POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

COVID-19 spread over Europe: A statistical study to detect regional contextual factors

Tesi di Laurea Magistrale in
Mathematical Engineering - Ingegneria Matematica

Author: **Francesca Anfossy**

Student ID: 963741
Advisor: Prof. Francesca Ieva
Co-advisors: Pietro Pinoli, Laura Savare'
Academic Year: 2021-22

Abstract

The objective of this research is to detect non-epidemiological regional factors that predict the covid-19 cases density during the first two waves of the pandemic over the European Union. In particular, we compare two approaches to do so: a pipeline of low complexity that provides intuitive results such as association rules, and a geostatistical approach that provides richer insights.

First, we propose a general pipeline that uses dichotomic data in order to evaluate different factors and their interactions to find association rules of a general risk level. We find that the significant rules found contain factors recognized by literature, while also discovering group effects. The relevant features are related to demography (population density and life expectancy), healthcare (available hospital beds and health personnel in the first wave, and long-term care beds in the second wave), economy (amount of hours worked in the first wave, and growth rate of regional GVA in the second wave), and mobility (stock of vehicles in the second wave).

Second, we perform a geostatistical analysis that considers the spatial factor of neighboring regions, by using variogram modelling and performing LISA clustering. We find that life expectancy, along with economical factors, such as growth rate of regional GVA and unemployment rate (the latter in the second wave only), and educational factors, such as participation in education, NEET rate and early leavers from education, can be associated with the development of COVID-19 spread. The model however explains regions with lower densities better than the most critical ones, and spatial dependency is not as evident as expected, possibly due to the amount and distribution of the observations.

We conclude that, in an application such as COVID-19 spread over a continent, a simple approach provides an easier understanding regarding interacting factors, but we need to consider the geographical factor, hence the two approaches studied are best used together in order to gain interpretable but also rich insights.

Keywords: covid-19, lasso selection, association rules, geostatistics, variogram modeling, lisa clustering

Abstract in lingua italiana

Questa ricerca ha l'obiettivo di trovare fattori regionali non epidemiologici che possano predire la densità di casi di COVID-19 nelle prime due ondate della pandemia nell'Unione Europea. Questo viene effettuato tramite due approcci: uno che mina regole di associazione, e uno che impiega strumenti di geostatistica.

Anzitutto, si generalizza una pipeline che usa dati dicotomici per valutare diversi fattori e le loro interazioni, per trovare regole di associazione di rischio. Le regole più significative comprendono fattori validati nella letteratura, mentre emergono anche interazioni di gruppo. I fattori significativi includono variabili demografiche (la densità di popolazione e la speranza di vita media), la robustezza del sistema sanitario (la quantità di letti disponibili e l'ammontare di lavoro del personale competente per la prima ondata, nonchè i letti di cura a lungo termine per la seconda), l'economia (la quantità di ore lavorate per la prima ondata, e la tassa di crescita del VAL regionale per la seconda), e mobilità.

Successivamente, viene compiuta un'analisi geostatistica che considera il fattore spaziale delle regioni vicine, usando modellazione di variogrammi e LISA clustering. Ne risulta che la speranza di vita media, fattori economici (il tasso di crescita del VAL e il tasso di disoccupazione nella seconda ondata) e fattori educazionali (partecipazione all'educazione, percentuale di NEET e quantità di giovani che abbandonano prematuramente istruzione e formazione) possono essere associati con lo sviluppo della diffusione del COVID-19. Il modello risultante spiega in modo migliore la risposta delle regioni con meno densità di casi rispetto alle regioni più critiche. Inoltre, la dipendenza spaziale è meno evidente di quanto atteso, possibilmente a causa della quantità e distribuzione spaziale delle osservazioni.

In conclusione, un approccio semplice offre maggiore comprensione sui fattori che intervengono. Tuttavia, non si può escludere la considerazione del fattore geografico. Si dimostra più conveniente adottare i due approcci studiati in combinazione, con l'obiettivo di trovare risultati tanto interpretabili quanto significativi.

Parole chiave: selezione lasso, regole di associazione, geostatistica, modelazione di variogramma, lisa clustering

Contents

Introduction

Motivation

COVID-19 is nowadays a wide topic to study from, which has provided the unprecedented advantage of having large amounts of available data, for example at the level of regions, where we consider non-epidemiological data, which is socio-economic information as complementary information. In the case of the European Union (EU), we can combine the long-term initiative of storing public social-demographic data at a high level of granularity, with the recent unified measure of sharing records of COVID-19 positive cases with detail over time and space.

Epidemiological data must be combined with other relevant factors, in order to try to understand the COVID-19 spread [28]. Moreover, these relevant factors must be correctly identified in order to detect control opportunities for either prevention or response to an outbreak. Some efforts have been made for global datasets where factors are related to intervention measures. This study focuses on the european continent and the factors considered are the pre-pandemic regional context.

The motivation of this study is to detect the role of complementary predispositional characteristics that can shed light on preemptive policies for a further preparation towards a potential future outbreak. Contextual vulnerability can be corrected if opportunities for action are detected, which can be more valuable when found early enough to perform long term measures for some necessary cases.

Objective

It is of particular interest, when it comes to regional information that could have a role in the pandemic, to predict a risk level that can help preparing the right measures. This has two important aspects, the first being the capacity to foresee the accurate risk in order to decide whether the situation will be influential or not, and the second being the capacity to detect important factors that explain and predict such risk in order to know how to manage the situation when it is necessary.

We consider the COVID@Lombardy dataset [8], that considers regional social-demographic data of the EU. In this case, we are interested in finding groups of interacting factors that can reduce or increase the risk of having a high infection rate in the region. There is evidence that sanitary, educational and demographic characteristics of the region can explain the difference of infection rates detected during the first and second waves in the 2020 outbreak.

The first approach considers using a proposition of a more general pipeline for dichotomic data that has the advantage of being intuitive, even if it is at the trade-off of predictive power. Intuition can be specially powerful to highlight clearly important characteristics in order to make political decisions, or investigate further in social sciences, since they are disciplines that can go further into qualitative analysis more than quantitative as in statistics.

The second approach, instead, considers geo-statistical analysis, which is more complex but attempts to make richer insights over the data. In this case, we add location information of the regions to the dataset, and hence we can evaluate a spatial factor as possible effect. Since this regards regional data, it is intuitive to take into account the distance between the regions, or the overall location of each region.

Through this work, we attempt to contrast both methods and their results analyzing this particular dataset. This way, we consider the points of strength and weakness of each, while we verify if their results are too similar or different, in case they are able to provide richer results when combined.

Outline

This work is presented as follows. Chapter 1 regards the original dataset description and exploration. Chapters 2 and 3 explain each the entire process for each approach, which covers the data pre-processing, explaining the methodology of each pipeline, and presenting the obtained results along with a discussion and final observations for further research opportunities. Chapter 4 presents the contrast between approaches and final conclusions of the overall work. The appendices show complementary results for further detail in each approach.

This work was performed using a Python notebook for the association rule mining approach, and an R script for the geo-statistical analysis approach. All code and saved files are available on GitHub: `https://github.com/fpjaa/geostats-covid`.

1 | Data exploration

1.1. Data presentation

The COVID@Lombardy dataset recovers data from 144 european regions, in 19 countries, in which the number of cases reported by the authorities could be considered reliable. The case density of each wave is registered, where we find:

- The first wave is considered between March 1st and August 20th in 2020.
- The second wave is considered between August 20th, 2020, and February 20th, 2021.

There are also measurements of 23 factors grouped in Education (5), Population (5), Healthcare (4), Mobility (2), Primary sector (2) and Economy (5). They are described by group order:

1. Early leavers from education and training: Percentage of people between 18 and 24 years old who left school, university or training classes after having started them, over the total number of people who joined schools, universities and training.
2. Students enrolled in tertiary education by education level programme orientation: Total number of students (independent from sex and age) enrolled in tertiary education schools.
3. Young people neither in employment nor in education and training: Percentage of people from 15 to 24 years old who are neither studying nor working over the total number of people of that age. Also abbreviated as NEET rate.
4. Participation in education and training: Percentage of people between 25 and 64 years old who, in the last 4 weeks, has participated in educational and training activities.
5. Pupils and students enrolled: Total number of students (independent from sex and age) enrolled in school.
6. Life expectancy: Years of life expectancy for a person.

7. Population density: Density people per square kilometer.
8. Population: Total population of the region.
9. Causes of death crude death rate: Deaths per 100,000 inhabitants, recorded in the population for a given period divided by population in the same period. Also abbreviated as death rate.
10. Deaths: Total number of deaths in each region.
11. Hospital discharges for respiratory diseases: Total number of people who left the hospital after having suffered from respiratory diseases.
12. Long term care beds: Relative number of long term beds available for every 100.000 inhabitants.
13. Health personnel: Total health care staff active in the health care sector (doctors, dentists, nurses, etc.).
14. Available hospital beds: Relative number of available hospital beds per 100.000 inhabitants.
15. Air passengers: Total number of passengers carried in the region in thousand scale.
16. Stock of vehicles: Total number of vehicles present in the region.
17. Farm labour force: Total agricultural labour force, expressed in persons and in Annual Work Units (AWU, corresponds to the work performed by one full-time worker).
18. Utilised agricultural area: Total land utilised for farming, occupied by the main agricultural land uses (arable land, permanent grassland and land under permanent crops).
19. Unemployment rate: Percentage of unemployment.
20. Employment thousand hours worked: Total sum of hours worked by employees in a certain area, in scale of 1.000.
21. Real growth rate of regional gross value added (GVA) at basic prices: Percentage change of GVA against the previous period. Also abbreviated as GVA growth rate.
22. Compensation of employees: Total sum of the compensations of all the employees of a certain area in million euros.
23. GDP: Gross Domestic Product (GDP) at current market prices in million euros.

1.2. Exploration analysis

First, we look at the behavior of the responses and features using box plots. Regarding the response (see Figure 1.1), which is the case density for each wave, we can find a more extreme behavior between regions on the first wave, that becomes more balanced on the second wave.

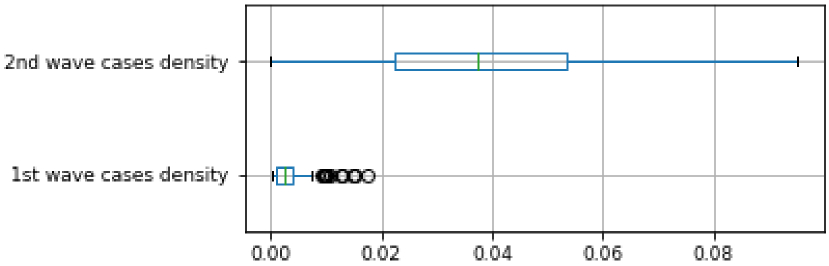


Figure 1.1: Box plot of case density for each wave.

When it comes to the features (see Figures 1.2, 1.3 and 1.4), we group them by magnitude to visualize their behavior, and notice that most of the features present only upper outliers, with few exceptions such as growth rate of regional GVA, life expectancy and crude death rate. Moreover, life expectancy is the only feature that presents only lower outliers.

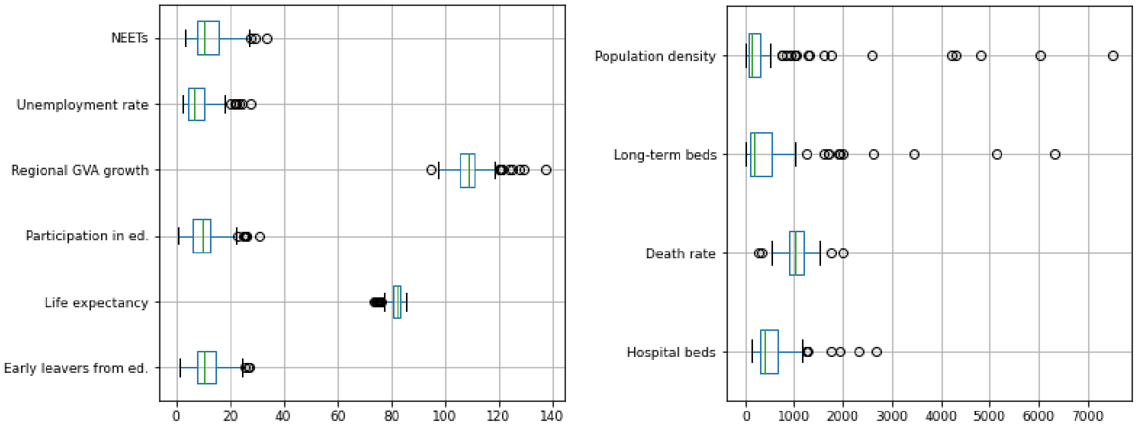


Figure 1.2: Box plots of features with smallest values.

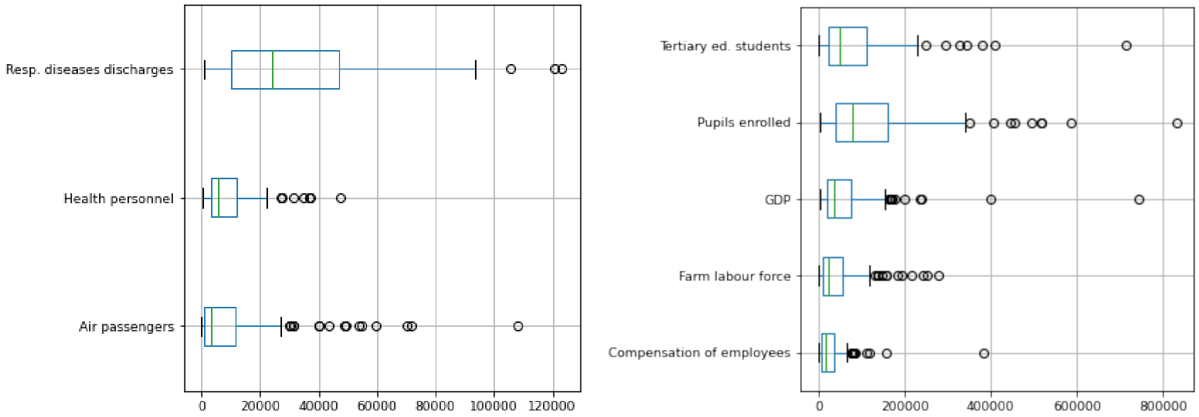


Figure 1.3: Box plots of features with intermediate values.

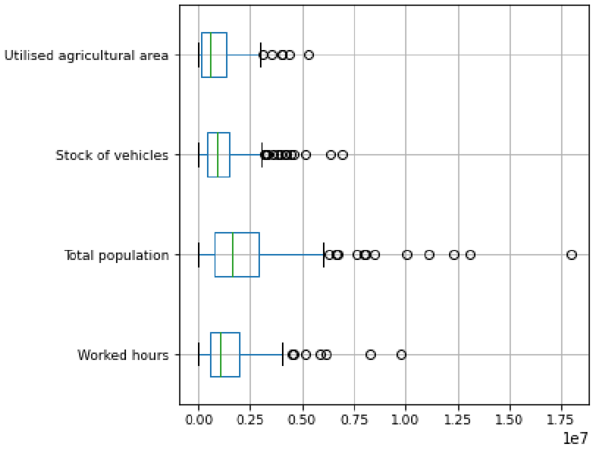


Figure 1.4: Box plots of features with largest values.

Second, we provide the summary statistics of the features, grouped by type, as presented in the previous section: Education, Population, Healthcare, Mobility, Primary sector, and Economy (see Table 1.1).

	Count	**Mean**	**Std. dev.**	**Min.**	**Median**	**Max.**
Early leavers from ed.	118	11.766	5.385	1.4	10.55	27.0
Ter. ed. stud.	115	80688.626	100353.430	521.0	50211.0	713715.0
NEET rate	122	12.3834	6.224	3.5	10.5	33.6
Particip. in ed.	122	10.351	6.384	0.6	9.8	30.8
Pupils enrolled	125	127157.56	137216.813	1978	78218	831044
Life expectancy	104	81.452	3.002	73.7	82.3	85.8
Pop. density	125	457.937	1106.534	3.4	118.3	7526.7
Tot. population	141	2541890.894	2784562.194	29884	1611621	17947221
Death rate	125	1038.286	260.940	286.69	1017.94	1998.32
Total deaths	141	26201.738	28023.403	266	16658	206479
Resp. diseases discharges	111	31544.640	26875.601	613.0	24310.0	123067.0
Long-term beds	92	543.871	987.408	0.0	177.97	6320.29
Health person.	118	8814.068	8790.370	222.0	5726.0	47481.0
Hospital beds	113	544.996	403.749	136.47	388.16	2688.020
Air passengers	114	11173.912	18187.845	3.0	3301.0	107991.0
Vehicles	120	1315559.383	1304883.749	0	949033	6967861
Farm lab. force	142	44581.127	53596.199	0	24080	279380
Util. agr. area	142	920012.394	1011654.138	0	592415	5295680
Unempl. rate	124	8.351	5.310	2.2	6.75	27.8
Worked hours	127	1544867.486	1538892.669	28751.89	1083673.4	9742156.29
GVA growth	125	109.002	6.262	94.7	108.7	137.6
Compensation of employees	125	27992.660	41466.651	719.57	15832.0	382297.39
GDP	125	61514.957	85544.500	1358.82	35255.67	742569.25

Table 1.1: Summary statistics of the data.

Third, we verify the linear correlation score among the features, and between the features and the responses. From Figure 1.5 we can note that, in general, there are few couples of factors that show significant linear correlation, mostly positive ans intuitive (such as total population and total number of deaths), but there is no strong linear correlation between the response of any wave and any feature. This can be considered in favor that there must be a combination of interacting factors that can explain the phenomenon, with more complex dynamics.

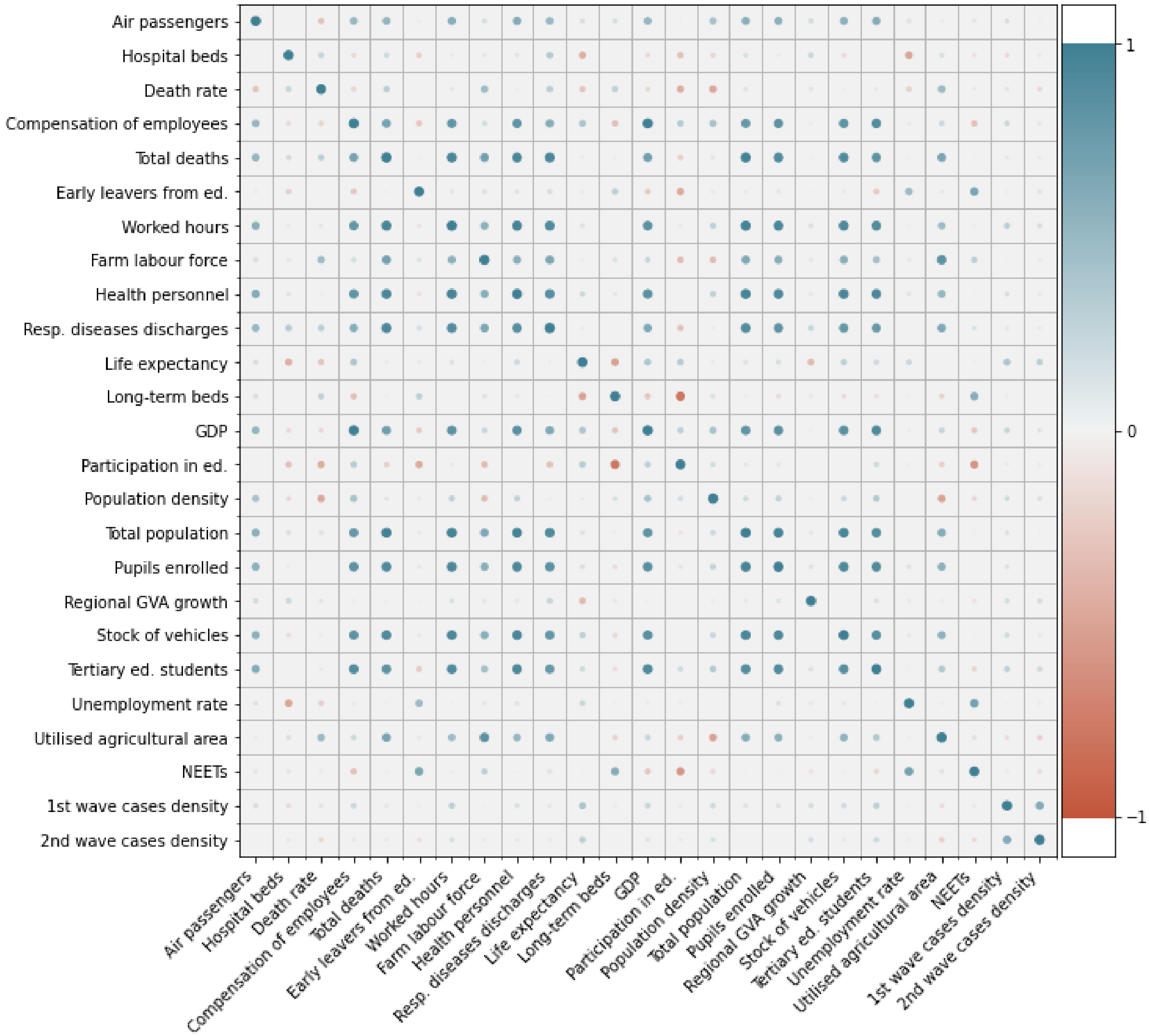


Figure 1.5: Correlation heat map of the data.

In addition, we keep in mind the amount of missing data in the features and response, where 3 regions have missing values for their case density (see Figure 1.6). However, the maximum amount of missing values for a feature reaches 36.1% (regions missing the value of the given feature).

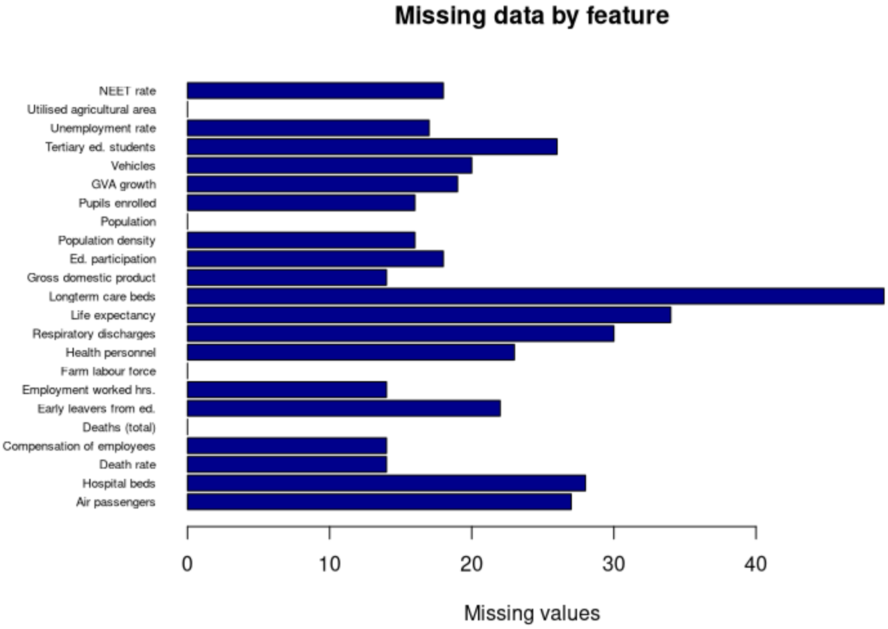


Figure 1.6: Missing values by features.

Making the analysis by regions instead of features, we note that 90 regions have at least 1 feature with a missing value, with its maximum being 96.67% of its data. We find that the 3 regions that have a missing value for their response were among the ones with most of their data missing. Later on, when discarding these regions for the analysis, we note that the maximum amount of missing data in a region reduces to 60.0%.

We also note that, in general, the lack of data follows from a lack of granularity by country, given that the "NUTS2" division fits well the division by regions present in Italy, but is less intuitive for german regions, which have a regional division that fits well the "NUTS1" logic, which is less granular. Hence, we present the maximum amount of missing features of the regions from each country, in order to visualize where the information is lacking most (see Figure 1.7). In this case, we note that there are some countries that have their regions with at most 1 feature missing (such as Bulgaria, Croatia, Denmark, Romania and Spain), while the largest amount corresponds to regions in Germany (18), followed by France (15) and Belgium (11).

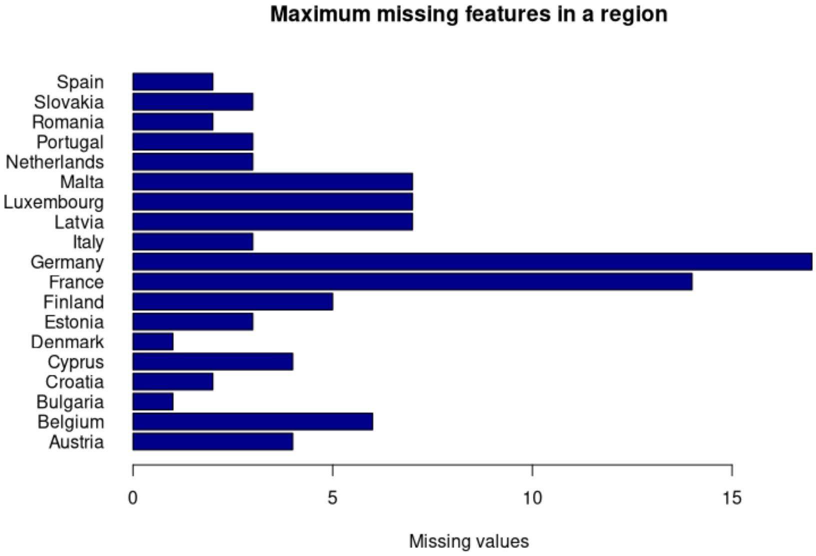


Figure 1.7: Maximum missing values by region.

Another thing to observe are the amount of regions by country that will be taken into analysis. When discarding the regions without information about the response, we can see the following amount of regions considered by country. In particular, we have some countries accounted by just one region (Cyprus, Estonia, Latvia, Luxembourg and Malta), while Italy has observations for all its regions, which are the largest amount (21), along with Spain (19), France (18) and Germany (16), that follow in number.

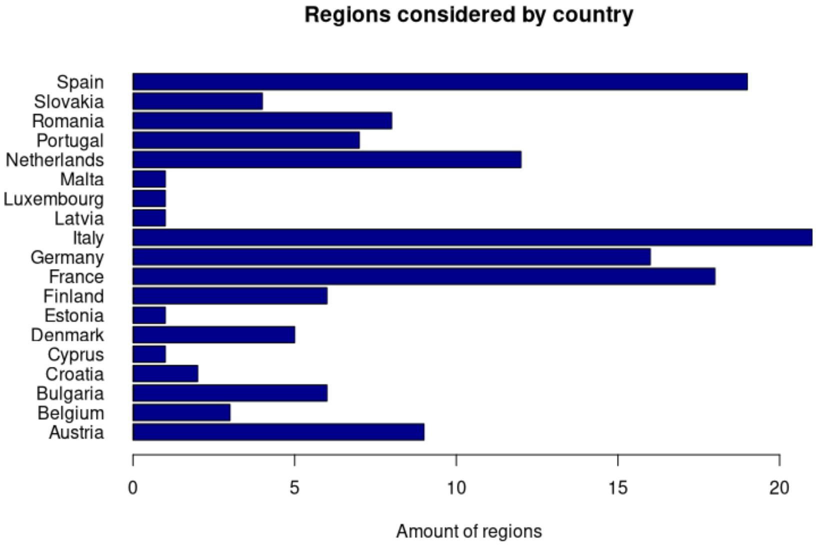


Figure 1.8: Amount of regions by country.

