UNIVERSITAT DE BARCELONA

BACHELOR’S DEGREE IN MATHEMATICS
BACHELOR’S DEGREE IN BUSINESS ADMINISTRATION

Bachelor’s thesis

Spatial statistics and Gravity models with an application to the analysis of migration flows in Catalonia using 2019 data

Alberto Murillo Auset

Mentors:

Dr. Josep Vives i Santa-Eulàlia

Department of Mathematics and Computer Science

Dr. Raul Ramos Lobo

Department of Econometrics, Statistics and Applied Economics

Barcelona, January 2023

Contents

Abstract **iii**

Acknowledgments **iv**

Introduction **v**

1 Random Field Modelling **1**

1.1 Gaussian Random Fields 1

1.2 Stationarity 3

1.3 Kriging 5

2 Models and Inference for Areal Unit Data **15**

2.1 Discrete Random Fields 15

2.2 Gaussian Autoregression Models 18

2.3 Markov Random Fields 22

3 Methods for the Analysis of Migration Flows **29**

3.1 What is a gravity model? 29

3.2 What is a Radiation Model? 31

3.3 Deep Gravity Model 33

3.3.1 Fundamentals 33

3.3.2 Method (Architecture of deep gravity) 37

4 Methods Implementation and Comparison **39**

4.1 Migration flows within Catalonia using Gravity Models 39

4.2 Main Issues 41

4.3 Migration flows within Catalonia using the Radiation Model 42

4.4 Comparison with Gravity Models 42

Conclusions **46**

References **47**

Appendixes **50**

Abstract

The aim of this work is to examine spatial statistics fundamentals and to study three different types of models used for flux migration forecasting: gravity models, radiation models and deep gravity models. The attention is drawn on Gaussian random fields, stationarity, kriging, Gaussian autoregression models and Markov random fields. Four gravity models, their respective four multi-linear regression models and a radiation model are implemented, all pulling the same data, to forecast migration flows within Catalonia (Spain) in 2019 between *comarques* by using data from these locations.

2020 Mathematics Subject Classification. 60G60, 60J10, 62H11, 62J04, 62M20, 62M30, 91B72, 68T07

Acknowledgments

Vull agrair als meus tutors Raul Ramos i Josep Vives pel suggeriment de la temàtica d'aquest treball, així com per l'ajuda i les correccions al llarg del semestre. María C., Mario, María R., Clara, Marc, Paula: gràcies de tot cor. Finalment, vull agrair als meus pares per haver fet possible el meu pas per la universitat.

"Flow generation" is the name given to the problem of studying migration flows without using historical data. There is no denying that we live in dynamic ecosystems and spatial flows are taking place constantly: commuters going to cities for work, young people leaving their hometowns to pursue a university degree, people going on vacation, escaping from natural disasters and wars, or people who change their location for a better opportunity and well-being. The motives are endless, and so are the variables that determine an exact flow prediction. Studying mobility patterns provides useful information to governments and entities that look after efficiency in big cities, counties and countries. Since it is not possible to know exactly the migration flows that are happening, we need to rely on mathematical models to generate these flows.

Gravity models were introduced in economics in the twentieth century. These follow Newton's gravity law to forecast economic flows such as trade, foreign investment or migration flows between two countries, considering the distance between them two, and the actual variable that is being studied. This had a projection towards the study of transport planning, spatial economics and epidemic spreading patterns. Following a physical law for obtaining a model is not an exclusive thing for gravity models: radiation models are based on how energetic particles travel through vacuum. However, these models do not capture the structure of real flows due to the fact that there is a large amount of information missing and not taken into consideration. The complexity of the geographical landscape requires of more data than just one variable and the distance between the origin and destination of the flow: more detailed data needs to be imputed and it is necessary to understand the characteristics of the spatial flows. This is why deep learning provides better results: they take into account more data and can generate realistic flows without information regarding historical data.

The same way that we are not able to register every migration and we use models to forecast them, many times, data scientists struggle to obtain data of any kind due to its expensiveness. This is not different to data related to locations. For instance, we cannot run air pollution tests in every squared kilometer to obtain a pollution map: it would take a lot of time and money. Once again, we need to rely on mathematical methods to sort this problem out. Spatial statistics is the area of study dedicated to statistical analysis of data with spatial information, adding uncertainty quantification. It gives a probabilistic frame to answer spatial-location questions. We find land surveying at least in 1400 B.C. in Egypt where the dimensions of taxable land plots were measured. Areas as botany and ethology have contributed to spatial statistics research through studying how plants distribute and animals migrate. This area still has many active issues that are part of the modern research's agenda: definition of objects of study, best analytical operations, how to present data spatially, presentation of results. In this dissertation the attention is dedicated to provide an introduction to spatial statistics.

Chapter 1

Random Field Modelling

In this chapter, random fields modelling is presented, setting off from Kolmogorov's work from 1933 [3] and the consistency theorem. Gaussian random fields and some stationarity concepts are revised using the textbook written by P. Billingsley [34] and Adler and Taylors' book [38]. Notation about covariance functions are followed using the same notation as Schlather [30] and Steins'[31] books. As for Kriging, simple and ordinary kriging are explained tracing N. Cressie [33] and G. Matheron [15], [16] and [17]'s work from the early 1960s. Some work follows the handbook of spatial statistics written by A. Gelfand, P. Diggle, M. Fuentes and P. Guttorp [1], and we suggest the reader to have a further revision using this book.

1.1 Gaussian Random Fields

Pollution, minimum temperatures or wave height on a given day, monthly precipitation an many other data are often represented on maps. These maps can be described by random quantities indexed by points in a region of interest, and the ensemble of these is called random field:

Definition 1.1. *Let (Ω, F, P) be a probability space. A random field is a family $X = \{X_t\}_{t \in T}$ of random variables X_t indexed by elements t in a subset $T \subseteq \mathbb{R}^d$ and defined on the same probability space.*

Let $\{t_1, \ldots, t_n\} \in T$ be a finite set of index values. The random vector $(X_{t_1}, \ldots, X_{t_n})$ has a well-defined probability distribution determined by its joint cumulative distribution function $F_{t_1,\ldots,t_n}(x_1, \ldots, x_n) = P(X_{t_1} \leq x_1; \ldots; X_{t_n} \leq x_n)$, where $x_i \in \mathbb{R}$

for $i = 1 \div n$. The Kolmogorov's Consistency Theorem proves that the probability distribution of X is uniquely defined by finite dimensional distributions, which are the set of all joint cumulative distribution functions given $t_1, \ldots, t_n$, and $n \in \mathbb{N}$. For more details about the theorem and its proof, consult the paper [3].

From now on we are going to assume that these joint distributions are normal, so only the mean and covariance function is required to be stated. This will ease our work and relies on the central limit theorem, which states that even when independent random variables with finite variance are not normally distributed, their distribution when they are summed up tends towards a normal distribution. This accounts for Gaussian models popularity.

Definition 1.2. *A random variable is normally (or Gaussian) distributed if its probability density function is*

$$f(x) = \frac{1}{\sigma(2\pi)^{1/2}} \exp\left\{ -\frac{1}{2} \left[\frac{x-\mu}{\sigma} \right]^2 \right\},$$

being μ the mean of the distribution and σ the standard deviation.

Formalizing the generalization of the one-dimensional normal distribution to higher dimensions, a random vector is n-variate normally distributed if every n-linear combination of its n components has a univariate normal distribution:

Definition 1.3. *A random vector $(X_1, \ldots, X_n)$ has a multivariate normal distribution with mean vector $m = (\mathbb{E}X_1, \ldots, \mathbb{E}X_n) \in \mathbb{R}^n$ and $n \times n$ covariance matrix Σ, being $\Sigma_{ij} = Cov(X_i, X_j)$, if any linear combination $a^\top X = \sum_{i=1}^n a_i X_i$, $a \in \mathbb{R}^n$, is normally distributed.*

Definition 1.4. *Let X be a random vector of dimension n with multivariate normal distribution. Its join density function is*

$$\phi(x) = \left(\frac{1}{2\pi} \right)^{n/2} |\Sigma|^{-1/2} \exp\left\{ -\frac{1}{2} (x-\mu)^\top \Sigma^{-1} (x-\mu) \right\}$$

where μ is its mean vector, Σ is its covariance matrix and $|\Sigma|$ is the determinant of Σ.

We can finally define a Gaussian random field:

Definition 1.5. *Let (Ω, F, P) be a probability space. A Gaussian random field is a family $X = \{X_t\}_{t \in T}$ of random variables X_t indexed by t in a subset $T \subseteq \mathbb{R}^d$ and defined on the same probability space that for any finite set $t_1, \ldots, t_n$, the random vector $(X_{t_1}, \ldots, X_{t_n})$ has a multivariate normal distribution.*

Proposition 1.6. *Let $m : T \to \mathbb{R}$; $m(t) = \mathbb{E}X_t$ be the mean of a Gaussian random field X, being $T \subseteq \mathbb{R}^d$. The function $\rho : T \times T \to \mathbb{R}$, $\rho(t_i, t_j) = Cov(X_{t_i}, X_{t_j})$ is its covariance, if and only if ρ is non-negative definite, this is, for any $t_1, \dots, t_n$, $n \in \mathbb{N}$ the matrix $(\rho(t_i, t_j))_{i,j=1}^n$ is non-negative definite. In different words, the matrix $(\rho(t_i, t_j))_{i,j=1}^n$ is symmetric and satisfies the property*

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i \rho(t_i, t_j) a_j \geq 0.$$

Proof. "$\Leftarrow$" We check the consistency of the finite dimensional distributions. Choosing $\mu_{t_1}, \dots, \mu_{t_n}$ a multivariate normal with the covariance matrix $\Sigma(t_1, \dots, t_n)$ filled with $\rho(t_i, t_j)$. $\mu_{t_1}, \dots, \mu_{t_n}$ is well defined since by hypothesis $\Sigma(t_1, \dots, t_n)$ is non-negative definite. Besides, $\mu_{t_1}, \dots, \mu_{t_n}$ are symmetric and their marginals are also normal with the covariance matrix since they are normal, so it is consistent. Finally, the Kolmogorov's consistency theorem is applied.

"$\Rightarrow$" by definition of covariance, it is non-negative definite. QED

1.2 Stationarity

We fix the index set to $T = \mathbb{R}^d$ in this section.

Definition 1.7. *A random field $X = (X_t)_{t \in \mathbb{R}^d}$ is strictly stationary if for all finite sets $t_1, \dots, t_n \in \mathbb{R}^d$, $n \in \mathbb{N}$, all $k_1, \dots, k_n \in \mathbb{R}$ and all $s \in \mathbb{R}^d$,*

$$P(X_{t_1+s} \leq k_1 \; ; \; \dots \; ; X_{t_n+s} \leq k_n) = P(X_{t_1} \leq k_1 \; ; \; \dots \; ; X_{t_n} \leq k_n)$$

Let now X be strictly stationary and its variance $\mathbb{E}X_t^2 < \infty$ for all $t \in \mathbb{R}^d$. Then

$$P(X_{t+s} \leq k) = P(X_t \leq k), \; \forall k,$$

so that both X_t and X_{t+s} have same distribution. Also, $\mathbb{E}X_t = \mathbb{E}X_{t+s}$, thus the mean funtion must be constant. Finally, before defining when a random field is weakly stationary, notice that if we now consider (X_{t_1}, X_{t_2}) and (X_{t_1+s}, X_{t_2+s}) satisfying

$$P(X_{t_1+s} \leq k_1 \; ; \; X_{t_2+s} \leq k_2) = P(X_{t_1} \leq k_1 \; ; \; X_{t_2} \leq k_2)$$

so they have same distributions, then their covariance are equal:

$$Cov(X_{t_1+s}, X_{t_2+s}) = Cov(X_{t_1}, X_{t_2})$$

and particularly if we set $s = -t_1$ we obtain

$$\rho(t_1, t_2) = \rho(t_1 + s, t_2 + s) = \rho(0, t_2 - t_1)$$

which is a function of $t_2 - t_1$.

Definition 1.8. *A random field $X = (X_t)_{t \in \mathbb{R}^d}$ is weakly stationary if*

1. $\mathbb{E}X_t^2 < \infty, \forall t \in \mathbb{R}^d$
2. $\mathbb{E}X_t \equiv m$ *is constant*
3. $Cov(X_{t_1}, X_{t_2}) = \rho(t_2 - t_1)$ *for some* $\rho : \mathbb{R}^d \to \mathbb{R}$.

Proposition 1.9. *Let $X = (X_t)_{t \in \mathbb{R}^d}$ be a Gaussian random field. If $X = (X_t)_{t \in \mathbb{R}^d}$ is weakly stationary, it is also strictly stationary.*

Proof. Let $(X_{t_1}, \dots, X_{t_n})$ a random vector, whose mean vector is $(m, \dots, m)$ and covariance matrix $\Sigma(t_1, \dots, t_n)$ filled with $\rho(t_j - t_i)$. The shifted vector is $(X_{t_{1+s}}, \dots, X_{t_{n+s}})$, which also follows a normal distribution like the random vector and has $(m, \dots, m)$ as mean vector and covariance matrix $\Sigma(t_1 + s, \dots, t_n + s)$ filled with $\rho(t_j + s - (t_i + s)) = \rho(t_j - t_i)$, no matter the s chosen. Hence, X is strictly stationary. QED

Proposition 1.10. *Let $\rho : \mathbb{R}^d \to \mathbb{R}$ be the covariance function of a weakly stationary (Gaussian) random field. The following holds:*

1. $\rho(0) \geq 0$
2. $\rho(t) = \rho(-t),\ \forall t \in \mathbb{R}^d$
3. $|\rho(t)| \leq \rho(0),\ \forall t \in \mathbb{R}^d$

Proof.

1. $\rho(0) = Cov(X_0, X_0) = Var(X_0) \geq 0$ by definition of variance.
2. $\rho(t) = Cov(X_0, X_t) = Cov(X_t, X_0) = \rho(-t)$
3. $|\rho(t)|^2 = |\mathbb{E}[(X_t - m)(X_0 - m)]|^2 \leq \mathbb{E}[(X_t - m)^2]\mathbb{E}[(X_0 - m)^2] = \rho(0)^2$ which proves the third claim by taking the square root on both sides. QED

Now let's consider a weakly stationary random field X and the increment $X_{t_1} - X_{t_2}$ for $t_1, t_2 \in T$ to define a even weaker form of stationarity. We can write the variance of the increment as the following, so it only depends on the spatial lag $t_2 - t_1$:

$$Var(X_{t_2} - X_{t_1}) = Var(X_{t_2}) + Var(X_{t_1}) - 2Cov(X_{t_2}, X_{t_1}) = 2\rho(0) - 2\rho(t_2 - t_1)$$

Definition 1.11. *A random field $X = (X_t)_{t \in \mathbb{R}^d}$ is intrinsically stationary if*

1. $\mathbb{E}X_t^2 < \infty, \forall t \in \mathbb{R}^d$
2. $\mathbb{E}X_t \equiv m$ *is constant*
3. $Var(X_{t_1} - X_{t_2}) = f(t_2 - t_1)$ *for some* $f : \mathbb{R}^d \to \mathbb{R}$

1.3 Kriging

The semi-variogram is often used in geostatistics instead of variance. It only requires the weaker assumption of intrinsic stationarity. The semi-variogram allows analizing the spatial behaviour of a variable in a defined area, obtaining an experimental variogram that displays the maximal distance and the way a point influences another depending on the distance. We can obtain the *scope*, this is the maximal distance where a sample influences another sample, and the vicinity where we can search for samples to estimate the value in a specific point. Another application is to use the kriging methodology through the data obtained by a theorical variogram.

Definition 1.12. *Let $X = (X_t)_{t \in \mathbb{R}^d}$ be intrinsically stationary. Then, the semi-variogram $\gamma : \mathbb{R}^d \to \mathbb{R}$ is defined by*

$$\gamma(t) = \frac{1}{2} Var(X_t - X_0), \quad t \in \mathbb{R}^d.$$

Note: For weakly stationary random fields we have $\gamma(t) = \rho(0) - \rho(t)$. Particularly, $\gamma(0) = \rho(0) - \rho(0) = 0$.

Now suppose that, given an intrinsically stationary random field X, we get the observations via the linear model

$$Y_i = X_{t_i} + E_i, \quad i = 1 \div n$$

with additional measurement error terms E_i that are independent, identically distributed and with mean of zero, with variance σ_E^2. Then using the definition of covariance used before with a γ,

$$\frac{1}{2} Var(Y_j - Y_i) = \gamma_X(t_j - t_i) + \frac{1}{2} Var(E_j - E_i) = \gamma_X(t_j - t_i) + \sigma_E^2 1\{i \neq j\}$$

so we have

$$\gamma_Y(t) = \begin{cases} \gamma_X(t) + \sigma_E^2 & , \quad t \neq 0 \\ \\ \gamma_X(t) & , \quad t = 0 \end{cases}$$

which has a discontinuity in $t = 0$. This is called *the nugget effect*. We can intuitively assume that as the distance between two sampled values increases, their dependence between them two diminishes. This can be seen as this limit when it exists: $lim_{\|t\|\to\infty}\rho(t) = 0$. If it exists, we define the *sill* as $lim_{\|t\|\to\infty}\gamma(t)$. The *partial sill*, considering the nugget effect, is defined as $lim_{\|t\|\to\infty}\gamma(t) - lim_{\|t\|\to 0}\gamma(t)$.

Considering the case where there is only one available single finite sample $X_{t_1}, \ldots, X_{t_n}$ for $n \in \mathbb{N}$ of the random field X and we want to implement statistical inference, we assume at least intrinsic stationarity to get an artificial replication. So, given lag t, all pairs of observations that are 'around' t apart and to average, are going to be considered. By doing this we get the following concept:

Definition 1.13. *Let $X_{t_1}, \ldots, X_{t_n}$ for $n \in \mathbb{N}$, be a finite sample available of a random field X. The Matheron estimator is defined as:*

$$\hat{\gamma}(t) = \frac{1}{2|N(t)|} \sum_{(t_i,t_j)\in N(t)} (X_{t_j} - X_{t_i})^2$$

where $N(t) = \{(t_i, t_j) : t_j - t_i \in B(t, \epsilon)\}$ is the t-neighbourhood, $|\cdot|$ denotes cardinality, and $B(t, \epsilon)$ is the closed ball of radius ϵ and center t.

