of variance means that exists interaction between two factors. So, it is necessary to introduce the interaction term in equation (2.30), obtaining the model with interaction

$$\log(m_{i,j}) = u + u_i^L + u_j^C + u_{i,j}^{LC}, \tag{2.32}$$

where

$$u_{i,j}^{LC} = \log(m_{i,j}) - \frac{\sum_{j=1}^{s} \log(m_{i,j})}{s} - \frac{\sum_{i=1}^{r} \log(m_{i,j})}{r} - u$$

represents the interaction effect between i-th level of factor A and j-th level of factor B. Furthermore, it is also verified that $\sum_{i=1}^{r} u_{i,j}^{LC} = 0$ and $\sum_{j=1}^{s} u_{i,j}^{LC} = 0$, which allow us to write that

$$u_{r,j}^{LC} = -\sum_{i=1}^{r-1} u_{i,j}^{LC} \qquad \text{and} \qquad u_{i,s}^{LC} = -\sum_{j=1}^{s-1} u_{i,j}^{LC}.$$

Referring to interaction terms, the hypothesis of statistical independence are

$$\begin{aligned} & H_{0LC}: \quad \forall i \in \{1,\dots,r\} \forall j \in \{1,\dots,s\},\ u_{i,j} = 0 \\ vs. & \\ & H_{1LC}: \quad \exists i \in \{1,\dots,r\} \exists j \in \{1,\dots,s\}:\ u_{i,j} \neq 0, \end{aligned} \tag{2.33}$$

which are equivalent to test if the adequate model for the data is model (2.30).

So, in model (2.32) we conclude that exists one parameter u, $r-1$ independent parameters u_i^L, $s-1$ independent parameters u_j^C and $(r-1)(s-1)$ independent parameters $u_{i,j}^{LC}$, which lead us to the fact that, in total, the model has $1 + r - 1 + s - 1 + (r-1)(s-1) = rs$ independent parameters, that is, a number equal to the amount of observed frequencies. As the expected values for the model are given by the observed frequencies, the model adjusts perfectly to the data, being then called a saturated model.

Estimate the interaction effect is useful to identify the categories that cause the lack of independence. In cases where it is not known whether independence exists, the first thing to do is test whether the model in expression (2.30) fits the data, using an independence test as described in subsection 2.1.2.

The main advantage of adjust log-linear models to the data is the possibility of obtain the estimates of the model's parameters that allow us to quantify the effects of the categories of the different factors and it's interactions.

2.4.3 Likelihood Equations for Log-linear Models

For the purpose of deriving the MLE for the two dimension loglinear model parameters, lets recall that $n_{i,j}$ represents the number of observations classified as being in j-th level of factor B, and in the i-th level of factor A, and let $p_{i,j}$ be the probability of an observation in the i-th level of factor A belongs to the j-th level of factor B, with $i = 1,\dots,r$ and $j = 1,\dots,s$. We will suppose that all $n_{i,j}$ are Poisson random variables with expected values $m_{i,j}$. Let us denote $n_{i,\cdot} = \sum_{j=1}^{s} n_{i,j}$, $n_{\cdot,j} = \sum_{i=1}^{r} n_{i,j}$ and $n = \sum_{i=1}^{r}\sum_{j=1}^{s} n_{i,j}$. So, the joint Poisson probability mass function (p.m.f.) of $n_{i,j}$ is

$$\prod_{i=1}^{r}\prod_{j=1}^{s} \frac{e^{-m_{i,j}} m_{i,j}^{n_{i,j}}}{n_{i,j}!}.$$

The kernel of the log-likelihood is, according to Agresti 1990, given by

$$L(m_{i,j}) = \sum_{i=1}^{r}\sum_{j=1}^{s} n_{i,j}\log(m_{i,j}) - \sum_{i=1}^{r}\sum_{j=1}^{s} m_{i,j}, \tag{2.34}$$

resulting from applying the logarithm to the last function. Considering the log-linear model

$$\log(m_{i,j}) = u + u_i^L + u_j^C + u_{i,j}^{LC},$$

we obtain the equivalent expression to the log-likelihood (2.34)

$$L(m_{i,j}) = nu + \sum_{i=1}^{r} n_{i,\cdot} u_i^L + \sum_{j=1}^{s} n_{\cdot,j} u_j^C + \sum_{i=1}^{r}\sum_{j=1}^{s} n_{i,j} u_{i,j}^{LC} - \sum_{i=1}^{r}\sum_{j=1}^{s} \exp\left\{u + u_i^L + u_j^C + u_{i,j}^{LC}\right\}, \tag{2.35}$$

which is the function we want to maximize. Let $\hat{m}_{i,j}$ be the fitted values for a model, that is the estimates for $m_{i,j}$, thus we have that these values are solutions to a set of likelihood equations. Each of the likelihood equations is obtained differentiating $L(m_{i,j})$ with respect to a parameter and and equaling the result to zero. So, we have

$$\begin{aligned}
\frac{\partial L}{\partial u} &= n - \sum_{i=1}^{r}\sum_{j=1}^{s} \exp\left\{u + u_i^L + u_j^C + u_{i,j}^{LC}\right\} = n - \hat{m}_{\cdot,\cdot} \\
\frac{\partial L}{\partial u_i^L} &= n_{i,\cdot} - \sum_{j=1}^{s} \exp\left\{u + u_i^L + u_j^C + u_{i,j}^{LC}\right\} = n_{i,\cdot} - \hat{m}_{i,\cdot} \\
\frac{\partial L}{\partial u_j^C} &= n_{\cdot,j} - \sum_{i=1}^{r} \exp\left\{u + u_i^L + u_j^C + u_{i,j}^{LC}\right\} = n_{\cdot,j} - \hat{m}_{\cdot,j} \\
\frac{\partial L}{\partial u_{i,j}^{LC}} &= n_{i,j} - \exp\left\{u + u_i^L + u_j^C + u_{i,j}^{LC}\right\} = n_{i,j} - \hat{m}_{i,j},
\end{aligned}$$

for all $i = 1,\ldots,r$ and $j = 1,\ldots,s$. Equaling the previous equations to zero, we will obtain the likelihood equations for our model, being those, for all $i = 1,\ldots,r$ and $j = 1,\ldots,s$,

$$\begin{aligned}
\frac{\partial L}{\partial u_0} = 0 &\Leftrightarrow n - \hat{m}_{\cdot,\cdot} = 0 \Leftrightarrow \hat{m}_{\cdot,\cdot} = n \\
\frac{\partial L}{\partial u_i^L} = 0 &\Leftrightarrow n_{i,\cdot} - \hat{m}_{i,\cdot} = 0 \Leftrightarrow \hat{m}_{i,\cdot} = n_{i,\cdot} \\
\frac{\partial L}{\partial u_j^C} = 0 &\Leftrightarrow n_{\cdot,j} - \hat{m}_{\cdot,j} = 0 \Leftrightarrow \hat{m}_{\cdot,j} = n_{\cdot,j} \\
\frac{\partial L}{\partial u_{i,j}^{LC}} = 0 &\Leftrightarrow n_{i,j} - \hat{m}_{i,j} = 0 \Leftrightarrow \hat{m}_{i,j} = n_{i,j},
\end{aligned}$$

from which we get the the estimates for the parameters models.

2.4.4 Model goodness of fit

As said in section 2.4.1, model (2.32) is called saturated model since it gives a perfect fit. However, that model is not helpful since it does not smooth the data or have the advantages that a simpler model has, such as been spared. Nonetheless, it serves as a base in obtaining simplest models, such as checking model fit.

To test model goodness of fit, we compare fitted cell counts with the observed cell counts. In order to do that, we use chi-squared statistics to test the null hypothesis that a given model fits well the data. In general, for a particular unsaturated model, we want to maximized log likelihood of that model and maximized log likelihood in the saturated case. From that we obtain the likelihood-ratio statistic for testing the null hypothesis that the model holds against the alternative that a more general model holds.

As can be seen in Agresti 1990, for Poisson two-way log-linear models with an intercept term, that is, with parameter u, the likelihood-ratio statistic called deviance simplifies to

$$G^2 = 2\sum_{i=1}^{r}\sum_{j=1}^{s} n_{i,j}\log\left(\frac{n_{i,j}}{\hat{m}_{i,j}}\right), \tag{2.36}$$

being this statistic the one used to test model goodness of fit. When the model holds, this statistic has central chi-square distribution when the sample size is large, that is, has asymptotically a central chi-square distribution.

These goodness of fit tests will have a number of degrees of freedom equal to the difference between the number of parameters in the saturated model and when the null hypothesis that the unsaturated model holds (null hypothesis). That is, in the first step, we will have in the saturated model rs parameters, and in the case when the model, for instance, with less 1 parameter, holds, we will have $1+r-1+s-1+(r-1)(s-1)-1 = rs-1$ parameters, from where the number of degrees of freedom (denoted by df) is obviously $\text{df} = rs-(rs-1) = 1$

For the model without interaction, that is model (2.30), we can use also G^2 statistic to test if the model fits well the data and the number of degrees of freedom is equal to the number of cells in the contingency table rs (equal to the number of observed frequencies) minus the number of linearly independent parameters in the model $(1+r-1+s-1)$, that is

$$\text{df} = rs-r-s+1 = (r-1)(s-1)$$

In a saturated model, the number of parameters is equal to the number of cells in the table, whereby its residual deviance degrees of freedom is equal to zero.

2.4.5 Backward Elimination Method

In general, after adjusting a regression model to the data, it may happen that some variables are not relevant to the study, which lead us to an unnecessary computational effort and time consuming. So, it is necessary to use a method that creates a sub-model that only contains the relevant parameters to the study. In order to do that, two of the most famous methods are Forward and Backward methods. Backward method consists in consider initially the complete model and step by step removes the variables that are not statistically significant to modeling the observations, creating then a sub-model with

less variables. Forward method will start with an empty model and step by step, adds the statistically significant variables to the model. In the end, only the not relevant variables are left out the model. For the scope of this thesis, we only present Backward elimination method since it is the one we will use. In particular, the backward elimination method (Agresti 1990) can be applied to a complete log-linear model to reduce the number of model parameters without significant loss of information. This method allows the selection of an alternative sub-model by eliminating the less significant parameters of the model.