2 | Association rule mining analysis

2.1. Data pre-processing

In order to dichotomize the data, we formulate couples of target-features, with the following proposed transformations for each:

1. The target variable, which is the case density of each wave by region, is proposed to be dichotomized in three possible ways:

    (a) Classified as either above the mean or under the mean, hence considering when the density of cases is high relative to the total average of the data as value 1, and 0 otherwise.

    (b) Classified as either above the median or under the median, hence considering when the density of cases is above the 50% quantile as value 1, and 0 otherwise.

    (c) Classified as either above the 90% quantile or under the 90% quantile, hence considering when the density of cases is exceptionally high relative to the 90% quantile as value 1, and 0 otherwise.

2. The factors, considering the social-demographic data, is proposed to be dichotomized in three possible ways:

    (a) Classified as either above the median or under the median, hence considering when each feature is above the 50% quantile with respect to the column as value 1, and 0 otherwise.

    (b) Classified as either out of the IQR or inside the IQR, hence considering when each feature (column) is outside of the inter-quantile range (IQR) with respect to the column as value 1, and value 0 when it is inside it.

    (c) Classified as either above the IQR or below the IQR, hence considering when each feature (column) is above the 75% quantile with respect to the column or when it is below the 25% quantile. Here we are formulating two columns instead of one, like in previous cases, where the first column contemplates being above the 75% quantile as value 1, and 0 otherwise, while the second column contemplates being below the 25% quantile as value 1, and 0 otherwise.

As we noticed in the section 1.2 of chapter 1, we have 3 regions without data on the response, so they will be discarded from the analysis. We can visualize the resulting division of the data in Figure 2.1, were we highlight that the amount of regions to consider is 141, hence there is no exact division considering the quantiles for cases 2 and 3 of the dichotomization of the response.

2.2. Methodology

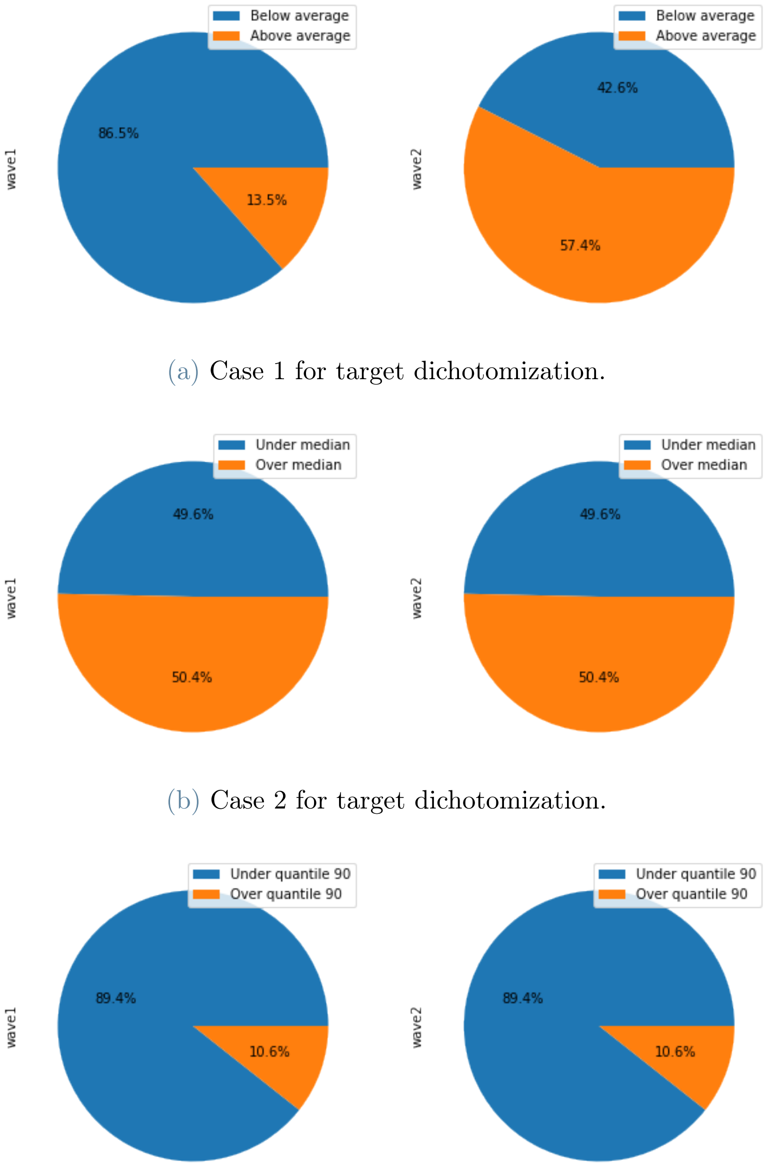
The overall pipeline for factor filtering and grouping is a generalization to the methodology proposed on the thesis of Antonio Esposito [11]. The overall process is described as follows:

1. We evaluate individual significance by performing feature selection through LASSO. Performing an LASSO penalized logistic regression, we select the factors that obtain significant weights, and thereby obtain severity factors with a positive weight, and mildness factors with a negative weight.

2. Depending on the case, we prepare the dataset by adding the complements of the significant factors, hence adding as regional factor if the significant variables present the contrary dichotomic state, with the assumption that they are significant factors for the contrary response.

3. We evaluate significant interactions between selected factors by performing association rules mining and selecting the most significant ones. These interactions are classified as severity rules if they associate with a high case density response, while mildness rules are the ones associated with a low case density response.

It is important to note that, for the first two cases of the features dichotomization, we can formulate its complement, since they are a binary partition of the observations (which we consider a “1-way division”). This means that when adding these complementary columns, the sum of the values for each pair of vectors result in a vector of only ones, which will have a clear interpretation in the association rule mining step. However, for the third case, we note that the complements of each column overlap, since these consider a partition between 3 states (where the middle state is having 0’s in both columns, hence we consider this a “no-middle” formulation). For this reason, we maintain the two columns of the third case as proposed, and no complements are considered for the association rule mining step.



(a) Case 1 for target dichotomization.

(b) Case 2 for target dichotomization.

(c) Case 3 for target dichotomization.

Figure 2.1: Data division for each dichotomization case for the response.

The methodology is applied differently considering the stated observation, where the "1-way" cases add the complementary column of the significant features for the AR mining stage, while "no-middle" cases do not go through this intermediate step. The scenarios to evaluate can be illustrated as in Figure 2.2.

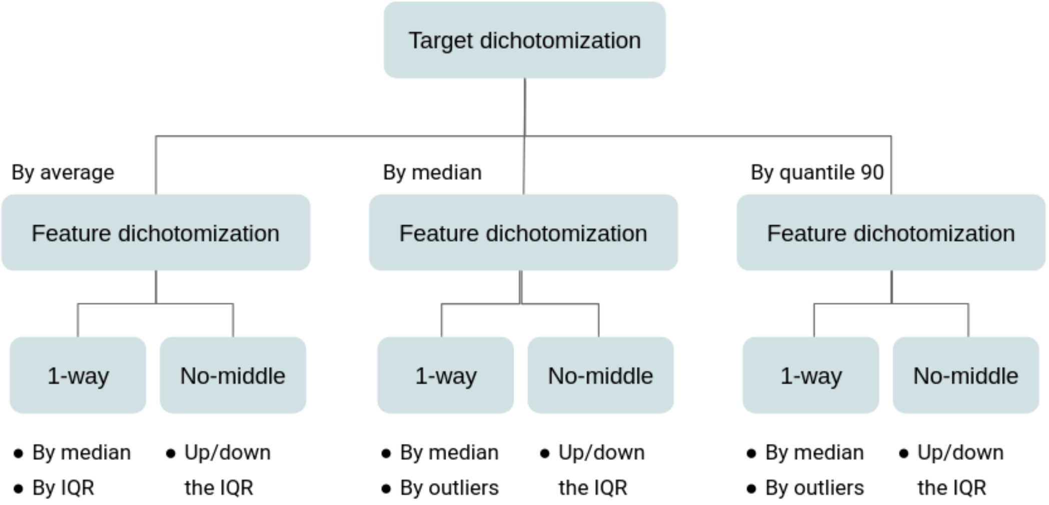


Figure 2.2: Cases diagram for the data pre-processing proposals.

2.2.1. LASSO selection

Lasso l1 regularized logistic regression is often used as an embedded feature selection for classification problems. Its main advantage is that shrinks the weights assigned to irrelevant features to 0, and it has been shown to provide strong generalization efficiency in the presence of many unrelated features. Considering m observations where $y^{(i)} \in \{0,1\}$ the dichotomic response and $x^{(i)} \in \{0,1\}^n$ the respective covariates vector, logistic regression models the probability distribution of the class mark $y^{(i)}$ as in equation 2.1.

$$p(y^{(i)} = 1|x^{(i)}; \Theta) = \sigma(\Theta^T x^{(i)}) = \frac{1}{1 + exp\{-\Theta^T x^{(i)}\}} \tag{2.1}$$

The Maximum a Posteriori estimate of the vector of coefficients of the logistic regression model Θ with a LASSO penalization $\lambda > 0$ is obtained by the optimization problem stated in 2.2.

$$\min_{\Theta \in \mathbb{R}^n} \sum_{i=1}^{m} -log(p(y^{(i)}|x^{(i)}; \Theta)) + \lambda|\Theta|_{l1} \tag{2.2}$$

First, in order to find the optimal penalization parameter, we tune it providing a grid of possible values and evaluate the performance of the feature selection using the average ROC-AUC score obtained through 10-fold cross validation. The ROC-AUC curve is obtained formulating a confusion matrix that groups group true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN), to later plot the True

Positive Rate (TPR, in equation 2.3a) against the False Positive Rate (FPR, in equation 2.3b), which gives a score between 0 and 1, where 1 is the best division of the data.

$$TPR = \frac{TP}{TP + FN}, \tag{2.3a}$$

$$FPR = \frac{FP}{FP + TN} \tag{2.3b}$$

Hence, we look for the $\lambda^* \in \{\lambda_i\}_{i \in \mathcal{I}}$ among the provided values of the grid search where the performance is closer to 1. The values provided are 76 equally spaced values between 10^{-3} and 10^2.

Second, we proceed to the selection of features, where we perform the penalized logistic regression using the chosen parameter, and select the factors with absolute weights bigger than 0.000001. This way, we consider only significant weights, where positive weights are considered as individual severity factors (associated with the target response 1 of having a high density of cases in a specific wave), or individual mildness factors (associated with the target response 0 of having a high density of cases in a specific wave, hence having a low density).

2.2.2. Association rules mining

Association rules mining is one of the most popular and well studied data mining techniques, which aims to extract interesting correlations, frequent patterns, associations, or casual structures within sets of items in transactional databases. Here, we perform class association rules mining, where we consider the association between a specific class item (the target, so the item is present where the response has value 1) and the rest of the items (the rest of the items present, hence all factors with value 1) within all transactions, which here are the covariates data vectors.

We are looking for associations that appear frequently enough and are strong enough, hence we take into account 2 main measures, which are support (associated with frequency) and confidence (associated with strength), that go over the observations counting the appearances of X an antecedent (that can be a set of 1 or more items) and Y a consequent (in our case, the response item). The measures are defined in equations 2.4a and 2.4b, where $\#\{X\}$ is the count of observations that include the item (or set) X and Ω is the total amount of observations.

$$s(X \cup Y) = \frac{\#\{X \cup Y\}}{\Omega}, \tag{2.4a}$$

$$c(X \cup Y) = \frac{s(X \cup Y)}{s(X)} \tag{2.4b}$$

In order to find significant groups of interacting factors, or association rules, we perform the following steps:

1. Determine the minimum support and confidence, in order to formulate all possible rules that are frequent and correlated enough. For this step, we perform a grid search providing different support-confidence pairs, and assess the obtained rules by performing an exact Fischer test on each, and computing the median of their p-value as a statistic for the entire group. Then we select the support-confidence pair with the lowest median p-value, in order to explore the group that should have the rules of most interest.

2. Rule mining with selected parameters, using the apriori algorithm. After mining the rules, we perform an exact Fischer test for each rule, and filter out the ones that surpass a maximum p-value threshold, which is obtained by performing the Bonferroni correction technique. The multiple testing of the rules is performed by the direct approach.

The apriori algorithm makes the identification of frequent itemsets and the generation of the rules more efficient by exploiting the anti-monotone properties of support and confidence. Apriori takes into consideration the itemsets identified as frequent at the previous pass, and by doing so it results in less generated itemsets, since if an itemset is frequent, then its subset must be frequent too, as states the property of support in 2.5.

$$\forall X, Y: \ X \subseteq Y \Rightarrow s(X) \geq s(Y) \tag{2.5}$$

Then, when all the frequent itemsets have been identified, the anti-monotone property of confidence can be used in order to reduce the number of candidate rules to be generated, given the property stated in 2.6, where it considers singular items A, B, C, D, but can be extended to any amount of items superior than 2 with the same logic.

$$\forall A, B, C, D: \ c(ABC \Rightarrow D) \geq c(AB \Rightarrow CD) \geq c(A \Rightarrow BCD) \tag{2.6}$$

Fisher's exact test is used to determine whether two categorical variables have nonrandom associations. It is used especially when the sample size is small, and it is based on building a contingency table, using Ω samples of two categorical variables X and Y, which under the null hypothesis of no association, the probability of obtaining the observed frequencies is as in equation 2.7 [15].

$$P_{cut} = \frac{Y!Y^C!X!X^C!}{\Omega!(Y \cap X)!(Y^C \cap X)!(Y \cap X^C)!(Y^C \cap X^C)!} \tag{2.7}$$

Bonferroni correction technique is used for dealing with multiple testing for Association Rules, the correction sets the significance level threshold to $\frac{\alpha}{n}$ with n the amount of tests to perform [16]. It is a conservative technique, hence the rule filtering is significant.

The most common and successful approaches used for multiple testing for association rules [20] are presented:

- The Direct Approach applies the correction for multiple testing directly on the extracted rules, by looking at the same set of transactions used for mining. A correction that is often used is Bonferroni correction. In this case, the α threshold is divided by the number of rules to be tested.

- The Permutation Based Approach recalculates the p-value of the rules by randomly shuffling the class labels of the data. Because the random shuffling breaks the link between patterns and class labels, the distribution of recalculated p-values is a close approximation of the null distribution, in which both sides of the rules are independent. The permutation-based approach preserves patterns' interactions, allowing it to find a more precise cut-off p-value threshold than the direct adjustment method. The permutation-based approach, on the other hand, is very expensive, having to generate the possible permutations [20].

- The Holdout Approach aims to address the shortcomings of the previous two approaches. It divides a dataset into two: an exploratory and an evaluation dataset. The exploratory dataset is used to extract association rules first. The set of rules with a p-value of less than are then validated using the evaluation dataset. The p-value of the rules on the evaluation dataset is adjusted using Bonferroni correction to control false positives at level α, but now the number of tests is the number of rules with a p-value no larger than α on the exploratory dataset, which is usually a number orders of magnitude smaller than the number of all rules mined, thus the holdout approach is expected to have a better chance of discovering rules with a moderately low p-value. The partitioning of the dataset may have an impact on the holdout approach's performance. If a rule only appears in the exploratory dataset or the evaluation dataset, it will not be discovered. On the other hand, it becomes harder for noise rules to turn out significant [29].

As mentioned at the beginning of the section, we choose the direct approach to perform multiple testing. This is due to the country hierarchy present in the data, which is a reason to discard having separate training data that even by shuffling could be missing country-level information.

2.2.3. Output evaluation

Subject analysis

After selecting the significant rules, we use as predictor of high risk that the subject satisfies one or more rules. According to this, we formulate a confusion matrix, in order to group true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN). As performance indexes, we use the metrics of precision (defined in equation 2.8a), recall (defined in equation 2.8b), accuracy (defined in equation 2.8c), and the F1 measure (defined in equation 2.8d).

$$Precision = \frac{TP}{TP + FP}, \tag{2.8a}$$

$$Recall = \frac{TP}{TP + FN}, \tag{2.8b}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}, \tag{2.8c}$$

$$F1 - measure = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{2.8d}$$

Feature analysis

After selecting the relevant features through LASSO selection and formulating the rules through AR mining, we compare the feature weights from the LASSO logistic regression model with the placing of said feature in the association rules. This means we are comparing the individual estimated risk type from the regression model and the estimated risk type of the feature within the group of each rule.

This assessment is represented as confusion matrices for a global overview, by either considering the counts of all the repetitions of the features, or counting each feature only once. Most generally, it is first assessed graphically by highlighting the features that have an opposite effect between its weight and the type of rule its found in, since it makes it able to detect the specific feature that presents a mismatch. In our case, we mark in bold the features that present a mismatch in the listing of the rule, and explicit it when presenting the confusion matrix.

2.3. Results

The LASSO selection procedure is done separately for each wave, hence obtaining a different penalization parameter and a different set of significant factors for each. The filtering behaves differently for each wave, since regional dynamics can change the relevance of the regional factors considered.

Here we present the case of the target with the first choice of dichotomization (using the mean) and the features with the second choice of dichotomization (using the inter-quantile range), since it shows the best set of performance measures during the subject analysis evaluation, which is displayed in Table 2.3. The rest of the results for each case are presented in Appendix A, along with the comparison of all metrics between the considered cases (some of them are discarded beforehand due to the lack of significant rules for one or both waves).

We obtain the following regarding this first step

- For the grid search results of the LASSO penalization parameter, see Figures 2.3 and 2.4. We show the performance of the model in terms of the ROC AUC score, for each penalization parameter candidate. The chosen parameter and its associated score are highlighted in red.

- For the feature selection for each wave, see Figures 2.5 and 2.6. Here we consider the weights of the regression model where a positive score classifies the feature as a risk factor (associated with response value 1, high density of cases), and a negative score classifies the feature as a mildness factor (associated with response value 0, low density of cases).

- A general summary of the process can be seen in Table 2.1. We notice that the second wave, with more sparse values for case density, takes into account more relevant factors to explain the response, even when the penalization parameter is higher.

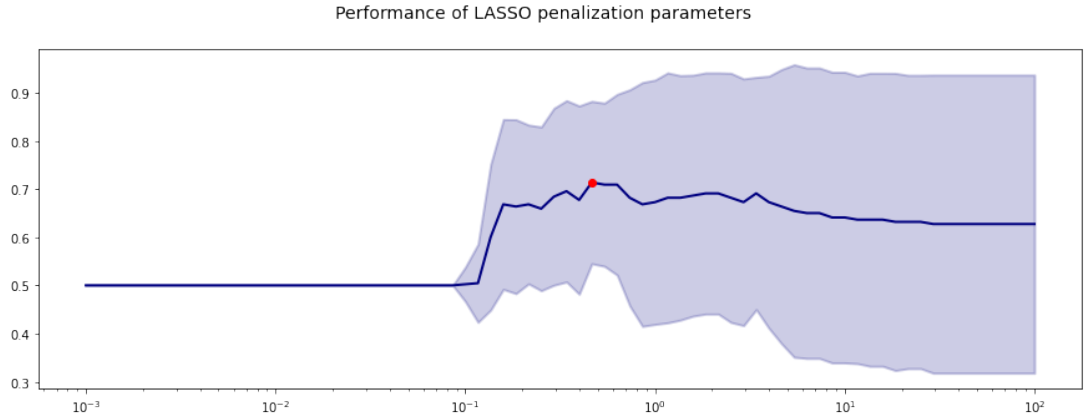


Figure 2.3: Chosen λ in grid search for wave 1.

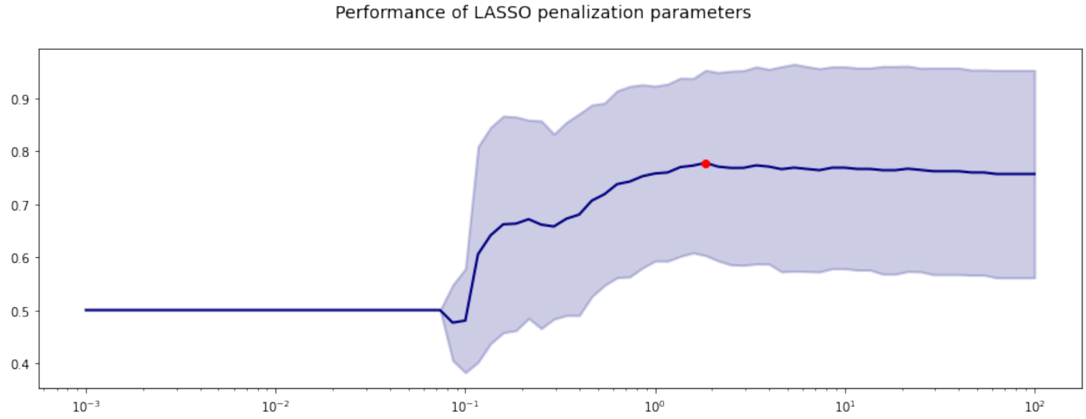


Figure 2.4: Chosen λ in grid search for wave 2.

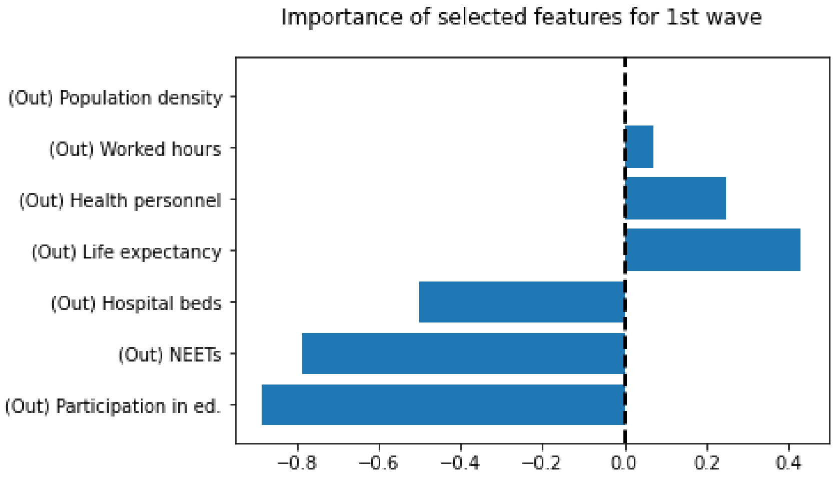


Figure 2.5: Weights of relevant features selected by LASSO in wave 1

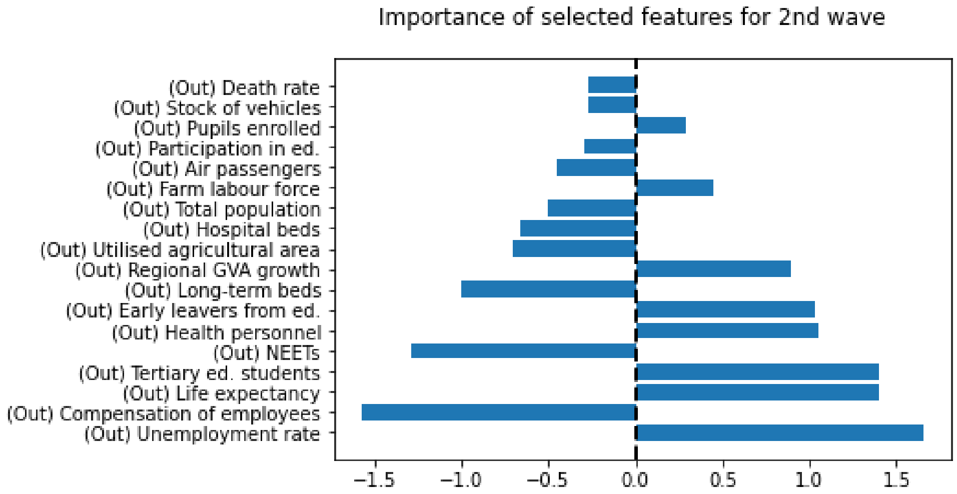


Figure 2.6: Weights of relevant features selected by LASSO in wave 1.

			Selected features	
	λ	**ROC AUC**	**Positive**	**Negative**
Wave 1	0.4642	0.7136	3	2
Wave 2	1.8478	0.7779	7	11

Table 2.1: LASSO feature selection summary.

As stated in section 2.1, the AR mining step considers the adding of complementary columns for the selected features. This way, all regions have always the same amount of items, where a single item can refer to the factor being outside the IQR or inside it.

The summary of statistics of the rules (for each type and each wave) can be seen in Table 2.2. Further details of the conformation of the rules are listed below, along with the description of the individual rules in a graphical format (see Fig. 2.7 and 2.8). We note that, given that the second wave has a more balanced amount of high and low cases density, the minimum support for the rules search is the same, and the case is opposite for the first wave, since severity rules are associated to higher-than-mean cases that are fewer.

- Wave 1

  1. life expectancy (outside IQR), population density (outside IQR), participation in education and training (inside IQR), young people neither in employment nor in education and training (inside IQR) → higher than average cases

	Wave 1		**Wave 2**	
	Severity	Mildness	Severity	Mildness
Amount	3	6	4	1
Support	0.0278 (min) 0.0347 (max)	0.2014 (min) 0.2986 (max)	0.2014 (min) 0.2153 (max)	0.2222
Confidence	0.625 (min) 0.7143 (max)	0.9767 (min) 1.0 (max)	0.8056 (min) 0.8529 (max)	0.7805
P-value	4.259e-04 (min) 2.829e-03 (max)	1.219e-03 (min) 4.499e-03 (max)	5.62e-05 (min) 5.009e-04 (max)	2.528e-08
Factors	4 (all)	2 (5) and 3 (1)	2 (all)	3 (all)

Table 2.2: AR mining results summary.