Two things deserve special mention here. First, the Matheron estimator is practically unbiased when $N(t)$ is not empty. Still considering X intrinsically stationary, $\mathbb{E}\hat{\gamma}(t)$ is the average value of $\gamma(t_j - t_i)$ over $N(t)$: we obtained this because

$$2|N(t)|\mathbb{E}\hat{\gamma}(t) = \sum_{(t_i,t_j)\in N(t)} \mathbb{E}\left[(X_{t_j} - X_{t_i})^2\right] = 2 \sum_{(t_i,t_j)\in N(t)} \gamma(t_j - t_i)$$

Second, the ϵ must be large enough to have a fair number of points in $N(t)$ to have a stable average, but also small enough to have $\gamma(t_j - t_i) \approx \gamma(t)$ for $t_j - t_i$ in the ball $B(t, \epsilon)$.

Furthermore, the Matheron estimator is not parametric but any family γ_θ that minimises

$$\sum_j w_j(\hat{\gamma}(h_j) - \gamma_\theta(h_j))^2$$

can fit, where w_j can either be, for instance, equal to $|N(h_j)|$ or $|N(h_j)|/\gamma_\theta(h_j)^2$ and the family of h_j is finite. As for the latter value for w_j, notice that the smaller the semi-variogram, the larger the weight for a pair of observations at that aproximated lag, to make up for their rare occurrence.

Focusing now on the goal of making a prediction of the value at a certain location t_0 where no measure has been made, having observed previously values $X_{t_1} = x_{t_1}, \dots, X_{t_n} = x_{t_n}$ of a random field $X = (X_t)_{t \in \mathbb{R}^d}$ at n locations $t_i \in \mathbb{R}^d, i = 1 \div n$, we need, though, the mean function m and the covariance ρ of X.

A linear predictor of X_{t_0} has the form

$$\hat{X}_{t_0} = c(t_0) + \sum_{i=1}^{n} c_i X_{t_i}$$

hence,

$$\mathbb{E}\hat{X}_{t_0} = c(t_0) + \sum_{i=1}^{n} c_i m(t_i)$$

then the predictor $\hat{X}_{t_0}$ is unbiased, this is $\mathbb{E}\hat{X}_{t_0} = m(t_0)$, if and only if

$$c(t_0) = m(t_0) - \sum_{i=1}^{n} c_i m(t_i)$$

Definition 1.14. *The mean squared error (mse) of $\hat{X}_{t_0}$ is defined as*

$$\mathbb{E}[(\hat{X}_{t_0} - X_{t_0})^2] = Var(\hat{X}_{t_0} - X_{t_0}) + (\mathbb{E}[\hat{X}_{t_0} - X_{t_0}])^2$$

Basically the mse is the sum of the variance and the bias. It is $mse(\hat{X}_{t_0}) = Var(\hat{X}_{t_0} - X_{t_0})$ when the predictor is unbiased.

Definition 1.15. *Let $X_{t_1} = x_{t_1}, \dots, X_{t_n} = x_{t_n}$ be a sample from a random field $X = (X_t)_{t \in \mathbb{R}^d}$ at n locations $t_i \in \mathbb{R}^d, i = 1 \div n$, and collect them in the vector Z. Let Σ be the covariance matrix of Z, existing and non-singular. Let $K = (K_i)_{i=1}^n$ be the n-vector with entries $K_i = \rho(t_i, t_0)$. The simple kriging estimator of X_{t_0} is defined as*

$$\hat{X}_{t_0} = m(t_0) + K^\top \Sigma^{-1}(Z - \mathbb{E}Z).$$

This was named after D.G. Kridge, statistician, mining engineer and pioneer in the field of geostatistics.

Theorem 1.16. *Let $X_{t_1} = x_{t_1}, \dots, X_{t_n} = x_{t_n}$ be a sample from a random field $X = (X_t)_{t \in \mathbb{R}^d}$ at n locations $t_i \in \mathbb{R}^d, i = 1 \div n$, and collect them in the vector Z. Let Σ be the* *covariance matrix of* Z*, existing and non-singular. Let* $K = (K_i)_{i=1}^n$ *be the n-vector with entries* $K_i = \rho(t_i, t_0)$*. Then, the simple kriging estimator of* X_{t_0}

$$\hat{X}_{t_0} = m(t_0) + K^\top \Sigma^{-1}(Z - \mathbb{E}Z)$$

is the best linear predictor of X_{t_0}*,* $t_0 \in \mathbb{R}^d$*, regarding the mean squared error. The mean squared prediction error is given by*

$$\rho(t_0, t_0) - K^\top \Sigma^{-1} K.$$

Proof. First we have $mse(\hat{X}_{t_0}) = Var(\hat{X}_{t_0} - X_{t_0})$, since the predictor is unbiased, and

$$\hat{X}_{t_0} - X_{t_0} = c(t_0) + \sum_{i=1}^{n} c_i X_{t_i} - X_{t_0}.$$

To use a simpler notation, we write $c^\top = (c_1, \ldots, c_n)$ and $Z^\top = (X_{t_1}, \ldots, X_{t_n})$, so we have $Var(\hat{X}_{t_0} - X_{t_0}) = Var(c^\top Z - X_{t_0}) = c^\top \Sigma c - 2c^\top K + \rho(t_0, t_0)$, being Σ an $n \times n$ matrix filled up with $\rho(t_i, t_j)$ and K a n-vector filled up with $\rho(t_i, t_0)$. The derivative with respect to ∂c is $2\Sigma c - 2K$. This is zero when $c = \Sigma^{-1} K$ given an invertible Σ. Nevertheless, notice that whenever Σ is singular, there would be a solution anyways since K is in the column of Σ.

Let $\tilde{c}$ be a solution of $\Sigma \tilde{c} = K$. We are going to verify that the null solution of the derivative is actually the minimiser of the mse. Rewriting the combination $c^\top Z$ as $(\tilde{c} + (c - \tilde{c}))^\top Z$. To make the notation simpler, let's call $d = c - \tilde{c}$, so we have

$$\begin{aligned} Var((\tilde{c} + d)^\top Z - X_{t_0}) &= \tilde{c}^\top \Sigma \tilde{c} + d^\top \Sigma d + 2\tilde{c}^\top \Sigma d - 2\tilde{c}^\top K - 2d^\top K + \rho(t_0, t_0) \\ &= \tilde{c}^\top \Sigma \tilde{c} - 2\tilde{c}^\top K + d^\top \Sigma d + \rho(t_0, t_0) \\ &= \rho(t_0, t_0) - \tilde{c}^\top K + d^\top \Sigma d \end{aligned}$$

using in the second and last equality that $K = \Sigma \tilde{c}$.

Notice that adding a scalar constant only affects the bias but not the variance. By adding d to $\tilde{c}$ it only adds a new non-negative term $d^\top \Sigma d$ due to the fact that the covariance matrix Σ is non-negative. QED

It is worth mentioning that the mean squared prediction error is smaller than the variance of X_{t_0}, and this reduction is explained by the fact that the estimator $\hat{X}_{t_0}$ considers information from locations around t_0.

The mean squared error is also called *Bayesian loss*. The Bayes estimator optimises the Bayes loss over all estimators that are functions of the sample $X_{t_1}, \ldots, X_{t_n}$.

Theorem 1.17. *Let $X = (X_t)_{t \in \mathbb{R}^d}$ be a random field, $X_{t_1}, \dots, X_{t_n}$ a sample from X, at n locations $t_i \in \mathbb{R}^d$, $i = 1 \div n$, collected in the n-vector Z. Then the Bayes estimator of X_{t_0}, $t_0 \in \mathbb{R}^d$, is given by $\hat{X}_{t_0} = \mathbb{E}[X_{t_0}|Z]$*

Proof. Based on the sample Z, suppose $\widetilde{X_{t_0}} = f(Z)$ and $M = \mathbb{E}[X_{t_0}|Z]$.

$$\mathbb{E}\left[(\widetilde{X_{t_0}} - X_{t_0})^2\right] = \mathbb{E}\left[(\widetilde{X_{t_0}} - M + M - X_{t_0})^2\right]$$

$$= \mathbb{E}\left[(\widetilde{X_{t_0}} - M)^2\right] + \mathbb{E}\left[(M - X_{t_0})^2\right] + 2\mathbb{E}\left[(\widetilde{X_{t_0}} - M)(M - X_{t_0})\right].$$

Since both M and $\widetilde{X_{t_0}}$ are functions of Z, we get

$$\mathbb{E}\left[(\widetilde{X_{t_0}} - M)(M - X_{t_0})\right] = \mathbb{E}\left(\mathbb{E}\left[(\widetilde{X_{t_0}} - M)(M - X_{t_0})\right] \mid Z\right)$$

$$= \mathbb{E}\left[(\widetilde{X_{t_0}} - M)(M - \mathbb{E}(X_{t_0}| \; Z))\right] = 0.$$

Finally,

$$\mathbb{E}\left[(\widetilde{X_{t_0}} - X_{t_0})^2\right] = \mathbb{E}\left[(\widetilde{X_{t_0}} - M)^2\right] + \mathbb{E}\left[(M - X_{t_0})^2\right] \geq \mathbb{E}\left[(M - X_{t_0})^2\right].$$

Note that it is an equality iif $\mathbb{E}\left[(M - X_{t_0})^2\right] = 0$.

QED

The Bayes estimator coincides in distribution with the best linear predictor whenever normality is provided. The Bayes estimator of a component is linear in Z given the other components, provided multivariate normally distributed random vectors and with $m(t_0) + K^\top \Sigma^{-1}(Z - \mathbb{E}Z)$ and the conditional variance is $\rho(t_0, t_0) - K^\top \Sigma^{-1} K$, which depends on Z (the covariances). In addition, we get the variance as follows

$$Var(X_{t_0}) = \mathbb{E}Var(X_{t_0}|Z) + Var(\mathbb{E}(X_{t_0}|Z)) = \rho(t_0, t_0) - (\rho(t_0, t_0) - K^\top \Sigma^{-1} K)$$

$$= K^\top \Sigma^{-1} K$$

Now let's see the case when we have an unknown global mean. This case is called ordinary kriging. Consider the model $X_t = \mu + E_t$, $t \in \mathbb{R}^d$, being $\mu \in \mathbb{R}$ the unknown global mean and E_t a zero mean random field, whose covariance function is $Cov(E_t, E_s) = \rho(t, s)$. We are searching a linear unbiased predictor that optimises the mse, given samples of X at $t_1, \dots, t_n$, $n \in \mathbb{N}$,

$$\hat{X}_{t_0} = c(t_0) + \sum_{i=1}^{n} c_i X_{t_i}$$

at a location $t_0 \in \mathbb{R}^d$. Using the same notation as before for Z and K. The simple kriging estimator would be

$$\hat{X}_{t_0} = \mu + K^\top \Sigma^{-1}(Z - \mathbb{E}_\mu Z)$$

which cannot be computed since we miss μ.

Theorem 1.18. *Let be $X_{t_1} = x_{t_1}, \dots, X_{t_n} = x_{t_n}$ sampled from a random field $X = (X_t)_{t \in \mathbb{R}^d}$ with unknown constant mean at n locations $t_i \in \mathbb{R}^d, i = 1 \div n$, and collect them in the vector Z. Let Σ be the covariance matrix of Z, existing and non-singular. Let $K = (K_i)_{i=1}^n$ be the n-vector with entries $K_i = \rho(t_i, t_0)$. Then,*

$$\hat{X}_{t_0} = K^\top \Sigma^{-1} Z + \frac{1 - \boldsymbol{1}^\top \Sigma^{-1} K}{\boldsymbol{1}^\top \Sigma^{-1} \boldsymbol{1}} \boldsymbol{1}^\top \Sigma^{-1} Z$$

is the best linear predictor of X_{t_0}, $t_0 \in \mathbb{R}^d$ in terms of mean squared error. The mean squared prediction error is equal to

$$\rho(t_0, t_0) - K^\top \Sigma^{-1} K + \frac{(1 - \boldsymbol{1}^\top \Sigma^{-1} K)^2}{\boldsymbol{1}^\top \Sigma^{-1} \boldsymbol{1}}$$

the last term accounts for the uncertainty regarding the mean.

Proof. Consider

$$\mu = \mathbb{E}_\mu \hat{X}_{t_0} = c(t_0) + \mu \sum_{i=1}^n c_i$$

unbiased, for all μ. Notice that when $\mu = 0 \implies c(t_0) = 0$, hence $\sum_{i=1}^n c_i = 1$. Being $c^\top = (c_1, \dots, c_n)$, we want to optimise the variance $Var_\mu(c^\top Z - X_{t_0})$ with the scale constraint on $c^\top$ by using the Euler-Lagrange method. Since E_{t_i} is a zero mean random field, $E_{t_i} = X_{t_i} - \mu$, then the expected value, considering

$$(\hat{X}_{t_0} - X_{t_0})^2 = \left(\sum_{i=1}^n c_i (X_{t_i} - \mu) - (X_{t_0} - \mu) \right)^2$$

$$= E_{t_0}^2 + \left(\sum_{i=1}^n c_i E t_i \right)^2 - 2E_{t_0} \sum_{i=1}^n c_i E_{t_i},$$

is

$$\mathbb{E}\left[(\hat{X}_{t_0} - X_{t_0})^2 \right] = \rho(t_0, t_0) + \sum_{i=1}^n \sum_{j=1}^n c_i c_j \rho(t_i, t_j) - 2 \sum_{i=1}^n c_i \rho(t_0, t_i).$$

To make the notation easier we write $\mathbf{1}^\top = (1, \dots, 1)$ so we have

$$\rho(t_0, t_0) + c^\top \Sigma c - 2c^\top K + \lambda(c^\top \mathbf{1} - 1)$$

the gradient equations (also called score or informant equations) follow as

$$\begin{cases} 0 = 2\Sigma c - 2K + \lambda \mathbf{1} \\ 1 = c^\top \mathbf{1} \end{cases}$$

Multiplying the first equation by $\mathbf{1}^\top \Sigma^{-1}$ and assuming that Σ is non-singular, we get

$$\begin{cases} 0 = \mathbf{2}\mathbf{1}^\top c - \mathbf{2}\mathbf{1}^\top \Sigma^{-1} K + \lambda \mathbf{1}^\top \Sigma^{-1} \mathbf{1} \\ 1 = c^\top \mathbf{1} \end{cases}$$

and the Lagrange multiplier is then

$$\lambda = 2\frac{\mathbf{1}^\top \Sigma^{-1} K - \mathbf{1}^\top c}{\mathbf{1}^\top \Sigma^{-1} \mathbf{1}} = 2\frac{\mathbf{1}^\top \Sigma^{-1} K - 1}{\mathbf{1}^\top \Sigma^{-1} \mathbf{1}}.$$

Now we substitute this value into the first equation and we obtain

$$c = \Sigma^{-1} K - \frac{\lambda}{2} \Sigma^{-1} \mathbf{1} = \Sigma^{-1} K + \frac{1 - \mathbf{1}^\top \Sigma^{-1} K}{\mathbf{1}^\top \Sigma^{-1} \mathbf{1}} \Sigma^{-1} \mathbf{1}$$

and the mean squared error is

$$\rho(t_0, t_0) + c^\top \Sigma c - 2c^\top K = \rho(t_0, t_0) - K^\top \Sigma^{-1} K + \frac{(1 - \mathbf{1}^\top \Sigma^{-1} K)^2}{\mathbf{1}^\top \Sigma^{-1} \mathbf{1}}.$$

Finally, let's see that it is indeed the optimised mse. Given an unbiased linear predictor $(c + d)^\top \mathbf{1} = 1$, that is, $d^\top \mathbf{1} = 0$, whose mse is

$$\rho(t_0, t_0) + c^\top \Sigma c - 2c^\top K + d^\top \Sigma d + 2c^\top \Sigma d - 2d^\top K.$$

We have $d^\top \Sigma c = d^\top K$ when it is not biased using the expression for c. Thus, we obtain that $d = 0$ makes the optimal mse.

QED

Note that in this case the mse is larger than when provided simple kriging.

Before getting into the last part of this section, the universal kriging, which relaxes the constant mean assumption, it is important to mention the case where Z is sampled from a Gaussian random field $(X_t)_{t \in \mathbb{R}^d}$. In this case we have a multivariate normally distributed with a mean μ that is constant and unknown to us, however we know the covariance matrix Σ which is non-singular.