Let us consider the log-linear model with interaction presented before

$$\log(m_{ij}) = u + \sum_{i=1}^{r} u_i^L + \sum_{j=1}^{s} u_j^C + \sum_{i=1}^{r}\sum_{j=1}^{s} u_{ij}^{LC},$$

which is a complete model and also saturated, thus it has $r \times s$ parameters equal to the number of cells in the contingency table. In consequence, the value of statistic G^2, expression (2.36), is equal to zero and has a number of degrees of freedom equal to 0. In the fist step, we start by testing the null hypothesis, for instance, that a group of parameters representing the interaction between the level j of category C with all levels of category L are all equal to zero, that is,

$$\begin{aligned} & H_0^{(j)}: \quad u_{1j}^{LC} = \ldots = u_{rj}^{LC} = 0 \\ vs. & \\ & H_1^{(j)}: \quad \exists i,k \in \{1,\ldots,r\} : u_{ij}^{LC} \neq u_{kj}^{LC} \end{aligned} \tag{2.37}$$

To test these hypotheses, we compute the statistic G^2, for the model without the r parameters, which has chi-square distribution with r degrees of freedom (df) corresponding to difference between the number of parameters in the complete model and in the simplest model. We reject H_0 if G^2 exceeds the chi-square quantile with r df and 5% of significant level alpha. Not rejecting H_0 indicates that we can eliminate those parameters from the model, since their are not significant. In the next steps we proceed in the same way by testing if the next group of parameters are significant in the model. We do this for all $j = 1,...,s$ of the LC interactions. An equivalent test could be performed for the parameters representing each level of the categories L and C, that is, testing $H_0 : u_i^L = 0$, for $i = 1,...,r$ and $H_0 : u_j^C = 0$, for $j = 1,...,s$. At the end, we obtain a simplest model containing only the parameters that are significant to explain the observations.

3

Study of annual precipitation patterns evolution in Alentejo region

3.1 Introduction

In this study we intend to realize if there have been statistically significant modifications in the annual cycles of rainfall in Alentejo region in the past 4 decades, that could be attributed to climate change. With this aim, we analyzed the number of rainfall cubic milliliters (mm^3) occurred monthly in Alentejo from 1979 to 2019. This study is part of a project that has as the final aim to assess the impact of climate change in the wine region of Alentejo using advanced statistical methods. We will use log-linear models with two dimensions fitted to contingency tables that account the number of rainfall mm^3 occurred in target months during periods of 10 years. Then, an analysis of variance is performed using the log-linear residual deviances, to test the significance of the factor/category year and month, as well as the interaction between both.

3.2 Data and methods

The data used in this study are precipitation data-sets retrieved from the European Centre for Medium-Range Weather Forecasts (ECMWF)[1] with 0.25 degrees of spatial resolution located over mainland Portugal. ERA5 | ECMWF is a dataset available for public use and provides hourly, daily and monthly estimates of a large number of atmospheric, land and oceanic climate variables, like precipitation and temperature. The data cover the Earth on a 30km grid and includes information since the year of 1979 to 2019.

Data set to be studied is composed by 34 locations (grid points) in Alentejo region selected from the grid referred above, where each amount of rainfall is expressed with seven decimal places. Figure 3.1 show all grid points located in Portugal and the selected

[1] More information in https://www.ecmwf.int/en/forecasts/dataset/ecmwf-reanalysis-v5

ones. Some locations in Alentejo were discarded because they caused problems in the

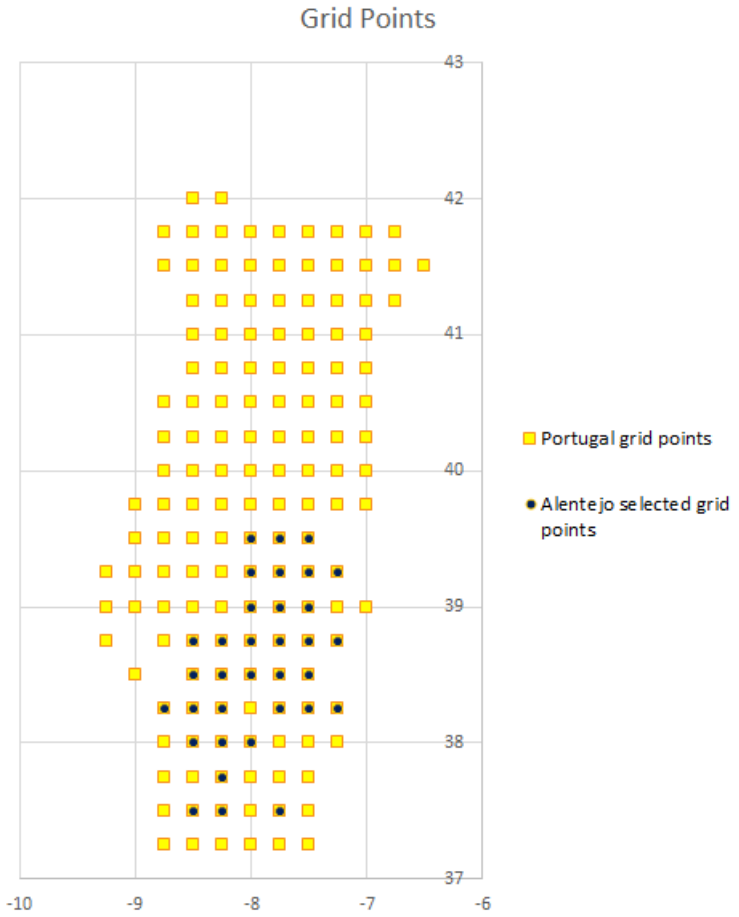

Figure 3.1: Portuguese grid points with the selected grid points in Alentejo highlighted.

fitting of log-linear models. However, since the grid resolution is high, we could discard some points and still remain with a representative sample of the Alentejo wine region for the purpose of the study objectives.

To each location were provided the number of mm^3 of rainfall for each month, between 1979 and 2019. In this study, we first compare the main differences between the fitted models for decades 1979-88, 1990-99, 2000-09 and 2010-19. We decided to not study the year of 1989 since we wanted to study the earliest and the latest decade available, and this year was the one selected do make no part in the study. Besides that, a different analyzes is done by comparing the relevance of each month in the model for more the recent decade (2010-19), contrasting with same information for the first 10 years of available data (1979-88).

The first task was to round each value to the closest integer. Only after this was possible to build the contingency tables for each location with the number of precipitation mm^3 occurred in each month for each year, and adjust saturated log-linear models to those tables. Each contingency table was build with 10 rows, each one corresponding to a year

of the decade in study, and with 9 columns, corresponding to the nine most important months of the year to agricultural crops, in particular to viticulture in Alentejo region, and by consequence, each table has 90 cells. June, July and August were not considered since they are months when it hardly rains in the Alentejo and their inclusion would result in contingency tables with many zeros, thus bringing problems in the adjustment of the log-linear models. Table 3.1 presents the contingency table for location 205 (Corval, Évora district) and the decade 1979-88, where each cell represents the number of mm^3 of rainfall occurred in month $j = 1,2,3,4,5,9,10,11,12$ of year $i = 1,..,10$. All 34 log-

Table 3.1: Contingency table for location 205 (Corval, Évora district), for the decade 1979-88.

Year \ Month	1	2	3	4	5	9	10	11	12
1979	10	33	19	178	50	13	35	100	264
1980	151	71	151	12	40	34	132	32	100
1981	55	19	133	70	18	96	56	137	151
1982	89	108	52	89	2	26	190	98	90
1983	68	71	45	26	19	10	151	25	41
1984	0	15	36	15	24	7	135	114	63
1985	40	63	100	42	1	48	204	171	45
1986	18	107	17	57	36	52	19	41	36
1987	77	95	28	148	50	33	37	35	53
1988	111	78	16	64	18	3	44	52	235

linear models fitted to these tables are composed by $9 \times 10 = 90$ independent parameters, which is equal to the number of cells in the contingency table, as mentioned in section 2.4.2. This process was accomplished using software R, following the method described in section 10.4.2 in Dunn and Smyth 2018. From the R output we obtained an ANOVA type table as in the example presented in Table 3.2.

Table 3.2: ANOVA type table for the decade 1979-88, location 205 (Corval, Evora district).

	Df	Deviance	Residual Df	Residual Deviance	p-value
NULL			89	3574.1	
Year	9	248.82	80	3325.3	$< 2.2e-16$
Month	8	643.69	72	2681.6	$< 2.2e-16$
Year:Month	72	2681.63	0	0.0	$< 2.2e-16$

In these tables we can find the residual deviances and correspondent degrees of freedom for the factor year, month and year:month interaction, as well as chi-square test p-values for each one.

3.3 Results and discussion

First, it should be noted that the results obtained for the fitted models indicate that both the year and month effects, as well as the year-month interaction are highly significant for all locations, that is, all parameters referring to the year, month and interaction

factors are extremely relevant in the model to explain the variability of precipitation. In particular, the interaction effect, which reveals the lack of independence between the year and month, indicates that over the years, there have been significant changes in the distribution of precipitation over the months of the year, that is, there has been climate change, otherwise the interaction residual deviance should be near zero. Since this happens in the four decades, we decided to compare graphically the values of residual deviance for the model without year:month interaction (see the expression for G^2 in (2.36)), to evaluate the differences between the four decades for each location. With this values, we produced the graph presented in Figure 3.2. The residual deviance values used to build this graphic are shown in Table I.1 of Annex I.

With the aim in verifying the affirmations previously stated, we performed using R a One-Way ANOVA and Tukey multiple comparison tests (Montegomery 2012) applied to the residual deviance values per decade regarding year:month interaction. To obtain a robust ANOVA, a balanced data situation must exist, that is, the number of observations per treatment (combination of factor levels) must be the same. This is the case, since for the one-way ANOVA we considered the same number of locations per decade (single factor having four levels). A balanced data situation allows still having a robust ANOVA for departures from usual ANOVA assumptions like normality, independence between observations and non-similar variances. In this case, the residual deviances are the observations, which are not normally distributed, however since we have a balanced data situation, we still can confidently apply ANOVA (Ito 1980, Scheffé 1959).