2. life expectancy (outside IQR), population density (outside IQR), young people neither in employment nor in education and training (inside IQR), available hospital beds (inside IQR) → higher than average cases

3. health personnel (outside IQR), population density (outside IQR), young people neither in employment nor in education and training (inside IQR), available hospital beds (inside IQR) → higher than average cases

4. participation in education and training (outside IQR), young people neither in employment nor in education and training (outside IQR) → lower than average cases

5. participation in education and training (outside IQR), employment thousands hours worked (inside IQR) → lower than average cases

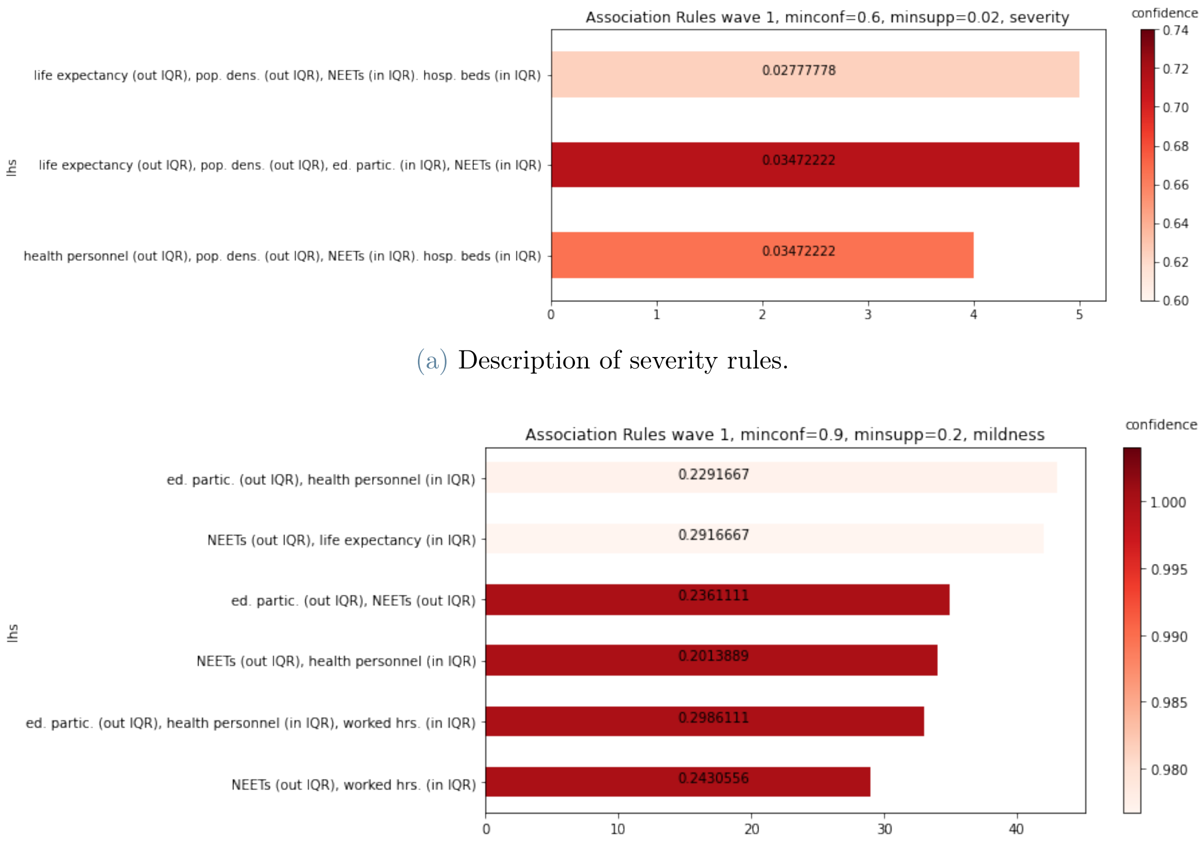
6. health personnel (inside IQR), young people neither in employment nor in education and training (outside IQR) → lower than average cases

7. life expectancy (inside IQR), young people neither in employment nor in education and training (outside IQR) → lower than average cases

8. participation in education and training (outside IQR), health personnel (inside IQR) → lower than average cases

9. participation in education and training (outside IQR), health personnel (inside IQR), employment thousands hours worked (inside IQR) → lower than average cases

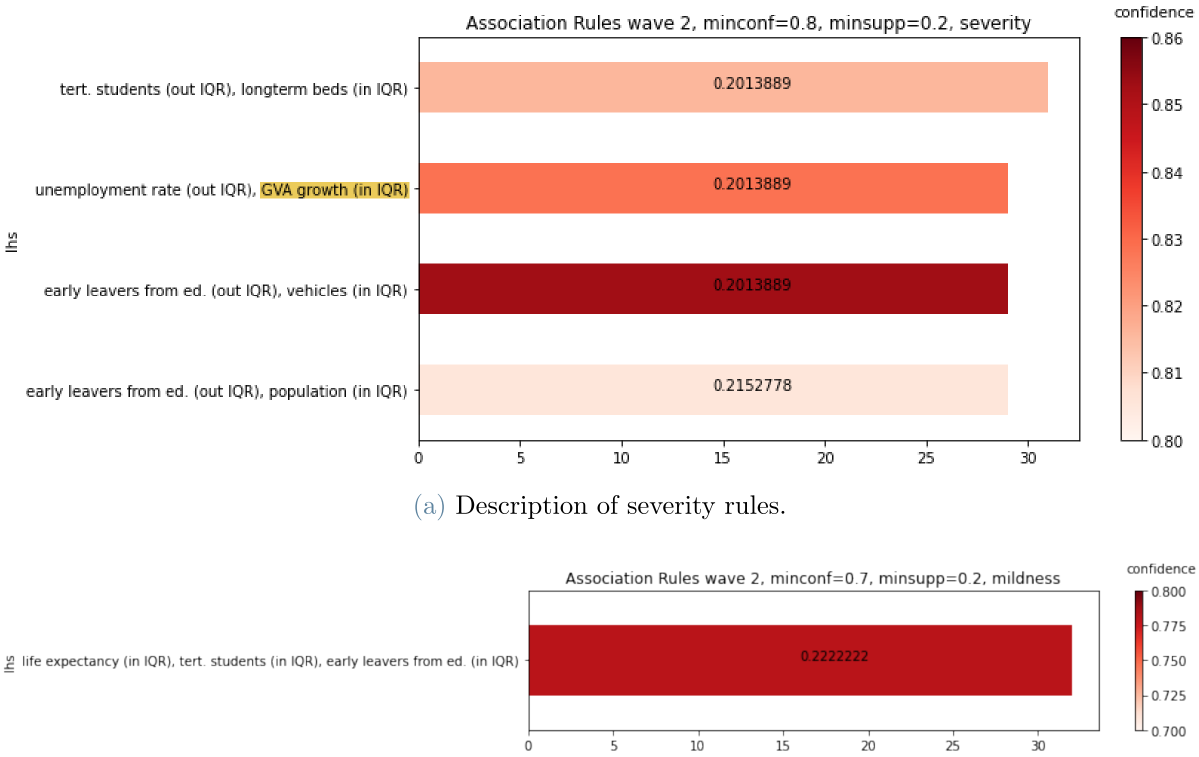
- Wave 2
    1. students enrolled in tertiary education (outside IQR), longterm care beds per hundred thousand (inside IQR) → higher than average cases
    2. early leavers from education and training (outside IQR), population (inside IQR) → higher than average cases
    3. early leavers from education and training (outside IQR), stock of vehicles (inside IQR) → higher than average cases
    4. unemployment rate (outside IQR), **real growth rate of regional GVA** (inside IQR) → higher than average cases
    5. life expectancy (inside IQR), students enrolled in tertiary education (inside IQR), early leavers from education and training (inside IQR) → lower than average cases



(a) Description of severity rules.

(b) Description of mildness rules.

Figure 2.7: Description of association rules for the first wave. We show the support count of the rules (number of regions for whom the rules hold) as bar width, with the support proportion inside the bar, and the confidence encoded by the color of the bar.



(a) Description of severity rules.

(b) Description of mildness rules.

Figure 2.8: Description of association rules for the second wave. We show the support count of the rules (number of regions for whom the rules hold) as bar width, with the support proportion inside the bar, and the confidence encoded by the color of the bar.

From these rules, the subject analysis are done for each type of rule for each wave, see Table 2.3 for a general summary, and the detail in Tables 2.4 and 2.5, where we see a generally good performance, except for the recall of the severity rules in wave 1, that is below 50%. Notice that for the mildness rules, satisfying 1 or more rules should predict a low amount of cases, while for severity cases, satisfying 1 or more rules should predict a high amount of cases. In particular, we note the difference between high case density regions and low case density regions by comparing the average amount of rules of each type they satisfy, and notice that it is coherent with the criteria, see Table 2.6. Additionally, we notice the average is generally low, which also suggests that different rules include different regions, rather than overlapping the same regions among many rules.

Feature analysis shows that most of the features coincide between their individual weight from the regression model and the type of rule they belong to (see Tables 2.7 and 2.8), with the exception of the real growth rate of regional GVA in the second wave, that the regression model classifies as a risk factor when it is outside the IQR, but participates in a severity rule when it is inside the IQR.

	Wave 1		Wave 2	
	Mildness	Severity	Mildness	Severity
Precision	97.47%	60.00%	78.05%	77.11%
Recall	63.11%	47.37%	53.33%	79.01%
Accuracy	67.36%	88.89%	74.31%	75.00%
F1-measure	76.62%	52.94%	63.37%	78.05%

Table 2.3: Subject analysis summary.

	Mildness		Severity	
	1+ rules	0 rules	1+ rules	0 rules
High	2	20	9	10
Low	77	45	6	119

Table 2.4: Subject analysis for wave 1.

	Mildness		Severity	
	1+ rules	0 rules	1+ rules	0 rules
High	9	75	64	17
Low	32	28	19	44

Table 2.5: Subject analysis for wave 2.

	Wave 1		Wave 2	
	Severity	Mildness	Severity	Mildness
High	0.737	0.091	1.457	0.107
Low	0.056	1.770	0.397	0.533

Table 2.6: Average satisfied rules.

2.4. Discussion

Velavan and Meyer [27], with Aabed and Lashin [1], had suggested to consider population density as a possible factor, where Ciotti et. al. [9] mention that a higher population reduces the possibility of maintaining safer distances to avoid the virus spread. Moreover, Roy and Ghosh [24] underlined that population density appears to be among the most

	Wave 1		Wave 2	
	Severity	Mildness	Severity	Mildness
Severity	12	0	7	1
Mildness	0	13	0	3

Table 2.7: Feature analysis counting all repetitions.

	Wave 1		Wave 2	
	Severity	Mildness	Severity	Mildness
Severity	6	0	6	1
Mildness	0	5	0	3

Table 2.8: Feature analysis counting unique appearances.

relevant factors. Here we note the following,

- Wave 1
  1. It is considered as a risk factor if the region is outside the IQR regarding population density.
  2. This factor is present in all severity rules.
  3. Severity rules for the first wave have a not very high precision and a low recall.
  4. Since this dichotomization for the features consider the half of the data with either lowest or highest values, we could consider that this effect is better understood with another dichotomization closer to what is proposed on literature.
  5. We note that in case 1-3 (see Annex A), population density is considered a risk factor when above quantile 75, for the only significant rule found, which has a higher precision, which supports the previous statement.
- Wave 2
  1. It is considered a risk factor if the region is inside the IQR regarding its overall population.
  2. This factor is present in only one of the four significant rules of this type.
  3. This rule alone is satisfied by almost half of high case density regions and a very low amount of low density regions.

4. We note that in other cases with a different feature dichotomization, there are no significant rules that involve this type of factor.
5. For this wave, we could consider that there is a more complex relation with this factor, possibly to sanitary measures taken by each region.

Farseev et al. [12] found that in EU regions, COVID-19 cases are positively correlated with factors usually associated with modern developed economies, such as high health system maturity. Allel et. al. [3] also stated that, among other factors, healthcare resources are to be considered relevant. Here we notice that health system factors are significant:

- Wave 1
  1. It is considered as risk factor to be inside the IQR regarding the amount of available hospital beds.
  2. This factor is in two of the three severity rules, and not present in any significant rule for other feature dichotomization case.
  3. It is also considered a risk factor to be outside the IQR regarding the amount of health personnel.
  4. It is considered a mildness factor for the region to be inside the IQR regarding the amount of health personnel .
  5. This factor is in only one of the three severity rules, but in half of the mildness rules, where precision and recall are higher, and they are not present in any significant rule for other feature dichotomization case.
  6. We can propose this effect for further analysis with literature, since they are considered relevant in these results.
- Wave 2
  1. The relevant factor in this area is the amount of long-term care beds per hundred thousand, where being inside the IQR with respect to the other regions is a risk factor available hospital beds, health personnel.
  2. This factor is present in one severity rule of 4.
  3. In case 1-1, it is considered a risk factor to be under the quantile 50 with respect to the amount of long-term care beds, in only one significant rule of 26.

4. This factor could be considered still a relatively weak relation with the phenomenon in this second wave, but can also be considered that it follows intuition from the literature.

An additional finding from these authors is a relevant positive correlation with older people and the illnesses associated with this demographic group. Allel et. al. [3] argues that high life expectancy increases the risk of infection, while Roy and Ghosh [24] confirmed that older populations tend to be directly correlated to the spread of the virus, and Kumar et. al. [19] also confirmed that old age is one of the main factors correlated to the spread of COVID-19. We notice in out results that life expectancy is always present as a factor for either wave, more in detail,

- Wave 1
    1. It is considered a mildness factor to be inside the IQR with respect to the rest of the regions.
    2. It is also considered a risk factor to be outside the IQR with respect to the rest of the regions.
    3. This factor appears in two of the three severity rules, but one of the six mildness rules.
    4. For case 1-1, it is considered a mildness factor to be below quantile 50, in half of the significant rules of this type and wave, where these present a prefect precision, but low performance for all other measures.
    5. For case 1-3, it is considered a risk factor to be below quantile 25, in the only significant severity rule found, presenting a similar precision from our case, but very low recall and F1-measure.
    6. We note inconsistencies among findings for this wave, where most regions present a high life expectancy, and hence rare cases are younger countries, hence we should make further inspections to find clearer views regarding this factor.
- Wave 2
    1. It is considered a mildness factor to be inside the IQR with respect to the rest of the regions.
    2. This factor is present in the only mildness rule found for this wave, that has a fairly high precision, but a recall barely over 50%.

3. In case 1-1, this factor is present in only one of the 26 significant severity rules, where being over quantile 50 is considered a risk factor.
4. We can consider that, even if this factor has a relatively low prevalence for this wave, it follows intuition from the literature, and can also raise research questions when limiting observations to mostly older regions.

Kumar et. al. [19] underline that a high level of economic wellness increases the possibilities for the virus to spread, because the community is more dynamic and people interact more with each other. For these type of factors, we find

- Wave 1
    1. The amount of hours worked is a mildness factor for regions that are inside the IQR.
    2. This factor appears in one third of the significant mildness rules.
    3. This factor does not appear in any other case when changing the factor dichotomization.
    4. We can consider that, even if this factor is not too recurrent, it follows intuition from the literature that not having an extreme amount of hours worked could be associated with a lower regional case density.
- Wave 2
    1. The growth rate of regional GVA is a risk factor for regions that are inside the IQR.
    2. This factor appears only in one of the 4 significant severity rules, while for LASSO it was expected to be the contrary condition to be of risk.
    3. In case 1-1, this factor appears in almost half of the significant mildness rules, where being under quantile 50 is associated with lower risk, and these rules have generally high values for all performance measures.
    4. Given that there is a low prevalence and a different effect from this factor when considered individually and in group, it should be further questioned how relevant is this factor when considering the rest of the covariates.
    5. The unemployment rate is a risk factor for regions that are outside the IQR.
    6. This factor appears only in one of the 4 significant severity rules.

We note that a general trend is to consider factors individually, and so we make some observations for the group dynamics found for each wave, in order to make further research questions.

During the first wave, we note that population and educational factors are recurring risk factors that can be further grouped with risk factors related to healthcare. Further research is proposed in order to detect if this triad of types of factors can characterize specific risk dynamics seen for this wave. In the other hand, we note that complementary educational factors are recurring mildness factors that can be further grouped with mildness factors related to economy, population or healthcare. Further research is proposed in order to detect if these types of factors can characterize specific protection dynamics seen for this wave.

During the second wave, we note that educational factors are recurring risk types that are grouped with risk factors of healthcare, population or economy. We also note that a single severity rule grouping only economical factors. Further research is proposed in order to detect if these three types of factors can characterize specific risk dynamics seen for this wave. In the other hand, there is a single mildness rule that considers a triad of education and population factors. Further research is proposed in order to detect if this pair of types of factors can characterize specific protection dynamics seen for this wave, where economy and healthcare do not seem to intervene.

2.5. Further developments

Since the pipeline is potentially applicable to any dataset with dichotomized data, it is important to highlight the importance of preprocessing the data, in case it needs transformations in order to prepare it as input. As seen in section 2.4, for our application it could be more convenient to perform different transformations to dichotomize the factors, depending on their type or their expected interaction according to what is seen in literature.

As we mentioned in section 2.2, we used the direct approach for multiple testing of the rules, due to the case of our dataset. However, as the methodology is formulated with the aim of becoming a general pipeline for other types of data, we have to highlight that we could either obtain optimistic performance estimates, or filter out too many rules if we start with a large amount. For this reason, the other approaches could perform better when there is no hierarchy over the subjects with a large amount of groups, as in our case with the countries of the regions.

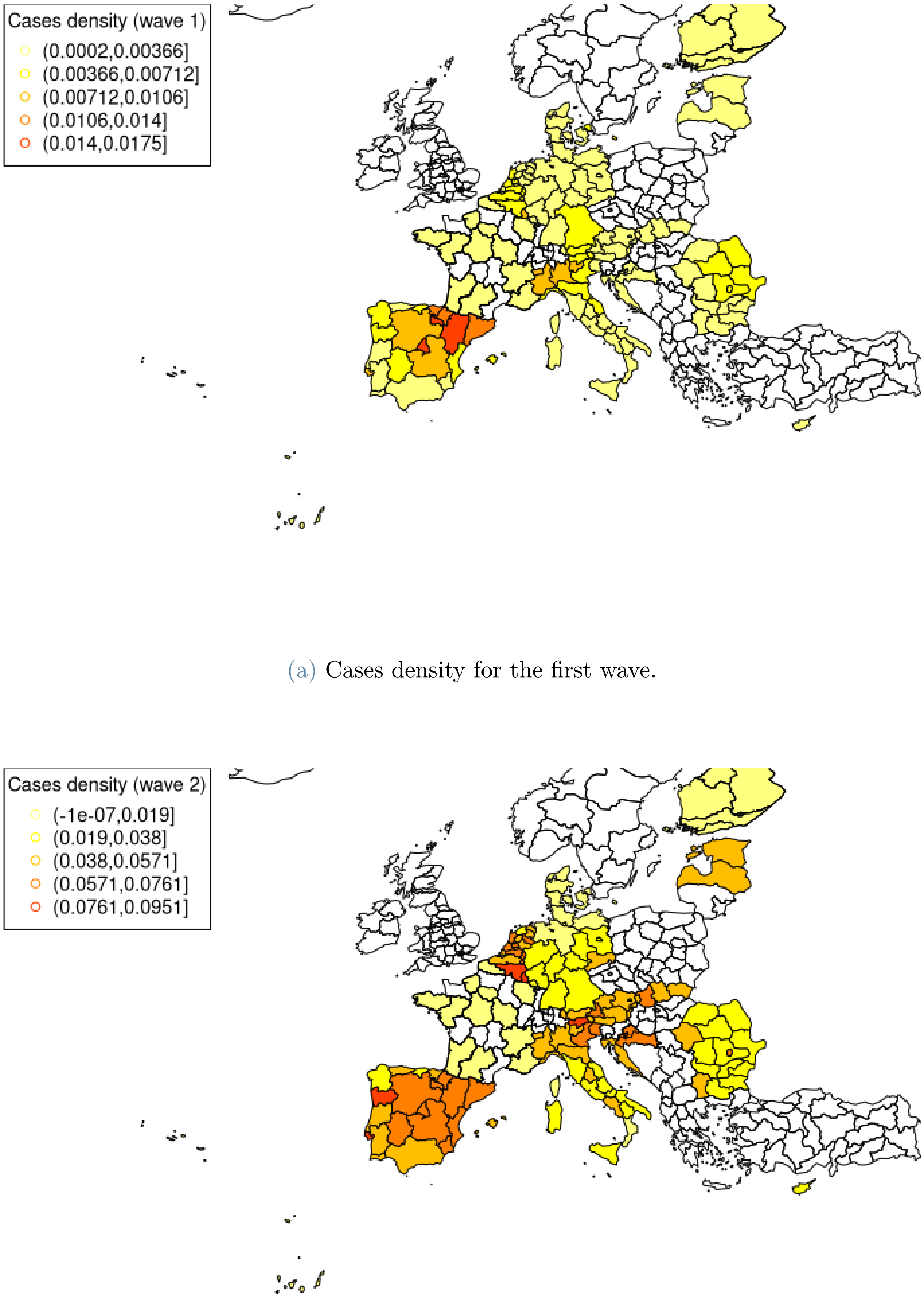
3 | Geo-statistical analysis

3.1. Data pre-processing

As a first step, we merge the current dataset with a locations dataset that contains latitude and longitude from Eurostat [10], in particular, we take the central point of each region, along with the polygons that delimit the entire region, and recover the complete amount completing with records from 2021, 2013 and 2006. Afterwards, we continue to discard the regions that do not have information about the cases density for each wave, which results in the 141 regions to analyze. We can visualize the behavior of the cases densities using geographical plots over Europe (see Figure 3.1), where the response is represented by the color.

As a second step, we perform the following changes in the available data:

1. We fill missing data from Belgium, which has data at the NUTS 1 region level, with data from their NUTS 2 sub-regions, for the following features:
   - Death rate: We multiply the sub-regional death rate with the total population, recovering the total amount of deaths, and divide between the sum of sub-regional deaths and the sum of sub-regional population amounts, recovering a NUTS 1 level death rate.
   - Compensation of employees: We sum the total amounts from the sub-regional amounts, to recover a NUTS 1 level amount of compensation.
   - Life expectancy: We take the average of the life expectancy of the sub-regions, to recover a NUTS 1 level life expectancy.
   - GDP: We take the sum of the sub-regional GDP, to recover a NUTS 1 level GDP.
2. We replace the available data for the GVA growth rate with the updated data from EUROSTAT for 2019, since the original data for this feature did not match current records for any year.



(a) Cases density for the first wave.

(b) Cases density for the second wave.

Figure 3.1: Visualization of the cases density for each region in the map.

As a third step, we fill the missing data considering the following approach:

1. We consider the following approximations:
    - The mean of the known values for the rest of the regions in the country.
    - The national value multiplied by the population proportion of the region.
    - The mean of the known values for the nearest 3 neighbours of the region.
2. For each feature, evaluate the absolute percentage of error of the regions that have known values regarding the usage of each approximation, and formulate an error triad taking the maximum, median and mean statistics.
3. We validate and choose the approach with smallest error triad, that allows partial or total completion of the data for each given feature. When the triad was not completely minimal, the triad with a minimum pair was chosen (usually median and mean, or maximum and mean), with an intermediate third value.
4. The features that refer to total amounts are returned to integers by rounding the approximation results.

Table 3.1 summarizes the chosen method to approximate the data regarding each feature, with its associated validation error triad and the amount of missing values approximated. For the features of total deaths, farm labour force, population, and utilised agricultural area, there were all 136 values, hence no approximation was needed. Instead, there were some cases where the data was not able to be approximated, due to the validation error range being too wide, and hence unreliable approximations were avoided.

In general, there was no trend on the validation errors that favored performing the approximation differently by country, hence the same method was used to approximate the entire column. In addition, validation errors that were extremely high (and hence showing a maximum error extremely high), constituted just few outliers, without a particular trend of belonging to a specific country, so we discard a high probability of an extremely bad approximation in these cases.

By this procedure, we reduce the amount of missing data as we visualize in Fig. 3.2. We notice how significant the change is by seeing that Germany, previously the country in which the maximum amount of missing data on any given region was 18, diminishes to a maximum of 2 missing values in any region, and so the most critical country becomes France, that still has a small amount of maximum 3 missing values in any region.

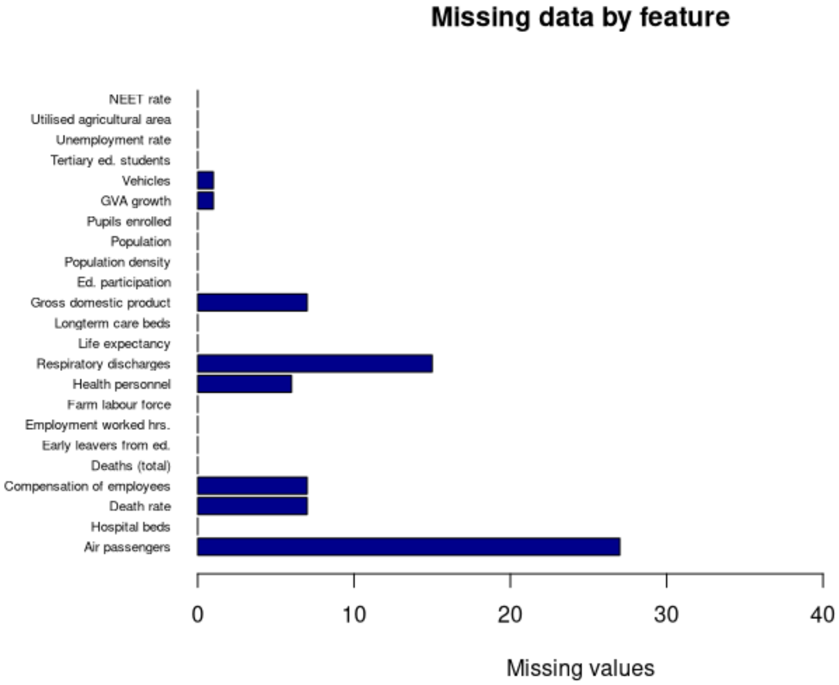
		Errors			
	Method	**Max.**	**Median**	**Mean**	**NAs filled**
Early leavers from ed.	1 - Mean	188.14%	16.71%	26.05%	22 (100%)
Ter. ed. stud.	2 - Population	399.81%	24.00%	50.20%	26 (100%)
NEET rate	1 - Mean	155.13%	18.37%	22.96%	18 (100%)
Particip. in ed.	1 - Mean	111.43%	10.15%	17.84%	18 (100%)
Pupils enrolled	2 - Population	91.08%	84.64%	84.77%	16 (100%)
Life expectancy	1 - Mean	2.84%	0.60%	0.79%	34 (100%)
Pop. density	2 - Population	100.00%	93.33%	85.69%	16 (100%)
Death rate	1 - Mean	52.52%	9.27%	11.95%	7 (50.0%)
Resp. diseases discharges	2 - Population	99.88%	99.78%	99.49%	15 (50.0%)
Long-term beds	3 - KNN	1001.41%	34.82%	61.39%	49 (100%)
Health person.	2 - Population	181.48%	10.18%	17.81%	17 (73.9%)
Hospital beds	3 - KNN	412.10%	15.54%	27.85%	28 (100%)
Air passengers	Not approximated				
Vehicles	2 - Population	335.09%	11.78%	21.19%	19 (95.0%)
Unempl. rate	1 - Mean	143.98%	16.95%	24.96%	17 (100%)
Worked hours	2 - Population	397.45%	8.19%	16.90%	14 (100%)
GVA growth	1 - Mean	11.94%	1.79%	2.48%	18 (94.7%)
Employ. compens.	2 - Population	86.31%	18.91%	24.66%	7 (50.0%)
GDP	2 - Population	69.56%	19.17%	21.80%	7 (50.0%)

Table 3.1: Summary of missing data approximation by feature.