The log likelihood at Z is the following:

$$-\frac{1}{2}(Z-\mu\mathbf{1})^\top\Sigma^{-1}(Z-\mu\mathbf{1}) = -\frac{1}{2}\sum_i\sum_j(X_{t_i}-\mu)\Sigma^{-1}_{ij}(X_{t_j}-\mu)$$

we derive it with respect to μ and we obtain:

$$-\frac{1}{2}\sum_i\sum_j\left[-\Sigma^{-1}_{ij}(X_{t_i}-\mu)-\Sigma^{-1}_{ij}(X_{t_j}-\mu)\right] = \mathbf{1}^\top\Sigma^{-1}(Z-\mu\mathbf{1})$$

now, this is equal to zero iif $\mathbf{1}^\top\Sigma^{-1}Z = \mu\mathbf{1}^\top\Sigma^{-1}\mathbf{1}$, thus

$$\hat{\mu} = \frac{\mathbf{1}^\top\Sigma^{-1}Z}{\mathbf{1}^\top\Sigma^{-1}\mathbf{1}}$$

Finally, we obtain the ordinary kriging predictor by substituting $\hat{\mu}$, which is the unique maximiser of the log likelihood because the second order derivative $-\mathbf{1}^\top\Sigma^{-1}\mathbf{1}$ is non-positive, in the simple kriging estimator:

$$\hat{X}_{t_0} = \hat{\mu} + K^\top\Sigma^{-1}(Z-\hat{\mu}\mathbf{1})$$

To finish this chapter, we are going to have a look through Universal Kriging. This model is suitable when we count on sampled values that depend linearly on p explanatory variables $m(t)_i$, for $i = 1 \div p$ and it eases off the constant mean that the ordinary kriging assumes, so we have a more general assumption which, given an unknown parameter vector $\beta \in \mathbb{R}^p$ and a known function $m : \mathbb{R}^d \to \mathbb{R}^p$, is that

$$\mathbb{E}X_t = m(t)^\top\beta$$

Considering the unbiasedness condition, a linear estimator $\hat{X}_{t_0} = c(t_0) + \sum_{i=1}^n c_i X_{t_i}$ is unbiased when the following equality satisfies for all β:

$$m(t_0)^\top\beta = c(t_0) + \sum_{i=1}^n c_i m(t_i)^\top\beta$$

since we have polynomials on both sides of the equality, coefficients must be equal, which means that $c(t_0) = 0$ and $m(t_0) = \sum_{i=1}^n c_i m(t_i)$. Under this constraint, the universal kriging looks for the optimisation of the mse

$$\mathbb{E}\left[\left(\sum_{i=1}^n c_i X_{t_i} - X_{t_0}\right)^2\right]$$

Since $M^\top\Sigma^{-1}M$ is regular, being M the nxp matrix with rows $m(t_i)$, its inverse $(M^\top\Sigma^{-1}M)^{-1}$ exists and, hence, the optimal linear coefficients are the components of the vector

$$c = \Sigma^{-1}\left[K + M(M^\top\Sigma^{-1}M)^{-1}(m(t_0) - M^\top\Sigma^{-1}K)\right]$$

and its mse is:

$$\rho(t_0, t_0) - K^\top\Sigma^{-1}K + (m(t_0) - M^\top\Sigma^{-1}K)^\top(M^\top\Sigma^{-1}M)^{-1}(m(t_0) - M^\top\Sigma^{-1}K)$$

It is important to mention that the covariance matrix Σ has to be of our knowledge, and also that, due to the fact that the field is neither weakly nor instrinsically stationary as the mean is not constant, the empirical semi-variogram cannot be used for the estimation. A solution of our interest is to estimate β in terms of least square, because we would need to know β if we approach this issue via the residual process $(E_t)_{t\in\mathbb{R}^d}$ instead. Having said that, we write:

$$Z = M\beta + E$$

where Z is again the sample X_{t_i} and the rows of M are the $m(t_i)^\top$ and E is the vector of residuals. Wanting to minimise

$$\sum_{i=1}^{n}(X_{t_i} - m(t_i)^\top\beta)^2 = (Z - M\beta)^\top(Z - M\beta)$$

over β, its gradient with respect to $\partial\beta$ is $-2M^\top(Z - M\beta)$, then equalising this to zero we obtain that

$$\hat{\beta} = \frac{M^\top Z}{M^\top M}$$

The vector $Z - M\hat{\beta}$ has a constant mean of zero when residuals provided also have mean zero, and its covariance might be estimated by using its empirical semi-variogram. This aproximation, though, may incur bias.

For more information about maximum likelihood methods and applications, consult chapter 14 of the book [42].

Chapter 2

Models and Inference for Areal Unit Data

In this chapter, three topics are going to be studied: discrete random fields, Gaussian autoregression models and Markov random fields. We use P. Whittle's work on Gaussian autorregression [35] and H. Rue [20]; and D. Brook's for Besag's factorization theorem [10], [22]. Up next, as for Markov random fields, work by P. Dobrushin is traced [40]. G. Grimmett paper is revised for Gibbs states. We present as well the Metropolis-Hasting algorithm with the form presented by Geyer and Thompson [7], having revised before Hastings's book about Monte Carlo sampling methods using Markov chains [43]. A brief introduction to Markov chains, following the textbook written by Meyn and Tweedie [41] is also provided.

2.1 Discrete Random Fields

In this section, random fields are going to count on a discrete index set this time. This allows us to use observations that have been gathered over areal units, for instance: census territory boundaries, tomographic bins or simply squares.

Definition 2.1. *Let $T \neq \emptyset$ be a finite collection of 'sites'. A random field X on L is a random vector $(X_i)_{i \in T}$ having L-valued components. If L is finite or countably infinite, the distribution of X is specified by the probability mass function*

$$\pi_X(x) = P(X = x) = P(X_i = x_i, i \in T), \quad x \in L^T.$$

Otherwise, $L \subseteq \mathbb{R}$ and X absolutely continuous with joint probability density π_X.

To have a better glimpse of the inference for areal unit data that is being revised in this chapter, let's have a look at the Ising Model and CAR models. As you will understand why in just a second, the Ising model was designed for studying ferromagnetism in statistical mechanics where the discrete variables show the dipole atomic moments of atomic spins, that can be either 1 or -1. CAR models are mainly used to have a description of spatial variations of explanatory variables (either latent variables or spatially varying random effects) and spatial relations among the data, as well as finding 'hot spots' or clusters.

Definition 2.2. *Let's suppose we have data records of a phenomenon of interest that when it is observed in a region represented by $i \in T$, the data record equals to 1, and 0 otherwise. Thus, our $L = \{0, 1\}$. Denoting $i \sim j$ when the i and j regions are adjacent. The Ising Model is defined by the probability mass function*

$$\pi_X(x) \propto exp\left[\alpha \sum_{i \in T} x_i + \beta \sum_{\{i,j\}:i\sim j} x_i x_j\right], \quad x \in L^T$$

where $\alpha, \beta \in \mathbb{R}$.

The α impacts the prevalence. When $\beta = 0$, the phenomenon of interest in each region is observed with probability $\frac{\exp\alpha}{1+\exp\alpha}$, independently of the rest of the regions. When $\beta > 0$ presence in a given region encourages presence in regions around, and when $\beta < 0$ presence in a given region discourages presence in regions around. This model is also used in statistical physics to study magnetisation, setting $L = \{-1, 1\}$.

Definition 2.3. *Let X be a random field, multivariate normally distributed, with mean zero, and covariance matrix $(I - B)^{-1}K$, where $K = \sigma^2 I$, being $I = I_n$ the identity matrix of size n, $\sigma^2 > 0$ unknown, $(I - B)$ non-singular, and $(I - B)^{-1}K$ symmetric and positive definite. We assume $b_{ii} = 0$. B is usually a sparse matrix (this is, most elements of the matrix are zero), for instance, $B = \phi W$, where ϕ is an unknown parameter, and $W = (w_{ij})$ a known "neighbourhood" matrix that is: nonnegative ($w_{ij} \geqslant 0$), symmetric, and $w_{ij} > 0 \Leftrightarrow i \sim j$, this is, i and j are neighbours, (otherwise $w_{ij} = 0$). We assure that $b_{ii} = 0$ assuming that the relation $\sim$ is non-reflexive (this is $i \not\sim i$, $\forall i \in T$. The matrix $(I - B)^{-1}K$ is said to follow a conditional autorregression (CAR) model.*

Note: the matrix W is also called adjacency matrix. The Gerschgorin disc theorem is useful to check that the matrix is positive definite.

Definition 2.4. *Let $T \neq \emptyset$ be a finite collection of sites. The local characteristics of a random field X on T with values in L are, whenever well-defined,*

$$\pi_i(x_i|x_{T\setminus\{i\}}), \quad i \in T, x \in L^T.$$

The local characteristics for the CAR model are Gaussian distributions with

$$\begin{cases} \mathbb{E}(X_i|X_j, j \neq i) = \sum_{j\neq i} b_{ij}X_j \\ Var(X_i|X_j, j \neq i) = \kappa_i \end{cases}$$

This explains why it is called 'conditional autoregression' model.

Considering the Ising model, we have

$$log\left[\frac{\pi_i(1|x_{T\setminus\{i\}})}{\pi_i(0|x_{T\setminus\{i\}})}\right] = \alpha + \beta\sum_{j\sim i} x_j$$

so, it is also known as first-order auto-logistic regression. Now,

$$\left[\frac{\pi_i(1|x_{T\setminus\{i\}})}{\pi_i(0|x_{T\setminus\{i\}})}\right] = \exp\left\{\alpha + \beta\sum_{j\sim i} x_j\right\};$$

$$\left[\frac{\pi_i(1|x_{T\setminus\{i\}})}{1-\pi_i(1|x_{T\setminus\{i\}})}\right] = \exp\left\{\alpha + \beta\sum_{j\sim i} x_j\right\};$$

$$\pi_i(1|x_{T\setminus\{i\}}) = \exp\left\{\alpha + \beta\sum_{j\sim i} x_j\right\} - \pi_i(1|x_{T\setminus\{i\}})\exp\left\{\alpha + \beta\sum_{j\sim i} x_j\right\};$$

$$\pi_i(1|x_{T\setminus\{i\}})\left(1 + exp\left[\alpha + \beta\sum_{j\sim i} x_j\right]\right) = exp\left[\alpha + \beta\sum_{j\sim i} x_j\right];$$

$$\pi_i(1|x_{T\setminus\{i\}}) = \frac{exp\left[\alpha + \beta\sum_{j\sim i} x_j\right]}{1 + exp\left[\alpha + \beta\sum_{j\sim i} x_j\right]}$$

Note that $\pi_X(\cdot|x_{T\setminus\{i\}})$ only depends on x_j, which are regions indexed by neighbours of i. Let's prove now that the local characteristics determine the distribution, given strictly positive distributions.

Theorem 2.5 (Besag's factorisation theorem - Brook's lemma)**.** *Let X be an L-valued random field on $T = \{1, \ldots, N\}, N \in \mathbb{N}$, such that $\pi_X(x) > 0$, $\forall x \in L^T$. Then, $\forall x, y \in L^T$,*

$$\frac{\pi_X(x)}{\pi_X(y)} = \prod_{i=1}^{N} \frac{\pi_i(x_i|x_1, \ldots, x_{i-1}, y_{i+1}, \ldots, y_N)}{\pi_i(y_i|x_1, \ldots, x_{i-1}, y_{i+1}, \ldots, y_N)}$$

Proof. Since we assume by hypothesis that $\pi_X(x)$ is strictly positive, the denominator is non-zero. We may write $\pi_X(x) = P(x_n|x_1, \ldots, x_{x_{n-1}})P(x_1, \ldots, x_{n-1})$ but $P(x_1, \ldots, x_{n-1})$cannot be usefully factorised because we might not been able to find $P(x_{n-1}|x_1, \ldots, x_{x_{n-2}})$ with the conditional distributions given. However, we can introduce y_n and get

$$\pi_X(x) = \frac{\pi_N(x_N|x_1, \ldots, x_{N-1})}{\pi_N(y_N|x_1, \ldots, x_{N-1})}\pi_X(x_1, \ldots, x_{N-1}, y_N).$$

Operating on x_{n-1} now,

$$\pi_X(x_1, \ldots, x_{N-1}, y_N) = \frac{\pi_{N-1}(x_{N-1}|x_1, \ldots, x_{N-2}, y_N)}{\pi_{N-1}(y_{N-1}|x_1, \ldots, x_{N-2}, y_N)}\pi_X(x_1, \ldots, x_{N-2}, y_{N-1}, y_N).$$

after an analogous introduction of y_{n-1}. We get to the desired equation by repeating the reduction process. QED

This theorem proves that the local characteristics determine the entire distribution for strictly positive distributions.

Corollary 2.6. *Let X be an L-valued random field on a finite collection $T \neq \emptyset$ of sites such that $\pi_X(x) > 0, \forall x \in L^T$. Then the local characteristics determine the whole distribution, that is, if Y is a random field having the same local characteristics as X, it only can be $\pi_Y \equiv \pi_X$.*

Proof. Let $a \in L$ be any element. Using Besag's factorisation theorem, $\frac{\pi_X(x)}{\pi_X(a,\ldots,a)}$ is determined by the local characteristics. Finally, we obtain the distribution by normalisation. QED

2.2 Gaussian Autoregression Models

Let's briefly define again the CAR model before presenting the SAR model:

Definition 2.7. *Let $K = diag(\kappa_i)$ be a diagonal $n \times n$ matrix with $\kappa_i > 0$ and B an NxN matrix whose diagonal elements $b_{ii} = 0$. Then, provided $I - B$ is invertible and $(I - B)^{-1}K$ is positive definite, a random field X that is normally distributed with zero-mean and covariance matrix $(I - B)^{-1}K$ is said to follow a conditional autoregression (CAR) model.*

Using the same notation from last definition and considering the normally-distributed matrix $E = (I - B)X$, with mean zero and covariance matrix

$$(I - B)(I - B)^{-1}K(I - B)^{\top} = K(I - B)^{\top},$$

hence, we have an autoregression $X = BX + E$ (the field E might be spatially correlated, though, and this is the random irregularity or randomness that can be found within a given data). In order to assume that this unexplained variation is independent, we let E be normally distributed with diagonal covariance matrix $D = diag(\lambda_i) = (\lambda_i)_i$, with $\lambda_i > 0$ for $i = 1 \div n$. Hence, $(I - B)^{-1}D(I - B^\top)^{-1}$ is the covariance matrix of $X = (I - B)^{-1}E$.

Definition 2.8. *Using the previous notation, the random field $X = (I - B)^{-1}E$ is a simultaneous autoregression (SAR) model.*

Proposition 2.9. *Any SAR model can be written as a CAR model.*

Proof. Let B be an $n \times n$ matrix such that $I - B$ is non-singular and $b_{ii} = 0, \ \forall i = 1 \div n$, D an $n \times n$ positive definite diagonal matrix. Thus, $(I - B)^{-1}D(I - B^\top)^{-1}$ is well-defined, positive definite and symmetric. Now, let's solve

$$(I - B)^{-1}D(I - B^\top)^{-1} = (I - M)^{-1}K$$

for $K = diag(\kappa_i)$ and M. This is the same as

$$(I - B^\top)D^{-1}(I - B) = K^{-1}(I - M).$$

Finally, we set $c_{ii} = 0$ and we only have to find the scale factors κ_i. Writing γ_i for the i-th element of the diagonal of D, $\forall i = 1 \div n$,

$$\frac{1}{\gamma_i} + \sum_{j=1}^{n} \frac{b_{ji}^2}{\gamma_i} = \frac{1}{\kappa_i}$$

and, thus, $\kappa_i > 0$.

QED

Let's now see the notion of interaction:

Definition 2.10. *Let $T \neq 0$ be a finite collection of sites, $L \subseteq \mathbb{R}$. An interaction potential A_n is a collection $\{V_A : A \subseteq T\}$ of functions $V_A : L^T \to \mathbb{R}$ such that $V_\emptyset(\cdot) \equiv 0$ and $V_A(x)$ only depends on the restriction x_A of $x \in L^T$ to sites in A. The interaction potential V is said to be normalised with respect to $a \in L$ if the property that $X_i = a$ for some $i \in A$ implies that $V_A(x) = 0$.*

A Gibbs state is a random field whose distribution is defined in terms of interaction potentials:

Definition 2.11. *Let X be an L-valued random field on a finite collection $T \neq \emptyset$ of sites and V an interaction potential. Then X is a Gibbs state with interaction potentials $V = \{V_A : A \subseteq T\}$, $V_A : L^T \to \mathbb{R}$, if*

$$\pi_X(x) = \frac{1}{Z} exp \left[\sum_{A \subseteq T} V_A(x_A) \right], \quad x \in L^T.$$

The constant Z is called partition function and is quite hard to deal with.

Now let's see the *Möbius inversion formula* that we will use to prove a theorem after.

Theorem 2.12 (Möbius inversion formula)**.** *Let T be a finite set and $f, g \to \mathbb{R}$ two functions defined on the power set of T, $P(T)$. Then $\forall A \subseteq T$,*

$$f(A) = \sum_{B \subseteq A} g(B) \iff g(A) = \sum_{B \subseteq A} (-1)^{|A \backslash B|} f(B).$$

This result helps to prove the previous theorem. In particular, it says that there is only one form to represent a function f like $f(A) = \sum_{B \subseteq A} g(B)$.