So, the tables resulting from One-Way ANOVA analysis and Tukey multiple comparison analysis can be seen in Tables 3.3 and 3.4.

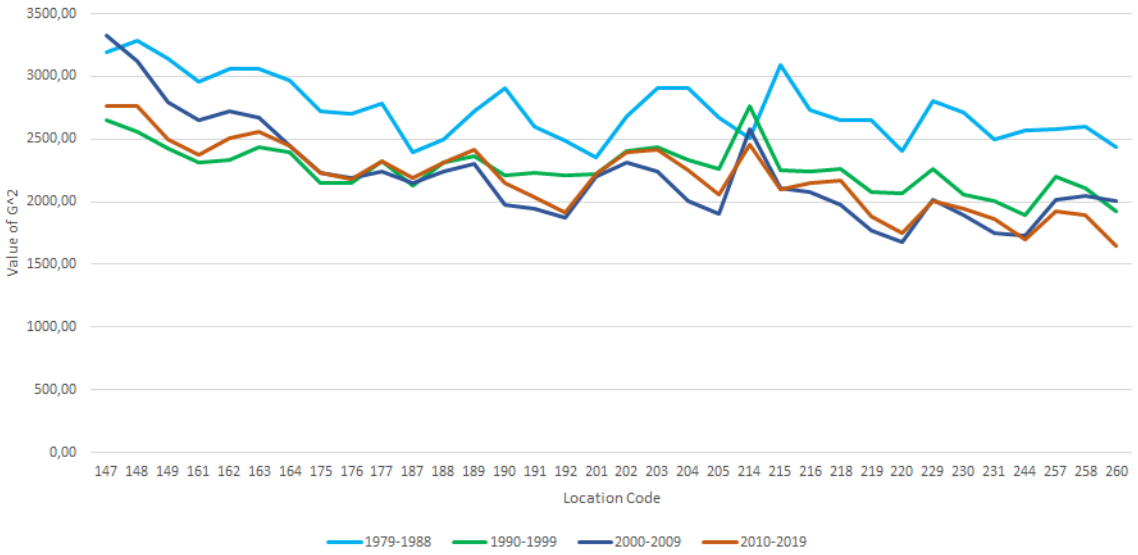

Figure 3.2: Residual deviance values for each location in each decade regarding year:month interaction.

Table 3.3: One-way ANOVA table applied to the residual deviance values per decade regarding year:month interaction

	Df	Sum Sq	Mean Sq	F	p-value
Decade	3	7010896	2336965	28.74	2,29E-14
Residuals	132	10734736	81324		

Table 3.4: Tukey multiple comparison table applied to the residual deviance values per decade regarding year:month interaction

	diff	lwr	upr	p-value adj
2-1	-478,633	-658,604	-298,663	0,000
3-1	-531,063	-711,033	-351,092	0,000
4-1	-552,325	-732,295	-372,354	0,000
3-2	-52,429	-232,400	127,541	0,873
4-2	-73,691	-253,662	106,279	0,711
4-3	-21,262	-201,233	158,709	0,990

As can be seen in Figure 3.2, the residual deviance for the model without interaction is noticeably higher for all locations in the decade starting in 1979 than in the others, with two exceptions. From these results we can draw that, in the first decade, the effect of the year changing, influenced more the distribution of precipitation over the months of the year, than in the most recent three decades, that is, the distribution of rainfall throughout the year has changed more in the 1980s than in the recent past. This fact is also confirmed by Table 3.3, where is shown that there are significantly differences between the decades, and by Table 3.4 that shows that the first decade has residual deviance values significantly different than the other decades, statistically speaking. These results may sound a little strange, since we could expect the opposite. However, this may also mean that there are less differences in precipitation between the months of the year in recent decades. In other words, climate change may be smoothing the differences in terms of precipitation between the months (intra-annual rainfall variability), making the rainfall more homogeneous throughout the months of September to May.

Now looking at the two locations that are the exceptions mentioned above, it is evident that location 147 (Belver, Portalegre district) has one of the highest values in each decade, and it is also remarkable that in location 214 (Carvalhal, Setúbal district) the deviance values are really close in all decades, which may indicate that there were no relevant changes in the intra-annual precipitation over the years of four decades.

With the residual deviations for the isolated contributions of the year and month factors obtained from the 3rd column of ANOVA table from R output, we built the two graphs shown in Figures 3.3 and 3.4. The deviance values used to build theses graphics are presented in Table I.2 and Table I.3, respectively, in Annex I. From the analysis of the graph for the year factor (Figure 3.3), we can see that the decade with the highest residual deviation is, in general, the 4th: 2010-2019, followed closely by the 2nd: 1990-1999. This

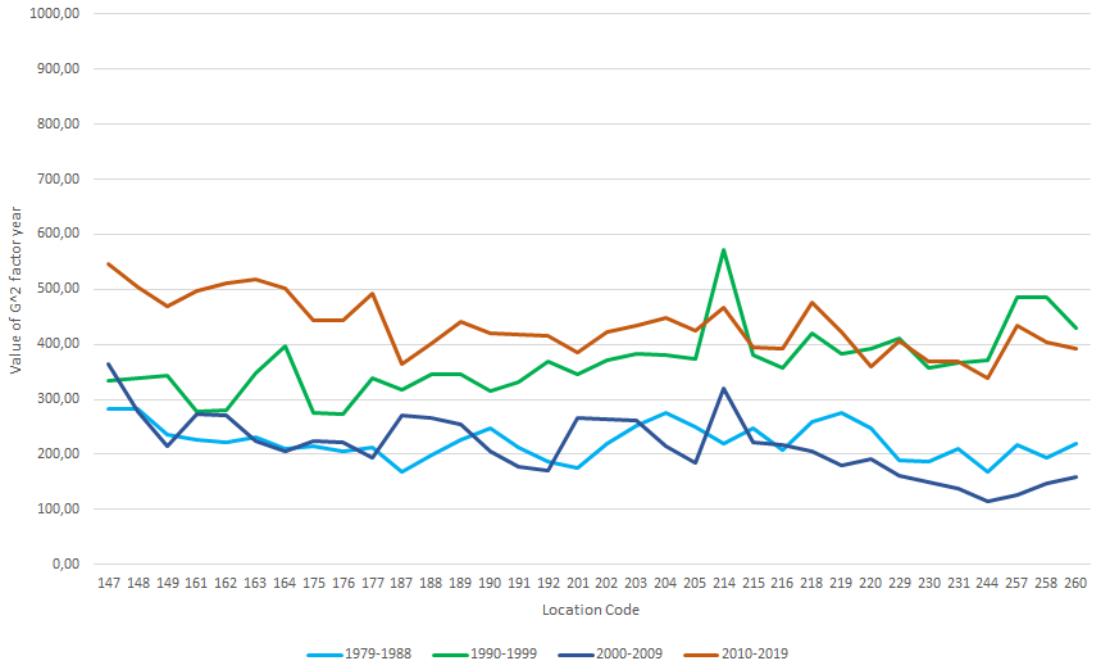

Figure 3.3: Deviation values in each decade regarding year factor.

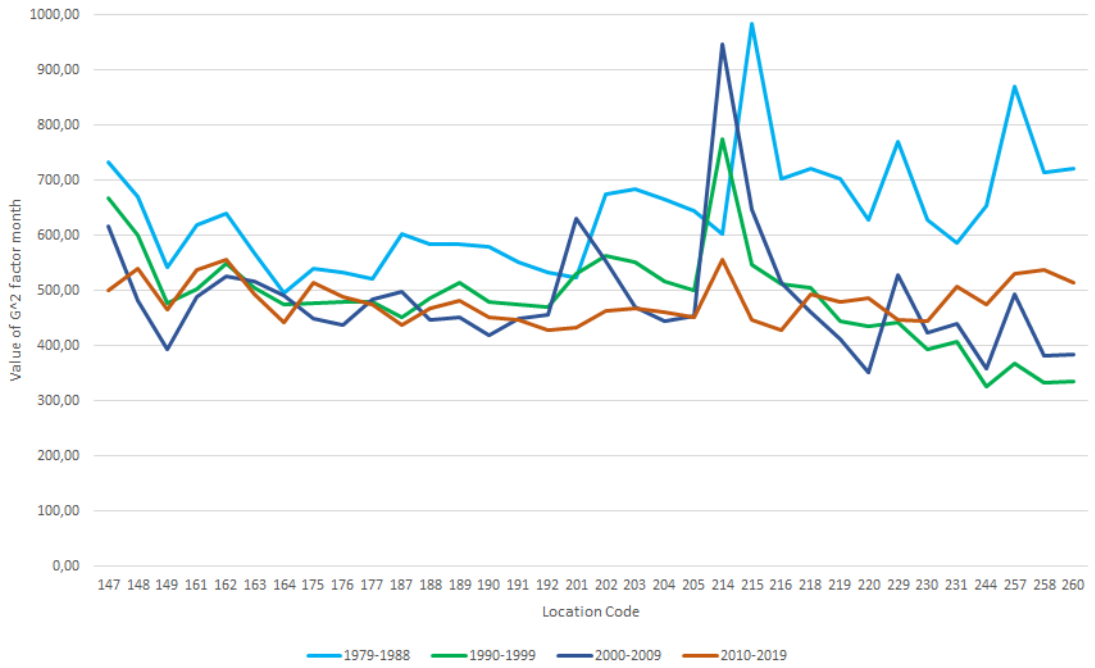

Figure 3.4: Deviation values in each decade regarding month factor

is indicative that in these decades, the year is more important for explaining the variability of rainfall than in the others and reveals a certain cyclical behavior, since the 1st and 3rd

decades are also close to each other, but with lower residual deviations. This results indicate that the significant differences in precipitation between the years (inter-annual variability) of the 4th and 2nd decades are higher than between the years of the 1st and 3rd decades. To prove statistically if this cyclical behaviour is verified, again we applied an One-Way ANOVA analysis followed by a Tukey multiple comparison analysis, which can be found in Tables 3.5 and 3.6. Again, as was expected, Table 3.5 shows that there are

Table 3.5: One-way ANOVA table applied to the residual deviance values per decade regarding year effect