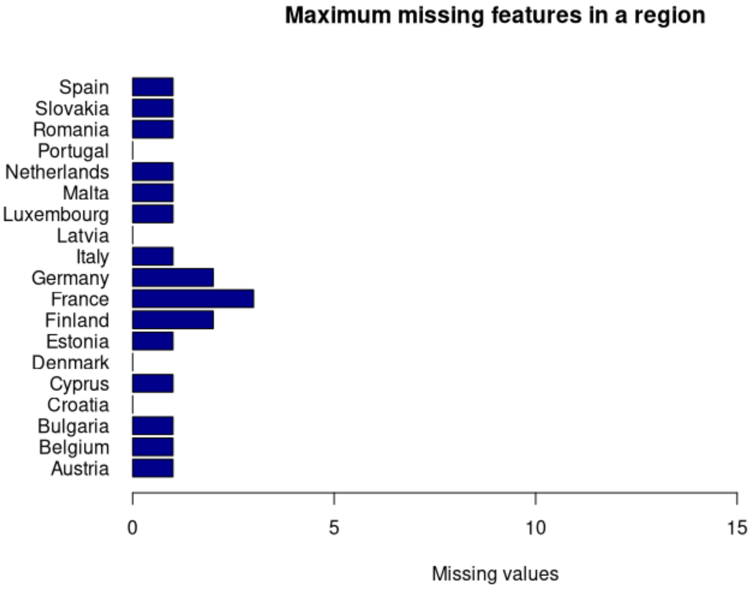
In order to perform proper analysis over the data, missing values can't be admitted. Hence, we work with a subset of the data obtained in the following way:

1. We first discard the column with the largest amount of missing data, since it can't be considered a reliable factor, which is “Air passengers”.
2. We discard all regions that are missing at least one value in the remaining columns.

By performing the mentioned selection, we consider 22 covariates over 112 regions. Figure 3.3 visualizes the cases density of each of them over the european map for each wave of the pandemic, while figure 3.4 shows the regions considered, distributed by country.

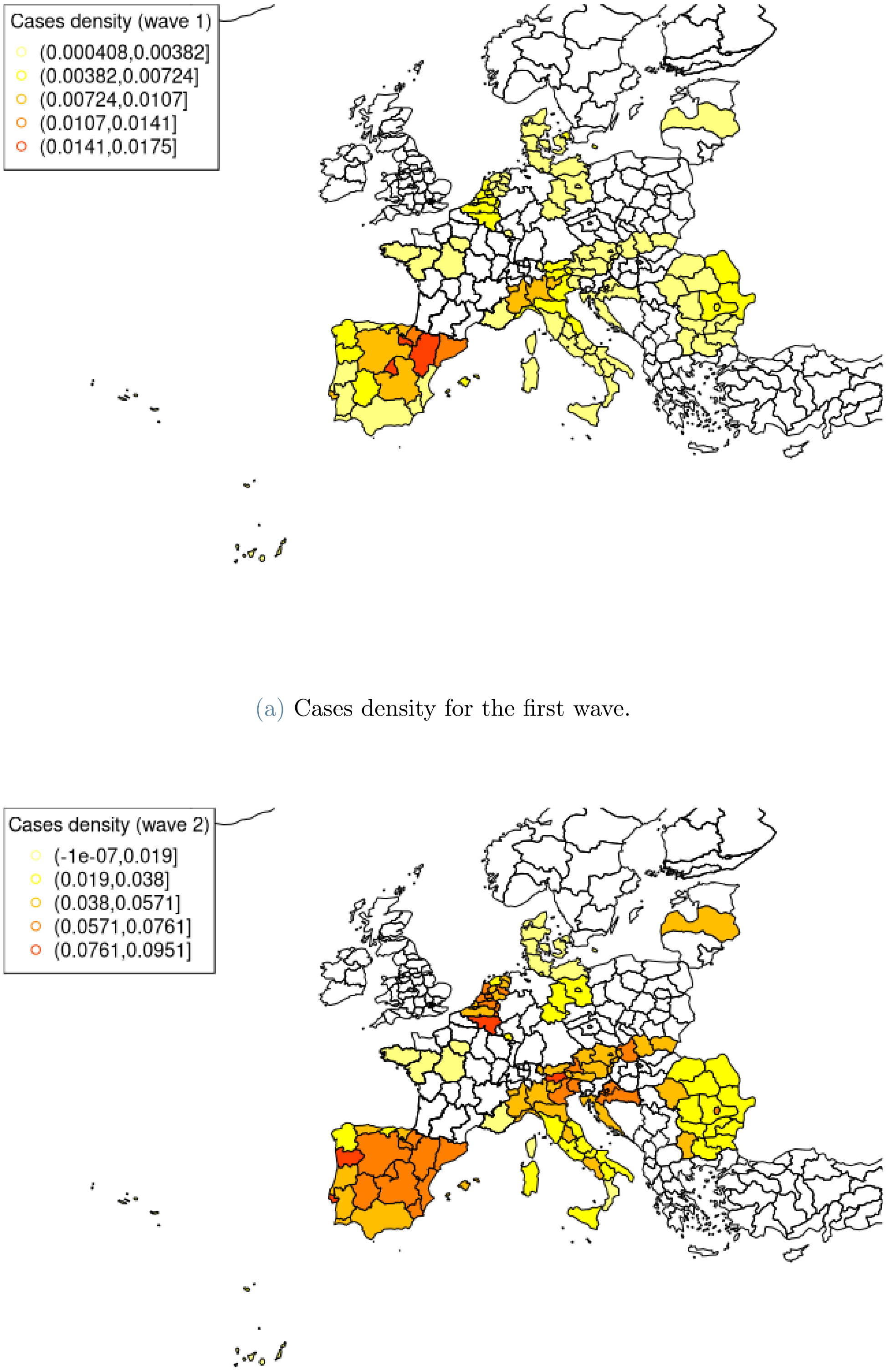


(a) Remaining missing values by feature.



(b) Remaining missing values by critical region of each country.

Figure 3.2: Amount of missing values remaining after filling by approximation.



(a) Cases density for the first wave.

(b) Cases density for the second wave.

Figure 3.3: Visualization of the cases density for each region of interest in the map.

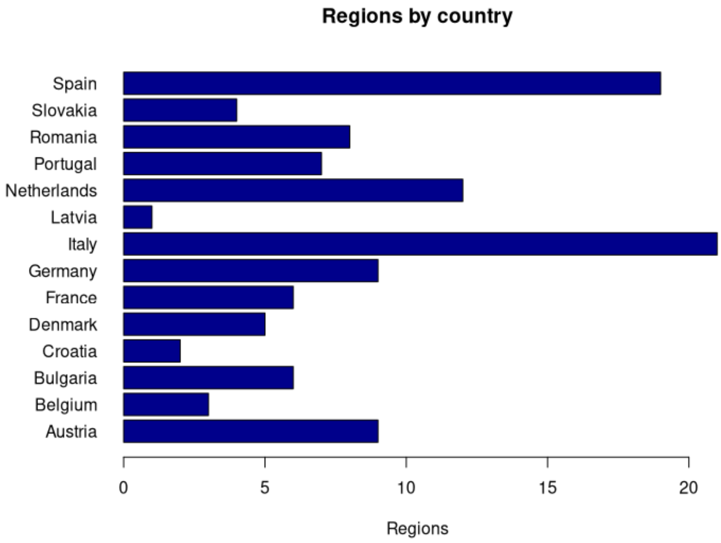


Figure 3.4: Regions considered for analysis by country.

3.2. Methodology

The overall process for the geo-statistical analysis of the data is described as follows:

1. We select a linear regression model transforming the response and performing feature selection through LASSO. Performing an LASSO penalized linear regression, we select the factors that obtain significant weights, and validate the model by verifying its assumptions through the Shapiro-Wilk test. We thereby obtain a multivariate linear model and its corresponding i.i.d. normal residuals for the specific transformation of the response that allows its corresponding model to be valid.

2. We analyze the empirical variogram of the residuals, in order to fit a reasonable model and estimate spatial dependency of the unexplained component of the linear model response.

3. We perform independently a LISA cluster analysis over the cases density, over the transformed response selected, and over the residuals of the formulated multivariate linear models, in order to find and compare spatial dynamics.

The methodology is applied separately for each wave, hence it is allowed for each wave's cases density to have a different transformation that will result in a specific valid multivariate linear model, with different significant factors and amount of residuals.

3.2.1. LASSO selection

Lasso l1 regularized linear regression is often used as an embedded feature selection for linear regression problems in which there are a large amount of covariates. Its main advantage is that shrinks the weights assigned to irrelevant features to 0, and it has been shown to provide strong generalization efficiency in the presence of many unrelated features. Considering m observations where $y^{(i)} \in \mathbb{R}$ the real response and $x^{(i)} \in \mathbb{R}^n$ the respective covariates vector, we compute the Maximum a Posteriori estimate of the vector of coefficients of the linear regression model β with a LASSO penalization $\lambda > 0$ by the optimization problem stated in 3.1.

$$\min_{\beta \in \mathbb{R}^n} \sum_{i=1}^{m} (y^{(i)} - \beta^T x^{(i)})^2 + \lambda |\beta|_{l1} \tag{3.1}$$

First, in order to find the optimal penalization parameter, we tune it providing a grid of possible values and evaluate the performance of the feature selection using the average MSE (mean squared error) obtained through 10-fold cross validation. Hence, we look for the $\lambda^* \in \{\lambda_i\}_{i \in \mathcal{I}}$ among the provided values of the grid search where the error is lowest. The values provided are 100 equally spaced values between 10^{-3} and 10^2. The observations considered are only the ones that do not have missing values in any of the covariates.

Second, we proceed to the selection of features, where we perform the penalized linear regression using the chosen parameter, and select the factors with absolute weights bigger than 0. This way, we consider only significant weights, where positive weights are considered as severity factors (associated with the target response being a higher density of cases in a specific wave), or mildness factors (associated with the target response being a lower density of cases in a specific wave).

As a third step, we proceed to analyze the residuals generated by the model, in order to verify the assumption of i.i.d. normal distributed residuals with the Shapiro-Wilk test. Its null hypothesis states that a sample $\{r_1, ..., r_m\}$ came from a normally distributed population. The test statistic is computed using $r_{(i)}$ the ith-smallest number in the sample $\bar{r}$ the mean of the sample, and a_i the ith coefficient of $a = \frac{c^T V^{-1}}{\|V^{-1} c\|}$, where c is the vector made of the expected values of the order statistics of independent and identically distributed random variables sampled from the standard normal distribution, and V is the covariance matrix of those normal order statistics. Finally, the statistic is stated in 3.2.

$$W = \frac{(\sum_{i=1}^{m} a_i r_{(i)})^2}{\sum_{i=1}^{m} (r_i - \bar{r})^2} \tag{3.2}$$

We validate the model by not rejecting the null hypothesis with confidence $\alpha = 0.1$, meaning we expect the p-value to be outside of critical values, in this case $0.5 < p < 0.95$. For this reason, a specific transformation of the response may be chosen.

We consider the following transformations for the cases density of each wave $y \in [0, 1]^n$:

- Logarithmic transformation: $f(y) = \log_{10}(y + 1)$
- Root transformations: $f(y) = y^p, p \in \frac{1}{3}, \frac{1}{2}$
- Logit transformation: $f(y) = \sigma^{-1}(y) = \ln \frac{y}{1-y}$
- Z transformation: $f(y) = \frac{y-\bar{y}}{sd(y)}$

3.2.2. Empirical variogram modeling

Definitions

A process of random variable $Z(s)$ over locations $s \in D$ is said second-order stationary if

1. $\mathbb{E}[Z(s)] = m \quad \forall s \in D$
2. $Cov[Z(s_i) - Z(s_j)] = C(s_i - s_j) \quad \forall s_i, s_j \in D$

Function C from the second condition is called the covariogram, which is characterized by the following properties:

1. Positive definiteness: $\sum_i \sum_j \lambda_i \lambda_j C(s_i - s_j) \quad \forall s_i, s_j \in D; \forall \lambda_i, \lambda_j \in \mathbb{R}$
2. Symmetry: $C(s_i - s_j) = C(s_j - s_i)$
3. Boundedness: $C(s_i - s_j) \leq C(0)$

A process of random variable $Z(s)$ over locations $s \in D$ is said intrinsically stationary if

1. $\mathbb{E}[Z(s)] = m \quad \forall s \in D$
2. $Var[Z(s_i) - Z(s_j)] = 2\gamma(s_i - s_j) \quad \forall s_i, s_j \in D$

Function γ from the second condition is called the semivariogram or variogram, which fulfills the following algebraic properties in order to be valid:

1. Conditional negative definiteness: $\sum_i \sum_j \lambda_i \lambda_j \gamma(s_i - s_j) \quad \forall s_i, s_j \in D; \lambda \in \mathbb{R}^{|D|} : \sum_i \lambda_i = 0$

2. Symmetry: $\gamma(s_i - s_j) = \gamma(s_j - s_i)$
3. Non-negativity: $\gamma(s_i - s_j) \geq 0$
4. Zero at the origin: $\gamma(0) = 0$
5. Sub-quadratic growth: $\lim_{\|s_i - s_j\| \to \infty} \frac{2\gamma(s_i - s_j)}{\|s_i - s_j\|^2} = 0$

We note that a second-order stationary process is also intrinsically stationary. In this case, the semivariogram is related to the covariogram via the identity 3.3.

$$\gamma(h) = C(0) - C(h) \quad \forall h \in \mathbb{R} \tag{3.3}$$

The main properties of a variogram, that are further used for modeling, are the following:

1. Nugget: The discontinuity value at the origin, if existent, denoted as $\tau^2 = \lim_{\|h\| \to 0} \gamma(h)$. It can be interpreted as the measurement error in the data or its micro-scale variability, when estimating γ.
2. Sill: The sum of the nugget and the partial sill, denoted as $\tau^2 + \sigma^2 = \lim_{\|h\| \to \infty} \gamma(h)$. The existence of a finite limit indicates that the process is second-order stationary, featured by a variance $C(0) = \tau^2 + \sigma^2$.
3. Range: The value where it reaches the sill $R : \gamma(R) = \tau^2 + \sigma^2$. It quantifies the range of influence of the process: for distances greater than the range, two elements of the process are uncorrelated. If the variogram range is infinite but the sill is reached asymptotically, one can define a practical range as $\tilde{R} : \gamma(\tilde{R}) = 0.95(\tau^2 + \sigma^2)$.

An intrinsic stationary process $\{Z(s), s \in D\}$ is said isotropic if its variogram is isotropic, $Var[Z(s_i) - Z(s_j)] = 2\gamma(\|s_i - s_j\|) \quad \forall s_i, s_j \in D$. This condition can be verified when the covariance structure is homogeneous over all the directions of $\mathbb{R}^d$, hence we inspect directional variograms, which are variograms computed for a number of fixed directions in $\mathbb{R}^d$. Directional variograms may reveal two main types of anisotropy [21]:

- Geometric anisotropy: It is found whenever the variogram slope near the origin varies over the explored directions, it is commonly associated with different ranges along different directions, and it can be corrected via a change of coordinates in the domain D.
- Zonal anisotropy: It is found when the asymptotes shown by the directional variograms (if stationary) are different along different directions, it typically has effect on the sills for different directions, so the variogram needs to be modelled in terms of separation vector h.

Models and estimation

To guarantee that the properties of a valid variogram are fulfilled, a number of parametric valid model are commonly employed, with properties and the physical interpretation of the corresponding parametrization known. We consider models with a sill (or transition models), except for the first one [2] [23]:

- Pure Nugget: The associated random field is a white noise of variance $\tau^2, \tau \in \mathbb{R}$. Usually a building block combined with another valid model, since the sum of valid models is a valid model. It provides discontinuity at the origin, hence its process presents a highly irregular behavior and it is not L^2-continuous. Since it is a model without a sill, a covariance function does not exist and only the variogram model is defined.

$$\gamma(h) = \begin{cases} \tau^2 & h > 0 \\ 0 & h = 0 \end{cases} \tag{3.4}$$

- Exponential model: The sill is $\sigma^2, \sigma \in \mathbb{R}$, the range is infinite, but one can define the practical range as $\tilde{R} = 3a, a \in \mathbb{R}$. It is linear at the origin, which is common in continuous but non-differentiable processes.

$$\gamma(h) = \begin{cases} \sigma^2(1 - e^{-h/a}) & h > 0 \\ 0 & h = 0 \end{cases} \tag{3.5}$$

- Spherical model: The range is $a \in \mathbb{R}$, and the sill is $\sigma^2, \sigma \in \mathbb{R}$. It is linear at the origin, which is common in continuous but non-differentiable processes.

$$\gamma(h) = \begin{cases} \sigma^2 & h \geq a \\ \sigma^2 \left[\frac{3}{2}\frac{h}{a} - \frac{1}{2}\left(\frac{h}{a}\right)^3 \right] & a > h > 0 \\ 0 & h = 0 \end{cases} \tag{3.6}$$

- Gaussian model: The sill is $\sigma^2, \sigma \in \mathbb{R}$, the range is infinite, but one can define the practical range as $\tilde{R} = \sqrt{3}a, a \in \mathbb{R}$. It is quadratic at the origin.

$$\gamma(h) = \begin{cases} \sigma^2(1 - e^{-(h/a)^2}) & h > 0 \\ 0 & h = 0 \end{cases} \tag{3.7}$$

In most analyses, the variogram estimate consists of two phases:

1. Raw estimate from data, it usually does not lead to a valid model.
2. Fitting of a parametric valid model via least squares (LS).

For the first step, a finite number of distances h are actually observed, which are the distances separating each couple of sampled locations, hence we can employ as estimate one of the following [21]:

- Semivariogram cloud: It is the cloud of points corresponding to the values of $\frac{1}{2}[Z(s_i) - Z(s_j)]^2$ which are actually observed. Such a cloud provides useful indications about the number of couples available for the estimation, and the dispersion of the squared increments in the plane (e.g., a triangular disposition that is denser in the bottom part is typical of a stationary variogram).
- Binned semivariogram: It is a discretization computed as average within a given amount of classes of distances. This estimator is often preferred to the semivariogram cloud for its convenience in highlighting the relevant features of the variogram. In this case, the dimension of the classes (i.e., the lag used) is key to provide sensible estimates, as a trade-off exists between overfitting and oversmoothing.

These give an unbiased estimate of γ if intrinsic stationarity holds true. We use the binned semivariogram $\hat{\gamma}(h) = (\hat{\gamma}(h_1), ..., \hat{\gamma}(h_K))^T$, with a discretization of $K = 15$ bins, and lag width assessed visually to find a reasonable pattern. Lag is computed from locations in latitude and longitude, using great circle distances in kilometers.

For the second step, we use the Weighted Least Squares optimization criterion, which consists of looking for the parameters θ from a valid model $\gamma(\cdot, \theta)$ which minimize expression 3.8. In our case, we use the weights $w_k = N_k/h_k^2$, with N_k the amount of observations on bin k.

$$\sum_{k=1}^{K} \frac{1}{w_k}(\hat{\gamma}(h_k) - \gamma(h_k, \theta))^2 \tag{3.8}$$

Model assessment and selection

After finding the best parameters of each model, we consider two types of assessment:

- Qualitative assessment: We verify that the fitted model shape is adequate when looking into the directional variogram. We consider directions 0 deg, 45 deg, 90 deg, 135 deg, in order to see the behavior over the main foci, without searching for overfit of potential noise.
- Quantitative assessment: We look into three error indicators, which are the resid-

ual sum of squares (SSErr), and the median and mean of the residuals of a GLS prediction of the model over the locations where the observations are located.

We select as best model the one with lowest triplet of errors, or the model with two indicators lowest, taking into account if the difference between the remaining indicator and its lowest value is significant or not.

3.2.3. Local Indicators of Spatial Association

Moran's I statistic is one of the most commonly used indicators of global spatial autocorrelation. It is a cross-product statistic between a variable and its spatial lag, with the variable expressed in deviations from its mean. For an observation of a variable x at location i, considering n observations, we formulate the deviations from the mean $z_i = x_i - \bar{x}$, the spatial weights w_{ij} such that $n = \sum_i \sum_j w_{ij}$. Hence, Moran's I statistic in this case is stated in 3.9, where we can see the decomposition of the global statistic into local statistics (or LISAs [4]) and simplify the expression using constant $k = (\sum_{i=1}^{n} z_i^2)^{-1}$. The local statistics are then the product of the deviation value at location i with its spatial lag, which is the weighted sum of the values at neighboring locations. They may be interpreted as indicators of local pockets of nonstationarity, or hot spots, or they may be used to assess the influence of individual locations on the magnitude of the global statistic and to identify “outliers”.

$$I = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} z_i^T z_j}{\sum_{i=1}^{n} z_i^2} = \sum_{i=1}^{n} I_i = \sum_{i=1}^{n} k \cdot z_i^T \sum_{j=1}^{n} w_{ij} z_j \tag{3.9}$$

The spatial weights conform a row-standardized matrix W where a non-zero weight establishes that locations i and j are neighbors, hence represent the possible spatial interactions between observations in space. By convention, the self-neighbor relation is excluded, so that the diagonal elements of W are zero. There are different criteria to establish this relation, among which are contiguity criteria [6] and based on distance [7]:

- Queen contiguity: The queen criterion, in analogy to the moves allowed for the queen piece on a chess board, defines neighbors as spatial units sharing a common edge or a common vertex.

- Rook contiguity: The rook criterion, in analogy to the moves allowed for the rook piece on a chess board, defines neighbors by the existence of a common edge between two spatial units.

- Distance-based criteria: It requires to compute a distance measure between each pair of spatial units, it defines neighbors by falling within a critical distance band between two spatial units. It requires to define the critical distance threshold, which can be chosen as the optimized distance threshold that guarantees that every observation has at least one neighbor, among other values.

- K-Nearest neighbor criteria: It requires to compute a distance measure between each pair of spatial units, it defines neighbors by being among the closest K units from a given spatial unit. Among the mentioned criteria, this is the only one providing a non-symmetric relation between spatial units.

We consider the queen contiguity criteria for the analysis, given that distance based criteria cannot provide intuitive insights and is not suited when considering the shape of the continent and the distribution of the considered regions. In the other hand, the rook criteria is discarded due to considering a limited type of interactions between regions.

The null hypothesis of Moran's I is spatial randomness, significance is based on a conditional permutation method [4], where we calculate a reference distribution for the statistic under the null hypothesis of spatial randomness by randomly permuting the observed values over the locations. The statistic is computed for each of these randomly reshuffled data sets, which yields a reference distribution. This approach is not sensitive to potential violations of underlying assumptions, which makes it robust but limited in generality to the actual sample. In the case of local statistics, the value of each z_i is held fixed at its location i, and the remaining $n-1$ values are then randomly permuted to yield a reference distribution for the local statistic (one for each location), hence the permutation is carried out for each observation in turn.

The reference distribution is used to calculate a pseudo p-value, where R is the number of times the computed statistic from the spatial random data sets (the permuted data sets) is equal to or more extreme than the observed statistic, while M is the number of permutations. The pseudo p-value is then $p = \frac{R+1}{M+1}$. Local spatial clusters, sometimes referred to as hot spots, may be identified as those locations or sets of contiguous locations for which the LISA is significant.

After obtaining the local statistics, we can elaborate a Moran scatter plot [5], where indication of significance is combined with the location of each observation, by classifying the values for the spatial lag above and below the mean as high and low. It provides an interpretable classification of spatial association into four categories, which are referred to as High-High, Low-Low, Low-High and High-Low, relative to the mean, which are the quadrants of the Moran scatterplot, with the mean in the center. Hence, we find positive spatial autocorrelation clusters (similar values at neighboring locations, which are the high-high and low-low categories) or negative spatial autocorrelation outliers (dissimilar values at neighboring locations, which are the high-low and low-high categories).

3.3. Results

3.3.1. First wave

We present the results regarding the logistic transformation for the response, see Fig. 3.5 for the comparison between the original cases density values and their transformation over the considered regions. By performing LASSO selection, 5 factors are filtered out by using an optimal penalization of 0.0057. Fig. 3.6a shows the decrease of the coefficients for each factor for an increasing penalization, where the upper axis shows the amount of considered factors, and the pointed vertical line shows the optimal penalization. Fig. 3.6b shows the cross-validation MSE evaluated for the considered grid for the penalization, where the left pointed vertical line shows the optimal penalization (and hence shows the minimum MSE), and the right pointed vertical line shows the lambda value with one standard deviation to the right of the optimal one.