Proof. "$\Rightarrow$" Let g be fixed and $f(A) = \sum_{B \subseteq A} g(B)$ for $A \subseteq T$.

$$\sum_{B \subseteq A} (-1)^{|A \backslash B|} f(B) = \sum_{B \subseteq A} (-1)^{|A \backslash B|} \left[\sum_{C \subseteq B} g(C) \right] = \sum_{B \subseteq A} \left[\sum_{C \subseteq B} (-1)^{|A \backslash C|} (-1)^{-|B \backslash C|} g(C) \right]$$

$$= \sum_{C \subseteq A} \left[\sum_{B: C \subseteq B \subseteq A} (-1)^{|B \backslash C|} \right] (-1)^{|A \backslash C|} g(C) = g(A),$$

where for the last equality we take into consideration the fact that, unless $A = C$,

$$\sum_{k=0}^{|A \backslash C|} \binom{|A \backslash C|}{k} (-1)^k = 0.$$

"$\Leftarrow$" Let f be fixed and $g(A) = \sum_{B \subseteq A} (-1)^{|A \backslash B|} f(B)$. Now, analogously,

$$\sum_{B \subseteq A} g(B) = \sum_{B \subseteq A} \left[\sum_{C \subseteq B} (-1)^{|B \backslash C|} f(C) \right]$$

$$= \sum_{C \subseteq A} \left[\sum_{B: C \subseteq B \subseteq A} (-1)^{|B \backslash C|} \right] f(C) = f(A).$$

QED

Theorem 2.13. *Let X be an L-valued random field on a finite collection $T \neq \emptyset$ of sites such that $\pi_X(x) > 0, \forall x \in L^T$. Then, X is a Gibbs state with respect to the canonical potential*

$$V_A(x) = \sum_{B \subseteq A} (-1)^{|A \backslash B|} \log \pi_X(x^B), \quad x \in L^T$$

where

$$x_i^B = \begin{cases} x_i & for \quad i \in B \\ \\ a \in L & otherwise \end{cases}$$

and a is a prefixed value. This is the unique normalised potential with respect to a. Moreover, for any element $i \in A$,

$$V_A(x) = \sum_{B \subseteq A} (-1)^{|A \backslash B|} \log \pi_i(x_i^B | x_{T \backslash i}^B), \quad x \in L^T.$$

Proof. Let $x \in L^T$. We can set $f_x(A) = \log \pi_X(x^A)$, being $A \subseteq T$, since $\pi_X(x) > 0$. We define now the interaction potential by: $V_A : L^T \to \mathbb{R}$,

$$V_A(x) = \sum_{B \subseteq A} (-1)^{|A \backslash B|} f_x(B), \quad x \in L^T.$$

Using the Möbius inversion formula we have

$$\pi_X(x) = exp\left[f_x(T)\right] = exp\left[\sum_{A \subseteq T} V_A(x)\right].$$

Now, unfixing x, we obtain that π_X is a Gibbs state, whose interaction potential is $\{V_A : A \subseteq T\}$. Let's see now that this interaction potential is normalised. For any $i \in A$:

$$\begin{aligned} V_A(x) &= \sum_{i \notin B \subseteq A} (-1)^{|A \backslash B|} \log \pi_X(x^B) + \sum_{i \in B \subseteq A} (-1)^{|A \backslash B|} \log \pi_X(x^B) \\ &= \sum_{B \subseteq A \backslash \{i\}} (-1)^{|A \backslash B|} \log \pi_X(x^B) - \sum_{B \subseteq A \backslash \{i\}} (-1)^{|A \backslash B|} \log \pi_X(x^{B \cup \{i\}}) \\ &= \sum_{B \subseteq A \backslash \{i\}} (-1)^{|A \backslash B|} \left[\log \pi_X(x^B) - \log \pi_X(x^{B \cup \{i\}})\right]. \end{aligned}$$

If $x_i = a \Rightarrow x^B = x^{B \cup \{i\}}$, $\forall B \subseteq A \backslash \{i\} \Rightarrow V_A = 0$. Hence, the interaction potential is normalised with respect to a, and $x_{T \backslash \{i\}}^B = x_{T \backslash \{i\}}^{B \cup \{i\}}$, $\forall B \subseteq A \backslash \{i\}$. Thus

$$\frac{\pi_X(x^B)}{\pi_X(x^{B \cup \{i\}})} = \frac{\pi_i(x_i^B \mid x_{T \backslash \{i\}}^B)}{\pi_i(x_i^{B \cup \{i\}} \mid x_{T \backslash \{i\}}^{B \cup \{i\}})}$$

Now we are going to see that $V_A \equiv U_A$, supposing π_X is a Gibbs state with respect to normalised potentials U_A. Let's fix $x \in L^T$ and note a_T for the realisation with only a-labels. We define the set function $h_x(A)$:

$$h_x(A) = log\frac{\pi_X(x^A)}{\pi_X(a_T)} = \sum_{B \subseteq A} [U_B(x) - U_B(a_T)] = \sum_{B \subseteq A} U_B(x),$$

assuming in the last equation that the interaction potential U is normalised. For all $A \neq \emptyset$, using the Möbius inversion formula we have

$$U_A(x) = \sum_{B \subseteq A} (-1)^{|A \backslash B|} h_x(B) = V_A(x) - \log \pi_X(a_T) \sum_{B \subseteq A} (-1)^{|A \backslash B|} = V_A(x).$$

Since it equals zero when $A = \emptyset$, we have finished. QED

2.3 Markov Random Fields

In this section we are going to consider a set of series, T, that has the property of presenting a symmetric relation $\sim$. Now the interaction potentials $V_A(x)$ are no longer present except for when A is a singleton or a pair i, j of $\sim$-related sites. Now, the conditional distribution of the label at site i, given those at all other sides, depends only on the labels of site i.

Definition 2.14. *Let $\sim$ be a symmetric relation on the finite set $T \neq \emptyset$ and define the boundary of $A \subseteq T$ by $\partial A = \{s \in T \backslash A : s \sim t, t \in A\}$. A random field X on T is a Markov random field with respect to $\sim$ if*

$$\pi_i(x_i | x_{T \backslash i}) = \pi_X(X_i = x_i | X_{\partial i} = x_{\partial i})$$

Definition 2.15. *Let $T \neq \emptyset$ be a finite collection of sites. Let $\sim$ be a symmetric relation on T. A clique with respect to $\sim$ is a subset $C \subset T$, for which $s \sim t, \forall s \neq t \in C$. The family of cliques is denoted $\mathcal{C}$.*

Theorem 2.16 (Hammersley-Clifford). *Let X be an L-valued random field on a finite collection $T \neq \emptyset$ of sites such that $\pi_X(x) > 0$ for all $x \in L^T$. Let $\sim$ be a symmetric relation on T. Then X is a Markov random field with respect to $\sim$ if and only if*

$$\pi_X(x) = \prod_{C \in \mathcal{C}} \varphi_C(x_C)$$

for some interaction functions $\varphi_C : L^C \rightarrow \mathbb{R}^+$ defined on cliques $C \in \mathcal{C}$.

This means that when given a Markov random field whose π_X is positive, it has a distribution that can be expressed by the interactions between neighbours, and it can be written as

$$\pi_X(x) = exp\left[\sum_{C \in \mathcal{C}} log\varphi_C(x_C)\right]$$

and X is a Gibbs state with interactions potentials $log\varphi_C$ that are not null and restricted to cliques.

Proof. "$\Rightarrow$" Suppose that X is a Markov random field with $\pi_X > 0$. Using the theorem 2.12 we have that X is a Gibbs state with canonical potential

$$V_A(x) = \sum_{B \subseteq A} (-1)^{|A \setminus B|} \log \pi_i(x_i^B | x_{T\setminus i}^B), \quad x \in L^T.$$

We claim that $V_A(x) = 0$ for all $A \notin \mathcal{C}$. If $A \subseteq T$ is not a clique, there are two sites, $s, t \in A, \quad s \neq t$, with $s \not\sim t$. Hence we have,

$$V_A(x) = \sum_{B \subseteq A} (-1)^{|A \setminus B|} \log \pi_s(x_s^B | x_{T\setminus s}^B), \quad x \in L^T.$$

We rewrite this as

$$\begin{aligned} &\sum_{B \subseteq A\setminus\{s,t\}} (-1)^{|A \setminus B|} \log \pi_s(x_s^B | x_{T\setminus s}^B) \\ &+ \sum_{B \subseteq A\setminus\{s,t\}} (-1)^{|A \setminus (B\cup\{s\})|} \log \pi_s(x_s^{B\cup\{s\}} | x_{T\setminus s}^{B\cup\{s\}}) \\ &+ \sum_{B \subseteq A\setminus\{s,t\}} (-1)^{|A \setminus (B\cup\{t\})|} \log \pi_s(x_s^{B\cup\{t\}} | x_{T\setminus s}^{B\cup\{t\}}) \\ &+ \sum_{B \subseteq A\setminus\{s,t\}} (-1)^{|A \setminus (B\cup\{s,t\})|} \log \pi_s(x_s^{B\cup\{s,t\}} | x_{T\setminus s}^{B\cup\{s,t\}}). \end{aligned}$$

Hence we have

$$V_A(x) = \sum_{B \subseteq A\setminus\{s,t\}} (-1)^{|A \setminus B|} log\left[\frac{\pi_s(x_s^B | x_{T\setminus s}^B)\pi_s(x_s^{B\cup\{s,t\}} | x_{T\setminus s}^{B\cup\{s,t\}})}{\pi_s(x_s^{B\cup\{t\}} | x_{T\setminus s}^{B\cup\{t\}})\pi_s(x_s^{B\cup\{s\}} | x_{T\setminus s}^{B\cup\{s\}})}\right]$$

but, since $s \not\sim t$, $\pi_s(x_s^B | x_{T\setminus s}^B) = \pi_s(x_s^{B\cup\{s,t\}} | x_{T\setminus s}^{B\cup\{s,t\}})$ and $\pi_s(x_s^{B\cup\{t\}} | x_{T\setminus s}^{B\cup\{t\}}) = \pi_s(x_s^{B\cup\{s\}} | x_{T\setminus s}^{B\cup\{s\}})$, we obtain that $V_A(x) = 0$. Thus, $\pi_X(x) = \prod_{C \in \mathcal{C}} \varphi_C(x_C)$ holds since the only non-null interaction potentials are for cliques.

"$\Leftarrow$" Let's prove that a distribution like $\pi_X(x) = \prod_{C \in \mathcal{C}} \varphi_C(x_C)$ presents the Markov property. Suppose L countable and denote $T_i^a x$ for the configuration, which X_i is replaced by a. Then we have

$$\pi_i(x_i | x_{T\setminus\{i\}}) = \frac{\prod_{i \in C} \varphi_C(x_C)}{\sum_{a \in L} \left[\prod_{i \in C} \varphi_C(T_i^a x_C)\right]}$$

where the side on the right only depends on x_i and $x_{\partial i}$. Note that for the absolutely continuous case, the sum is an integral. QED

Corollary 2.17. *Let X be an L-valued random field on a finite collection $T \neq \emptyset$ of sites such that $\pi_X(x) > 0, \forall x \in L^T$. Then, the spatial Markov property*

$$\pi(X_A = x_A | X_{T\setminus A} = x_{T\setminus A}) = \pi(X_A = x_A | X_{\partial A} = x_{\partial A})$$

holds $\forall A \subseteq T$, being A nonempty set.

Proof. Using the Hammersley-Clifford theorem,

$$\pi(X_A = x_A | X_{T\setminus A} = x_{T\setminus A}) = \frac{\prod_{A \bigcap C \neq \emptyset} \varphi_C(x_C)}{\sum_{y \in L^A} \prod_{A \bigcap C \neq \emptyset} \varphi_C((T_A^y)_C)}$$

where we have replaced x_A by $y(x_A, y \in L^A)$ on the set $A \subseteq T$. QED

The Markov chain is a discreet stochastic process where the probability that an events happens only depends on the previous event happened, no matter other passed events. This helps us to use the Monte Carlo maximum likelihood estimation method, since this needs samples from the model of interest. But before talking about it, let's define the Markov chain:

Definition 2.18. *A Markov chain is a stochastic process of discreet time $\{X_n : n = 0, 1, \dots\}$ with space of discreet states S where, $\forall n \geqslant 0$ and $\forall x_0, \dots, x_{n+1} \in S$ satisfies*

$$P(X_{n+1} = x_{n+1} | X_0 = x_0, X_1 = x_1, \dots, X_n = x_n) = P(X_{n+1} = x_{n+1} | X_n = x_n).$$

This is called the Markov's property.

In other words, it is an infinite sequence $x_1, x_2, \dots, x_k, x_{k+1}, \dots$ of connected variables such that when x_k is known, for all k, x_{k+1} is independent of $x_1, x_2, \dots, x_{k-1}$. The following definition is more flexible and used because it gives enough conditions for convergence:

Definition 2.19. *Given a sequence $M_0, M_1, \ldots$ of random variables. We say it is a Markov chain with transition kernel* [1] *$p(\cdot,\cdot)$ if*

$$P(M_t \in A_t; M_{t-1} \in A_{t-1}; \cdots ; M_1 \in A_1 | M_0 = m_0)$$

$$= \int_{A_1} \cdots \int_{A_{t-1}} \int_{A_t} p(m_0, m_1) \ldots p(m_{t-2}, m_{t-1}) p(m_{t-1}, m_t) dm_1 \ldots dm_{t-1} dm_t$$

$\forall t \in \mathbb{N}$ and all measurable $A_i \subseteq \Omega$, $i = 1 \div t$. The fixed starting state m_0 can be replaced by any probability distribution on Ω. Of course, for countable state spaces the integral is changed for a sum.

We might find tricky to work with the joint distribution π_X for random fields X, but we can use the local characteristics $\pi_i(x_i | x_{T \setminus \{i\}})$ since that will ease the work. Having said that, we are going to implement the Metropolis-Hastings algorithm to define transitions by making the components X_i one at a time. The Metropolis-Hasting algorithm is a Monte Carlo method of Markov's chain used to obtain a sequence of random samples given a probability distribution that is difficult to do a direct sampling. It is also used for numerical integration, and mainly for multidimensional distributions with a high number of dimensions. Finally, we'll see that the concept of periodicity can also be defined for Markov chains.

Definition 2.20. *The Metropolis-Hastings algorithm has the following steps:*

1. *The current state is $M_t = x \in L^T$. To sample a site $i \in T$ and, given some probability density $q(x, y)$ a new label $l \in L$ to yield state*

$$y = (y_j)_{j \in T} = \begin{cases} y_j = x_j for \quad j \neq i \\ \\ y_j = l for \quad j = i \end{cases}$$

2. *Accept the proposal that has probability*

$$A(x,y) = \begin{cases} 1 \quad if \quad \pi_X(y) q(y,x) \geq \pi_X(x) q(x,y) \\ \\ \frac{\pi_X(y) q(y,x)}{\pi_X(x) q(x,y)} \quad otherwise \end{cases}$$

[1] It is a way of moving randomly from a given position to a new position in space.

The $A(x,y)$ are called the acceptance probabilities. Notice that the acceptance probabilities only depend on

$$\frac{\pi_X(y)}{\pi_X(x)} = \frac{\pi_i(l \mid x_{T\setminus\{i\}})}{\pi_i(x_i \mid x_{T\setminus\{i\}})}.$$

The transition kernel can be obtained by combining both proposal and acceptance probabilities: $for\ x \neq y,\ p(x,y) = q(x,y)A(x,y).$

Proposition 2.21. *Let X be an L-valued random field on a finite collection $T \neq \emptyset$ of sites. Then the Metropolis-Hastings algorithm satisfies the following properties:*

i. so-called 'detailed balance'

$$\pi_X(x)p(x,y) = \pi_X(y)p(x,y)$$

ii. π_X is an invariant measure, this is, $\forall$ measurable $A \subseteq L^T$,

$$\pi_X(X \in A) = \int P(M_1 \in A | M_0 = x)\pi_X(x)dx$$

when $L = \mathbb{R}$, and when L is countable:

$$\pi_X(X \in A) = \sum_x P(M_1 \in A | M_0 = x)\pi_X(x)$$

Proof. Suppose $x \neq y$ and $\pi_X(x)q(x,y) < \pi_X(y)q(y,x)$. Then

$$\pi_X(x)p(x,y) = \pi_X(x)q(x,y) = \frac{\pi_X(x)q(x,y)}{\pi_X(y)q(y,x)}\pi_X(y)q(y,x)$$

$$= A(y,x)\pi_X(y)q(y,x) = \pi_X(y)p(y,x)$$

This property ('detailed balance') implies the invariance:

$$\int P(M_1 \in A | M_0 = x)\pi_X(x)dx = \int \left(\int_A p(x,y)\pi_X(x)dy \right) dx$$

$$= \int \left(\int_A p(y,x)\pi_X(y)dy \right) dx = \int_A \pi_X(y) \left(\int p(y,x)dx \right) dy = \int_A \pi_X(y)dy.$$

QED

Definition 2.22. *A Markov chain $(M_t)_{t\in\mathbb{N}_0}$ on a countable state space Ω is irreductible if $\forall x,y \in \Omega$ there exists some $t \in \mathbb{N}$ such that $P(M_t = y | M_0 = x) > 0$. A Markov chain $(M_t)_{t\in\mathbb{N}_0}$ with state space $\Omega = \mathbb{R}^T$ is π_X-irreductible if $\forall x \in \Omega$ and all Borel sets $A \subset \mathbb{R}^T$ for which $\pi_X(A) > 0$ there exists some $t \in \mathbb{N}$ such that $P(M_t = y | M_0 = x) > 0$.*

Note: due to the fact that the probability of returning a single state will be zero when $\Omega = \mathbb{R}^T$, the restriction to sets with positive probability is necessary.

Definition 2.23. *A π_X-irreductible Markov chain $(M_t)_{t\in\mathbb{N}_0}$ is aperiodic if there is no partition into non-empty measurable sets $B_0, \dots, B_{r-1}$, $r \geqslant 2$, such that $\forall t \in \mathbb{N}$,*

$$P(M_t \in B_{t\, mod\, r} | M_0 = x \in B_0) = 1$$

and the union of $B_0, \cdots, B_{r-1}$ has π_X-mass one.

Theorem 2.24 (Fundamental convergence theorem)**.** *If π_X is an invariant probability measure for a Markov chain $(M_t)_{t\in\mathbb{N}_0}$ that is π_X-irreductible and aperiodic, then M_t converges to π_X in total variation from π_X-almost all initial states, this is,*

$$\lim_{t\to\infty} \sup_A |P(M_t \in A | M_0 = x) - \pi_X(A)| = 0$$

for π_X-almost all x. The supremum is taken over all measurable sets.

Theorem 2.25. *Let X be an L-valued random field on a finite collection $T \neq \emptyset$ of sites and $(M_t)_{t\in\mathbb{N}_0}$ a Metropolis-Hastings chain on $D_\pi = \{x \in L^T : \pi_X(x) > 0\}$. If the Markov chain governed by q is π_X-irreductible and $q(x,y) = 0 \Leftrightarrow q(y,x) = 0$, then $(M_t)_{t\in\mathbb{N}_0}$ is π_X-irreductible.*

Proof. First, the condition of $q(x,y) = 0$ when $q(y,x) = 0$ implies that the acceptance probabilities on D_π are strictly positive. Denoting the Markov chain by $(Q_t)_{t\in\mathbb{N}_0}$, ruled by the $q(x,y)$, q^t its t-step transition kernel and p^t the t-step transition kernel of M_t. Let's see by induction that $q^t(x,y) > 0 \Rightarrow p^t(x,y) > 0$. For $t = 1$, suppose $q(x,y) > 0$, $x, y \in D_\pi$. Since $A(x,y) > 0$, $p(x,y) \geq q(x,y)A(x,y) > 0$. Now, for the step $t+1$, suppose $q^{t+1}(x,z) > 0$ for $x, z \in D_\pi$, let's denote $S_p^t(x)$ for the support of $p^t(x,\cdot)$ and $S_q^t(x)$ for the support of $q^t(x,\cdot)$, and suppose that $S_q^t(x) \subseteq S_p^t(x)$. Since by assumption $z \in S_q^{t+1}(x)$,

$$\int_{S_p^t(x)} q^t(x,y)q(y,z)dy \geq \int_{S_q^t(x)} q^t(x,y)q(y,z)dy > 0.$$

If $z \notin S_p^{t+1}(x)$, then the support of the function $y \mapsto p^t(x,y)q(y,z)$ would be $\emptyset$. The support of $q^t(x,\cdot)q(\cdot,z)$ would have measure zero by the induction assumption, which contradicts the inequality. To finish, given an A with positive π_X-mass, since q is π_X-irreductible, a $t \geq 1$ such that

$$P(Q_t \in A | Q_0 = x) = \int_A q^t(x,y)dy > 0$$

can be found. We obtain, hence, that $P(M_t \in A | M_0 = x) > 0$. QED

Chapter 3

Methods for the Analysis of Migration Flows

In this chapter, gravity and radiation models are presented and studied. There is a large amount of literature available. Regarding gravity models, Head and Mayer's paper [24] has been revised for a better glimpse of their applications in International Economics, as well as its economic foundations presented by Colwell [9] and J. Bergstrand [5]. As for radiation models, recent papers by Inho Hong et al [21] as well as C. Kang et al, [23] are traced. The first one provides a good perspective regarding using the models with population data. We recommend revising papers [11], written by Filippo Simini et al, for an elaborated study on migration patterns. Literature regarding a comparison with radiation models is followed through the paper written by A. Amini et al [2]. Finally, deep gravity model is explained. The literature used for this section is mainly provided by its authors F. Simini, G. Barlacchi, M. Luca, L. Pappalardo [13].