	Df	Sum Sq	Mean Sq	F	p-value
Decade	3	1187473	395824	150.7	2E-16
Residuals	132	346629	2626		

Table 3.6: Tukey multiple comparison table applied to the residual deviance values per decade regarding year effect

	diff	lwr	upr	p-value adj
2-1	144,472	112,132	176,812	0,000
3-1	-7,057	-39,397	25,283	0,941
4-1	210,316	177,976	242,656	0,000
3-2	-151,529	-183,869	-119,190	0,000
4-2	65,844	33,504	98,184	0,000
4-3	217,373	185,033	249,713	0,000

significant differences between the decades, and performing Tukey multiple comparison test (whose result is presented in Table 3.6), we can conclude that there are no significant differences between 1979-88 and 2000-09 decades. However, Tukey multiple comparison test show also significant differences between 1990-1999 and 2010-2019 decades, which contradicts partially the hypothesis of a cyclical behaviour between the decades. Looking to Figure 3.3, we can see that there are a set locations where those differences are more obvious, and another set where those differences are less evident. From the analysis of the graph related to the isolated month factor (Figure 3.4), we can see that again the 1st decade has higher residual deviance values compared with the other three, meaning that in the 1st decade there were bigger significant differences between the months of the year. When comparing the deviations of the month factor with those of the year factor, it is verified that the month generally has greater deviations, therefore it has more responsibility to explain the changes in precipitation occurred in each of the decades, thus has also more weight in the year-month interaction. As previously, we performed One-Way ANOVA and Tukey multiple comparison tests applied to the residual deviance values per decade regarding month effect, from where were produced Tables 3.7 and 3.8. As was verified to the residual deviance values per decade regarding year:month interaction, Table 3.7 shows that there are significantly differences between the decades, and Table 3.8 shows that the first decade has residual deviance values significantly different than the

Table 3.7: One-way ANOVA table applied to the residual deviance values per decade regarding month effect

	Df	Sum Sq	Mean Sq	F	p-value
Decade	3	619408	206469	26.29	2.13E-13
Residuals	132	1036627	7853		

Table 3.8: Tukey multiple comparison table applied to the residual deviance values per decade regarding month effect

	diff	lwr	upr	p-value adj
2-1	-152,806	-208,732	-96,879	0,000
3-1	-154,794	-210,720	-98,867	0,000
4-1	-159,645	-215,571	-103,719	0,000
3-2	-1,988	-57,914	53,939	1,000
4-2	-6,839	-62,766	49,087	0,989
4-3	-4,851	-60,778	51,075	0,996

other more recent decades. From all this analysis, we can conclude that there seems to be a certain trend towards a decrease of the intra-annual variability of precipitation, that is, within each year (the differences in precipitation between months may be fading) which corroborate the conclusion drawn from the analysis year-month interaction (Figure 3.2). On the other hand, inter-annual variability seems to have a more cyclical behavior since decades with less inter-annual variability alternate with decades with more inter-annual variability. Nevertheless, the decades with the higher inter-annual variability are the 2nd and the 4th, which may be interpreted as decades having some years with high levels of precipitation and other years with very low levels.

Despite the reflection performed in the previous paragraphs, we also try to get an explanation for the values presented in Figures 3.3 and 3.4 with locations 214 (Carvalhal, Setúbal district) and 215 (Santiago, Setúbal district), showing abnormal high deviances. As was shown before, in location 214 the values for the deviation are really close in all decades, and location 215 has the highest deviance values in 1979-88 decade. Considering deviance values for each month, these locations stand out due to high values in September and November. In order to understand what is different for these locations, we performed an analysis on the amount of rainfall per month and per year, considering the mean of the values in each year. To analyse rainfall values for each month in these locations, we calculated some statistics, whose values are exposed in Tables 3.9 and 3.10. Figure 3.4 suggests that location 214 has high residual deviation values in every decade, when compared with the other locations in this decades, except the first one, and the decade with the higher residual deviation value is 2000-09. In Table 3.9 it can be seen that, despite the fact that 2000-09 decade has an amount of rainfall for the first quartile greater then the other decades, the value for the second quartile is one of the lowest, what shows that 25% of the rainfall registers are between 8.966 and 35.894. By the other hand, 25% of the rainfall observations are higher than the third quartile, that is, higher than

Table 3.9: Values of some statistics about the amount of rainfall in each month for location 214 (Carvalhal, Setúbal district).

	Global	1979-88	1990-99	2000-09	2010-19
Minimum	0.173	0.173	0.224	0.238	0.311
1st Quartile	7.598	7.276	7.610	8.966	7.497
2nd Quartile	36.371	35.671	39.943	35.894	36.278
3rd Quartile	73.675	74.821	65.721	72.342	80.891
Maximum	376.231	232.638	376.231	263.926	226.307
Mean	52.387	52.334	50.570	52.036	51.760

Table 3.10: Values of some statistics about the amount of rainfall in each month for location 215 (Santiago, Setúbal district)

	Global	1979-88	1990-99	2000-09	2010-19
Minimum	0.040	0.040	0.094	0.079	0.057
1st Quartile	7.102	6.597	10.324	7.762	5.322
2nd Quartile	33.169	32.036	33.751	33.045	33.562
3rd Quartile	65.284	69.246	59.572	64.460	72.270
Maximum	269.141	201.114	269.141	212.991	196.976
Mean	45.460	45.207	44.501	45.232	44.480

72.342. This fact may help explaining the reason for the higher residual deviation in this location, however the values for the other statistics do not stand out from the others decades, namely the mean and the range between maximum and the minimum, which is not particularly high, as it could be expected. The decade of 1990-99 is the one where is verified the highest range in the rainfall amounts in this location. Here, we verify that we have a mean value closer to the third quartile median, what suggests that the data are distributed through all the range.

Changing our focus to the other location, Figure 3.4 suggests that location 215 has a higher residual deviation value in 1979-88 decade than the others. Analysing Table 3.10 it can be seen that half of rainfall amount is lower than 32.036 (which is the lowest median) and that the third quartile is one of the highest, higher than the global value, which lead us to the conclusion that there was a greater disparity of the rainfall amounts in this decade. In contrast the range between maximum and the minimum rainfall amounts is one of the lowest. From this analysis, it is not clear the reason why these two locations stands out from the others. Perhaps an analysis of the same type for each year within each decade could give us more clues.

In addition to the monthly analysis to the rainfall amount, Figure 3.3 shows that it is also important to perform an analysis to location 214 in an annual perspective. There, it is possible to see that residual deviation values are higher for 1990-99 and 2000-09 decades. Thence, we performed a study similar to the previous one. Table 3.11 presents the same basic statistics about the amount of rainfall in each year for location 214. As it is possible to observe, in 1990-99 decade we have a very small difference between first and second quartiles, and there is a bigger difference between the third quartile and the

Table 3.11: Values of some statistics about the amount of rainfall in each year for location 214 (Carvalhal, Setúbal district).

	Global	1979-88	1990-99	2000-09	2010-19
Minimum	32.290	38.732	35.115	34.798	32.290
1st Quartile	40.830	46.095	41.138	41.025	39.076
2nd Quartile	52.259	50.331	44.660	56.485	53.704
3rd Quartile	62.117	61.787	52.373	61.907	55.839
Maximum	88.801	67.913	88.801	64.011	82.692
Mean	52.387	52.334	50.570	52.036	51.760

maximum rainfall value than between the third quartile and the minimum rainfall value. This decade has also the lowest mean when compared to the other decades in study and the highest range between maximum and minimum values, indicating large disparity between the years of this decade. For 2000-09 decade, location 214 does not stand out as much in terms of deviance, but it is possible to verify that the difference between the mean and the maximum value is way lower that the difference between the mean and the minimum value. These facts may constitute an explanation to the high residual deviation values.

In a different analysis, we compared the decades of 1979-88 and 2010-19, in order to evaluate the importance that each month has in the saturated model. These decades were chosen because they were the first and last decade with records and because we want to see if there were major differences regarding the importance of each month in the models between the most recent and the oldest decade. We decided to eliminate from the models at a time the parameters relative to a month, using the backward elimination method mentioned in section 2.4.5. For instance, the parameters relative to February were all considered zero, the model were refitted and a new residual deviance was obtained. Then, we performed chi-square test to see if the sub-model was not rejected. We did that for each one of the nine months considered in the models and all the sub-models were rejected, meaning that the parameters relative to each month are all relevant in the models. The correspondent residual deviances were retained and compared between months and also between the same months of the first and last decade. Figures B.1 to B.9 presented in Appendix B show the comparison between the deviance values for each month in both decades, from January to December. When comparing the deviances resulting from eliminating each month considered, we could not concluded clearly about which months have more relevance in the models. In general, all of them seem to have, some more than others, high contribution in the model to explain rainfall variability, variating of course with the location. February is the one with less contribution in both decades and December the one with higher contribution. When comparing the same month between the two decades, deviance values are generally higher in the older decade than in the most recent one, which is in line with the results obtained from the analysis of Figure 3.2. That is, for the most part of the locations, the precipitation intra-annual

variability was higher in the elder years. We could also conclude that the months with bigger differences between decades are December, October and March, meaning also that the climate change was higher in these months.

Now doing an analysis on the locations showing stage behavior, it is possible to see that location 177 (Monforte, Portalegre district) had an atypical May month in the most recent decade. In Table 3.12 are shown the rainfall values observed in May. As can be

Table 3.12: Rainfall values observed in May for location 177 (Monforte, Portalegre district) for 2010-19 decade.

Year	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Value	40	74	43	14	23	5	131	37	59	8

seen, 2015 and 2019 had almost no rainfall in May and 2016 the had a significantly above average rainfall. This can explain the high deviance value verified. In what concerns to April, we can observe really irregular values in each decade. In September, it can be seen two highlights: locations 203 (Nossa Srª da Tourega, Évora district) and 215 (Santiago, Setúbal district) with very high deviance values for the recent decade. Tables 3.13 and 3.14 show the rainfall values in September for both locations. In both locations

Table 3.13: Rainfall values observed in September for location 203 (Nossa Srª da Tourega, Évora district) for 2010-19 decade.

Year	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Value	8	36	45	46	101	13	14	0	4	11

Table 3.14: Rainfall values observed in September for location 215 (Santiago, Setúbal district) for 2010-19 decade.