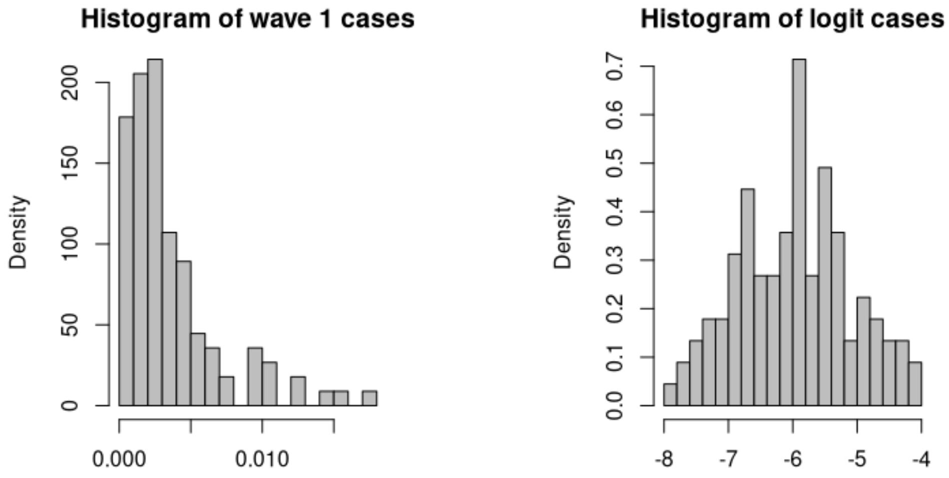
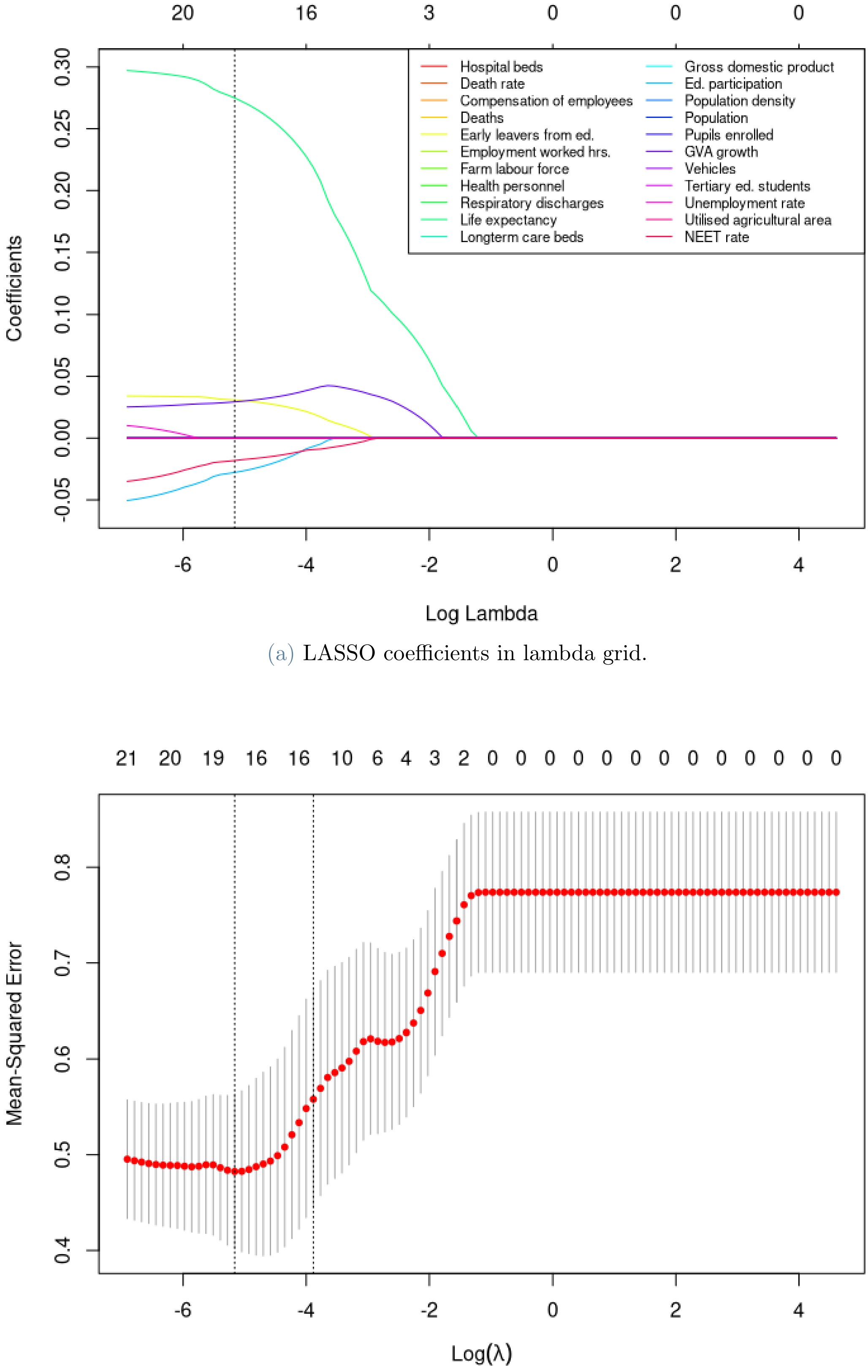


Figure 3.5: Histograms of original response and chosen transformation for wave 1.



(a) LASSO coefficients in lambda grid.

(b) Cross-validation lambda tuning.

Figure 3.6: Lambda tuning and model selection for wave 1.

Afterwards, we verify the performance of the model by looking into its R^2 value, which is 0.578, hence it barely performs better than using the sample mean to explain further variability of the response. Note, however, that minimum MSE reached during the cross validation stage is around 0.49, which is fairly high given the range of values for the response, and so is coherent with the R^2 value.

Looking into the coefficients with highest magnitude (see Fig. 3.7 for the visualization of the coefficients for all factors), we find the following:

- Risk factors: Those with positive coefficient, hence leading to a higher response (which is equivalent to a higher cases density)
  - Life expectancy.
  - Early leavers from education.
  - GVA growth rate.
- Mildness factors: Those with negative coefficient, hence leading to a lower response (which is equivalent to a lower cases density)
  - Education participation.
  - NEET rate.

We then verify the assumptions of the model by performing the Shapiro-Wilk test, where we obtain a p-value of 0.135, hence we do not reject the hypothesis of having normally distributed residuals, and hence the model is verified, see Fig. 3.8 for the visualization of some of the diagnostic plots of the residuals. We note that the normality test is diputable due to the spread of the tails of the residuals, which can be seen both in the histogram and the QQ plot. In addition, inspecting both the residuals of the model and the response over the european map, in general we find that original response values and residual values do not necessarily show the same dynamics (see Fig. 3.9).

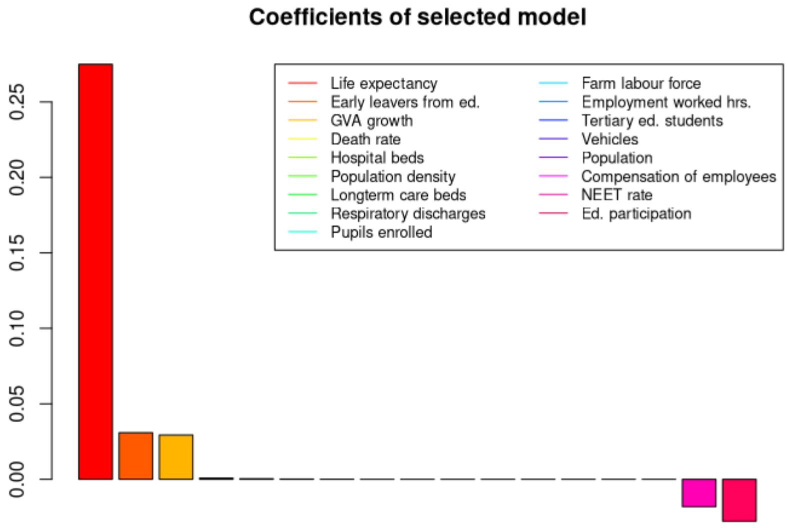
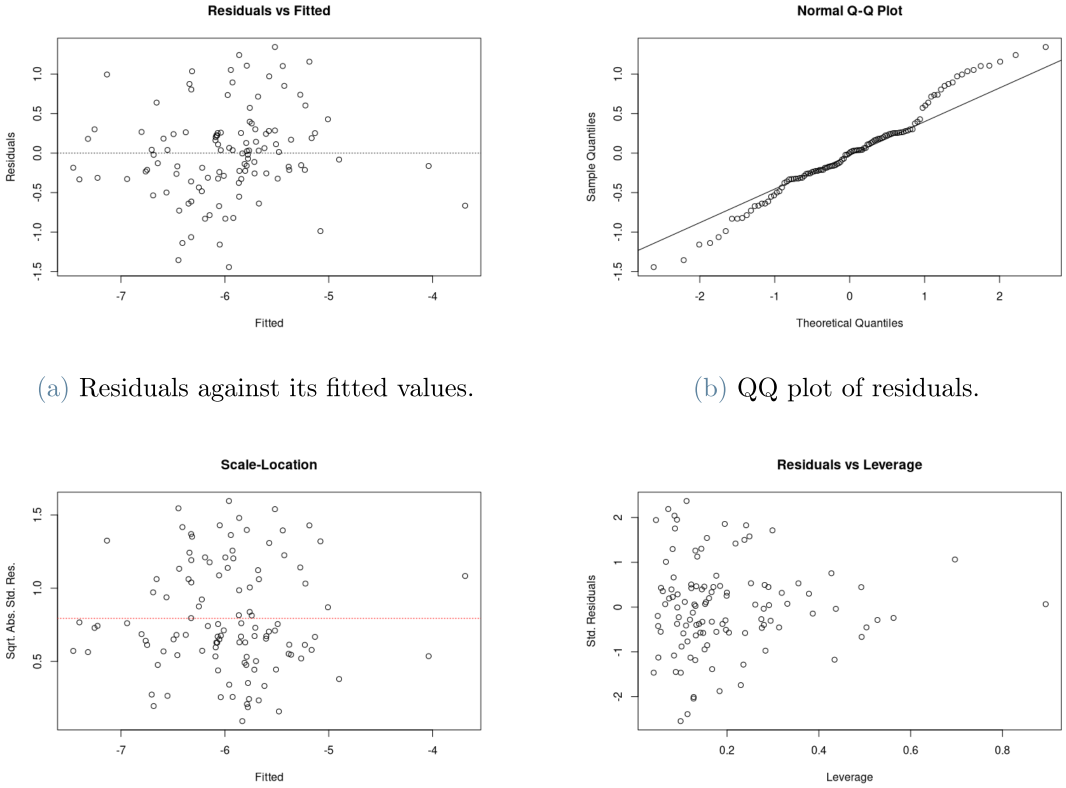


Figure 3.7: Factor coefficients in selected model.



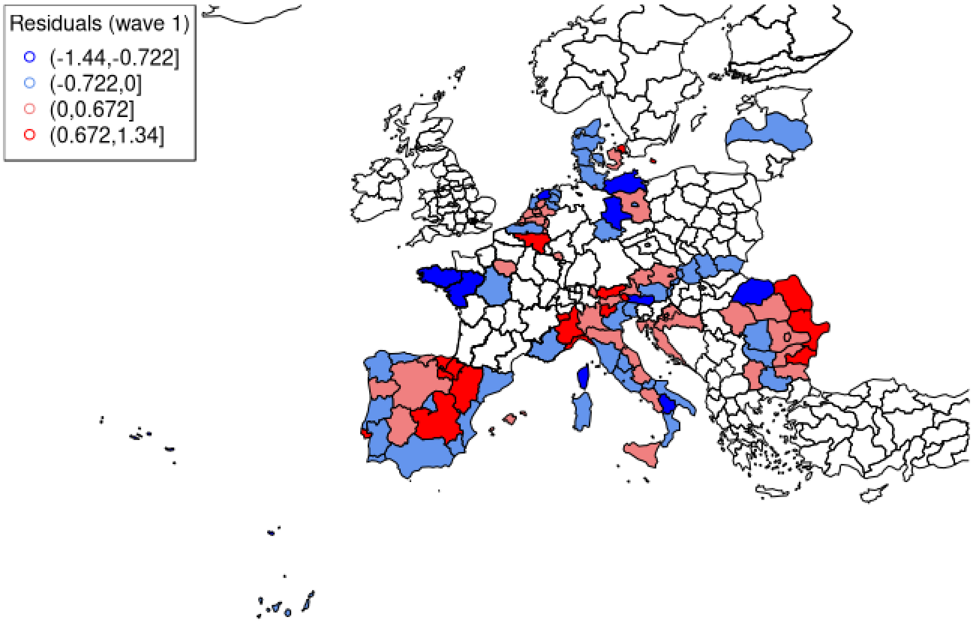
(a) Residuals against its fitted values.

(b) QQ plot of residuals.

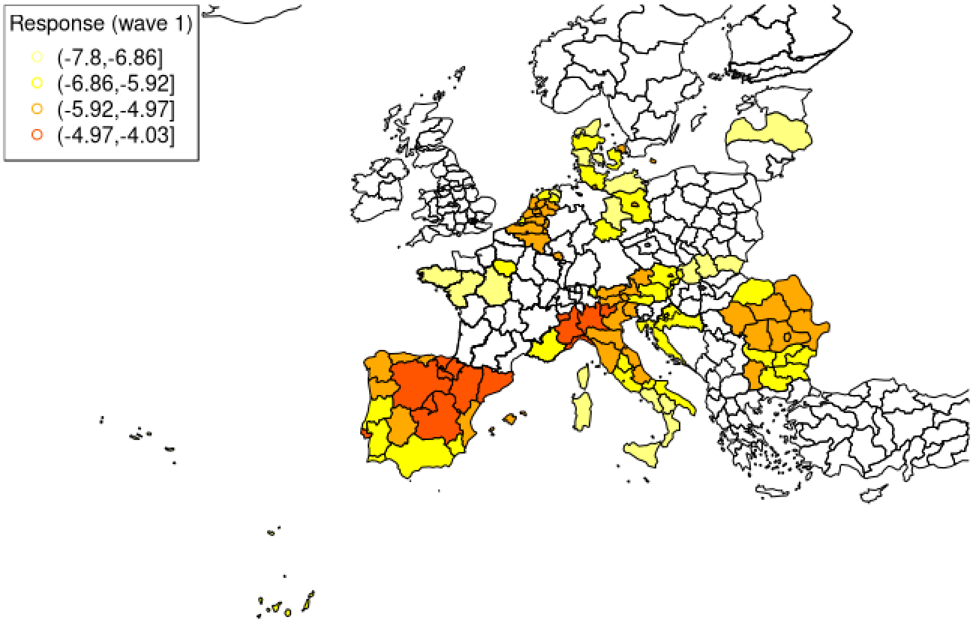
(c) Spread-Location plot.

(d) Residuals against leverage values.

Figure 3.8: Residuals visualization and diagnostics for the model of the first wave.



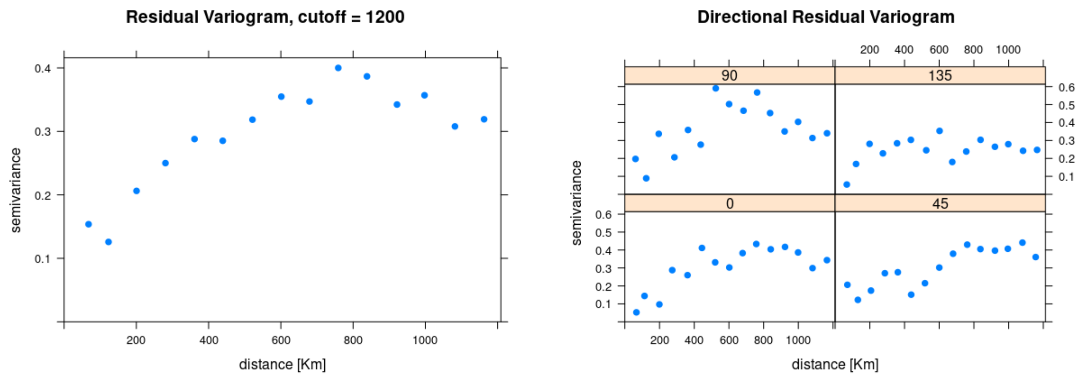
(a) Model residuals.



(b) Response (logit transform of the cases density).

Figure 3.9: Residual and response values for each region over the european map.

After this procedure, we continue to analyse what the model was not able to explain, and hence we investigate the possibility of spatial dependence in the residuals. As a first step, we formulate the empirical variogram to study. We first find that there is a dependence on longitude, and so we search for a variogram model over the residuals from a linear dependence on this coordinate. We then find that a reasonable lag limit can be set to 1200 Km, given that it is close to a third of the maximum distance between any pair of regions of interest (around 3200 Km), and we find a variogram shape that is intuitive with having neglectable correlation between regions with a larger distance, since it is harder to maintain significant interactions between them. Fig. 3.11 shows the residual variogram following the mentioned trend, both neglecting direction and considering four main directions.



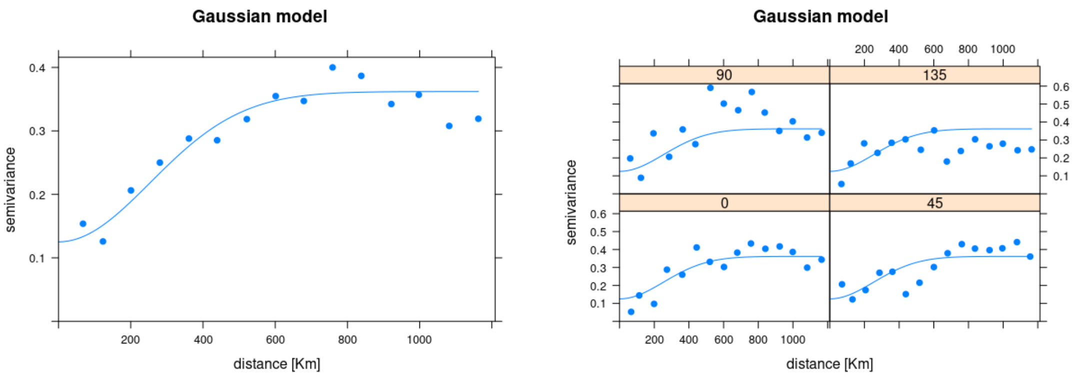
(a) Residual variogram.

(b) Directional variogram.

Figure 3.10: Empirical variogram of residuals from a linear trend depending on longitude.

We find that the gaussian model fits best the behavior found (more details in Annex B.1), which has the following features:

- It has a nugget of 0.129325, the highest from all models, which means that the discontinuity jump is relatively high.
- It has a partial sill of 0.230249, the lowest from all models, however we note that it conforms 76% of the total sill.
- It has a practical range 362.2699 Km, the lowest from all models, and so we consider that the residuals are fairly uncorrelated at a relatively short distance, even though the sill is reached asymptotically.
- It is discontinuous at the origin, and quadratic near it, so we can consider that the residuals have a quite continuous behavior when discarding the white noise.



(a) Residual variogram.

(b) Directional variogram.

Figure 3.11: Fitted variogram of residuals from a linear trend depending on longitude.

Finally, we inspect the LISA clusters over the european map. First, considering the cases density (see Fig. 3.12a), we note the following:

- Spain is an important cluster with high-high dynamics, hence we detect a near country-level behavior of high infections.
- Lombardy is not part of a significant cluster, however Piedmont is identified as a high-high cluster, which also showed a high cases density in northern Italy.
- Northern Denmark has a low-low cluster, which is expected given their low cases density over the country.
- Southern Italy shows low-low clusters in Apulia, Basilicata and Campania, which is intuitive given that in this regions there were low values of cases density.
- Slovakia has low-low clusters extending until Austria, highlighting an interaction between both countries.

Second, considering the response, which is the logit transformation, that has a higher concentration of values towards a center to resemble a bell shape (see Fig. 3.12b), we note the following:

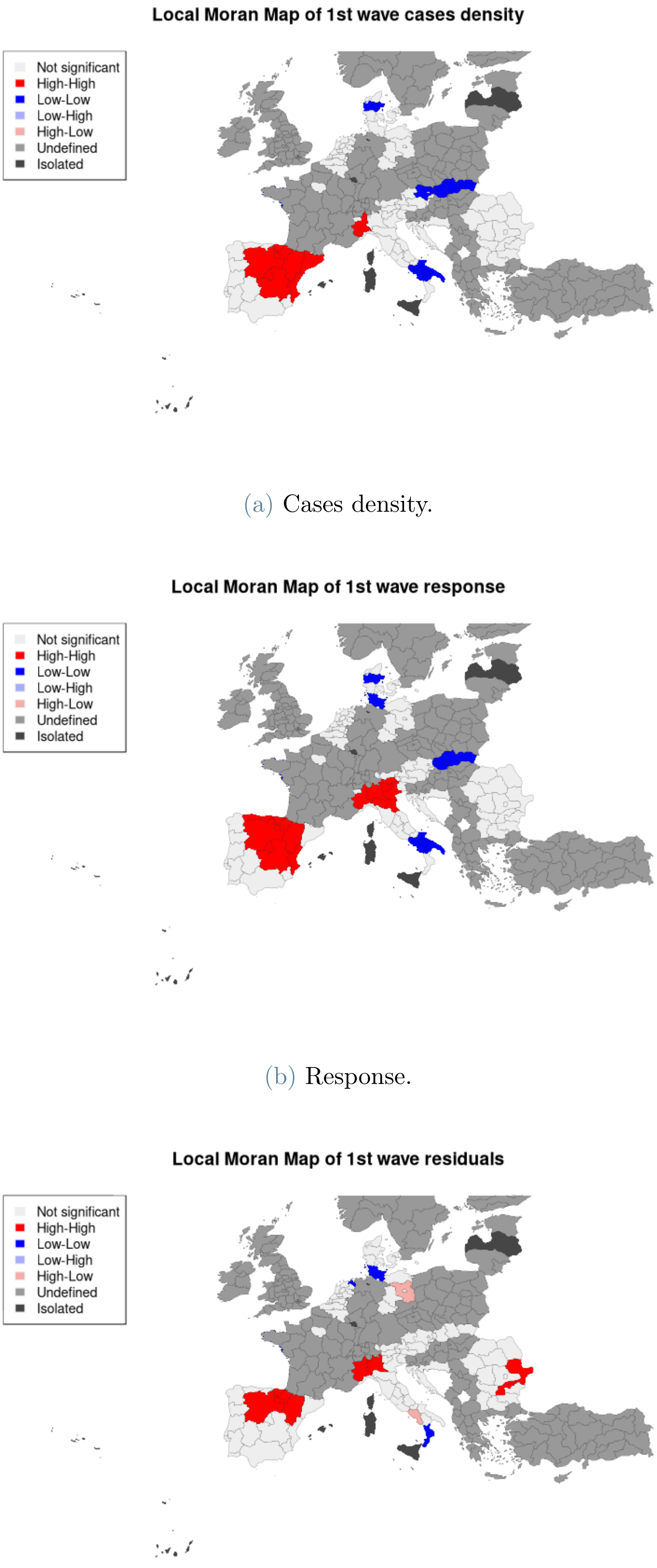
- Spain has even wider high-high dynamics, adding a northern region, since its response value is nearer to the rest of the region (its cases density originally was not high, hence it was not considered in the previous case).
- Many northern regions of Italy conform a high-high cluster, including Lombardy, which was expected given the national experience.

- Denmark has an additional region within its low-low clusters, so their dynamics become more generalized.
- Southern Italy maintains the same dynamics
- There is no low-low cluster in Austria, only Slovakia remains with its dynamics.

Third, considering the residuals, we find several differences within dynamics considering single regions (see Fig. 3.12c), in particular:

- Spain has a smaller area with high-high dynamics.
- Less northern regions of Italy conform a high-high cluster, but Lombardy remains included.
- Southern Denmark maintains its region as low-low cluster, while the northern regions loses significance.
- In southern Italy, the low-low cluster shrinks and moves to the single region of Calabria, while Campania from a low-low cluster to a high-low outlier.
- Slovakia's low-low clusters lose significance.
- Germany shows a high-low outlier.
- Romania shows high-high clusters.

.



Local Moran Map of 1st wave cases density

(a) Cases density.

Local Moran Map of 1st wave response

(b) Response.

Local Moran Map of 1st wave residuals

(c) Residuals.

Figure 3.12: LISA clusters over considered regions for first wave.

3.3.2. Second wave

We present the results regarding the normalization (Z transformation) for the response, see Fig. 3.13 for the comparison between the original cases density values and their transformation over the considered regions. By performing LASSO selection, 4 factors are filtered out by using an optimal penalization of 0.0115. Fig. 3.14a shows the decrease of the coefficients for each factor for an increasing penalization, where the upper axis shows the amount of considered factors, and the pointed vertical line shows the optimal penalization. Fig. 3.14b shows the cross-validation MSE evaluated for the considered grid for the penalization, where the left pointed vertical line shows the optimal penalization (and hence shows the minimum MSE), and the right pointed vertical line shows the lambda value with one standard deviation to the right of the optimal one.

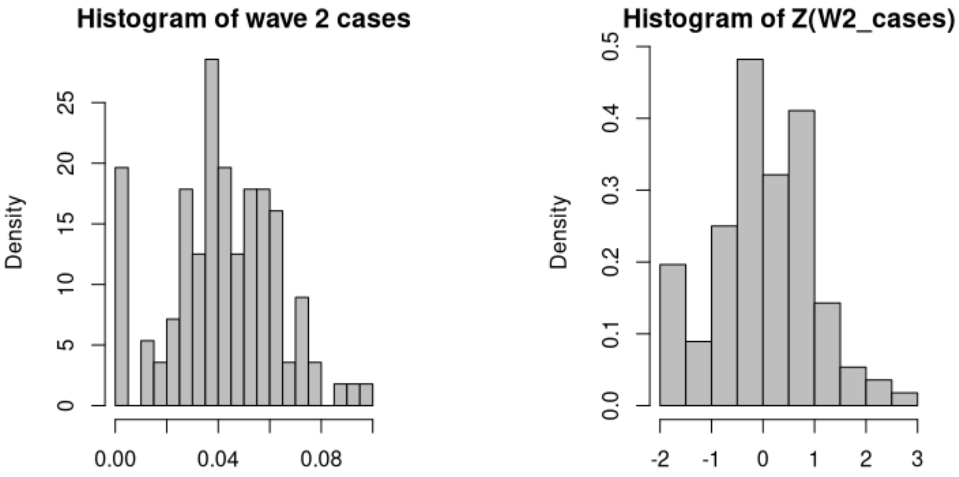
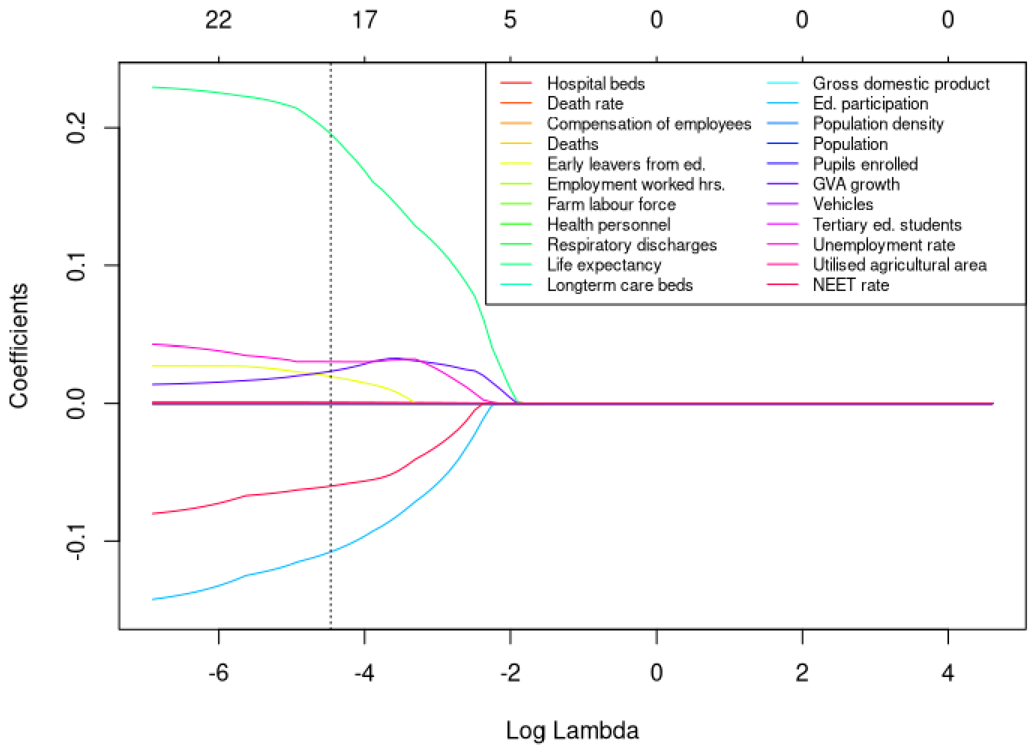
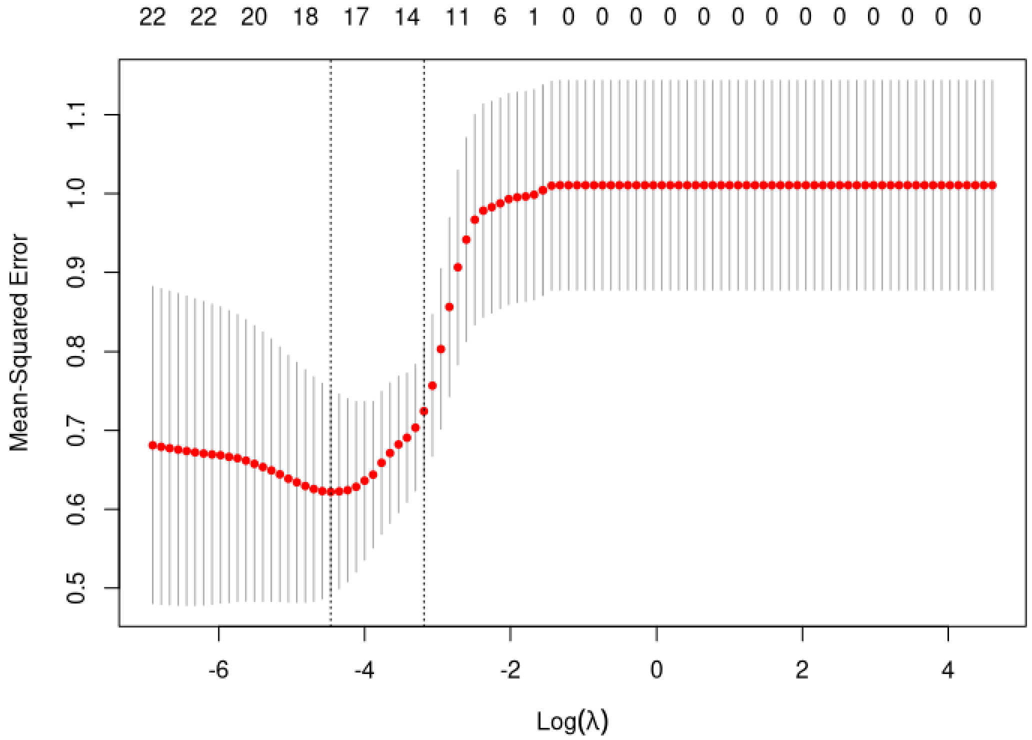


Figure 3.13: Histograms of original response and chosen transformation for wave 2.