3.1 What is a gravity model?

Gravity models are models used to forecast and study certain conducts in social sciences, based on Isaac Newton's law of gravity[1]. They were introduced in economics in the twentieth century by Willian J. Reilly (1931) with the Reilly's law of retail gravitation and George K. Zipf (1946) with his Zipf's law, but it is Tinbergen (1962), Poyhonen (1963), Pulliainen (1963), Geraci and Prewo (1977)

[1] *Every particle attracts every other particle in the universe with a force that is directly proportional to the product of their masses and inversely proportional to the square of the distance between their centers [14].*

and Abrams (1980) who estimate gravity equations in economics, even though some authors reference Carey (1858) and Ravenstein (1885). For more information about it, consult [5] and [29].

Similar to the gravitational interaction of two planets, gravity models use the distance between two or more countries and their gross domestic product (GDP), population, certain type of goods and services or any other apropiate variable to try to estimate migration and trade flows, access to certain services (such as health care), foreign direct investment, alliances on trade or traffic flow, among other behaviours. The applications are endless to city planners, transportation analysts, location firms, social scientists and many others to analyse migration, commuting, vacationing, shopping, collecting and distributing [24][25].

The main most basic model is the following:

$$F_{ij} = G \cdot \frac{M_i \cdot M_j}{D_{ij}}$$

where G is a constant, M_i and M_j are the economic dimensions of each country studied, D_{ij} the distance between the two and F_{ij} is the forecast. Intuitively, it is fair to think that this model makes sense when considering the fact that the bigger the distance, the bigger the costs related to the trade, migration or commuting and thus the smaller the flow.

Given that the gravity model does not hold precisely, it is common to express

$$F_{ij} = G \cdot \frac{M_i^{\beta_1} \cdot M_j^{\beta_2}}{D_{ij}^{\beta_3}} \cdot \eta_{ij}$$

where η represents the error term, expected to be 1. β_1 is called *emissivity* and it is the potential to make the movement happen, oftentimes related to an overall welfare. β_2 is called *atractiveness* and it is the potential to attract movements. β_3 is a parameter of *transport friction* that justifies the efficiency of the transport between the two locations. Since the friction of distance is greater when the destination is further, the friction is rarely linear, and the justification is analogous for the other two betas [9]. For instance: if there is a highway between locations i and j or any improvement in the transport infrastructure then β_3 will be lower compared to the case of having an uncomfortable road; β_1 will be higher if the rate of unemployment is high; β_2 will be higher if the salaries are higher in that location or it is more economically active.

The conventional method to approach this model is by using logarithms on both sides, so we get a log-log model:

$$\ln F_{ij} = \beta_0 + \beta_1 \cdot \ln M_i + \beta_2 \cdot \ln M_j - \beta_3 \cdot \ln D_{ij} + \epsilon_{ij}$$

where β_0 is $ln\ G$ and ϵ_{ij}, the error term, is $ln\ \eta_{ij}$.

A *logistic regression* is the estimation of the parameters of a logistic model. A *logistic model*, also called *logit model*, models the probability of an event to happen, considering a linear combination of independent variables. We talk about a *multinomial logistic regression* when we generalize the logistic regression to a multiclass problem, which is a problem with multiple possible discrete outcomes [42], so it can be used to calculate the probabilities of each possible outcome of a categorically distributed dependent variable. This process is also called *calibration*.

Focusing on migration flows from now on, these are not just determined by the population of both countries and the distance: there are other factors that influence them, such as the linguistic and cultural proximity, better opportunities in terms of employment and salary, safety, political freedom or even climate, among many others [37]. All these pull and push factors are represented in variables that can be added to the models.

3.2 What is a Radiation Model?

Physics first used the radiation model to understand the process by which waves and energetic particles move through vacuum: particles are released in a certain location and there is a probability p to be absorbed by locations nearby.

In the social science it is applied once again for the study of flows between different locations. Firstly, the traveler assigns a number to every destination according to their fitness, x, chosen from some distribution $p(x)$, representing the quality of the opportunity presented in that location. Secondly, the traveler rates again all location but this time considering the distance from the original location. Then, the closest location with the highest fitness rate and, of course, higher than the traveler's fitness threshold, taken as well from the fitness distribution $p(x)$ will be the chosen one [11]. Hence, T_{ij} the average number of travelers from location i to j follows this expression:

$$T_{ij} = T_i \cdot \frac{1}{1-\frac{m_i}{M}} \cdot \frac{m_i \cdot m_j}{(m_i + s_{ij}) \cdot (m_i + m_j + s_{ij})}$$

where $T_i = \sum_{j \neq i} T_{ij}$ are the departures from i, m_i and m_j are the number of opportunities (or population, as explained later) in i and j respectively, $M = \sum_i m_i$ is the total number of opportunities and s_{ij} is the total opportunities in the circle with center i and radius $|i - j|$ excluding the origin and the destination population. The destination of the O_i trips are sampled following a distribution of probabilities of a trip from i finishes in j. It is necessary to normalize this conditional probability to make the probability that a journey beginning in the area of interest and ending there is equal to one. When the finite system is given, this equals $1 - \frac{m_i}{M}$ [4]. In the original version of the model, the number of opportunities is estimated by the population in that location (the more population, the more opportunities might arise in that location), or can also be estimated by the total inflows.

In case a infinite system was given, the number of commuters would be the following:

$$T_{ij}^{\infty} = T_i \frac{m_i \cdot m_j}{(m_i + s_{ij}) \cdot (m_i + m_j + s_{ij})}$$

Note that $T_{ij} \rightarrow T_{ij}^{\infty}$ when $M \rightarrow \infty$. According to [4], T_{ij}^{∞} is a fit approximation for large systems.

Radiation model, the same way gravity models do, also have different versions regarding the information needed to run the model. Here [8] it is shown that better performing models do not even consider population but amenities, since amenities already provide the information we would take from population itself. It is worth having a look at papers [36], [27], [45], [23], [26], [2] to check how these models have been use to replicate observed changes in different cities' population and see that it is less effective in developing countries than in developed countries.

The same paper [8] proposes a generalized radiation model for human migration where the proxys are called urbanization indexes U, which are a weighted sum of component factors f_k:

$$U = \sum_k w_k f_k,$$

where f_k can be any feature of a locality. Then the model looks like this:

$$T_{ij}^{\infty} = T_i \frac{U_i \cdot U_j}{(U_i + v_{ij}) \cdot (U_i + U_j + v_{ij})},$$

where U_i and U_j are the urbanization index at i and j, respectively, and v_{ij} is the total urbanization index in the circle with ratio $|i - j|$ excluding source and destinations' indexes. Notice that due to the fact that f_k have different scales, the value of each feature must be normalized so they can be comparable to each other. This can be done either by using the *min-max* method, *adjusted z-score*, *logistic z-score* or *percentile*.

3.3 Deep Gravity Model

We finally arrive to our last model: deep gravity model. As you will see, this model relies in neural networks and it does not need historical data. It provides good results and makes us wonder to what extent are statistics sufficient for forecasting, now that machine learning and AI are promising fast-growing fields. This chapter is fully based on paper [13]. Nevertheless, since the study of spatial networks is out of our scope, we strongly recommend the reader to have a look at the paper [28], where the model is implemented and its results are compared to gravity models in the UK, Italy and New York State. Due to the complexity of the programming behind the model, this one is not implemented to forecasting migration flows in Catalonia, nonetheless is worth revising.

3.3.1 Fundamentals

The Deep Gravity model is a mobility flows generation model proposed by F. Simini, G. Barlacchi, M. Luca and L. Pappalardo. The main purpose of this model is to generate information about mobility flows in any specific region when there is no information available, setting up probabilities based on geographic data such as land use, road network, food, health facilities, education, retail facilities and transport, among others extracted from OpenStreetMap [2], finding non-linear correlations between those characteristics and mobility flows training deep neural networks. Compared to deep-learning approaches, the Deep Gravity Model does not rely on migration flow historical data.

[2] OpenStreetMap is an international project to create a free map of the world. It is a public and voluntary geographic information system. https://www.openstreetmap.org

Gravity models, due to its restricted set of variables, disregards variables that might be essential to forecast and explain a migration flow, while a deep gravity model takes into consideration more detailed data that responds to how diverse are the points of interest, the transportation network and overall how complex is the geographical landscape.

In the first chapter we mentioned that gravity models are a multinomial logistic regression equivalent to a linear neural network with one softmax layer. A neural network is a collection of nodes called neurons, whose interconnections are called edges. Each neuron can signal other neurons, by sending real numbers, and the output of each neuron is computed by some non-linear function of the sum of its inputs. Neurons and edges have *weighs* that go adjusting as learning goes on. A softmax layer is a function that converts real values of a k-dimension vector into a probability distribution of K possible outcomes, this is, to normalize the output of a network to a probability distribution: $\sigma : \mathbb{R}^K \to (0,1)^K,\ K \geq 1,$

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}},\ for\ i = 1 \div K,\ and\ z = (z_1, \dots, z_K) \in \mathbb{R}^K.$$

The method presented in the deep gravity model adapts a nonlinear variant of the multinomial logistic regression by adding hidden layers, hence, nonlinearities and constructing more complex representations of the input features.

Definition 3.1. *Let R be a surface. A tessellation T is a set of polygons l_i called locations, satisfying the following properties:*

- The number of locations l_i contained in the tessellation T is finite: $T = \{l_i : i = 1, ..., n\}$.

- No location overlaps any other one: $l_i \cap l_j = \emptyset, \forall i \neq j$

- All locations cover the whole surface: $\cup_{i=1}^{n} l_i = R$

Definition 3.2. *The flow between locations l_i and l_j, $y(l_i, l_j)$, is the total number of people moving from l_i and l_j for any reason per unit time. The total outflow, O_i is the total number of trips per unit time with departure from location l_i, i.e., $O_i = \sum_j y(l_i, l_j)$.*

Over a region of interest R and given a tessellation T and the total outflow from locations in T , we aim to forecast the flows between any two locations in T, having in mind that we don't consider flows as input data. This is, we are not to consider any historical information to generate flows in our region, therefore our model tested to work on region R must have been trained on another nonoverlapped region.

To control the performance of the flow generation models by calculating the similarity between real and generated flows, the following index will be used.

Definition 3.3. *Let y^r and y^g be the real flows and the generated flows, respectively. The Sørensen-Dice index or Common Part of Commuters (CPC) is the quotient*

$$CPC = \frac{2\sum_{i,j} min(y^g(l_i,l_j), y^r(l_i,l_j))}{\sum_{i,j} y^g(l_i,l_j) + \sum_{i,j} y^r(l_i,l_j)}$$

Note: CPC indicates the accuracy of the prediction and CPC$\in [0,1]$, where 1 indicates a perfect match and 0 shows a bad performance.

Definition 3.4. *The Pearson's correlation coefficient (PCC) is the covariance of two variables X, Y divided by the product of their standard deviations:*

$$PCC = \rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y},$$

used for the linear correlation.

Definition 3.5. *Let O_i and S_i be the observed values and estimated values, respectively. The normalized root mean squared error or NRMSE, also called scatter index, is a statistical error indicator*

$$NRMSE = \frac{\sum(S_i - O_i)^2}{\sum O_i^2}$$

Definition 3.6. *The Kullback-Leibler divergence, or KL divergence:*

$$D_{KL}(q\|p) = -E_{q(x)}\left[log\frac{p(x)}{q(x)}\right] = -\int q(x)\cdot\left[log\frac{p(x)}{q(x)}\right]dx$$

measures how a probability distribution q is different from a reference probability distribution p.

The KL divergence has the following properties:

1. It is not a symetric metric, thus it cannot be used as a distance metric.

2. When $KL = 0$, a similar or even the same behaviour of the two distributions is expected. When $KL = 1$, the expectation, given the first distribution, approaches 0, since both distributions behave differently.

Definition 3.7. *The Jensen-Shannon divergence (JSD), also known as information radius (IRad) or total divergence to the average:*

$$D_{JS}(q\|p) = \frac{1}{2} \cdot \left[D_{KL}\left(p\|\frac{p+q}{2}\right) + D_{KL}\left(q\|\frac{p+q}{2}\right)\right]$$

Note. The JS divergence is also used for studying the dissimilarity between the distributions of the real and the generated flows.

In the gravity equation with logarithms being used, the negative of loglikelihood is proportional to the cross entropy loss $H = -\sum_i \sum_j \frac{y(l_i,l_j)}{O_i} \ln p_{i,j}$ of a shallow neural network with an input of dimension two and a single linear layer followed by a softmax layer. The cross entropy between two probability distributions p and q under the same set of events provides the average number of bits needed to identify an event from the set, if the coding scheme is optimised with a probability distribution q, and the true distribution is p. Let's see this relation: let $q_\theta(X = i)$ the estimated probability of outcome i, to be optimised with the parameter θ. Let $p(X = i)$ be the frequency (empirical probability) of outcome i in the training set. Let N be the conditionally independent samples. Having that the likelihood of the parameter θ is

$$L(\theta) = \prod_{i \in X} (estimated\ probability\ of\ i)^{(number\ of\ ocurrences\ of\ i)} = \prod_i q_\theta(X = i)^{Np(X=i)},$$

so, dividing the log-likelihood by N, we have

$$\frac{1}{N} log(L(\theta)) = \frac{1}{N} log \prod_i q_\theta(X = i)^{Np(X=i)} = \sum_i p(X = i) log q_\theta(X = i) = -H(p,q),$$

hence minimizing the cross-entropy is the same as maximizing the likelihood with respect to θ.

This suggests to project this flow generation goal as a classification problem: each trip (or the unit flow chosen) from each location of origin is assigned to the actual location of destination (correct class) among all the given classes, which are the locations of the tessellation T.

The model gives, for each destination in the tessellation, the probability of an individual from any location l_i would go to that destination. These probabilities are multiplied by the origin's total outflow to calculate the average flows from an origin.

3.3.2 Method (Architecture of deep gravity)

The model computes a n-dimensional vector of probabilities $p_{i,j}$ for $j = 1, ..., n$ from a given location of origin l_i. The probabilities are computed the following way:

Firstly, we concatenate the two feature vectors x_i and x_j, of the origin and destination locations l_i and l_j, respectively, and the geographic distance $r_{i,j}$ between them two. The distance is measured along the earth's surface between the centroids of the polygons that represent the locations. We have then the input vectors $x(l_i, l_j) = concat[x_i, x_j, r_{i,j}]$ for $i, j = 1, ..., n$. The location feature vectors x_i give properties of the location and a spatial depiction of the area, i.e., the number of hospitals, the length of roads or the number of restaurants. The total number of features taken into consideration is its dimension d. In this case $d = 18$, hence each flow using this model is computed by 39 features: 18 per origin, 18 per destination, the distance and the population of them two.

The location features considered in this model are the following [13]:

- Population size (1 feature)
- Land-use areas (5 features): total area in km^2 for each possible land-use class (residential, commercial, industrial, natural and retail).
- Road network (3 features): total length in km^2 for different types of roads (residential, main and other).
- Transport facilities (2 features): total number of Points of Interest or POIs and buildings linked to the transports facilities (bus or train stations, stops and parkings)
- Food facilities (2 features): total number of POIs and buildings related to food facilities (bars, cafes, restaurants).
- Health facilities (2 features): total number of POIs and buildings related to health facilities (clinics, hospitals, pharmacies).

- Education facilities (2 features): total number of POIs and buildings related to education facilities (schools, colleges and kindergarten).
- Retail facilities (2 features): total number of POIs and buildings related to retail facilities (supermarkets, malls, department stores).

It must be mentioned that all features excluding area are divided to the location's area so they are normalized.

Secondly, the feed-forward neural network [3] receives in parallel the input vectors $x(l_i, l_j)$. The network used has 15 hidden layers, this is, 15 layers between the input and the output: the bottom six are of dimensions 256 and the other of dimensions 128, using an activation fuction a, LeakyReLu (rectified linear unit) activation function to be precised, which prevents from becoming saturated at 0. The output of hidden layer h is the vector

$$z^{(0)}(l_i, l_j) = a(W^{(0)} \cdot x(l_i, l_j))$$

for $h = 0$, and

$$z^{(h)}(l_i, l_j) = a(W^{(h)} \cdot z^{(h-1)}(l_i, l_j))$$

for $h > 0$, where W are matrices with parameters learned on the training. The last layer output is a scalar number that is called *score*: $s(l_i, l_j) \in [-\infty, +\infty]$. The higher this number is, the higher the probability that a trip from l_i to l_j takes place.

The third step, a softmax function transforms the scores into probabilities, summing up to one:

$$p_{i,j} = \frac{e^{s(l_i,l_j)}}{\sum_k e^{s(l_i,l_k)}}$$

Finally, the origin's total outflow is multiplied by the model's probability to obtain the generated flow between l_i and l_j.

[3] a feed-forward neural network is a neural network wherein connections between the nodes do not form a cycle.

Chapter 4

Methods Implementation and Comparison

In this chapter, four gravity models are run with 2019 data from Catalonian comarques along with a radiation model. Predictability patterns are studied in papers [44] by Xiao-Yong Yan et al, [26] by Marshall et al, [39] by Robert M. Beyer et al and [45] by Yang et al. We suggest to revise the book [42] for further econometric analysis and a good perspective for designing models. Appendixes A and B are provided at the end: appendix A shows the results of the models run in *RStudio* and appendix B displays a table of the Catalonian data from 2019 used for the models. All 2019 data used in the models is extracted from *Idescat.cat*. The flow matrix used for the models has a dimension of 1764 registers, as Catalonia has 42 *comarques*.