Year	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Value	7	16	30	40	108	8	12	0	2	11

is possible to verify that 2014 was a rainy month in both locations, and in several months the precipitation values were lower than 20, having inclusive registers of no precipitation in 2017. In what concerns to November, despite the fact that both decades have close values for deviance, location 231 (Mombeja, Beja district) has a higher deviance value than the others in 1979-88 decade. Table 3.15 show the rainfall values for location 231. As can be seen, in 1979 and 1981 the amount of rainfall was way lower than in the other

Table 3.15: Rainfall values observed in November for location 231 (Mombeja, Beja district) for 1979-88 decade.

Year	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988
Value	7	60	5	65	150	89	67	47	49	153

years, what can cause the high deviance value. In 2010-19 similar phenomena were verified in three other locations: 177 (Monforte, Portalegre district), 190 (Santa Maria,

Portalegre district) and 215 (Santiago, Setúbal district). Regarding December, it can be seen that there is a considerable difference between deviance values for both decades. Location 258 (Santana da Serra, Beja district) has the biggest value in 1979-88 decade, and the rainfall values for this month are presented in table 3.16. As can be seen, there

Table 3.16: Rainfall values observed in December for 258 (Santana da Serra, Beja district) for 1979-88 decade.

Year	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988
Value	14	4	144	23	79	62	58	19	168	5

are five years with less than 25 mm^3 of rainfall (1979, 1980, 1982, 1986 and 1988), which is in contrast very high rainfall amounts in 1981 and 1987. This variability may explain the deviance value observed.

4

Conclusions

In this research study, log-linear models were used to model monthly rainfall data for the last 40 years and led us to obtain some interesting findings about the changes that occurred in this period regarding this climate indicator. Contingency tables accounting the number of mm^3 of precipitation recorded for each month during periods of 10 years, were built and log-linear models with two dimensions were fitted to them. This approach shown to be very useful in obtaining results about the possible effects of climate change in Alentejo region.

During the course of this work, we encountered some difficulties. After we build the contingency tables for the 49 locations available in Alentejo and fitting the log-linear models, when we were trying to apply Backward elimination method in order to eliminate from the models the less relevant parameters, for 15 of the locations were not possible to apply the method for, at least, one of the decades. As a result, we had to discard these 15 locations, but as we had a high resolution grid of locations, the remaining sample of 34 locations was more than enough to carry out a reliable study on Alentejo region. We also excluded the months of June, July and August from the study since, due to the traditional low levels of precipitation, those were the less relevant months to grapes development and also to avoid too many zeros in the contingency tables, thus problems in the fitting of the models.

The interpretation of the results was a challenge, but from the first analysis done on the results, we could conclude that there were bigger differences of precipitation between months of the year in the first decade than the in the more recent ones. This suggest that climate change may be smoothing the differences in terms of precipitation between the months (intra-annual rainfall variability). When analyzed the differences between the rainfall through the years for each decade for all the selected locations, results suggested that the variability of precipitation has a cyclic behaviour, where a decade with less rainfall variability between the years, is followed by a decade with more rainfall variability between the years. When we studied the rainfall variability just between the months of the years in each decade for all the selected locations, we could conclude

that, the first decade of the data has in general higher deviance values than the others, suggesting again that the differences between the months of the year tend to be lower for the recent decades. We can also highlight the fact that month factor has more importance in explaining rainfall variability than the year factor.

With a different objective, we analysed 1979-88 and 2010-19 decades in order to evaluate the importance of each month in the model. The first and immediate conclusion is that, for both decades we could not discard any month from the model. However, February seem to be the one with lower importance and December with higher importance in both decades. In addition, we were able to conclude that the months of December, October and March are the ones with the greatest differences between the two decades, therefore, those were the months where there was more climate change.

In terms of future work, we intend to go deeper in the analysis to try find other type of changes in the annual precipitation cycles that was not possible to reach with the present approach. Namely, if there are a shift of the annual and seasonal cycles of precipitation. We would also like to perform a more in-depth analysis of the locations, namely those close to the Alqueva Dam (the largest artificial lake in Europe) in order to find some evidence of its influence on the climate of these locations. We also intent to carry out a similar study on maximum and minimum temperatures recorded over and bellow critical values, to find significant changes in temperatures distribution over the months for the past 40 years.

Bibliography

Agresti, A. (1990). *An Introduction to Categorical Data Analysis*. Hoboken, New Jersey (cit. on pp. 3, 5, 7, 19, 20, 22, 23).

Dunn, P. K. and G. K. Smyth (2018). *Generalized Linear Model With Examples in R*. New York, NY, ISBN: 978-1-4419-0117-0: Springer (cit. on pp. 9, 27).

Everitt, B. (1977). *The analysis of Contingency Tables*. London: Chapman & Hall (cit. on pp. 5, 7).

Fraga, H. et al. (2017). "Viticulture in Portugal: A review of recent trends and climate change projections". In: *OENO One* 51(2). DOI: ttps://doi.org/10.20870/oeno-one.2017.51.2.1621 (cit. on p. 1).

Ito, K. (1980). "Robusteness of ANOVA and MANOVA test procedures". In: *Krishnaiah PR (ed) Handbook of statistics* 1, pp. 199–236. DOI: https://doi.org/10.1016/S0169-7161(80)01009-7 (cit. on p. 28).

Lourenço, J. M. (2021). *The NOVAthesis LaTeX Template User's Manual*. NOVA University Lisbon. URL: https://github.com/joaomlourenco/novathesis/raw/master/template.pdf (cit. on p. iii).

Montegomery, D. C. (2012). *Design and Analysis of Experiments, Pag. 188-189*. Hoboken, NJ, ISBN: 978-1118-14692-7: John Wiley & Sons, Inc. (cit. on pp. 11, 12, 14, 15, 17, 28).

Moreira, E., J. Mexia, and L. Pereira (2012). "Are drought occurrence and severity aggravating? A study on SPI drought class transitions using log-linear models and ANOVA-like inference". In: *Hydrology and Earth System Sciences* 16. DOI: 10.5194/hess-16-3011-2012 (cit. on p. 1).

— (2013). "Assessing homogeneous regions relative to drought class transitions using an ANOVA-like inference. Application to Alentejo, Portugal". In: *Stochastic Environmental Research and Risk Assessment* 27(1). DOI: 10.1007/s00477-012-0575-z (cit. on p. 1).

Moreira, E., C. Pires, and L. Pereira (2016). "SPI drought class prediction driven by the Northern Atlantic Oscillation index using log-linear modelling". In: *Water* 43.8. DOI: 10.3390/w8020043 (cit. on p. 1).

Moreira, E., A. Russo, and R. Trigo (2018). "Monthly Prediction of Drought Classes Using Log-Linear Models under the Influence of NAO for Early-Warning of Drought and Water Management". In: *Water* 65.10. DOI: 10.3390/w10010065 (cit. on p. 1).

Pearson, K. (1904). *On the theory of contingency and its relation to association and normal correlation*. 37 Soh Square, W., ISBN: 978-1313368902: Dulau & Co. (cit. on p. 3).

Santos, J. et al. (2020). "A Review of the Potential Climate Change Impacts and Adaptation Options for European Viticulture". In: *Applied Sciences* 3092.10(2). DOI: 10.3390/app10093092 (cit. on p. 1).

Scheffé, H. (1959). *The analysis of variance*. New York: Wiley (cit. on p. 28).

A

Proof of equivalence of expressions (2.26) and (2.30)

We want to prove that the expression (2.26) is equivalent to (2.30), that is, that

$$\log(m_{i,j}) = \log(m_{i,\cdot}) + \log(m_{\cdot,j}) - \log(n)$$

is equivalent to

$$\log(m_{i,j}) = u + u_i^L + u_j^C,$$

with

$$\begin{aligned} u &= \frac{\sum_{i=1}^{r}\sum_{j=1}^{s}\log(m_{i,j})}{rs}, \\ u_i^L &= \frac{\sum_{j=1}^{s}\log(m_{i,j})}{s} - u, \\ u_j^C &= \frac{\sum_{i=1}^{r}\log(m_{i,j})}{r} - u. \end{aligned}$$

So, we have that

$$\begin{aligned} \log(m_{i,j}) &= u + u_i^L + u_j^C \\ &= \frac{\sum_{i=1}^{r}\sum_{j=1}^{s}\log(m_{i,j})}{rs} + \left\{\frac{\sum_{j=1}^{s}\log(m_{i,j})}{s} - \frac{\sum_{i=1}^{r}\sum_{j=1}^{s}\log(m_{i,j})}{rs}\right\} + \\ &\quad + \left\{\frac{\sum_{i=1}^{r}\log(m_{i,j})}{r} - \frac{\sum_{i=1}^{r}\sum_{j=1}^{s}\log(m_{i,j})}{rs}\right\} \\ &= \frac{\sum_{j=1}^{s}\log(m_{i,j})}{s} + \frac{\sum_{i=1}^{r}\log(m_{i,j})}{r} - \frac{\sum_{i=1}^{r}\sum_{j=1}^{s}\log(m_{i,j})}{rs}. \end{aligned}$$

Multiplying both sides of the expression by rs, we have

$$rs\log(m_{i,j}) = r\sum_{j=1}^{s}\log(m_{i,j}) + s\sum_{i=1}^{r}\log(m_{i,j}) - \sum_{i=1}^{r}\sum_{j=1}^{s}\log(m_{i,j}).$$

Now, applying the conditions (2.27) to (2.29), we will have

$$\begin{aligned} rs\log(m_{i,j}) &= r\left\{s\log(m_{i,\cdot}) + \sum_{j=1}^{s}\log(m_{\cdot,j}) - s\log(n)\right\} + \\ &\quad + s\left\{\sum_{i=1}^{r}\log(m_{i,\cdot}) + r\log(m_{\cdot,j}) - r\log(n)\right\} - \\ &\quad - \left\{s\sum_{i=1}^{r}\log(m_{i,\cdot}) + r\sum_{j=1}^{s}\log(m_{\cdot,j}) - rs\log(n)\right\} \\ &= rs\log(m_{i,\cdot}) + rs\log(m_{\cdot,j}) - rs\log(n) \\ &= rs\{\log(m_{i,\cdot}) + \log(m_{\cdot,j}) - \log(n)\}, \end{aligned}$$

which is equivalent to write

$$\log(m_{i,j}) = \log(m_{i,\cdot}) + \log(m_{\cdot,j}) - \log(n),$$

which proves that the expression (2.26) is equivalent to (2.30).