Afterwards, we verify the performance of the model by looking into its R^2 value, which is 0.619, hence it performs better than using the sample mean to explain further variability of the response. Note, however, that minimum MSE reached during the cross validation stage is around 0.62, which is high given the range of values for the response, and so is coherent with not having a higher R^2 value.



(a) LASSO coefficients in lambda grid.



(b) Cross-validation lambda tuning.

Figure 3.14: Lambda tuning and model selection for wave 2.

Looking into the coefficients with highest magnitude (see Fig. 3.15 for the visualization of the coefficients for all factors), we find the following:

- Risk factors: Those with positive coefficient, hence leading to a higher response (which is equivalent to a higher cases density)
  - Life expectancy.
  - Unemployment rate.
  - GVA growth rate.
  - Early leavers from education.
- Mildness factors: Those with negative coefficient, hence leading to a lower response (which is equivalent to a lower cases density)
  - Education participation.
  - NEET rate.

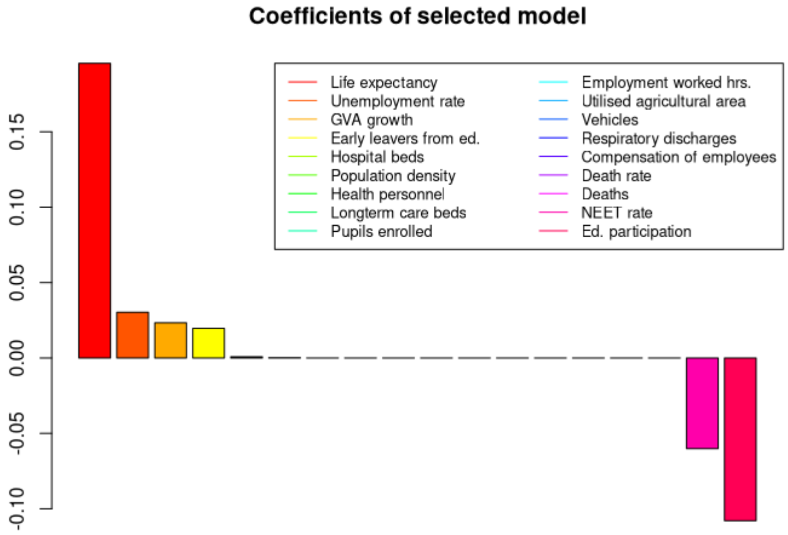
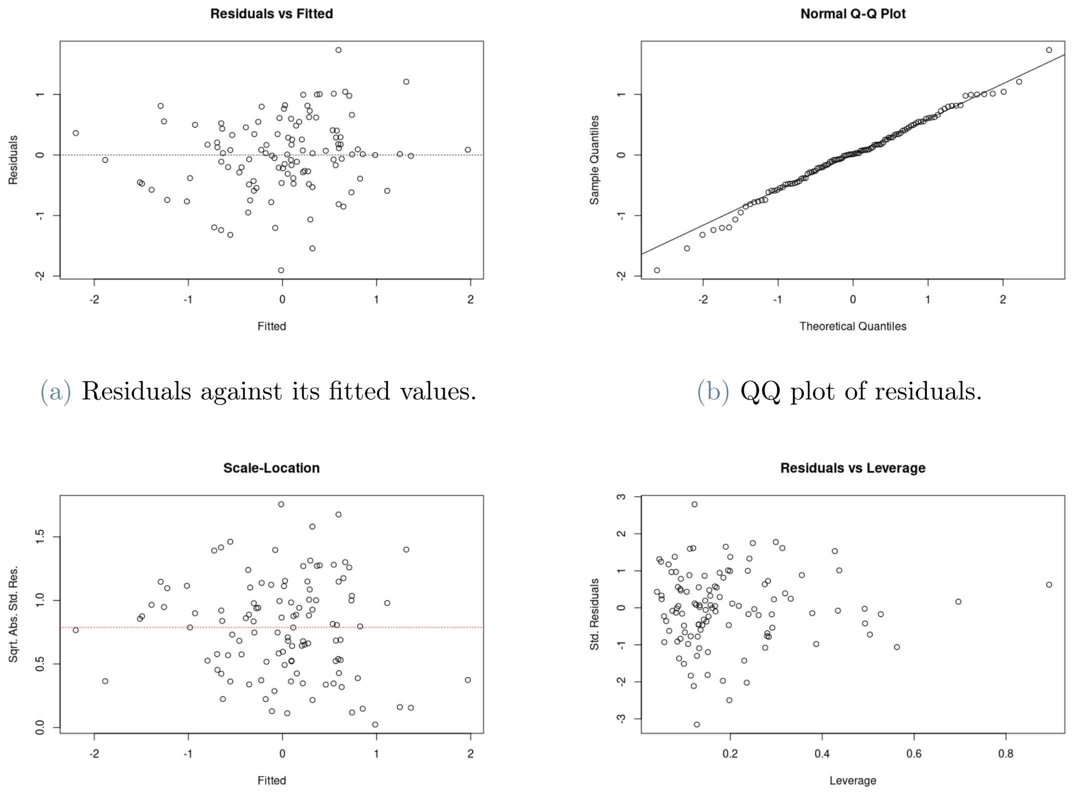


Figure 3.15: Factor coefficients in selected model.



(a) Residuals against its fitted values.

(b) QQ plot of residuals.

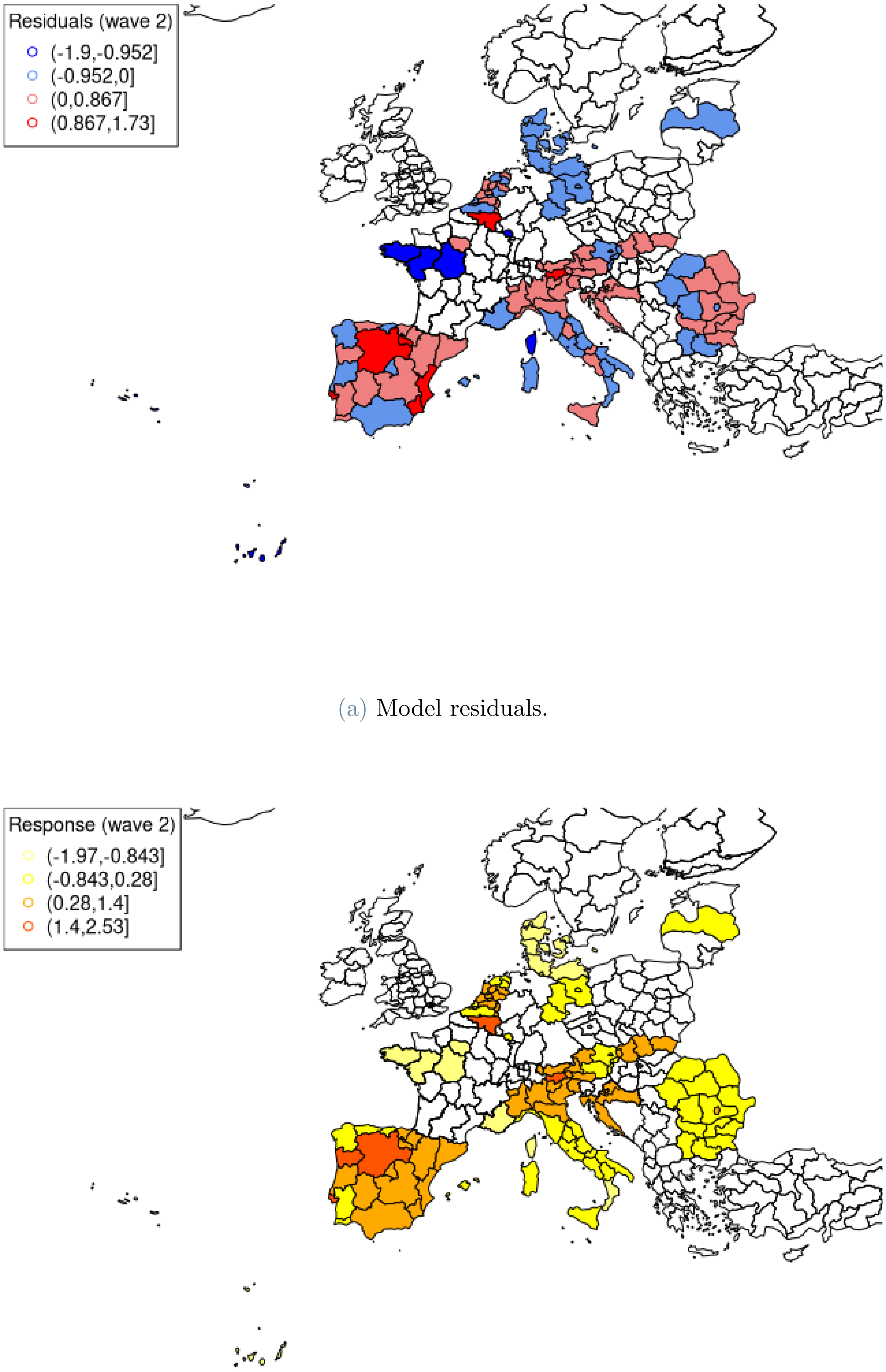
(c) Spread-Location plot.

(d) Residuals against leverage values.

Figure 3.16: Residuals visualization and diagnostics for the model of the second wave.

We then verify the assumptions of the model by performing the Shapiro-Wilk test, where we obtain a p-value of 0.673, hence we do not reject the hypothesis of having normally distributed residuals, and hence the model is verified (see Fig. 3.16). By inspecting both the residuals of the model and the response over the european map, in general we find that original response values and residual values show very similar dynamics (see Fig. 3.17).

We continue towards analyzing what the model was not able to explain, and hence we investigate the possibility of spatial dependence in the residuals. As a first step, we formulate the empirical variogram to study. We first find that there is a dependence on latitude, and so we search for a variogram model over the residuals from a linear dependence on this coordinate.

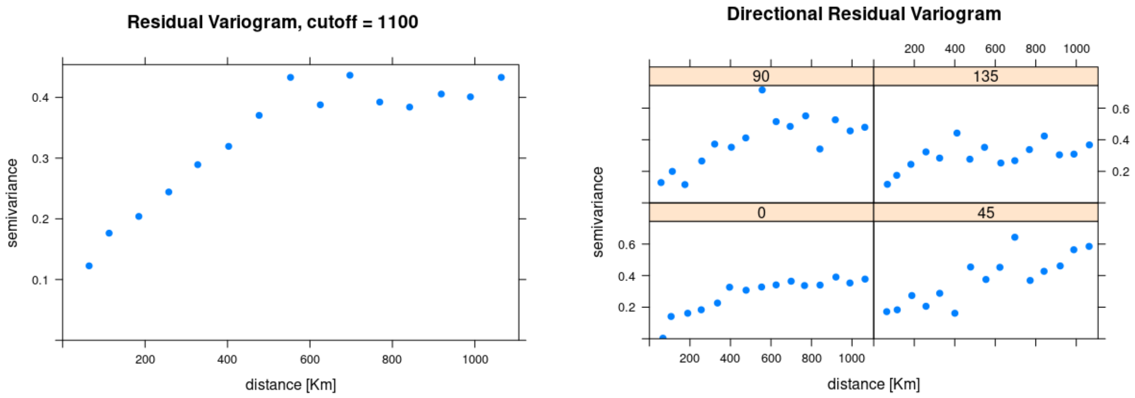


(a) Model residuals.

(b) Response (Z transform of the cases density).

Figure 3.17: Residual and response values for each region over the european map.

We then find that a reasonable lag limit can be set to 1100 Km, given that it is close to a third of the maximum distance between any pair of regions of interest (around 3200 Km), and we find a variogram shape that is intuitive with having neglectable correlation between regions with a larger distance, since it is harder to maintain significant interactions between them. Fig. 3.19 shows the residual variogram following the mentioned trend, both neglecting direction and considering four main directions.



(a) Residual variogram.

(b) Directional variogram.

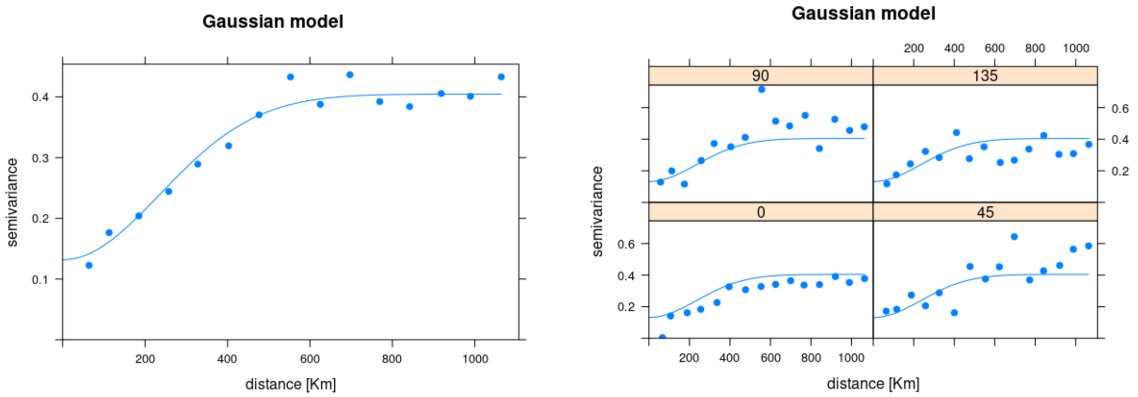
Figure 3.18: Empirical variogram of residuals from a linear trend depending on latitude.

We find that the gaussian model fits best the behavior found (more details in Annex B.2), which has the following characteristics:

- It has a nugget of 0.13183, the highest from all models, which means that the discontinuity jump is relatively high.
- It has a partial sill of 0.27286, the lowest from all models, we note however that it conforms 81% of the total sill.
- It has a practical range 332.7346 Km, the lowest from all models, and so we consider that the residuals are fairly uncorrelated at a relatively short distance, even though the sill is reached asymptotically.
- It is discontinuous at the origin, and quadratic near it, so we can consider that the residuals have a quite continuous behavior when discarding the white noise.

Finally, we inspect the LISA clusters over the european map. First, considering the cases density (see Fig. 3.20a), we note the following:

- All Denmark and northern Germany conform a group of low-low clusters, which is intuitive given their low values on cases density.



(a) Residual variogram.

(b) Directional variogram.

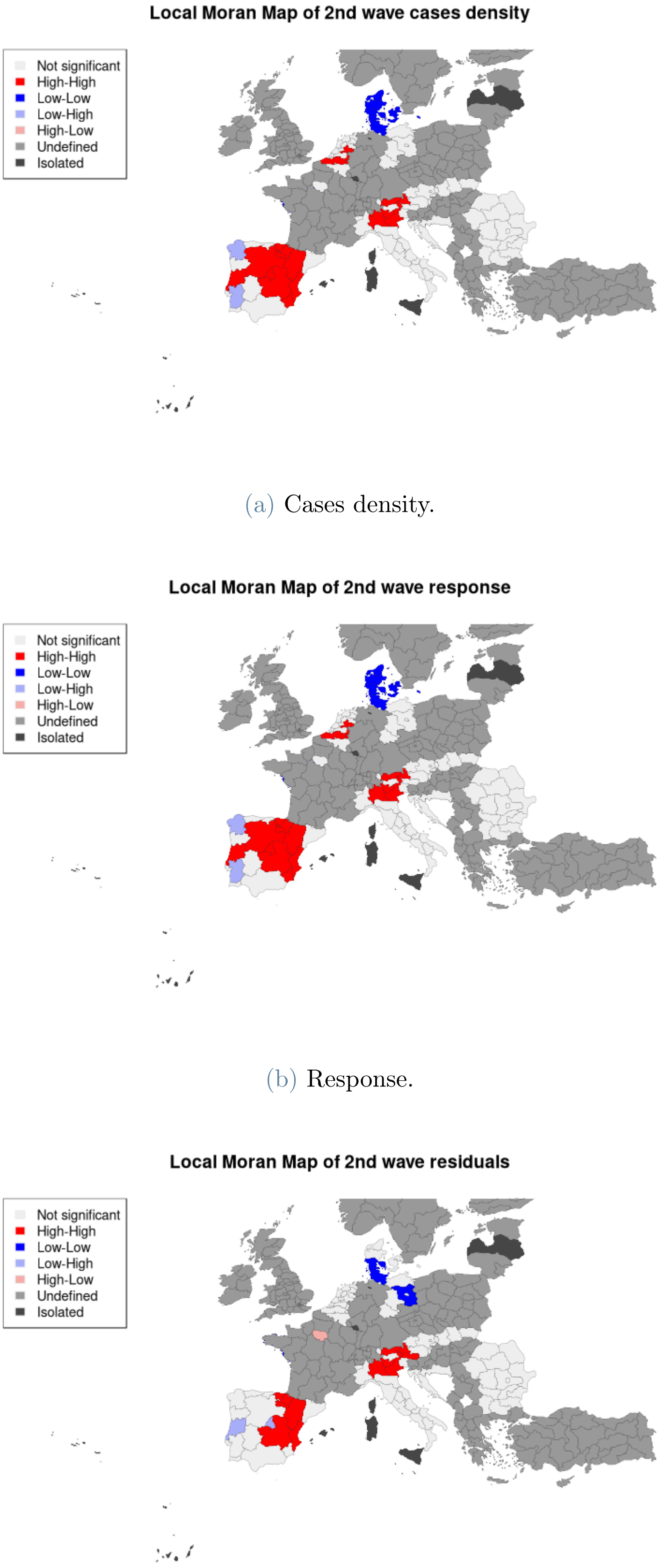
Figure 3.19: Fitted variogram of residuals from a linear trend depending on latitude.

- Northern Belgium and southern Netherlands conform a group of high-high clusters, which is of interest given that they do not have exceptionally high values of cases density, but their immediate surrounding regions do have lower values.
- Most of Spain, and a region of Portugal, make a large group of high-high clusters, which is expected due to the high levels of infections in the entire area.
- Contiguous regions of north-west Spain and southern Portugal conform low-high outliers, which are the regions with least cases compared to their neighbors.
- Northern Italy (Lombardy, Veneto and Trentino South Tyrol), along with some regions of Austria, show high-high clusters, and high cases density.

Second, considering the response, which is its normalization (see Fig. 3.20b), we note that the dynamics are the same. This is natural given that it is a translation and rescale using the exact same parameters for all values, so relative effects should not change.

Third, considering the residuals, we find that only dynamics between Italy and Austria are maintained (see Fig. 3.20c), while we find the following differences:

- Only southern Denmark and a region of east Germany have low-low clusters.
- Northern Belgium and southern Netherlands lose significance.
- Spain has a shrunken group of high-high clusters toward the east.
- A region of central Spain and northern Portugal conform low-high outliers.
- France has a north-east region considered a high-low outlier.



(a) Cases density.

(b) Response.

(c) Residuals.

Figure 3.20: LISA clusters over considered regions for second wave.

3.4. Discussion

As a first remark, we note that the regression models for both the first and second wave are between 55% and 65%. These values, even if not considerably high, are still notable given that there is no clinical data used for the covariates. This means that contextual information regarding local socio-economic factors has a role when analyzing the spread dynamics of the pandemic.

For both the first and second wave, we found that the following factors were most influential to their respective responses:

- Risk factors: Those with positive coefficient, hence leading to a higher response (which is equivalent to a higher cases density)
  - Life expectancy: We can expect that an older population is more vulnerable to contagion, since high life expectancy has been found to be positively correlated with COVID-19 spread [3] [24], while we also expect it to be one of the main factors [19].
  - Early leavers from education: We find this correlation intuitive, given that earlier potential workers could potentially add to people exposed to contagion.
  - GVA growth rate: We can expect that higher economical development is related to more infections, since it is linked to higher activity and exposure [19].
- Mildness factors: Those with negative coefficient, hence leading to a lower response (which is equivalent to a lower cases density)
  - Education participation: We find this intuitive and coherent with the overall model, given that more young people studying are linked to less early leavers.
  - NEET rate: We find this intuitive and coherent with the overall model, given that more young people not studying nor working are less active or exposed.

In addition, for the second wave, we found that an important risk factor was the unemployment rate, which can be associated with social fragility, and hence a signal of vulnerability that could indicate certain dynamics that can lead to higher infection rates. This factor is generally considered within the Social Vulnerability Index, where one of the 15 factors considers unemployment [14], and has been studied by Karaye and Horney [18], that found that overall SVI was associated with increased COVID-19 case counts. In our case, we can find that during the second wave, a high unemployment rate is an important risk factor towards having a high regional cases density, as a specific factor instead of considering a mixed index.

In general, we find that educational factors interact consistently with what is found in literature. Kashem et. al. [17] found that lower education levels are positively correlated with cumulative case rates, since a limited formal education plays an important role in the virus' prevalence and is also related to occupation and income, since lower levels of education or training generally result in lower-paid work where remote work is often not possible, and is more likely to require face-to-face interactions with few safety measures. In our case, we can find that specific indicators related to education are important when predicting risk, and that not any education indicator is selected as relevant.

As we have seen from the LISA clusters and their difference when verifying spatial dependency among the regression model residuals, we are able to find that the linear model explains better the spatial dependency on regions with lower cases density. This can be reasonable since we see that more critical regions are concentrated in specific areas that the model ignores, and so higher residuals remain concentrated in specific zones.

3.5. Further developments

This approach considers a geo-spatial model that is constructed by parts, instead of performing a direct spatial regression. As we explained in the methodology, we first formulate a multivariate linear model that considers as covariates regional factors, without including any variable that is explicitly related to geography, and only later we consider the spatial dependency over the residuals, that are the unexplained part of the linear model. Hence, further research can be made by considering an integrated model.

In the other hand, spatial dependency emerges when knowing the response in neighboring regions, however, we started with an incomplete dataset in terms of observed regions of the European Union, and furthermore, we discarded regions with missing values in the covariates. Hence, the model has regions in which the response is not best explained, but further investigation could lead to more interesting and complete results when adding missing observations and values to the dataset.

4 | Conclusions

4.1. Contrast between approaches

The motivation of this research was to study two different approaches where each had specific advantages to their use. We first describe them for each approach:

- Association rule mining approach:

  1. Association rules consider conceptually simple types of covariates, which are modeled as attributes that are represented by a present or absent item in a transaction. In this case, the transaction is an observation that has as items the characteristics of the region.

  2. Association rules have a direct interpretation, since the antecedent is directly correlated to the consequent. In this case, the antecedent is the group of interacting items, while the consequent is either having high or low cases density with respect to the mean.

  3. Association rules provide specific insights about how groups of factors interact between them.

  4. There is a lower impact when there are missing values in the dataset where this approach is applied. Since the association rules are descriptive of the dataset, this approach can work with a small amount of observations.

  5. There are efficient algorithms for association rule mining in case of large datasets, so this approach is not necessarily expensive in terms of resources [25].

- Geo-statistical analysis approach:

  1. This approach considers spatial effects and patterns [22], which is critical in a study about a virus diffusion within the EU territory. It has been observed in previous studies that every country and region is affected differently by COVID-19 spread [13].

2. This approach works well with continuous data since we use linear models and variogram models, hence it is not necessary to dicotomize them as in the other approach. This is an important advantage given that the dataset has only continuous variables, and so this approach has no information loss from data dichotomization.

In the other hand, each approach has particular weaknesses, which we now enumerate:

- Association rule mining approach:
    1. This approach requires specifically dichotomic variables, so that values equal to one can represent a present item and zero an absent one. However, given that the dataset has only continuous variables, this approach leads to information loss from data dichotomization.
    2. Association rules are only descriptive of the current dataset, since they show patterns of co-occurrence rather than a specific relationship between attributes. With this approach, it is not possible to make inference nor we gain predictive power for new data [26].
- Geo-statistical analysis approach:
    1. There is a higher impact when there are missing values, in terms of robustness and confidence [22]:
        - When approximating missing values over the covariates, this method is particularly sensible to measurement and/or approximation errors within the data.
        - When removing observations due to having missing values in the covariates that are not possible to approximate, this method is particularly sensible to the distribution of locations considered, since we lose information of neighboring regions.
    2. This method is expensive in terms of resources [22], since it is complex even when working with a small dataset, and, moreover, it requires a large amount of observations in order to have significant results.

Note that, however, the points of strength of one approach consider a compensation to the points of weakness of the other. Hence, we state that combining both approaches can potentially take advantage of the strong points of each and compensate the points of weakness, if formulated well enough.

4.2. Contrast between results

In order to compare the results between both approaches, we verify the most important factors selected by each, which are:

- For the association rule mining approach, the factors that appear in the significant rules found.
- For the geo-statistical approach, the factors selected by LASSO with an absolute weight of order equal or higher than e-4.