4.1 Migration flows within Catalonia using Gravity Models

In this section three different gravity models are going to be tested with R to forecast the migration flow between *comarques*[1] of Catalonia in 2019. The first model is going to strictly follow the most basic model and the second one is going to consider some extra variables.

Since the denominator of the formula is the distance between origin and destination, and the counties are one next to the other, the distance considered is the

[1] A *comarca* is a group of municipalities, roughly equivalent to a county in the USA or a district in the UK. *https://en.wikipedia.org/wiki/Comarques_of_Catalonia*

distance between their capitals, so it is never zero. Another alternative would have been the distance between the center points, also called geographical center or centroid. Ignoring the problems of the curvature or the Earth's surface by radially projecting it to the sea level or geoid surface[2]. Note that the comarca *Vallès Occidental* has two capitals and *Sabadell* has been chosen for computing the distances. As for the flows within comarques, we have set 10km as the distance between their origin and destination to avoid cutting registers that might me useful.

Model 1: $\ln F_{ij} \sim \beta_1 \cdot \ln M_i + \beta_2 \cdot \ln M_j - \beta_3 \cdot \ln D_{ij} + c_{ij}$

being M_i the population of origin i and M_j the population of location j, both in 2019, D_{ij} is the distance between locations i and j, and c_{ij} is a constant that depends as well on origin i and destination j.

Model 2: $\ln F_{ij} \sim \beta_1 \cdot \ln G_i + \beta_2 \cdot \ln G_j - \beta_3 \cdot \ln D_{ij} + c_{ij}$

being G_i the GDP of origin i and G_j the GDP of location j, both in 2019, D_{ij} is the distance between locations i and j, and c_{ij} is a constant that depends as well on origin i and destination j.

Model 3: $\ln F_{ij} \sim \beta_1 \cdot \ln M_i + \beta_2 \cdot \ln M_j + \beta_3 \cdot \ln G_i + \beta_4 \cdot \ln G_j - \beta_5 \cdot \ln D_{ij} + c_{ij}$

where we add the following variable G_k representing the GDP in 2019.

Model 4: $\ln F_{ij} \sim \beta_1 \cdot \ln M_i + \beta_2 \cdot \ln M_j + \beta_3 \cdot G_i + \beta_4 \cdot G_j + \beta_5 \cdot E_i + \beta_6 \cdot E_j + \beta_7 \cdot U_i + \beta_8 \cdot U_j + \beta_9 \cdot Y_i + \beta_{10} \cdot Y_j + \beta_{11} \cdot R_i + \beta_{12} \cdot R_j + \beta_{13} \cdot T_i + \beta_{14} \cdot T_j - \beta_{15} \cdot \ln D_{ij} + c_{ij}$

here it has been added: E_k representing the number of bachelor's degrees of more than 240 ECTS achieved, U_k representing the number of unemployed population, Y_k representing the population between 25 and 44 (associated with increased mobility), R_k representing the industrial waste, T_k representing the number of cultivated lands, all in 2019, from both the location of origin and destination.

A summary of these models run in *RStudio* can be found in Annex I. When using logs, there is a much more distinguished and or adjusted linear regression function through the base of the data points, resulting in a better prediction model. This is shown by adjusted r-squared. Adding new variables does not lower this

[2] The *geoid* is the shape that the ocean surface would adopt if winds and tides weren't there, under the effect of Earth's gravity, including gravitational attraction and Earth's rotation. This surface is extended through the continents. *https://en.wikipedia.org/wiki/Geoid*

indicator, it can only stay the same or make it higher because more information is being taken into consideration: as you can see, model 4 is better. However, not always having a low r-squared implies having a bad model.

4.2 Main Issues

When using logarithms, the presence of zeros and negative values is a problem. Some researches decide to exclude these registers from the sample. This is not convenient since we are not considering information that might be relevant. Another solution is to give them a very small value instead. According to [37], there are two alternatives: the first one consists of using Poisson, negative binomial and zero-inflated models; the second is to use Heckman's selection model. However, these don't seem to be fully convincing: Poisson helps us to not have to use logarithms but, on the other hand, over-weighs high values; as for Heckman's selection model, it is not easy task to find a variable that accounts for null values in our sample.

It is important to understand as well the data input that is needed. When studying migration flows we must have a precise definition of which conditions determine a migrant and their origin, and check that data sources have the same definition. Regarding our study in Catalonia, migration restrictions between *comarques* do not exist, but it is worth mentioning that, when studying international migration flows, visa restrictions do exist and are considered. Furthermore, the extension of territory in Catalonia is not as big as when considering migration flows between continents, therefore migration costs might be not as high as in those situations. For instance, you might find a job in a different *comarca* but you won't move there because you prefer commuting.

We notice the inability to accurately capture the structure of the real flows and a greater variability of real flows than expected. Gravity models generate flows without considering information that is essential to account for the complexity of the geographical landscape, such as land use, the diversity of points of interest and the transportation. More detailed input data is needed along with more flexible models to generate more realistic mobility flows.

The paper [39] proves that gravity models are not statistically supported and do not present the quality of understanding the variation of flows across time as a response to changes and, therefore, cannot be used to predict migration flows. We strictly recommend to have a look at the case studied in the paper, as well as revising the book *Econometric Analysis* by William H. Greene [42] to have a look through concepts that build the regression analysis fundamentals, such as dummy variables, seasonal dummies, indicator variables, semilog functions, lagged independent variables, instrumental variables, fixed effects or heteroskedasticity, to obtain a better knowledge of econometric data analysis.

4.3 Migration flows within Catalonia using the Radiation Model

To implement the radiation model, we need again the number of population and commuters, available on Idescat.cat. The data needed for the variable s_{ij} has been extracted by using the SEDAC Population Service by NASA[3] and making the appropiate calculations. Regarding the distance between origin and destination, we use the distance between the capitals of the *comarques*.

A summary of this model run in *RStudio* can be found in Annex I. Radiation model forecasts flows considering the influence of variables in the proximity of the flows' origin, as well as the number of migrations that also originate there to other *comarques*, unlike gravity models. Nevertheless, we are still missing a lot of factors that definitely affect the migration. Once again, as we have seen already with gravity models, migration flows are not just explained by population and distance, for instance one might consider the salary or the amount of close friends in the area. Results obtained with the Radiation model are analogous to the gravity models run earlier in terms of significance only when logs are used, otherwise the prediction is bad. Literature revised also asserts this statement, adding that gravity models have a better overall performance but radiation models do give competitive results especially for large scales.

4.4 Comparison with Gravity Models

The paper [21] performs a transportation flow comparison between both types of model and conclude that although gravity models provide an overall better performance, radiation models are also competitive especially at a large scale. How-

[3] https://sedac.ciesin.columbia.edu/mapping/popest/pes-v3/

ever, researches state that radiation models have several advantages compared to gravity models such as clear theoretical background and universality due to the absence of parameters to be estimated. Besides, prediction for long-distance travels is better with radiation models, despite some unresolved issues like relatively poor predictability on short-distance travels. Radiation models requires additional information on T_i, compared to gravity models. The variants of the radiation models: a population-weighted opportunities model and a radiation model with an additional scaling exponent have also been studied and can be found in this paper [44]. Finally, another difference is related to what we mentioned early about the fact that the number of opportunities is estimated by the population in that location, which can also be estimated by the total inflows: radiation models do not depend on the distance between locations, compared to gravity models.

Gravity Models		
	Model 1	Model 1 without logs
Residual standard error	0.4366 on 1760 DF	890.7 on 1760 DF
Multiple R-squared	0.7484	0.1782
Adjusted R-squared	0.748	0.1768
F-statistic	1745 on 3 and 1760 DF	127.2 on 3 and 1760 DF
NSMRE	0.19028502	0.80285202
	Model 2	Model 2 without logs
Residual standard error	0.4426 on 1760 DF	893.4 on 1760 DF
Multiple R-squared	0.7415	0.1731
Adjusted R-squared	0.741	0.1717
F-statistic	1683 on 3 and 1760 DF	122.8 on 3 and 1760 DF
NSMRE	0.20324222	0.80784107
	Model 3	Model 3 without logs
Residual standard error	0.4367 on 1758 DF	890 on 1758 DF
Multiple R-squared	0.7486	0.1804
Adjusted R-squared	0.7479	0.178
F-statistic	1047 on 5 and 1758 DF	77.38 on 5 and 1758 DF
NSMRE	0.18753912	0.80076393
	Model 4	Model 4 without logs
Residual standard error	0.4205 on 1748	891.5 on 1748 DF
Multiple R-squared	0.7683	0.1823
Adjusted R-squared	0.7663	0.1753
F-statistic	386.3 on 15 and 1748 DF	25.97 on 15 and 1748 DF
NSMRE	0.18692208	0.79891033

Table 4.1: Summary of the results applying gravity models and multi-linear regressions. p-value: < 2.2e-16.

Radiation Model	
	Model 1
Residual standard error	0.5725 on 1757 DF
Multiple R-squared	0.5682
Adjusted R-squared	0.5667
F-statistic	385.3 on 6 and 1757 DF
NSMRE	0.72484092

Table 4.2: Summary of the radiation model. p-value: < 2.2e-16.

Conclusions

Spatial statistics offer formal techniques to analyze any entity by using their topological, geometric or geographic properties. This allows us to gather substantiated data that might be useful to solve real-life problems, such as using high-quality data when running forecasting models. Gravity and radiation models enhance forecasting results based on a relation between variables following their respective formulas by running multi-linear regressions. We confirm that, econometrically, these are methods minimally useful to study migration or economic policies, as contrasted with different literature. However, despite the fact that they offer enough satisfaction, they present limitations in terms of accuracy. But nothing could be further than the truth, its formulas do not consider a pandemic directly, but indirectly through the GDP or population. The deep gravity model clearly shows a path towards AI alternatives such as machine learning for more accurate predictions. Its results rely on more information, since they are trained on a large set of different data. As next steps for this research, the following three problems are propounded: to implement these tools to study commuting flows to work; to study adequate migration policies in the USA for European citizens; the creation of a deep gravity model specifically for internal migration within a country trained with migration from many different countries with live data, including China when the coronavirus pandemic started - would the detection of non-migration in a territory alert flows forecasts for other territories?

References

[1] A. Gelfand, P. Diggle, M. Fuentes, P. Guttorp (2010). *Handbook of Spatial Statistics*. Chapman Hall/CRC.

[2] Amini, A., Kung, K., Kang, C., Sobolevsky, S. Ratti, C. (2014). *The impact of social segregation on human mobility in developing and industrialized regions*. EPJ Data Sci. 3.

[3] Andrei N. Kolmogoroff (1956). *Foundations of the Theory of Probability*. Second English Edition.

[4] A. P. Masucci, J. Serras, A. Johansson, M. Batty. *Gravity versus radiation models: On the importance of scale and heterogeneity in commuting flows* [paper]. Physical Review E 88 (2013) 022812.

[5] Bergstrand, Jeffrey H. *The Gravity Equation in International Trade: Some Microeconomic Foundations and Empirical Evidence*. The Review of Economics and Statistics 67, no. 3 (1985): 474–81. https://doi.org/10.2307/1925976.

[6] Brian D: Ripley (1981). *Spatial Statistics*. John Wiley Sons.

[7] Charles J. Geyer, Elisabeth A. Thompson (1969). *Constrained maximum likelihood for dependent data*. Journal of the Royal Statistical Society.

[8] Christian Alis, Erika Fille Legara Christopher Monterola (2021). *Generalized radiation model for human migration.* Scientific Reports 11.

[9] Colwell, P. F. (1982) *Central place theory and the simple economic foundations of the gravity model*, Journal of Regional Science, Vol. 22, No. 4, pp. 541-546.

[10] D. Brook (1964). *On the distinction between the conditional probability and the joint probability approaches in the specification of nearest-neighbour systems*. Biometrika.

[11] Filippo Simini, Marta C. Gonzales, Amos Maritan, Albert-Laszlo Barabasi (2012) *A universal model for mobility and migration patterns* [paper]. Nature 2012. 7392. 484: 96-100.

[12] Frank Nielsen (2021). *On a variational definition for the Jensen-Shannon symmetrization of distances based on the information radius.* Entropy. MDPI. 23 (4): 464.

[13] F. Simini, G. Barlacchi, M. Luca, L. Pappalardo. *A Deep Gravity model for mobility flows generation.* [paper]. Nat Commun 12, 6576 (2021)

[14] Fritz Rohrlich (1989) *From Paradox to Reality: Our Basic Concepts of the Physical World* [paper]. Cambridge University Press. pp. 28.

[15] G. Matheron (1969). *Le krigeage universel*. Cahiers du Centre de Morphologie Mathematique no 1.

[16] G. Matheron (1973). *The Intrinsic Random Functions and their Applications*. Advances in Applied Probability, Volume 5. Cambridge University Press.

[17] G. Matheron (1962). *Traité de geostatistique appliquée. Memoires de Recherches Géologiques et Miniéres no 14* and no 24. Editions Technip.

[18] G. R. Grimmett (1973). *A theorem about random fields*. Bulletin of the London Mathematical Society.

[19] Greene, William H. (2012). *Econometric Analysis (Seventh ed.)*. Boston: Pearson Education. pp. 803–806.

[20] H. Rue, L. Held (2005). *Gaussian Markov Random Fields. Theory and Applications*. Chapman Hall/CRC.

[21] Inho Hong,Woo-Sung Jung,Hang-Hyun Jo (2019). *Gravity model explained by the radiation model on a population landscape.* Plos One Journals.

[22] Julian E. Besag (1969). *Spatial Interaction and the Statistical Analysis of lattice Systems*. Journal of the Royal Statistical Society.

[23] Kang, C., Liu, Y., Guo, D. Qin, K. (2015). *A generalized radiation model for human mobility: Spatial scale, searching direction and trip constraint.* PLoS ONE 10.

[24] Keith Head and Thierry Mayer (2014) *Gravity Equations: Workhorse, Toolkit, Cookbook* [paper]. Elsevier's Handbook of International Economics Vol. 4

[25] Kingsley E. Haynes, A. Stewart Fotheringham (1984) *Gravity and Spatial Interaction Models* [paper]. Regional Research Institute, West Virginia University.

[26] Marshall, J. M. et al (2018). *Mathematical models of human mobility of relevance to malaria transmission in Africa.* Sci. Rep. 8.

[27] Masucci, A. P., Serras, J., Johansson, A. Batty, M. (2013). *Gravity versus radiation models: On the importance of scale and heterogeneity in commuting flows*. Phys. Rev. E88.

[28] M. Barthélemy. *Spatial Networks.* [paper]. Physics Reports, Volume 499, Issues 1-3, pages 1-101, ISSN 0370-1573 (2011)

[29] Michel Beine, Simone Bertoli, Fernández-Huertas Moraga (2014): *A practitioners' guide to gravity models of international migration* [paper]. Fundación de Estudios de Economía Aplicada (FEDEA).

[30] M. Schlather (2012). *Construction of covariance functions and unconditional simulation of random fields*. Chapter 2. Springer Lecture Notes in Statistics.

[31] Michael L. Stein (1999). *Interpolation of Spatial Data. Some Theory for Kriging*. Springer Series in Statistics.

[32] M. N. M. van Lieshout (2019). *Theory of Spatial Statistics. A Concise Introduction*. Texts in Statistical Science.

[33] N. Cressie (1990). *The Origins of kriging. Mathematical Geology 22.*

[34] Patrick Billingsley (1995). *Probability and Measure*. Wiley series in Probability and Mathematical Statistics.

[35] P. Whittle (1954). *On Stationary Processes in the Plane*. Biometrika.

[36] Piovani, D., Arcaute, E., Uchoa, G., Wilson, A. Batty, M. (2018). *Measuring accessibility using gravity and radiation models*. R. Soc. Open Science 5.

[37] Raul Ramos (2016) *Gravity models: A tool for migration analysis* [paper]. IZA World of Labor 2016: 239

[38] Robert J. Adler, J. E. Taylor (2007). *Random Fields and Geometry*. Springer Monographs in Mathematics.

[39] Robert M. Beyer, Jacob Schewe, Hermann Lotze-Campen. *Gravity models do not explain, and cannot predict, migration flows.* [paper]. Humanities and Social Siences Communications (2022).

[40] Roland. L. Dobrushin (1968). *The description of a random field by means of conditional probabilities and conditions of its regularity*. The theory of Probability and its Applications.

[41] S. Meyn, R. L. Tweedie (2009). *Markov Chains and Stochastic Stability*. Cambridge University Press.

[42] W. Greene (2010) *Maximum Simulated Likelihood Methods and Applications (Advances in Econometrics)*, Ilustrated edition, Chapter 14.