B

Deviance graphs for every month

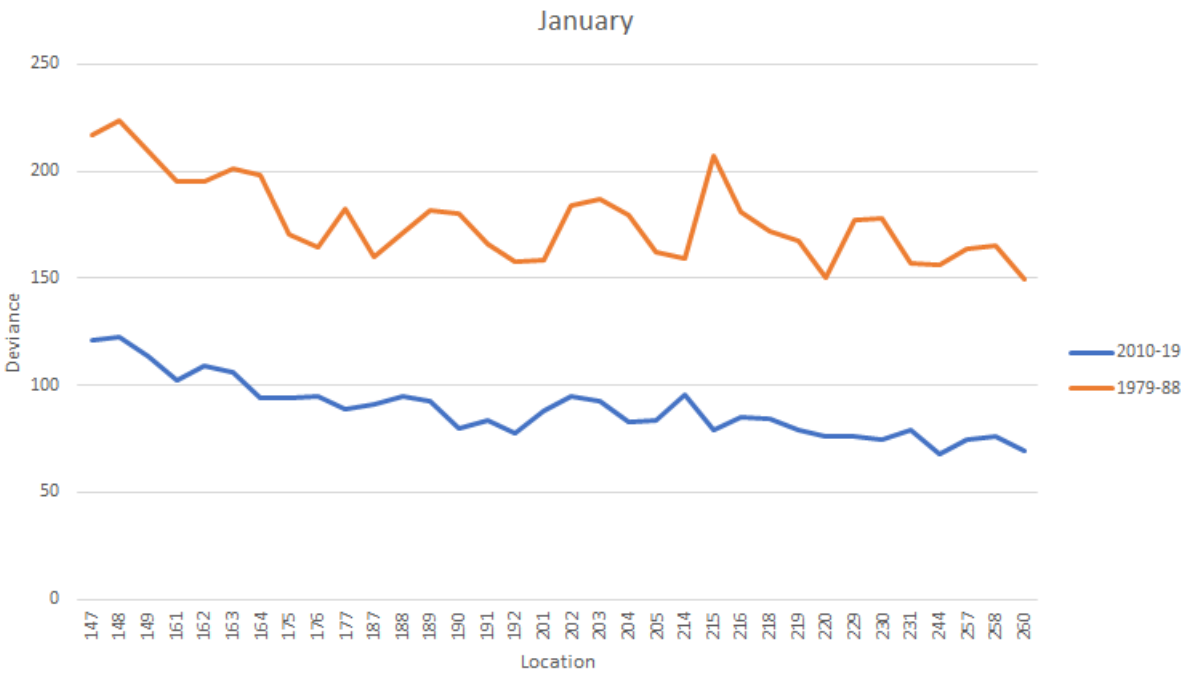

Figure B.1: Deviance values for every location resulting from eliminating January

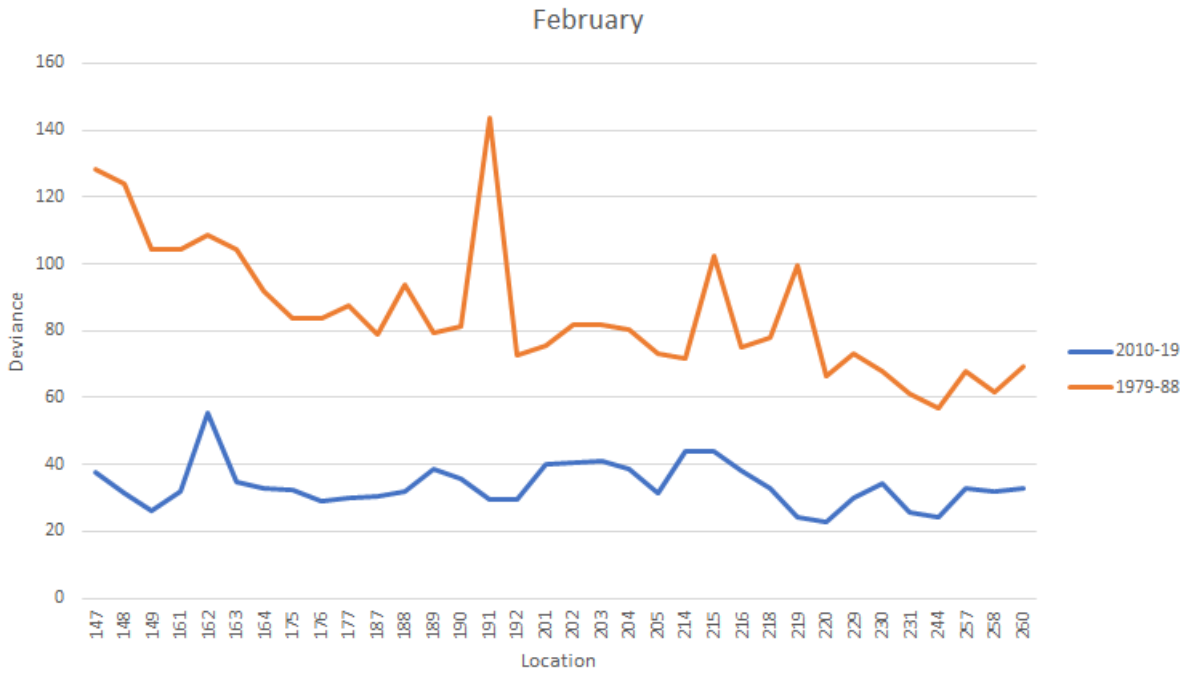

Figure B.2: Deviance values for every location resulting from eliminating February