We first find that both approaches have a reasonable amount of factors that are considered important by each, which are the following:

- First wave
    - Severity
        - Life expectancy: In the association rules, we find that regions with life expectancy values outside the inter-quantile range are associated with a high cases density (moreover, we find that also inside values are associated with lower densities). In the geo-statistical approach, we find that the linear model considers a positive weight for this variable, which means that a higher life expectancy is linearly related to a higher cases density.
        - Available hospital beds: In the association rules, we find that regions with an amount of available hospital beds inside the inter-quantile range are associated with a high cases density. In the geo-statistical approach, we find that the linear model considers that a higher amount of beds is linearly related to a higher cases density.
        - Population density: In the association rules, we find that regions with a population density outside the inter-quantile range are associated with a high cases density. In the geo-statistical approach, we find that the linear model considers a positive weight for this variable.
    - Mildness
        - Education participation: In the association rules, we find that regions with participation values outside the inter-quantile range are associated with a low cases density (moreover, we find that also inside values are associated with higher densities). In the geo-statistical approach, we find that the linear model considers a negative weight for this variable, which means

        that a higher participation is linearly related to a lower cases density.

        - NEET rate: In the association rules, we find that regions with an amount of available hospital beds outside the inter-quantile range are associated with a low cases density (moreover, we find that also inside values are associated with higher densities). In the geo-statistical approach, we find that the linear model considers that a higher rate is linearly related to a lower cases density.

- Second wave: Severity

    - Early leavers from education: In the association rules, we find that regions with an amount of early leavers outside the inter-quantile range are associated with a high cases density (moreover, we find that also inside values are associated with lower densities). In the geo-statistical approach, we find that the linear model considers a positive weight for this variable, which means that a having more early leavers is linearly related to a higher cases density.

    - GVA growth rate: In the association rules, we find that regions with growth rates inside the inter-quantile range are associated with a high cases density. In the geo-statistical approach, we find that the linear model considers a positive weight for this variable.

    - Life expectancy: In the association rules, we find that regions with life expectancy values inside the inter-quantile range are associated with a low cases density. In the geo-statistical approach, we find that the linear model considers that a higher life expectancy is linearly related to a higher cases density.

    - Unemployment rate: In the association rules, we find that regions with growth rates outside the inter-quantile range are associated with a high cases density. In the geo-statistical approach, we find that the linear model considers a positive weight for this variable.

In the other hand, each approach has a reasonable amount of factors that are valued differently, which are the following:

- First wave

    - Severity

        - Health personnel: Only the association rules state that regions with an amount of personnel outside the inter-quantile range are associated with a high cases density. Moreover, here we find that also inside values are associated with lower densities.

- Early leavers from education: Only the geo-statistical approach considers a positive weight for this variable in the linear model, which means that a having more early leavers is linearly related to a higher cases density.
- GVA growth rate: Only the geo-statistical approach considers that a higher rate is linearly related to a higher cases density.
- Death rate: Only the geo-statistical approach considers a positive weight for this variable in the linear model.

- Mildness
    - Worked hours: Only the association rules state that regions with an amount of worked hours inside the inter-quantile range are associated with a low cases density.

- Second wave
    - Severity
        - Long-term care beds: Only the association rules state that regions with an amount of beds inside the inter-quantile range are associated with a high cases density.
        - Population: Only the association rules state that regions with a population inside the inter-quantile range are associated with a high cases density.
        - Stock of vehicles: Only the association rules state that regions with a stock inside the inter-quantile range are associated with a high cases density.
        - Available hospital beds: Only the geo-statistical approach considers a positive weight for this variable in the linear model, which means that a having more beds available is linearly related to a higher cases density.
        - Population density: Only the geo-statistical approach considers that a higher population density is linearly related to a higher cases density.
    - Mildness
        - Students in tertiary education: Only the association rules state that regions with an amount of students inside the inter-quantile range are associated with a low cases density. Moreover, here we find that also outside values are associated with higher densities.

    - ∗ Education participation: Only the geo-statistical approach considers a negative weight for this variable in the linear model, which means that a higher participation is linearly related to a lower cases density.
    - ∗ NEET rate: Only the geo-statistical approach considers that a higher rate is linearly related to a lower cases density.

If we consider the type of factors that are selected as relevant by each approach, we note the following:

- For the first wave, both approaches consider the same type of features, that are demographic, educational, economical and healthcare factors.
- For the second wave, we find that both approaches mostly coincide, by considering the same types of factors than the first wave, with the only difference that the association rule mining approach includes an additional type, which is a factor related to mobility.

4.3. Final considerations

We conclude that both methods should be considered, given that

- Different factors are found to be important on each stage, so the results can be used as complementary in order to study further from more insights.
- There are common important factors on both stages, so we should consider that there is a mutual validation between approaches.
- Simpler methods are a convenient first approach that can handle better additional observations with faulty data, such as our dataset with missing values.
- More complex methods provide a richer analysis on a limited amount of observations where missing values can be approximated.

In an application such as COVID-19 spread over a continent, a simple approach provides an easier understanding of different factors and their interactions. However, we need to consider the geographical factor, since it is an important aspect of the phenomenon. Hence, the two approaches studied are best used together in order to gain interpretable but also rich insights, so we can understand how the covariates behave and also predict the spatial spread of COVID-19.

Bibliography

[1] K. Aabed and M. M. A. Lashin. An analytical study of the factors that influence covid-19 spread. *Saudi Journal of Biological Sciences*, 28(2):1177–1195, 2021.

[2] M. Al-Abdallah. A special analytical methodology for variogram modeling and interpolation of terrain elevation data by kriging method. *Damascus University Journal For The Engineering Sciences*, 34(1), 2018. URL `http://journal.damascusuniversity.edu.sy/index.php/engj/article/download/383/334/1413`.

[3] K. Allel, T. Tapia-Muñoz, and W. Morris. Country-level factors associated with the early spread of covid-19 cases at 5, 10 and 15 days since the onset. *Global Public Health*, 15(11):1589–1602, 2020.

[4] L. Anselin. *Geographical Analysis*, chapter 27, page 93–115. 1995. URL `https://dces.webhosting.cals.wisc.edu/wp-content/uploads/sites/128/2013/08/W4_Anselin1995.pdf`.

[5] L. Anselin. Local spatial autocorrelation: Lisa and local moran, 2020. URL `https://geodacenter.github.io/workbook/6a_local_auto/lab6a.html`.

[6] L. Anselin and S. J. Rey. *Modern Spatial Econometrics in Practice, a Guide to Geoda, Geodaspace and Pysal*, chapter 3, pages 39–80. The name of the publisher, 2014. URL `https://sergerey.org/giasp16/pdfs/anselin_rey_weights.pdf`.

[7] L. Anselin and S. J. Rey. *Modern Spatial Econometrics in Practice, a Guide to Geoda, Geodaspace and Pysal*, chapter 4. 2014.

[8] F. Capelli. Kaggle: Covid@lombardy dataset. URL `https://www.kaggle.com/federicocapello/covidlombardy`.

[9] M. Ciotti, M. Ciccozzi, A. Terrinoni, W. Jiang, C. bin Wang, and S. Bernardini. The covid-19 pandemic. *Critical reviews in clinical laboratory sciences*, 57(6):365–388, 2020.

[10] E. Commission. Nuts - gisco - eurostat. URL `https:`

//ec.europa.eu/eurostat/web/gisco/geodata/reference-data/administrative-units-statistical-units/nuts.

[11] A. Esposito. Associations of mutated genes explain the clinical course of covid-19. Master's thesis, Politecnico di Milano, Department of Electronics, Information and Bioengineering, 2021.

[12] A. Farseev, Y.-Y. Chu-Farseeva, Q. Yang, and D. B. Loo. Understanding economic and health factors impacting the spread of covid-19 disease. *medRxiv*, 2020.

[13] M. Fatima, K. J. O'Keefe, W. Wei, S. Arshad, and O. Gruebner. Geospatial analysis of covid-19: A scoping review. *International journal of environmental research and public health*, 18(5), 2021. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7956835/.

[14] A. for Toxic Substances and D. Registry. Cdc/atsdr social vulnerability index. URL https://www.atsdr.cdc.gov/placeandhealth/svi/index.html.

[15] J. V. Freeman and M. J. Campbell. The analysis of categorical data: Fisher's exact test. *Scope*, 33(5):11–12, 06 2007.

[16] M. Goldman. Statistics for bioinformatics. URL https://www.stat.berkeley.edu/~mgoldman/Section0402.pdf.

[17] I. M. Karaye and J. A. Horney. The impact of social vulnerability on covid-19 in the u.s.: An analysis of spatially varying relationships. *American Journal of Preventive Medicine*, 59(3), 2020. URL https://www.ajpmonline.org/article/S0749-3797(20)30259-2/fulltext.

[18] I. M. Karaye and J. A. Horney. The impact of social vulnerability on covid-19 in the u.s.: An analysis of spatially varying relationships. *American Journal of Preventive Medicine*, 59(3), 2020. URL https://www.ajpmonline.org/article/S0749-3797(20)30259-2/fulltext.

[19] R. Kumar, A. Pandey, R. G. Ibsa, D. Sinwar, and V. S. Dhaka. Study of social and geographical factors affecting the spread of covid-19 in ethiopia. *Journal of Statistics and Management Systems*, 24(1):99–113, 2021.

[20] G. Liu, H. Zhang, and L. Wong. Controlling false positives in association rule mining. *arXiv*, 2011.

[21] A. Menafoglio. Geostatistical analysis of spatially dependent data: from real to hilbert-space valued random fields. 2021.

[22] C. Paramasivam. Merits and demerits of gis and geostatistical techniques. *GIS and Geostatistical Techniques for Groundwater Science*, pages 17–21, 2019. URL https://doi.org/10.1016/B978-0-12-815413-7.00002-X.

[23] E. J. Pebesma. Gstat user's manual, 2014. URL http://www.gstat.org/gstat.pdf.

[24] S. Roy and P. Ghosh. Factors affecting covid-19 infected and death rates inform lockdown-related policymaking. *PloS one*, 15(10), 2020.

[25] D. Talia, P. Trunfio, and F. Marozzo. Chapter 1 - introduction to data mining. *Data Analysis in the Cloud*, pages 1–25, 2016. URL https://www.sciencedirect.com/science/article/pii/B9780128028810000019.

[26] C. M. Teng. Data, data, everywhere: Statistical issues in data mining. *Philosophy of Statistics*, 7:1099–1117, 2011. URL https://www.sciencedirect.com/science/article/pii/B9780444518620500344.

[27] T. P. Velavan and C. G. Meyer. The covid-19 epidemic. *Tropical medicine & international health : TM & IH*, 25(3):278–280, 2020.

[28] O. Wahltinez, A. Cheung, R. Alcantara, D. Cheung, M. Daswani, A. Erlinger, M. Lee, P. Yawalkar, P. Lê, O. P. Navarro, M. P. Brenner, and K. Murphy. Covid-19 open-data a global-scale spatially granular meta-dataset for coronavirus disease. *Scientific Data*, 9(162), 2022. URL https://www.nature.com/articles/s41597-022-01263-z.

[29] G. I. Webb. Discovering significant rules. page 434–443. ACM SIGKDD, 8 2006.

A | Appendix A

A.1. Results for each case

A.1.1. Case 1-1

Here we present the case of the target with the first choice of dichotomization (using the mean) and the features with the first choice of dichotomization (using the median). As stated in section 2.1, the AR mining step considers the adding of complementary columns for the selected features. This way, all regions have always the same amount of items, where a single item can refer to the factor being over the median or under the median.

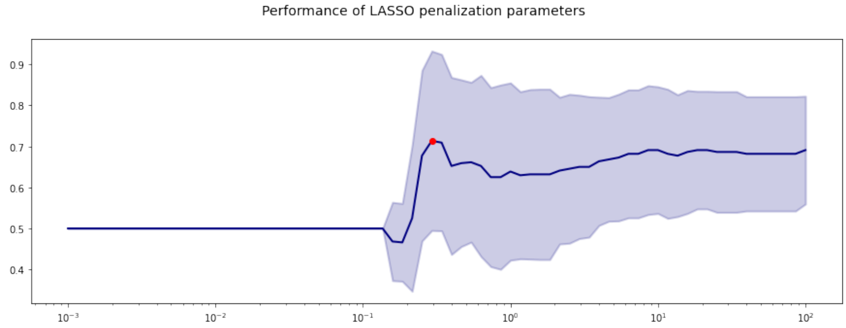


Figure A.1: 1-1: Chosen λ in grid search for wave 1.

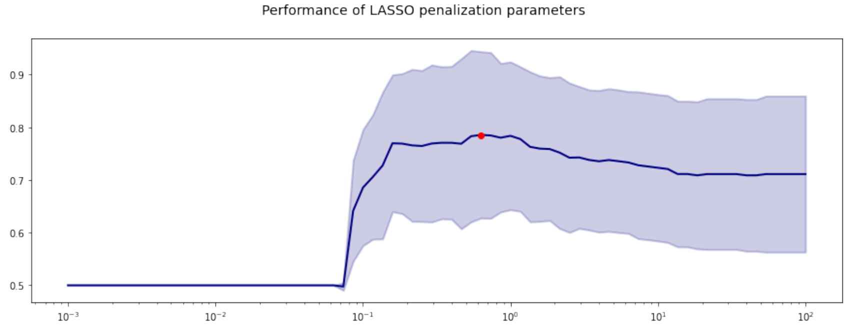


Figure A.2: 1-1: Chosen λ in grid search for wave 2.

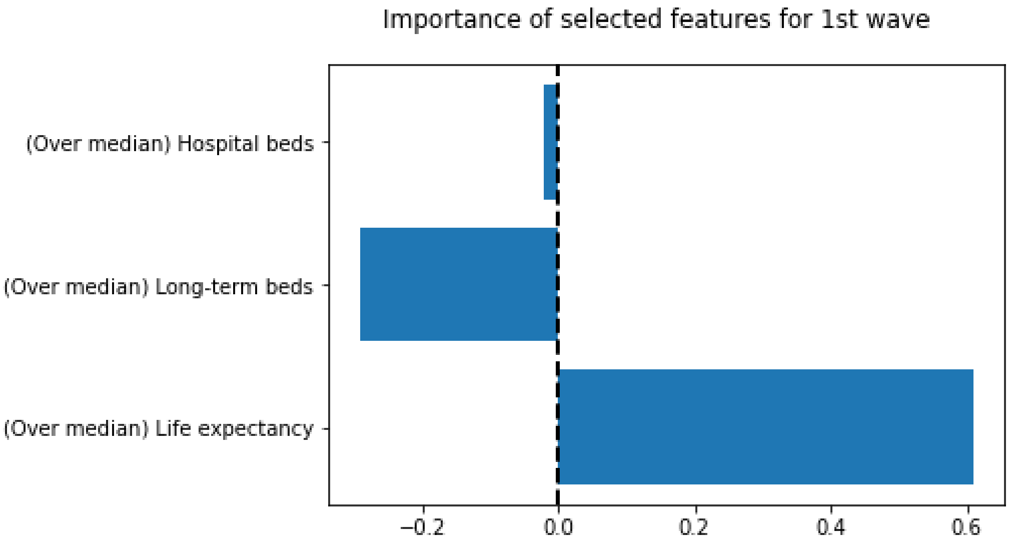


Figure A.3: 1-1: Weights of relevant features selected by LASSO in wave 1

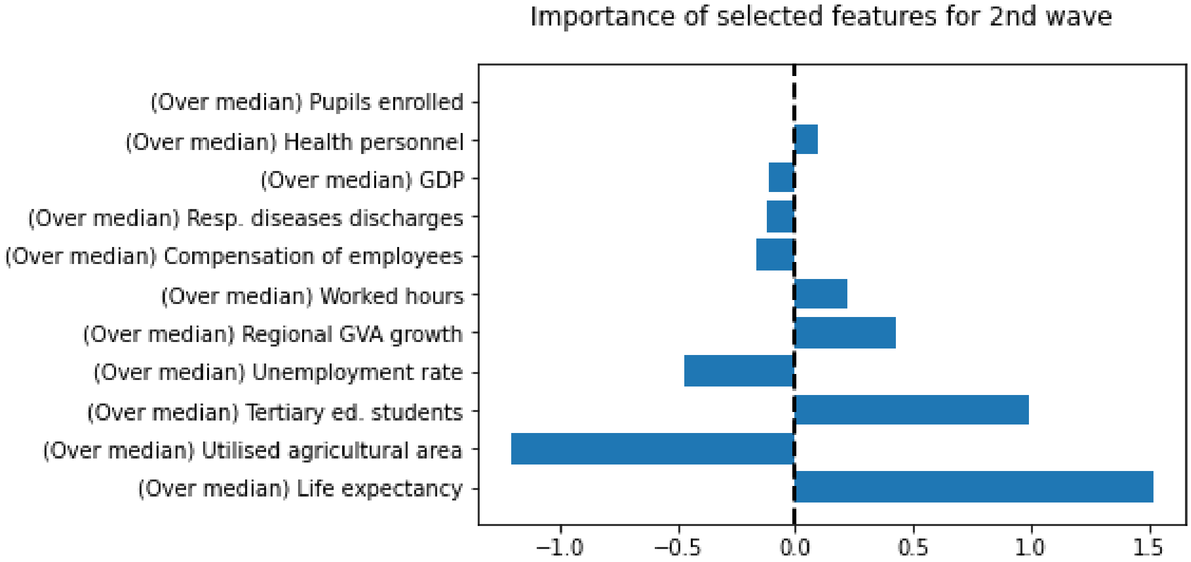


Figure A.4: 1-1: Weights of relevant features selected by LASSO in wave 2.

	λ	**ROC AUC**	**Selected features**
Wave 1	0.2929	0.7136	3
Wave 2	0.631	0.7855	11

Table A.1: 1-1: LASSO feature selection summary.

	Wave 1		Wave 2	
	Severity	Mildness	Severity	Mildness
Amount	2	6	26	602
Support	0.1042 (min) 0.1111 (max)	0.3121 (min) 0.4043 (max)	0.2014 (min) 0.2569 (max)	0.1064 (min) 0.1489 (max)
Confidence	0.8333 (min) 0.7143 (max)	0.9661 (min) 1.0 (max)	0.8 (min) 0.8409 (max)	0.9 (min) 1.0 (max)
P-value	1.276e-15 (min) 3.391e-14 (max)	4.419e-04 (min) 1.9856e-03 (max)	4.545e-06 (min) 5.009e-04 (max)	1.725e-07 (min) 9.252e-06 (max)
Factors	2 (all)	2 (4) and 3 (2)	2-6 (all)	2-9 (all)

Table A.2: 1-1: AR mining results summary.

	Wave 1		Wave 2	
	Severity	Mildness	Severity	Mildness
Precision	86.36%	100.00%	73.47%	81.13%
Recall	100.00%	18.03%	88.89%	71.67%
Accuracy	97.92%	30.56%	75.69%	80.85%
F1-measure	92.68%	30.56%	80.45%	76.11%

Table A.3: 1-1: Subject analysis summary.

	Mildness		Severity	
	1+ rules	0 rules	1+ rules	0 rules
High	2	17	19	0
Low	76	46	3	122

Table A.4: 1-1: Subject analysis for wave 1.

	Mildness		Severity	
	1+ rules	0 rules	1+ rules	0 rules
High	10	71	72	9
Low	43	17	26	37

Table A.5: 1-1: Subject analysis for wave 2.

	Wave 1		Wave 2	
	Severity	Mildness	Severity	Mildness
High	1.632	0.316	10.025	8.642
Low	0.048	2.492	2.984	164.266

Table A.6: 1-1: Average satisfied rules.

	Wave 1		Wave 2	
	Severity	Mildness	Severity	Mildness
Severity	2	0	70	22
Mildness	2	12	1262	2552

Table A.7: 1-1: Feature analysis counting all repetitions.

	Wave 1		Wave 2	
	Severity	Mildness	Severity	Mildness
Severity	2	0	11	5
Mildness	2	5	8	13

Table A.8: 1-1: Feature analysis counting unique appearances.

A.1.2. Case 1-3

Here we present the case of the target with the first choice of dichotomization (using the mean) and the features with the third choice of dichotomization (above and below the IQR). As stated in section 2.1, the AR mining step does not add complementary columns for the selected features. This way, regions have items only where they are above or below the IQR, and not when they are inside it.

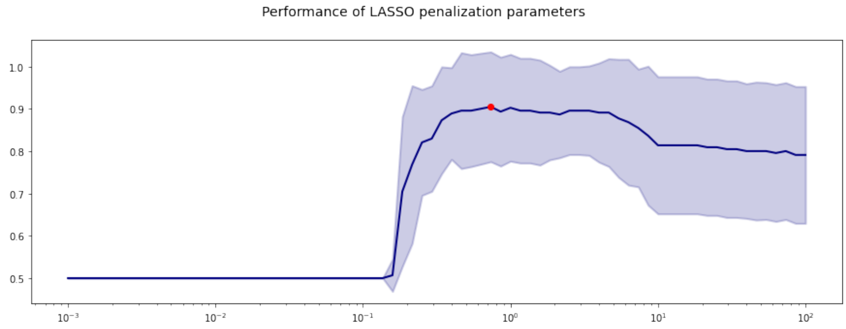


Figure A.5: 1-3: Chosen λ in grid search for wave 1.

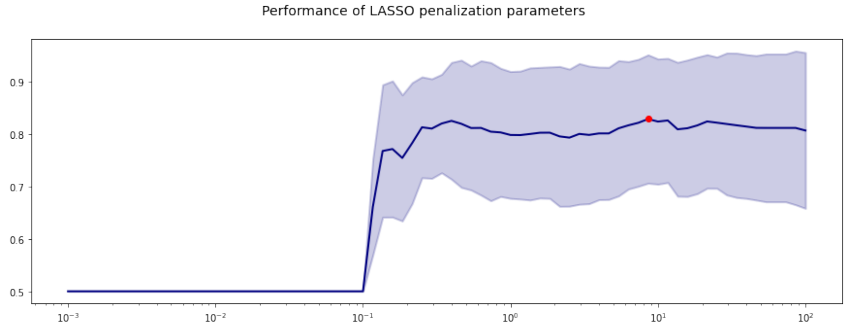


Figure A.6: 1-3: Chosen λ in grid search for wave 2.

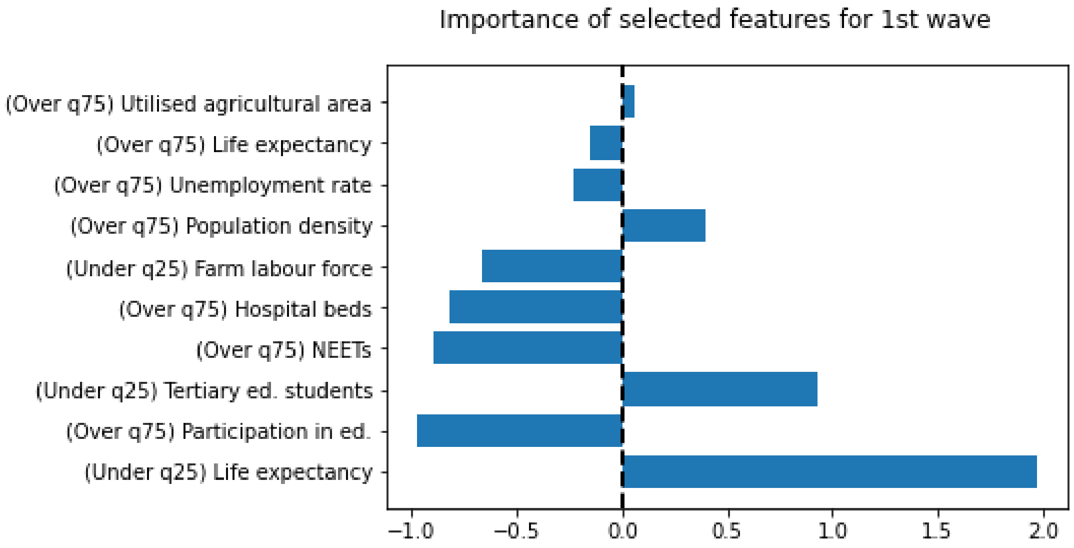


Figure A.7: 1-3: Weights of relevant features selected by LASSO in wave 1

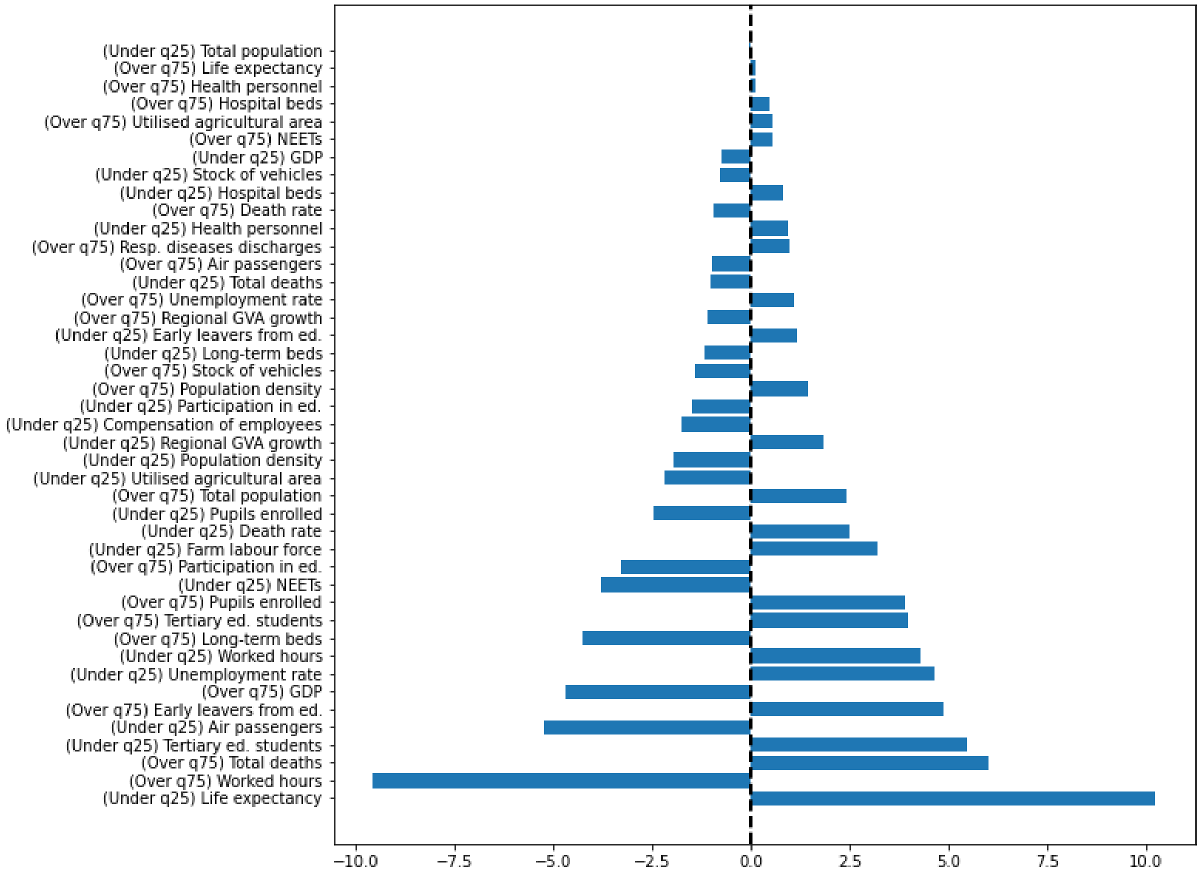


Figure A.8: 1-3: Weights of relevant features selected by LASSO in wave 2.