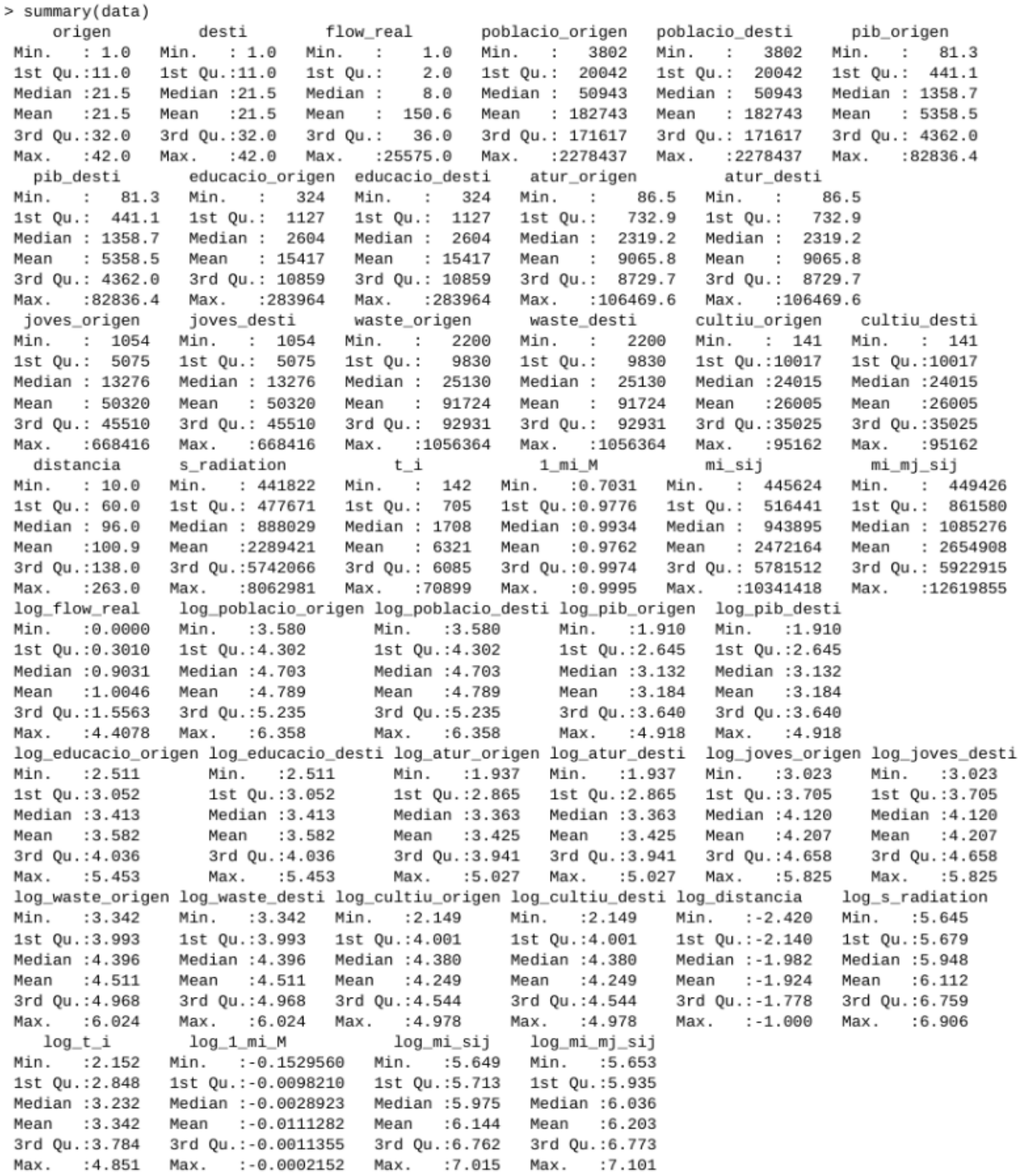
[43] W. K. Hastings (1970). *Monte Carlo sampling methods using Markov chains and their applications*. Biometrika.

[44] Xiao-Yong Yan, Chen Zhao, Ying Fan, Zengru Di and Wen-Xu Wang (2014). *Universal predictability of mobility patterns in cities.* The Royal Society Publishing.

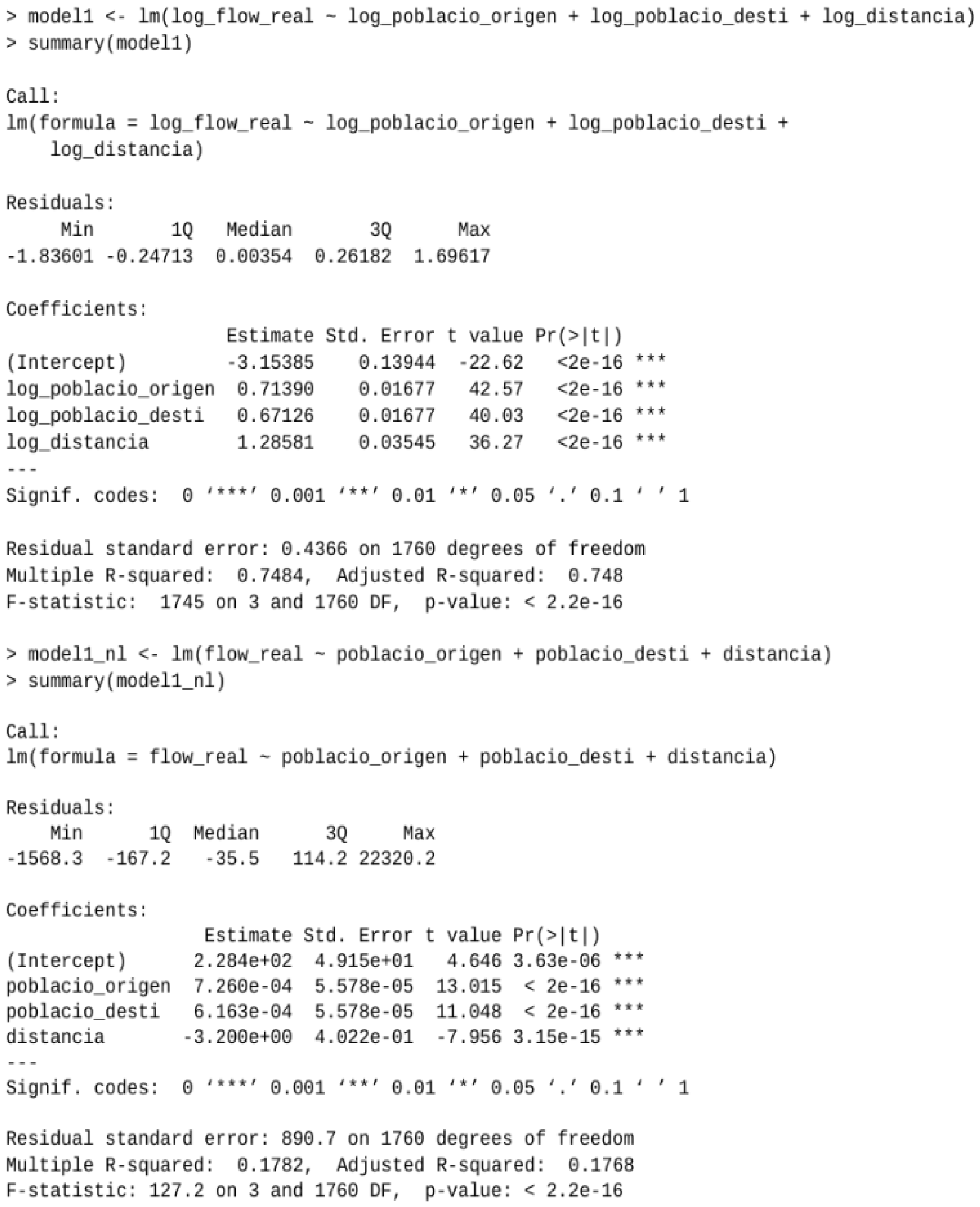
[45] Yang, Y., Herrera, C., Eagle, N. González, M. C. (2014). *Limits of predictability in commuting flows in the absence of data for calibration.* Sci. Rep. 4, 1–9.

Appendixes

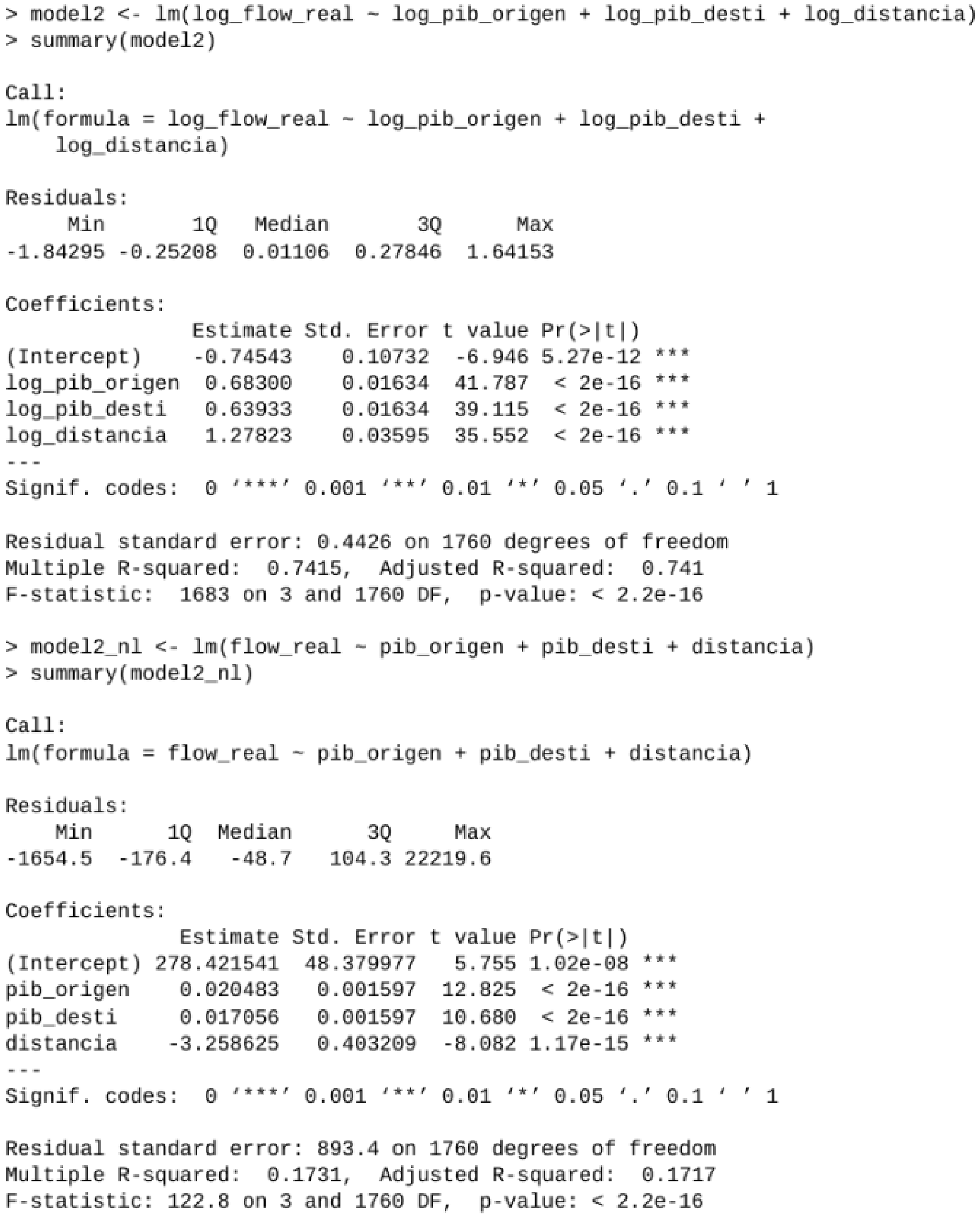
Appendix A: models run in R



```
> summary(data)
     origen          desti         flow_real        poblacio_origen   poblacio_desti      pib_origen
 Min.   : 1.0   Min.   : 1.0   Min.   :    1.0   Min.   :   3802   Min.   :   3802   Min.   :   81.3
 1st Qu.:11.0   1st Qu.:11.0   1st Qu.:    2.0   1st Qu.:  20042   1st Qu.:  20042   1st Qu.:  441.1
 Median :21.5   Median :21.5   Median :    8.0   Median :  50943   Median :  50943   Median : 1358.7
 Mean   :21.5   Mean   :21.5   Mean   :  150.6   Mean   : 182743   Mean   : 182743   Mean   : 5358.5
 3rd Qu.:32.0   3rd Qu.:32.0   3rd Qu.:   36.0   3rd Qu.: 171617   3rd Qu.: 171617   3rd Qu.: 4362.0
 Max.   :42.0   Max.   :42.0   Max.   :25575.0   Max.   :2278437   Max.   :2278437   Max.   :82836.4
   pib_desti        educacio_origen  educacio_desti    atur_origen          atur_desti
 Min.   :   81.3   Min.   :   324   Min.   :   324   Min.   :    86.5   Min.   :    86.5
 1st Qu.:  441.1   1st Qu.:  1127   1st Qu.:  1127   1st Qu.:   732.9   1st Qu.:   732.9
 Median : 1358.7   Median :  2604   Median :  2604   Median :  2319.2   Median :  2319.2
 Mean   : 5358.5   Mean   : 15417   Mean   : 15417   Mean   :  9065.8   Mean   :  9065.8
 3rd Qu.: 4362.0   3rd Qu.: 10859   3rd Qu.: 10859   3rd Qu.:  8729.7   3rd Qu.:  8729.7
 Max.   :82836.4   Max.   :283964   Max.   :283964   Max.   :106469.6   Max.   :106469.6
  joves_origen      joves_desti       waste_origen      waste_desti       cultiu_origen    cultiu_desti
 Min.   :  1054   Min.   :  1054   Min.   :   2200   Min.   :   2200   Min.   :  141   Min.   :  141
 1st Qu.:  5075   1st Qu.:  5075   1st Qu.:   9830   1st Qu.:   9830   1st Qu.:10017   1st Qu.:10017
 Median : 13276   Median : 13276   Median :  25130   Median :  25130   Median :24015   Median :24015
 Mean   : 50320   Mean   : 50320   Mean   :  91724   Mean   :  91724   Mean   :26005   Mean   :26005
 3rd Qu.: 45510   3rd Qu.: 45510   3rd Qu.:  92931   3rd Qu.:  92931   3rd Qu.:35025   3rd Qu.:35025
 Max.   :668416   Max.   :668416   Max.   :1056364   Max.   :1056364   Max.   :95162   Max.   :95162
   distancia       s_radiation            t_i             1_mi_M              mi_sij            mi_mj_sij
 Min.   : 10.0   Min.   : 441822   Min.   :  142   Min.   :0.7031   Min.   :  445624   Min.   :  449426
 1st Qu.: 60.0   1st Qu.: 477671   1st Qu.:  705   1st Qu.:0.9776   1st Qu.:  516441   1st Qu.:  861580
 Median : 96.0   Median : 888029   Median : 1708   Median :0.9934   Median :  943895   Median : 1085276
 Mean   :100.9   Mean   :2289421   Mean   : 6321   Mean   :0.9762   Mean   : 2472164   Mean   : 2654908
 3rd Qu.:138.0   3rd Qu.:5742066   3rd Qu.: 6085   3rd Qu.:0.9974   3rd Qu.: 5781512   3rd Qu.: 5922915
 Max.   :263.0   Max.   :8062981   Max.   :70899   Max.   :0.9995   Max.   :10341418   Max.   :12619855
 log_flow_real    log_poblacio_origen log_poblacio_desti log_pib_origen  log_pib_desti
 Min.   :0.0000   Min.   :3.580       Min.   :3.580      Min.   :1.910   Min.   :1.910
 1st Qu.:0.3010   1st Qu.:4.302       1st Qu.:4.302      1st Qu.:2.645   1st Qu.:2.645
 Median :0.9031   Median :4.703       Median :4.703      Median :3.132   Median :3.132
 Mean   :1.0046   Mean   :4.789       Mean   :4.789      Mean   :3.184   Mean   :3.184
 3rd Qu.:1.5563   3rd Qu.:5.235       3rd Qu.:5.235      3rd Qu.:3.640   3rd Qu.:3.640
 Max.   :4.4078   Max.   :6.358       Max.   :6.358      Max.   :4.918   Max.   :4.918
 log_educacio_origen log_educacio_desti log_atur_origen log_atur_desti  log_joves_origen log_joves_desti
 Min.   :2.511       Min.   :2.511      Min.   :1.937   Min.   :1.937   Min.   :3.023    Min.   :3.023
 1st Qu.:3.052       1st Qu.:3.052      1st Qu.:2.865   1st Qu.:2.865   1st Qu.:3.705    1st Qu.:3.705
 Median :3.413       Median :3.413      Median :3.363   Median :3.363   Median :4.120    Median :4.120
 Mean   :3.582       Mean   :3.582      Mean   :3.425   Mean   :3.425   Mean   :4.207    Mean   :4.207
 3rd Qu.:4.036       3rd Qu.:4.036      3rd Qu.:3.941   3rd Qu.:3.941   3rd Qu.:4.658    3rd Qu.:4.658
 Max.   :5.453       Max.   :5.453      Max.   :5.027   Max.   :5.027   Max.   :5.825    Max.   :5.825
 log_waste_origen log_waste_desti log_cultiu_origen log_cultiu_desti log_distancia    log_s_radiation
 Min.   :3.342    Min.   :3.342   Min.   :2.149     Min.   :2.149    Min.   :-2.420   Min.   :5.645
 1st Qu.:3.993    1st Qu.:3.993   1st Qu.:4.001     1st Qu.:4.001    1st Qu.:-2.140   1st Qu.:5.679
 Median :4.396    Median :4.396   Median :4.380     Median :4.380    Median :-1.982   Median :5.948
 Mean   :4.511    Mean   :4.511   Mean   :4.249     Mean   :4.249    Mean   :-1.924   Mean   :6.112
 3rd Qu.:4.968    3rd Qu.:4.968   3rd Qu.:4.544     3rd Qu.:4.544    3rd Qu.:-1.778   3rd Qu.:6.759
 Max.   :6.024    Max.   :6.024   Max.   :4.978     Max.   :4.978    Max.   :-1.000   Max.   :6.906
    log_t_i         log_1_mi_M            log_mi_sij     log_mi_mj_sij
 Min.   :2.152   Min.   :-0.1529560   Min.   :5.649   Min.   :5.653
 1st Qu.:2.848   1st Qu.:-0.0098210   1st Qu.:5.713   1st Qu.:5.935
 Median :3.232   Median :-0.0028923   Median :5.975   Median :6.036
 Mean   :3.342   Mean   :-0.0111282   Mean   :6.144   Mean   :6.203
 3rd Qu.:3.784   3rd Qu.:-0.0011355   3rd Qu.:6.762   3rd Qu.:6.773
 Max.   :4.851   Max.   :-0.0002152   Max.   :7.015   Max.   :7.101
```