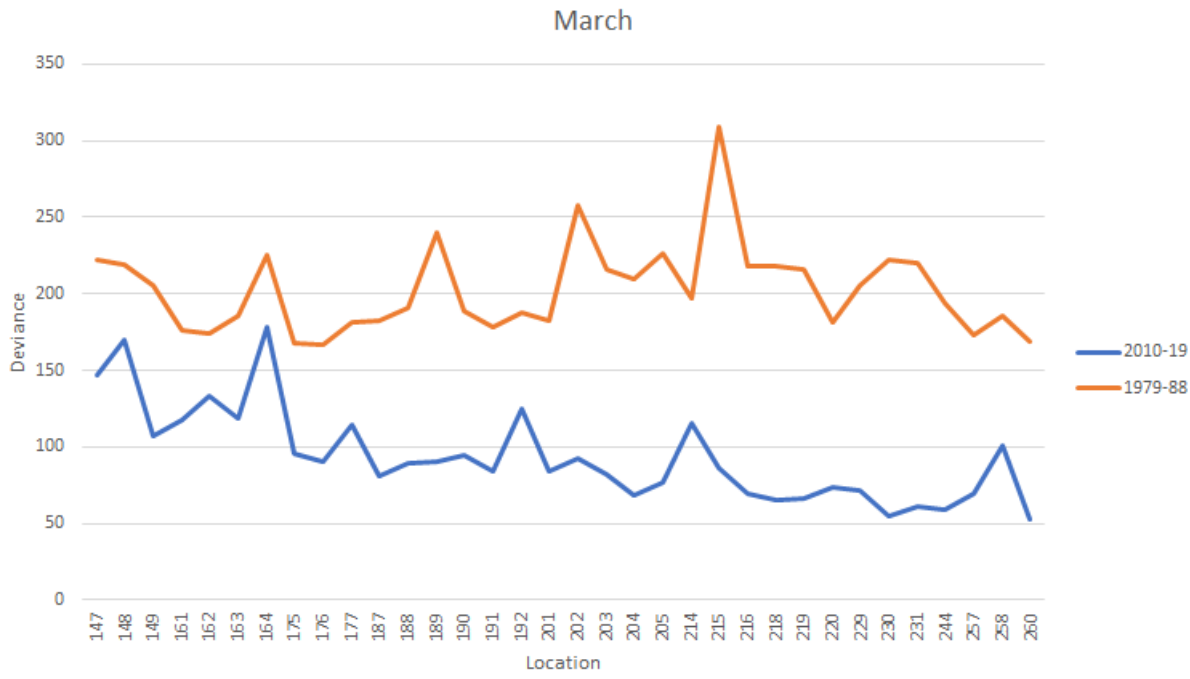

Figure B.3: Deviance values for every location resulting from eliminating March

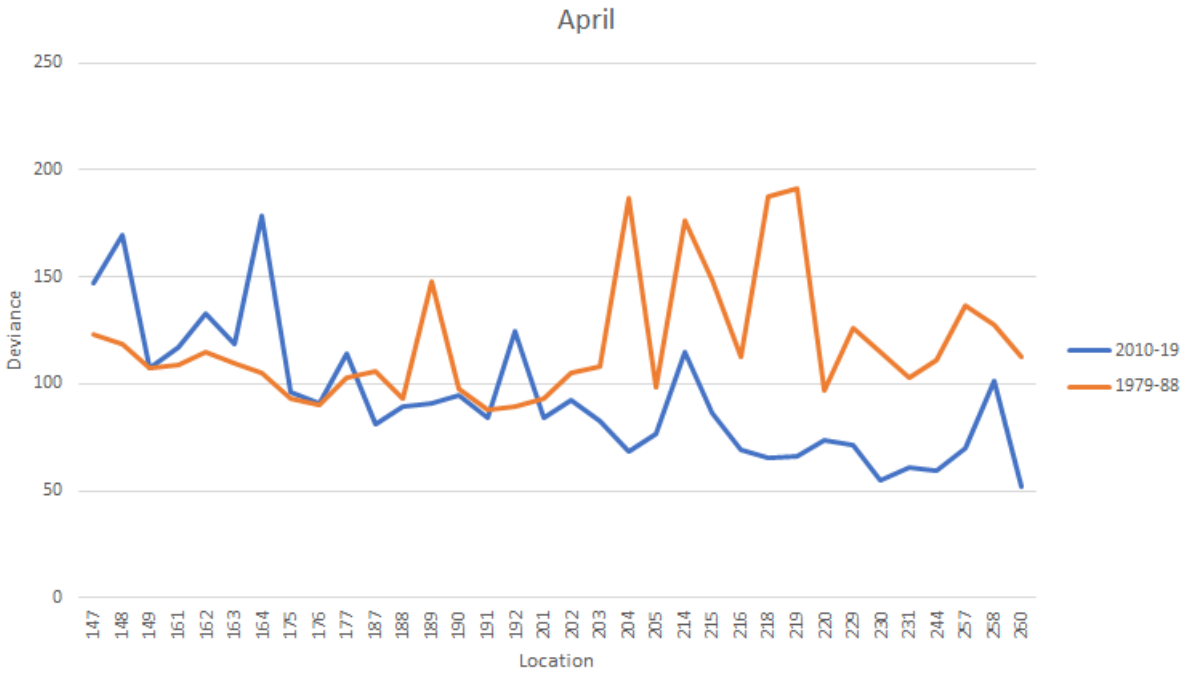

Figure B.4: Deviance values for every location resulting from eliminating April

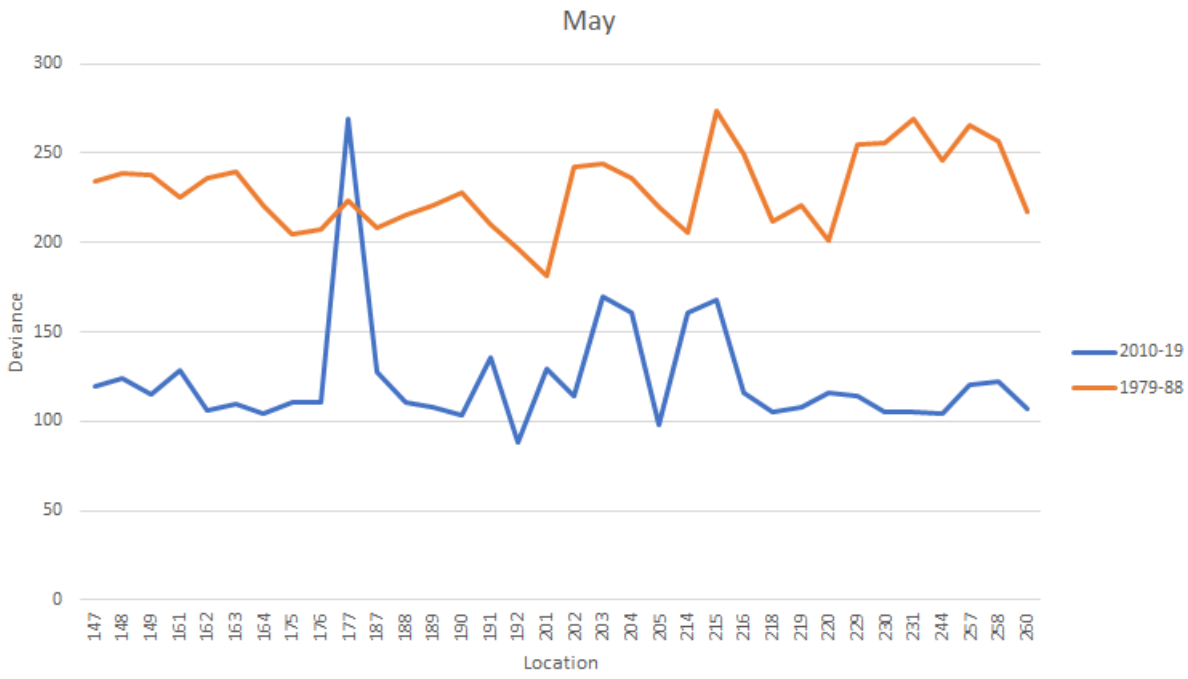

Figure B.5: Deviance values for every location resulting from eliminating May

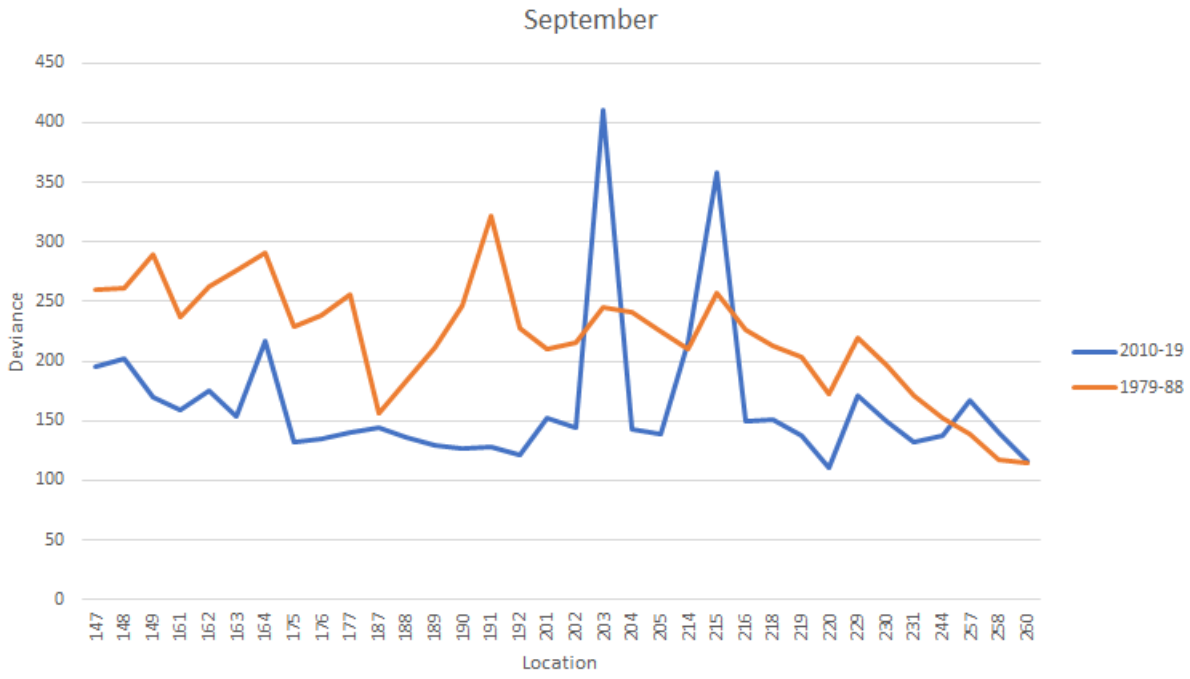

Figure B.6: Deviance values for every location resulting from eliminating September

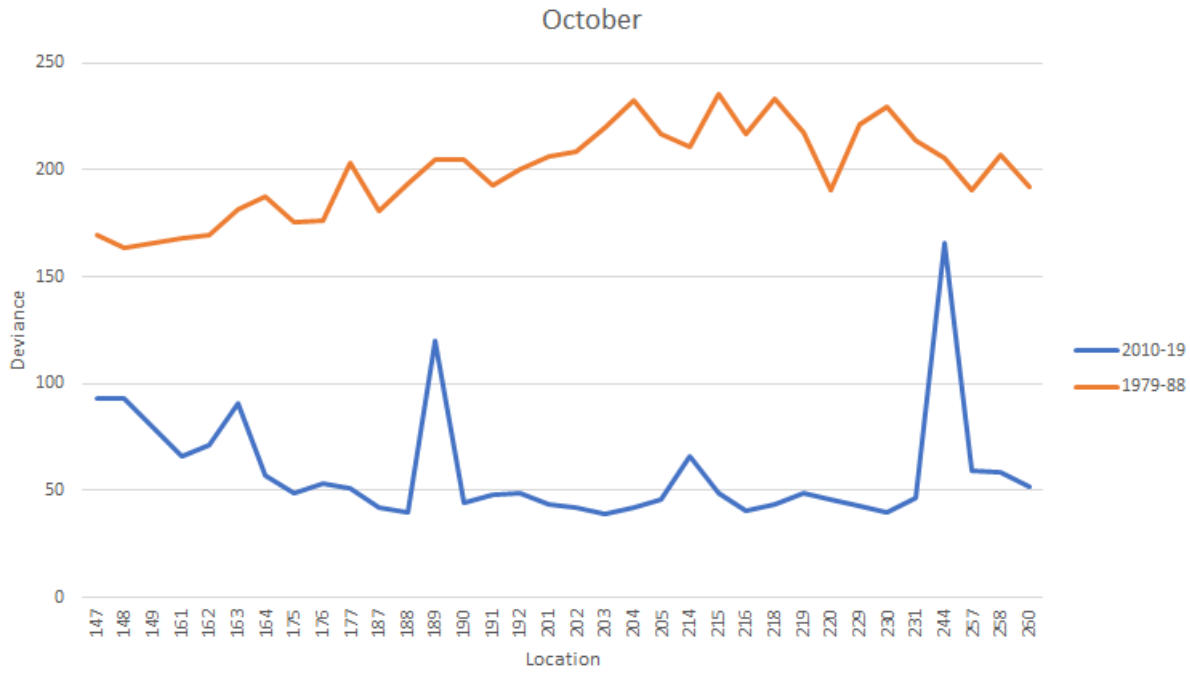

Figure B.7: Deviance values for every location resulting from eliminating October

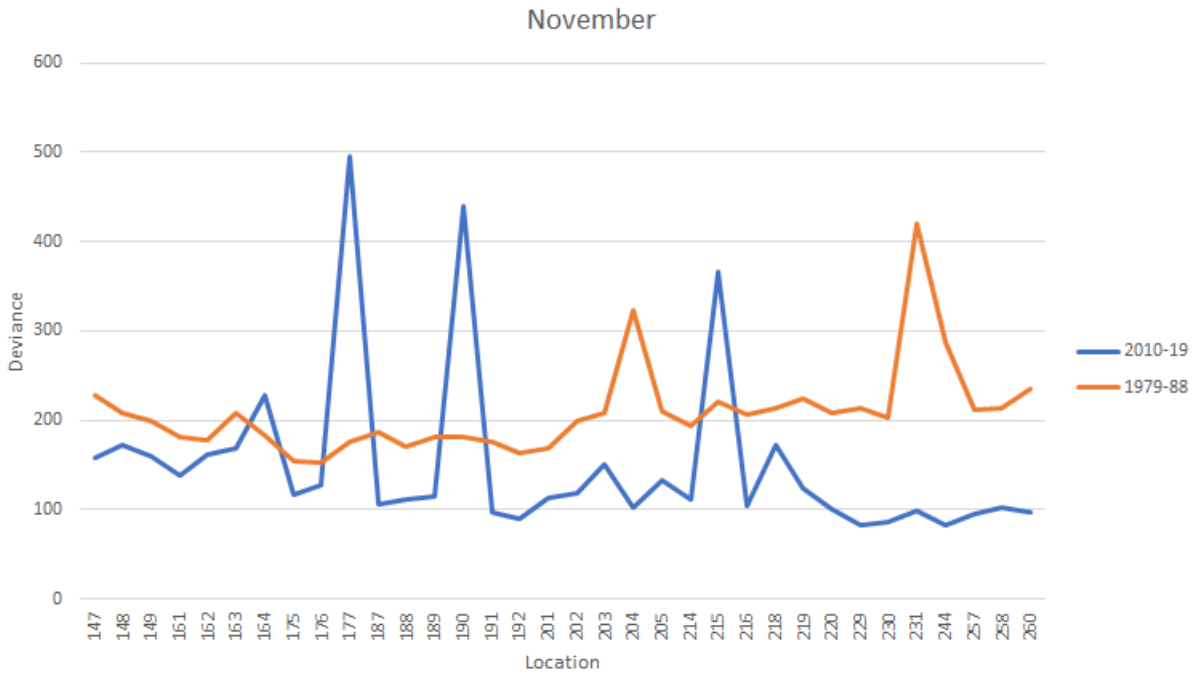

Figure B.8: Deviance values for every location resulting from eliminating November

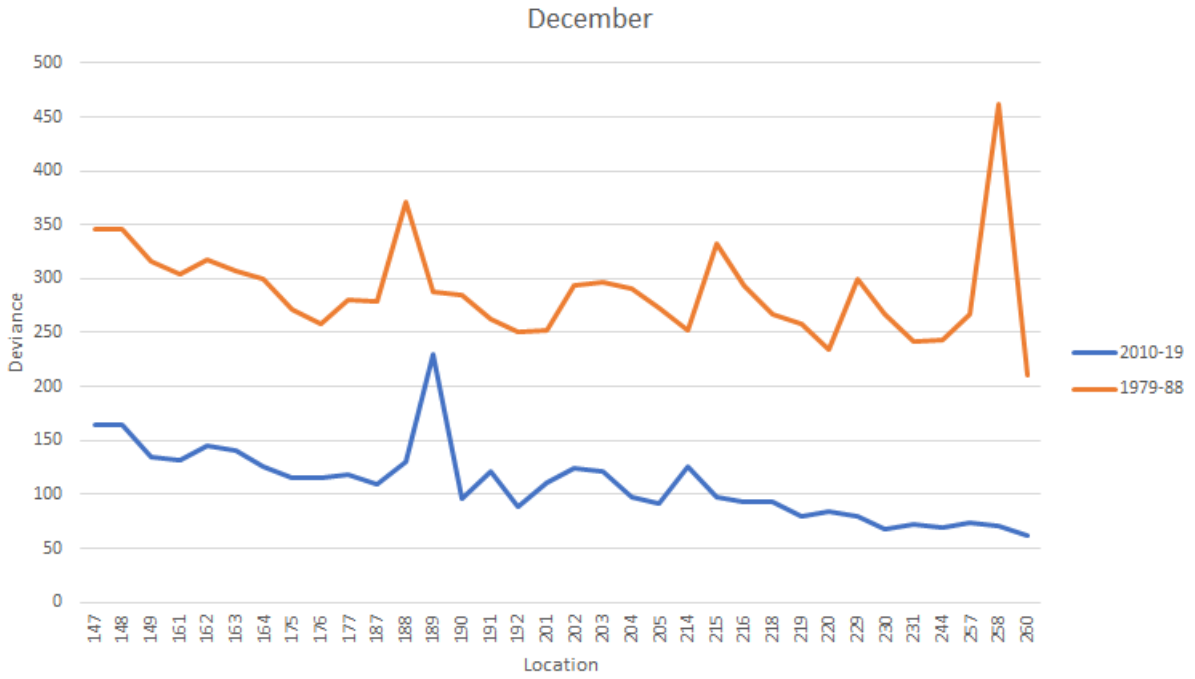

Figure B.9: Deviance values for every location resulting from eliminating December

I

Residual deviance values tables

Table I.1: Residual deviance values for model without interaction per location in each decade

Location	147	148	149	161	162	163	164	175	176	177	187	188
1979-88	3193.80	3282.70	3142.35	2958.55	3062.17	3055.11	2968.29	2719.59	2701.66	2780.86	2397.76	2493.35
1990-99	2652.69	2561.49	2421.32	2316.55	2333.92	2431.53	2391.37	2149.81	2144.76	2322.25	2128.12	2310.54
2000-09	3327.00	3120.44	2791.81	2649.76	2724.96	2666.42	2442.41	2228.43	2193.99	2245.14	2147.91	2236.98
2010-19	2758.34	2765.20	2495.12	2374.98	2510.22	2558.69	2447.11	2232.16	2181.72	2321.94	2186.92	2310.85

Location	189	190	191	192	201	202	203	204	205	214	215	216
1979-88	2720.11	2907.26	2603.68	2487.38	2356.40	2677.76	2906.36	2903.40	2671.18	2504.40	3093.80	2734.47
1990-99	2363.12	2212.26	2230.14	2210.75	2216.87	2401.11	2437.98	2334.17	2266.06	2761.69	2248.43	2238.69
2000-09	2301.25	1979.22	1946.16	1873.14	2203.33	2311.75	2236.90	2006.83	1906.97	2573.89	2106.34	2074.96
2010-19	2417.67	2151.57	2033.97	1910.43	2219.09	2389.82	2413.92	2249.61	2056.17	2452.42	2095.96	2149.24

Location	218	219	220	229	230	231	244	257	258	260		
1979-88	2646.30	2654.49	2404.80	2799.03	2709.65	2494.05	2570.63	2581.44	2598.73	2439.28		
1990-99	2256.47	2074.88	2071.98	2258.69	2060.91	2001.55	1896.82	2204.40	2110.46	1925.48		
2000-09	1971.48	1770.21	1680.02	2010.93	1896.43	1747.21	1726.00	2016.53	2044.33	2005.53		
2010-19	2165.07	1885.73	1749.20	2002.73	1940.03	1857.74	1694.67	1924.48	1895.06	1643.92		

Table I.2: Deviance values for the year factor per location in each decade

Location	147	148	149	161	162	163	164	175	176	177	187	188
1979-88	282.70	283.60	234.89	227.38	220.96	231.60	209.32	214.07	205.73	212.62	168.61	197.72
1990-99	334.66	337.89	343.35	277.50	281.37	347.77	396.86	275.80	272.67	337.89	316.52	346.29
2000-09	364.30	278.88	215.52	272.57	271.38	223.92	205.29	225.38	221.50	193.41	270.69	266.46
2010-19	545.92	503.20	468.49	496.09	510.55	519.11	501.54	442.81	444.67	491.35	364.77	400.60

Location	189	190	191	192	201	202	203	204	205	214	215	216
1979-88	227.68	246.53	213.77	185.93	176.25	219.43	251.26	275.54	248.82	220.11	247.44	209.09
1990-99	345.16	314.73	332.50	368.69	346.13	370.29	383.70	379.42	374.67	570.95	381.66	357.67
2000-09	254.09	204.90	178.77	171.47	265.60	264.17	260.90	214.60	184.37	321.02	221.86	216.83
2010-19	440.80	421.08	417.49	414.39	385.70	421.60	434.70	449.28	425.25	467.18	395.55	392.55

Location	218	219	220	229	230	231	244	257	258	260		
1979-88	258.99	276.12	247.01	188.38	187.79	210.14	167.36	217.65	195.06	220.50		
1990-99	421.00	381.97	391.13	409.91	356.67	366.28	370.78	484.58	485.69	429.96		
2000-09	204.99	180.93	190.72	161.61	149.42	138.87	114.86	125.94	146.44	158.45		
2010-19	477.05	422.94	360.14	406.26	369.55	370.01	339.42	433.67	404.64	392.45		

Table I.3: Residual deviance values for the month factor per location in each decade

Location	147	148	149	161	162	163	164	175	176	177	187	188
1979-88	733,70	669,80	541,89	619,77	640,89	566,89	494,40	539,70	532,73	520,42	602,39	583,26
1990-99	666,94	601,23	478,03	503,23	549,97	504,26	475,25	475,91	478,68	479,51	451,44	485,56
2000-09	617,20	480,59	392,39	488,22	526,92	516,10	491,65	449,64	438,44	483,41	498,76	447,24
2010-19	501,23	539,00	464,46	538,40	556,68	492,42	442,45	513,90	488,32	474,18	436,75	466,53

Location	189	190	191	192	201	202	203	204	205	214	215	216
1979-88	585,10	578,33	552,17	531,81	523,56	675,75	682,82	665,01	643,69	601,91	983,66	702,42
1990-99	513,64	478,49	474,04	469,68	529,66	564,03	551,41	515,47	500,07	774,16	546,41	512,10
2000-09	451,61	419,27	448,78	456,79	631,11	554,76	470,97	444,51	454,22	947,74	645,84	514,44
2010-19	481,59	452,14	445,62	427,93	432,39	462,41	468,56	460,44	450,49	556,34	446,04	427,25

Location	218	219	220	229	230	231	244	257	258	260		
1979-88	720,82	702,28	627,87	769,64	627,21	586,87	654,73	869,94	713,50	720,10		
1990-99	504,09	444,41	436,08	441,60	392,90	407,92	326,82	368,56	333,35	334,73		
2000-09	461,15	411,62	350,79	527,71	423,25	439,71	358,36	492,90	381,96	383,99		
2010-19	492,15	479,55	486,06	446,64	444,32	506,58	474,81	530,13	537,28	514,06