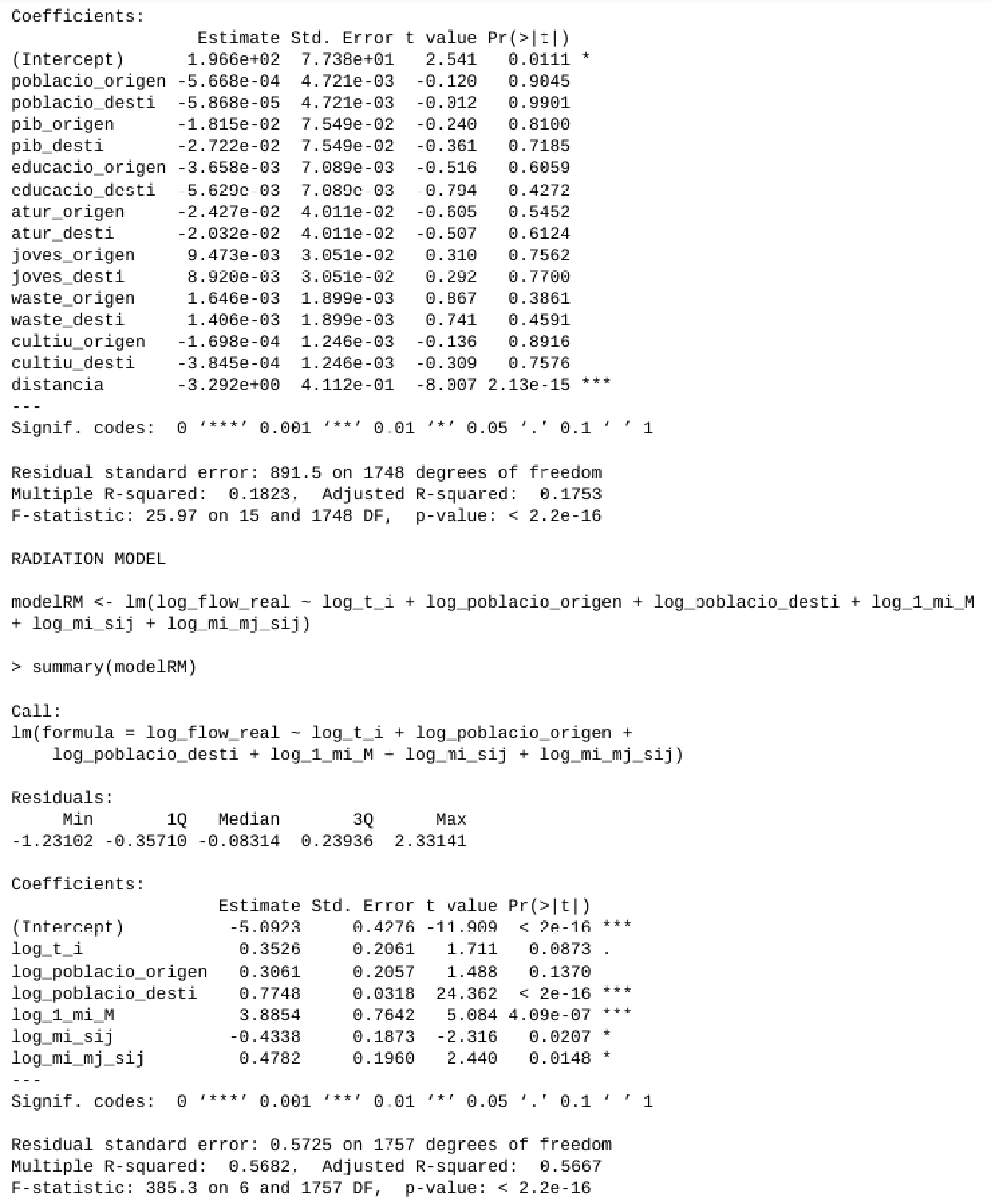
```
> model1 <- lm(log_flow_real ~ log_poblacio_origen + log_poblacio_desti + log_distancia)
> summary(model1)

Call:
lm(formula = log_flow_real ~ log_poblacio_origen + log_poblacio_desti +
    log_distancia)

Residuals:
     Min       1Q   Median       3Q      Max
-1.83601 -0.24713  0.00354  0.26182  1.69617

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         -3.15385    0.13944  -22.62   <2e-16 ***
log_poblacio_origen  0.71390    0.01677   42.57   <2e-16 ***
log_poblacio_desti   0.67126    0.01677   40.03   <2e-16 ***
log_distancia        1.28581    0.03545   36.27   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4366 on 1760 degrees of freedom
Multiple R-squared:  0.7484,  Adjusted R-squared:  0.748
F-statistic:  1745 on 3 and 1760 DF,  p-value: < 2.2e-16

> model1_nl <- lm(flow_real ~ poblacio_origen + poblacio_desti + distancia)
> summary(model1_nl)

Call:
lm(formula = flow_real ~ poblacio_origen + poblacio_desti + distancia)

Residuals:
    Min      1Q  Median      3Q     Max
-1568.3  -167.2   -35.5   114.2 22320.2

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      2.284e+02  4.915e+01   4.646 3.63e-06 ***
poblacio_origen  7.260e-04  5.578e-05  13.015  < 2e-16 ***
poblacio_desti   6.163e-04  5.578e-05  11.048  < 2e-16 ***
distancia       -3.200e+00  4.022e-01  -7.956 3.15e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 890.7 on 1760 degrees of freedom
Multiple R-squared:  0.1782,  Adjusted R-squared:  0.1768
F-statistic: 127.2 on 3 and 1760 DF,  p-value: < 2.2e-16
```



```
> model2 <- lm(log_flow_real ~ log_pib_origen + log_pib_desti + log_distancia)
> summary(model2)

Call:
lm(formula = log_flow_real ~ log_pib_origen + log_pib_desti +
    log_distancia)

Residuals:
     Min       1Q   Median       3Q      Max
-1.84295 -0.25208  0.01106  0.27846  1.64153

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    -0.74543    0.10732  -6.946 5.27e-12 ***
log_pib_origen  0.68300    0.01634  41.787  < 2e-16 ***
log_pib_desti   0.63933    0.01634  39.115  < 2e-16 ***
log_distancia   1.27823    0.03595  35.552  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4426 on 1760 degrees of freedom
Multiple R-squared:  0.7415,  Adjusted R-squared:  0.741
F-statistic:  1683 on 3 and 1760 DF,  p-value: < 2.2e-16

> model2_nl <- lm(flow_real ~ pib_origen + pib_desti + distancia)
> summary(model2_nl)

Call:
lm(formula = flow_real ~ pib_origen + pib_desti + distancia)

Residuals:
    Min      1Q  Median      3Q     Max
-1654.5  -176.4   -48.7   104.3 22219.6

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 278.421541  48.379977   5.755 1.02e-08 ***
pib_origen    0.020483   0.001597  12.825  < 2e-16 ***
pib_desti     0.017056   0.001597  10.680  < 2e-16 ***
distancia    -3.258625   0.403209  -8.082 1.17e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 893.4 on 1760 degrees of freedom
Multiple R-squared:  0.1731,  Adjusted R-squared:  0.1717
F-statistic: 122.8 on 3 and 1760 DF,  p-value: < 2.2e-16
```

```
> model3 <- lm(log_flow_real ~ log_poblacio_origen + log_poblacio_desti + log_pib_origen +
log_pib_desti + log_distancia)
> summary(model3)

Call:
lm(formula = log_flow_real ~ log_poblacio_origen + log_poblacio_desti +
    log_pib_origen + log_pib_desti + log_distancia)

Residuals:
     Min       1Q   Median       3Q      Max
-1.82353 -0.24769  0.00251  0.25803  1.68484

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         -3.03574    0.34288  -8.854  < 2e-16 ***
log_poblacio_origen  0.58121    0.13239   4.390 1.20e-05 ***
log_poblacio_desti   0.73391    0.13239   5.544 3.41e-08 ***
log_pib_origen       0.12858    0.12728   1.010    0.313
log_pib_desti       -0.06068    0.12728  -0.477    0.634
log_distancia        1.28524    0.03549  36.217  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4367 on 1758 degrees of freedom
Multiple R-squared:  0.7486,	Adjusted R-squared:  0.7479
F-statistic:  1047 on 5 and 1758 DF,  p-value: < 2.2e-16

> model3_nl <- lm(flow_real ~ poblacio_origen + poblacio_desti + pib_origen + pib_desti +
distancia)
> summary(model3_nl)

Call:
lm(formula = flow_real ~ poblacio_origen + poblacio_desti + pib_origen +
    pib_desti + distancia)

Residuals:
    Min      1Q  Median      3Q     Max
-1518.2  -173.4   -28.8   122.3 22476.6

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      1.910e+02  5.337e+01   3.578 0.000355 ***
poblacio_origen  9.253e-04  4.653e-04   1.988 0.046911 *
poblacio_desti   1.586e-03  4.653e-04   3.407 0.000671 ***
pib_origen      -5.722e-03  1.328e-02  -0.431 0.666645
pib_desti       -2.786e-02  1.328e-02  -2.098 0.036058 *
distancia       -3.163e+00  4.025e-01  -7.859 6.71e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 890 on 1758 degrees of freedom
Multiple R-squared:  0.1804,	Adjusted R-squared:  0.178
F-statistic: 77.38 on 5 and 1758 DF,  p-value: < 2.2e-16
```

```
> model4 <- lm(log_flow_real ~ log_poblacio_origen + log_poblacio_desti + log_pib_origen +
log_pib_desti + educacio_origen + educacio_desti + atur_origen + atur_desti + joves_origen +
joves_desti + waste_origen + waste_desti + cultiu_origen + cultiu_desti + log_distancia)
> summary(model4)

Call:
lm(formula = log_flow_real ~ log_poblacio_origen + log_poblacio_desti +
    log_pib_origen + log_pib_desti + educacio_origen + educacio_desti +
    atur_origen + atur_desti + joves_origen + joves_desti + waste_origen +
    waste_desti + cultiu_origen + cultiu_desti + log_distancia)

Residuals:
     Min       1Q   Median       3Q      Max
-1.91407 -0.23493  0.00194  0.23475  1.95589

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)         -1.552e+00  4.615e-01  -3.363 0.000787 ***
log_poblacio_origen  3.813e-01  1.651e-01   2.310 0.021026 *
log_poblacio_desti   6.947e-01  1.651e-01   4.208 2.70e-05 ***
log_pib_origen       5.402e-02  1.529e-01   0.353 0.723949
log_pib_desti       -2.013e-01  1.529e-01  -1.316 0.188217
educacio_origen     -4.662e-06  2.444e-06  -1.907 0.056647 .
educacio_desti      -2.772e-06  2.444e-06  -1.134 0.256975
atur_origen          1.581e-05  1.160e-05   1.363 0.172948
atur_desti           3.718e-06  1.160e-05   0.321 0.748571
joves_origen        -4.915e-06  3.014e-06  -1.631 0.103133
joves_desti         -2.014e-06  3.014e-06  -0.668 0.504155
waste_origen         3.681e-06  8.068e-07   4.562 5.42e-06 ***
waste_desti          2.351e-06  8.068e-07   2.914 0.003614 **
cultiu_origen        3.169e-06  6.054e-07   5.234 1.85e-07 ***
cultiu_desti         2.342e-06  6.054e-07   3.868 0.000114 ***
log_distancia        1.318e+00  3.455e-02  38.162  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4205 on 1748 degrees of freedom
Multiple R-squared:  0.7683,  Adjusted R-squared:  0.7663
F-statistic: 386.3 on 15 and 1748 DF,  p-value: < 2.2e-16

> model4_nl <- lm(flow_real ~ poblacio_origen + poblacio_desti + pib_origen + pib_desti +
educacio_origen + educacio_desti + atur_origen + atur_desti + joves_origen + joves_desti +
waste_origen + waste_desti + cultiu_origen + cultiu_desti + distancia)
> summary(model4_nl)

Call:
lm(formula = flow_real ~ poblacio_origen + poblacio_desti + pib_origen +
    pib_desti + educacio_origen + educacio_desti + atur_origen +
    atur_desti + joves_origen + joves_desti + waste_origen +
    waste_desti + cultiu_origen + cultiu_desti + distancia)

Residuals:
    Min      1Q  Median      3Q     Max
-1536.6  -174.2   -25.5   126.0 22460.1
```



```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      1.966e+02  7.738e+01   2.541   0.0111 *
poblacio_origen -5.668e-04  4.721e-03  -0.120   0.9045
poblacio_desti  -5.868e-05  4.721e-03  -0.012   0.9901
pib_origen      -1.815e-02  7.549e-02  -0.240   0.8100
pib_desti       -2.722e-02  7.549e-02  -0.361   0.7185
educacio_origen -3.658e-03  7.089e-03  -0.516   0.6059
educacio_desti  -5.629e-03  7.089e-03  -0.794   0.4272
atur_origen     -2.427e-02  4.011e-02  -0.605   0.5452
atur_desti      -2.032e-02  4.011e-02  -0.507   0.6124
joves_origen     9.473e-03  3.051e-02   0.310   0.7562
joves_desti      8.920e-03  3.051e-02   0.292   0.7700
waste_origen     1.646e-03  1.899e-03   0.867   0.3861
waste_desti      1.406e-03  1.899e-03   0.741   0.4591
cultiu_origen   -1.698e-04  1.246e-03  -0.136   0.8916
cultiu_desti    -3.845e-04  1.246e-03  -0.309   0.7576
distancia       -3.292e+00  4.112e-01  -8.007 2.13e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 891.5 on 1748 degrees of freedom
Multiple R-squared:  0.1823,  Adjusted R-squared:  0.1753
F-statistic: 25.97 on 15 and 1748 DF,  p-value: < 2.2e-16

RADIATION MODEL

modelRM <- lm(log_flow_real ~ log_t_i + log_poblacio_origen + log_poblacio_desti + log_1_mi_M
+ log_mi_sij + log_mi_mj_sij)

> summary(modelRM)

Call:
lm(formula = log_flow_real ~ log_t_i + log_poblacio_origen +
    log_poblacio_desti + log_1_mi_M + log_mi_sij + log_mi_mj_sij)

Residuals:
     Min       1Q   Median       3Q      Max
-1.23102 -0.35710 -0.08314  0.23936  2.33141

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          -5.0923     0.4276 -11.909  < 2e-16 ***
log_t_i               0.3526     0.2061   1.711   0.0873 .
log_poblacio_origen   0.3061     0.2057   1.488   0.1370
log_poblacio_desti    0.7748     0.0318  24.362  < 2e-16 ***
log_1_mi_M            3.8854     0.7642   5.084 4.09e-07 ***
log_mi_sij           -0.4338     0.1873  -2.316   0.0207 *
log_mi_mj_sij         0.4782     0.1960   2.440   0.0148 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5725 on 1757 degrees of freedom
Multiple R-squared:  0.5682,  Adjusted R-squared:  0.5667
F-statistic: 385.3 on 6 and 1757 DF,  p-value: < 2.2e-16
```

Appendix B: 2019 data by *comarques*

Comarca	Population	GDP	Education	Unemployment	Youth	Waste	Cultivated lands
Alt Camp	44.296,00	1.548,60	2.278	2.569	11.810	21.633	20.372
Alt Empordà	141.339,00	3.576	7.328	7.169	37.197	117.724	44.278
Alt Penedès	108.411,00	3.286	6.670	5.277	28.716	59.798	22.751
Alt Urgell	20.177,00	489	1.309	770	5.168	10.430	27.535
Alta Ribagorça	3.802,00	104,5	324	86,5	1.054	2.200	10.095
Anoia	120.738,00	2.791,30	6.116	7.208	32.484	61.492	30.674
Aran	10.093,00	374	754	275	3.121	8.636	31.466
Bages	178.885,00	4.888,40	10.859	9.499	46.161	86.596	28.454
Baix Camp	190.973,00	4.632	11.831	11.232	51.760	108.152	15.412
Baix Ebre	77.596,00	1.849	3.655	4.120	19.671	45.552	35.025
Baix Empordà	134.359,00	3.179	8.224	6.181	34.365	109.545	21.484
Baix Llobregat	825.963,00	28.297	55.668	38.744	222.175	385.545	4.283
Baix Penedès	104.991,00	2.046	5.117	7.346	27.525	72.325	7.523
Barcelonès	2.278.437,00	93.256	283.964	106.470	668.416	1.056.364	141
Berguedà	39.446,00	1.035	2.307	1.817	9.563	17.157	33.721
Cerdanya	18.192,00	473	1.477	452	5.075	13.372	32.276
Conca de Barberà	20.042,00	636,9	1.201	738	4.988	10.560	25.683
Garraf	150.887,00	3.043,50	14.083	8.167	38.470	92.931	1.839
Garrigues	18.833,00	442	875	690	4.681	7.317	39.718
Garrotxa	57.590,00	1.721,60	3.682	2.069	14.743	28.627	14.580
Gironès	193.908,00	6.333	15.760	9.097	54.303	85.620	13.845
Maresme	452.690,00	10.222	36.180	24.291	116.917	253.106	3.362
Moianès	13.603,00	304	984	497	3.400	7.736	9.370
Montsià	67.436,00	1.336,30	2.900	3.555	18.311	34.679	38.719
Noguera	38.770,00	974	1.892	1.790	9.801	16.360	73.595
Osona	160.821,00	4.897	10.221	7.364	42.627	71.642	40.278
Pallars Jussà	13.080,00	288	905	571	3.153	6.913	39.228
Pallars Sobirà	6.932,00	192,3	611	216,8	1.796	5.252	54.464
Pla d'Urgell	36.693,00	1.114,20	1.705	1.401	10.133	16.489	25.279
Pla de l'Estany	32.293,00	884	2.210	1.043	8.542	15.889	9.564
Priorat	9.245,00	175	593	378,7	2.238	4.241	8.951
Ribera d'Ebre	21.865,00	1.087,90	1.020	1.058	5.256	9.830	19.753
Ripollès	25.087,00	626,4	1.557	857	5.906	14.000	32.503
Segarra	23.052,00	801,3	1.127	733	6.184	9.822	46.971
Segrià	209.818,00	6.274,90	13.685	10.842	56.559	88.215	95.162
Selva	171.617,00	4.715	8.299	8.730	45.510	114.380	11.118
Solsonès	13.469,00	360	784	451	3.472	6.909	33.136
Tarragonès	256.730,00	9.011,80	18.238	13.985	74.100	156.946	8.294
Terra Alta	11.490,00	276	496	368	2.561	4.256	22.490
Urgell	36.693,00	1.045,40	1.952	1.719	9.637	17.801	43.131
Vallès Occidental	925.237,00	29.662	72.070	49.475	255.317	400.801	5.680
Vallès Oriental	409.638,00	13.171	26.622	21.463	110.593	195.562	10.017

Table 4.3: 2019 data by *comarques* extracted from Idescat.cat