	λ	ROC AUC	Selected features
Wave 1	0.7356	0.9045	10
Wave 2	8.577	0.8292	43

Table A.9: 1-3: LASSO feature selection summary.

	Wave 1		**Wave 2**	
	Severity	Mildness	Severity	Mildness
Amount	1	1	1	8
Support	0.03546	0.1528	0.111	0.0625
Confidence	0.7143	1.0	1.0	0.9
P-value	4.72e-4	1.883e-02	4.883e-05	1.716e-03
Factors	2 (all)	2 (all)	2 (all)	2-5 (all)

Table A.10: 1-3: AR mining results summary.

	Wave 1		**Wave 2**	
	Severity	Mildness	Severity	Mildness
Precision	71.43%	100.00%	100.00%	90.00%
Recall	26.32%	18.03%	19.75%	15.00%
Accuracy	88.65%	30.56%	54.86%	63.89%
F1-measure	38.46%	30.56%	32.99%	25.71%

Table A.11: 1-3: Subject analysis summary.

	Mildness		**Severity**	
	1+ rules	0 rules	1+ rules	0 rules
High	0	22	5	14
Low	22	100	2	120

Table A.12: 1-3: Subject analysis for wave 1.

	Mildness		Severity	
	1+ rules	0 rules	1+ rules	0 rules
High	1	83	16	65
Low	9	51	0	63

Table A.13: 1-3: Subject analysis for wave 2.

	Wave 1		Wave 2	
	Severity	Mildness	Severity	Mildness
High	0.263	0.000	0.198	0.095
Low	0.016	0.180	0.000	1.200

Table A.14: 1-3: Average satisfied rules.

	Wave 1		Wave 2	
	Severity	Mildness	Severity	Mildness
Severity	2	0	2	0
Mildness	0	2	16	12

Table A.15: 1-3: Feature analysis counting all repetitions.

	Wave 1		Wave 2	
	Severity	Mildness	Severity	Mildness
Severity	2	0	2	0
Mildness	0	2	3	2

Table A.16: 1-3: Feature analysis counting unique appearances.

A.1.3. Case 2-1

Here we present the case of the target with the second choice of dichotomization (using the median) and the features with the first choice of dichotomization (using the median). As stated in section 2.1, the AR mining step considers the adding of complementary columns for the selected features. This way, all regions have always the same amount of items, where a single item can refer to the factor being over the median or under the median.

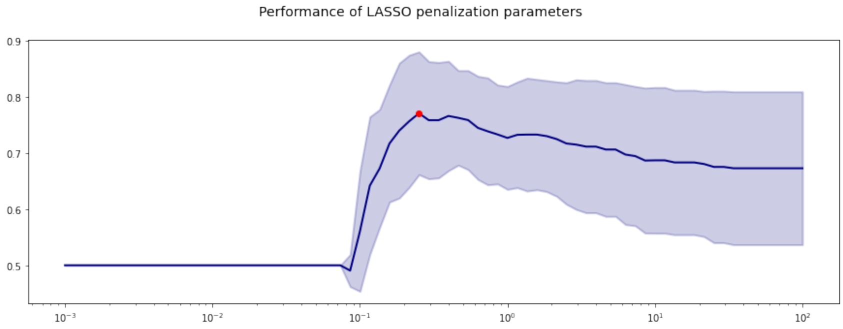


Figure A.9: 2-1: Chosen λ in grid search for wave 1.

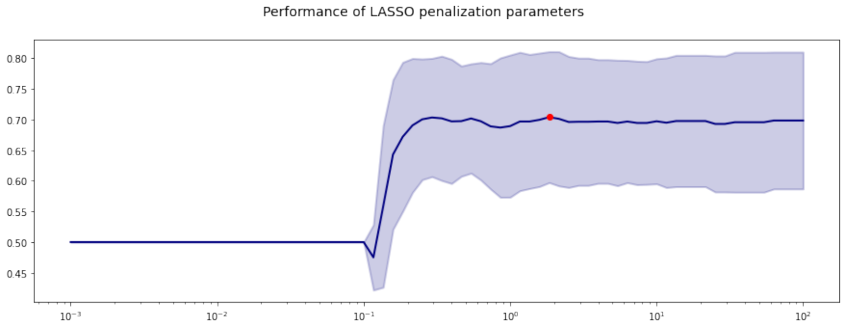


Figure A.10: 2-1: Chosen λ in grid search for wave 2.

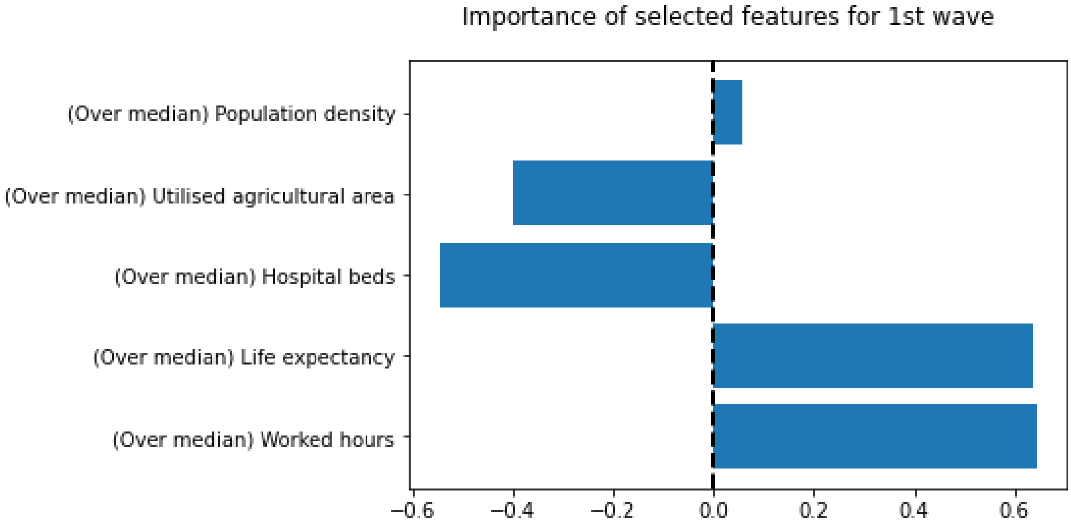


Figure A.11: 2-1: Weights of relevant features selected by LASSO in wave 1

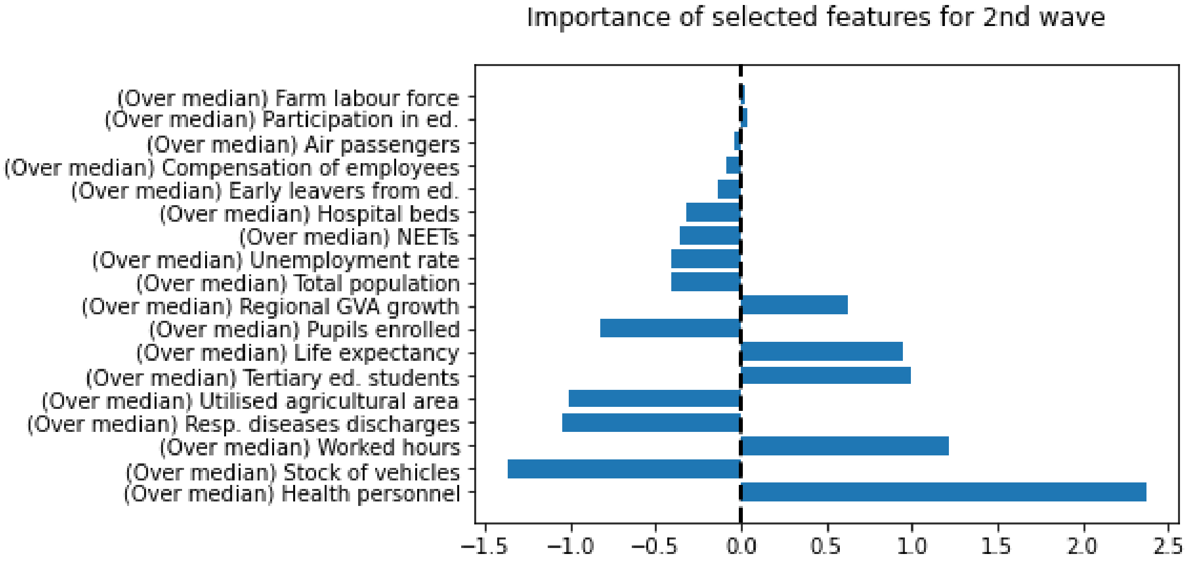


Figure A.12: 2-1: Weights of relevant features selected by LASSO in wave 2.

	λ	**ROC AUC**	**Selected features**
Wave 1	0.2512	0.7704	5
Wave 2	1.8478	0.704	18

Table A.17: 2-1: LASSO feature selection summary.

	Wave 1		Wave 2	
	Severity	Mildness	Severity	Mildness
Amount	4	1	7	1
Support	0.234 (min) 0.2482 (max)	0.2128	0.1064 (min) 0.1348 (max)	0.2057
Confidence	0.7234 (min) 0.7727 (max)	0.9091	0.9 (min) 0.9412 (max)	0.8056
P-value	1.344e-05 (min) 1.906e-04 (max)	2.028e-08	4.063e-05 (min) 1.395e-04 (max)	1.331e-05
Factors	2 (all)	3 (all)	3-5 (all)	3 (all)

Table A.18: 2-1: AR mining results summary.

	Wave 1		Wave 2	
	Severity	Mildness	Severity	Mildness
Precision	68.29%	90.91%	89.74%	80.56%
Recall	78.87%	42.86%	49.30%	41.43%
Accuracy	70.92%	69.50%	71.63%	65.96%
F1-measure	73.20%	58.25%	63.64%	54.72%

Table A.19: 2-1: Subject analysis summary.

	Mildness		Severity	
	1+ rules	0 rules	1+ rules	0 rules
High	3	68	56	15
Low	30	40	26	44

Table A.20: 2-1: Subject analysis for wave 1.

	Mildness		**Severity**	
	1+ rules	0 rules	1+ rules	0 rules
High	7	64	35	36
Low	29	41	4	66

Table A.21: 2-1: Subject analysis for wave 2.

	Wave 1		**Wave 2**	
	Severity	Mildness	Severity	Mildness
High	1.915	0.042	1.592	0.414
Low	0.671	0.429	0.129	0.099

Table A.22: 2-1: Average satisfied rules.

	Wave 1		**Wave 2**	
	Severity	Mildness	Severity	Mildness
Severity	8	0	23	0
Mildness	0	3	0	3

Table A.23: 2-1: Feature analysis counting all repetitions.

	Wave 1		**Wave 2**	
	Severity	Mildness	Severity	Mildness
Severity	4	0	8	0
Mildness	0	3	0	3

Table A.24: 2-1: Feature analysis counting unique appearances.

A.1.4. Case 2-2

Here we present the case of the target with the second choice of dichotomization (using the median) and the features with the second choice of dichotomization (using the inter-quantile range). As stated in section 2.1, the AR mining step considers the adding of complementary columns for the selected features. This way, all regions have always the same amount of items, where a single item can refer to the factor being outside the IQR or inside it.

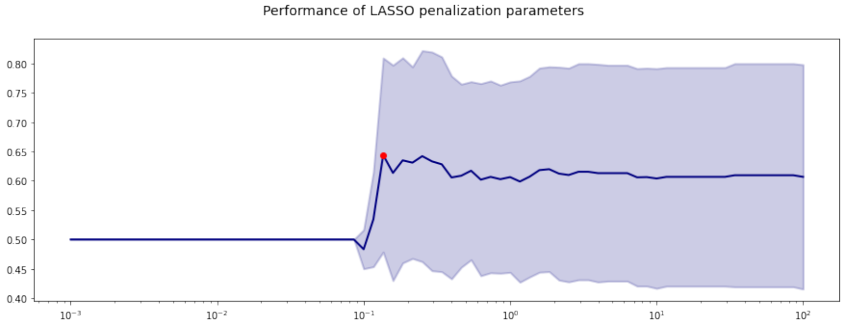


Figure A.13: 2-2: Chosen λ in grid search for wave 1.

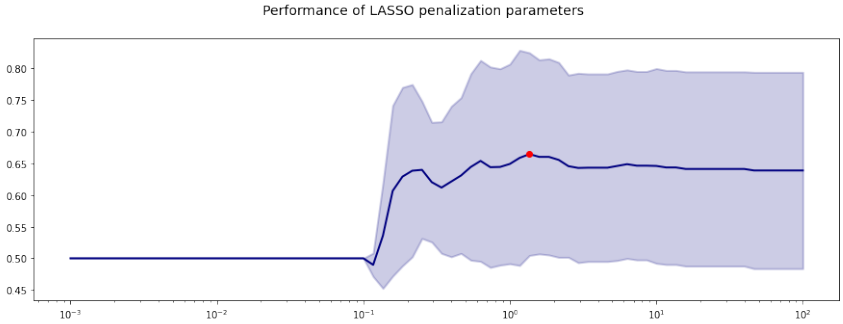


Figure A.14: 2-2: Chosen λ in grid search for wave 2.

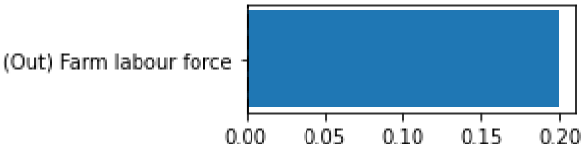


Figure A.15: 2-2: Weights of relevant features selected by LASSO in wave 1

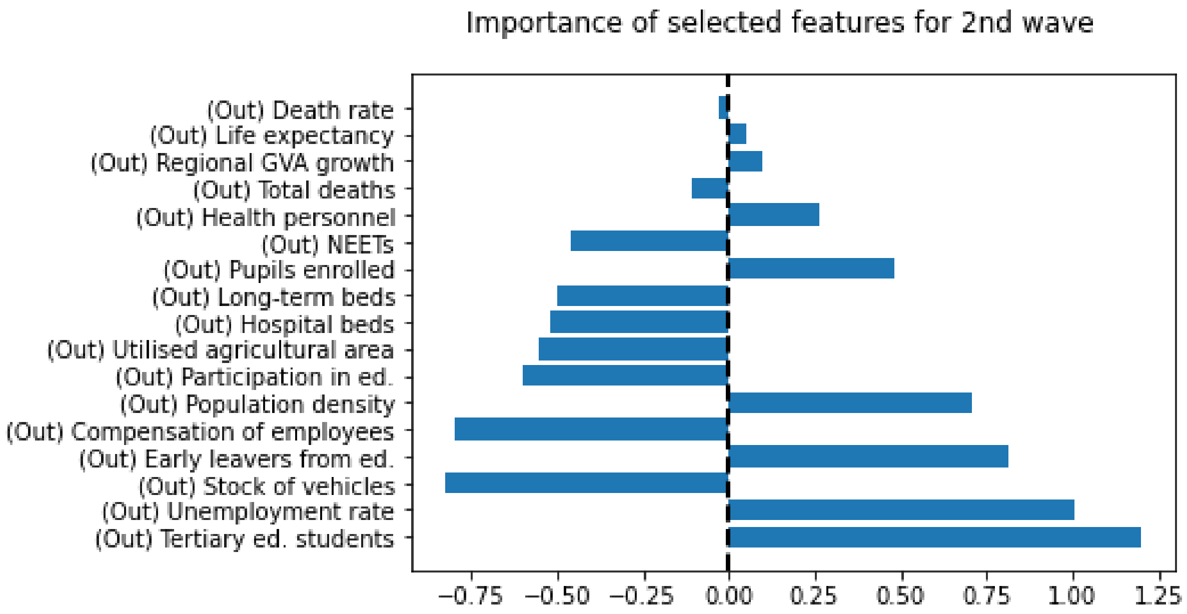


Figure A.16: 2-2: Weights of relevant features selected by LASSO in wave 2.

	λ	**ROC AUC**	**Selected features**
Wave 1	0.1359	0.644	1
Wave 2	1.3594	0.6643	17

Table A.25: 2-2: LASSO feature selection summary.

	Wave 1		**Wave 2**	
	Severity	Mildness	Severity	Mildness
Amount	0	0	7	3
Support			0.1064 (min) 0.1206 (max)	0.2057 (min) 0.2340 (max)
Confidence			0.9375 (min) 1.0 (max)	0.8286 (min) 0.8684 (max)
P-value			1.495e-05 (min) 1.395e-04 (max)	3.868e-08 (min) 4.116e-06 (max)
Factors			3 (5) and 4 (2)	3 (all)

Table A.26: 2-2: AR mining results summary.

A.1.5. Case 2-3

Here we present the case of the target with the second choice of dichotomization (using the median) and the features with the third choice of dichotomization (above and below the IQR). As stated in section 2.1, the AR mining step does not add complementary columns for the selected features. This way, regions have items only where they are above or below the IQR, and not when they are inside it.

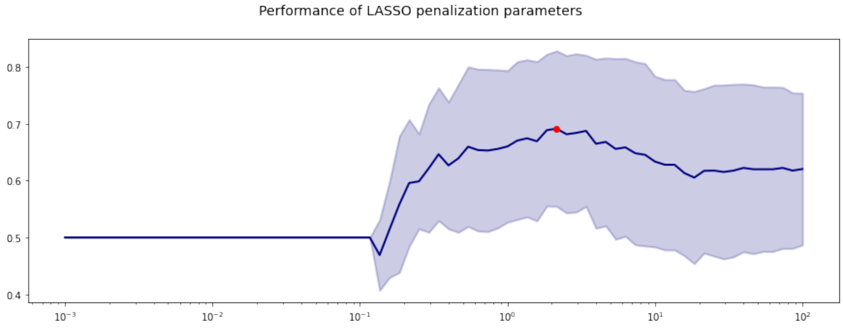


Figure A.17: 2-3: Chosen λ in grid search for wave 1.

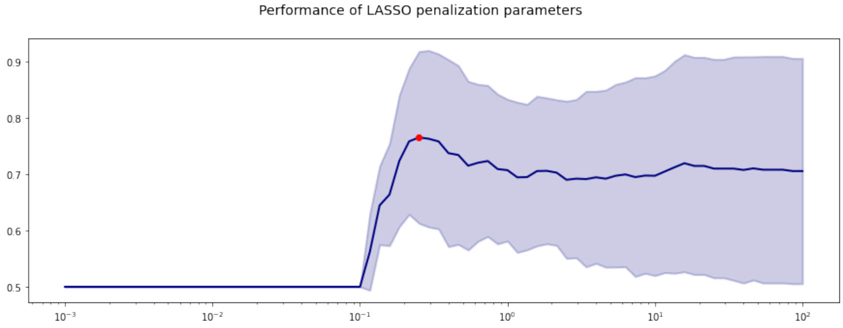


Figure A.18: 2-3: Chosen λ in grid search for wave 2.

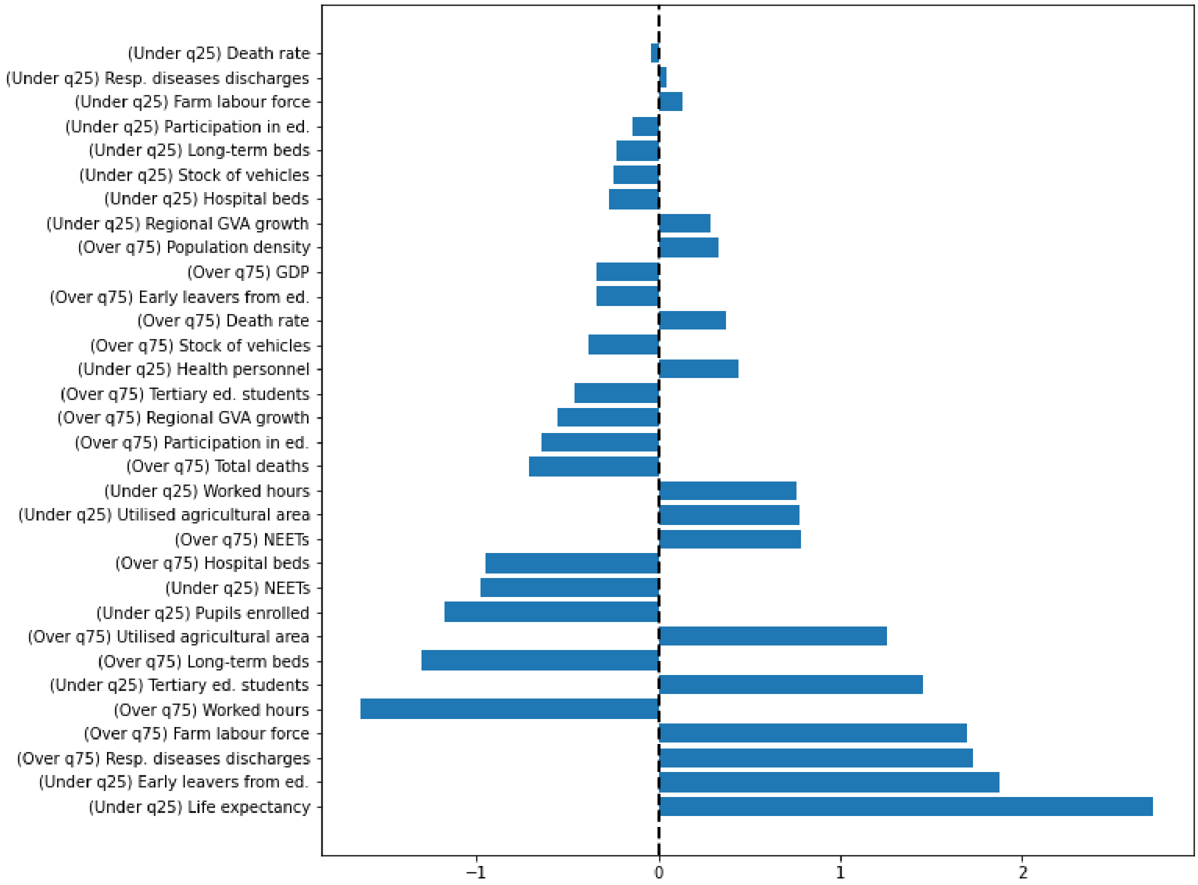


Figure A.19: 2-3: Weights of relevant features selected by LASSO in wave 1

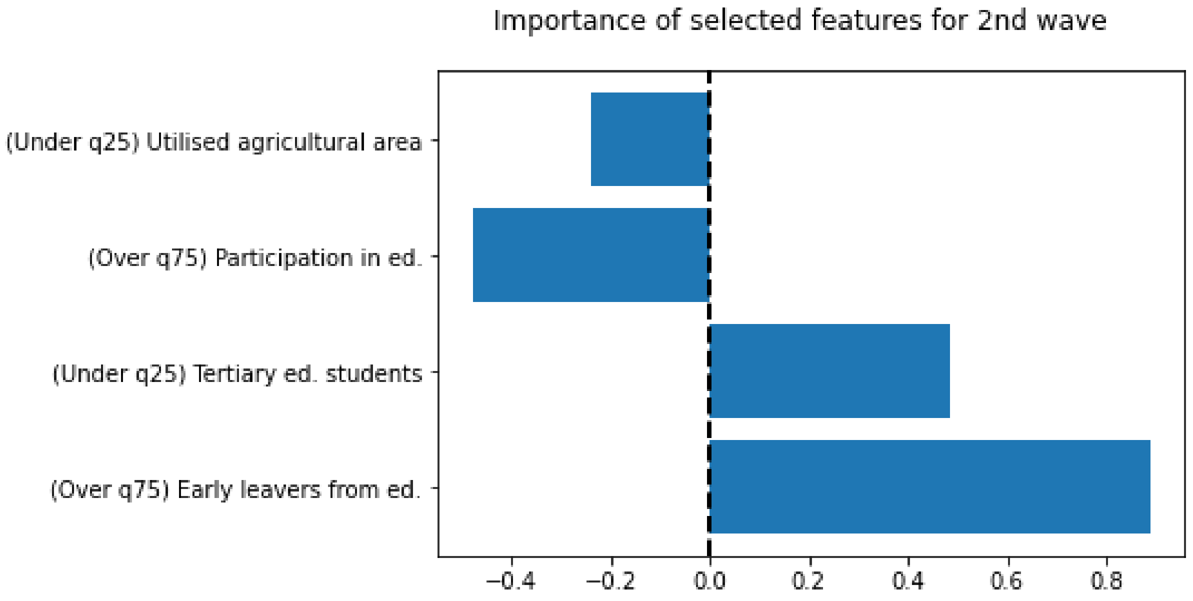


Figure A.20: 2-3: Weights of relevant features selected by LASSO in wave 2.

	λ	ROC AUC	Selected features
Wave 1	2.1544	0.6913	32
Wave 2	0.2512	0.7653	4

Table A.27: 2-3: LASSO feature selection summary.

	Wave 1		**Wave 2**	
	Severity	Mildness	Severity	Mildness
Amount	5	1	1	0
Support	0.1135 (min) 0.1277 (max)	0.08511	0.03546	
Confidence	0.8 (min) 0.8947 (max)	0.8	1.0	
P-value	1.726e-04 (min) 3.718e-03 (max)	1.21e-02	3.012e-02	
Factors	2-4 (all)	2 (all)	2 (all)	

Table A.28: 2-3: AR mining results summary.

A.1.6. Case 3-1

Here we present the case of the target with the third choice of dichotomization (using the quantile 90) and the features with the first choice of dichotomization (using the median). As stated in section 2.1, the AR mining step considers the adding of complementary columns for the selected features. This way, all regions have always the same amount of items, where a single item can refer to the factor being over the median or under the median.

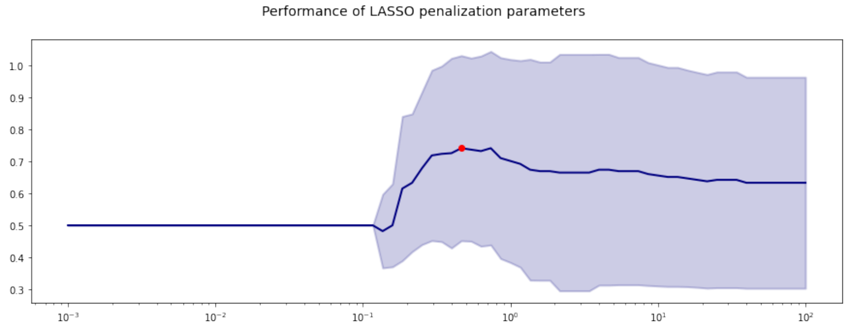


Figure A.21: 3-1: Chosen λ in grid search for wave 1.

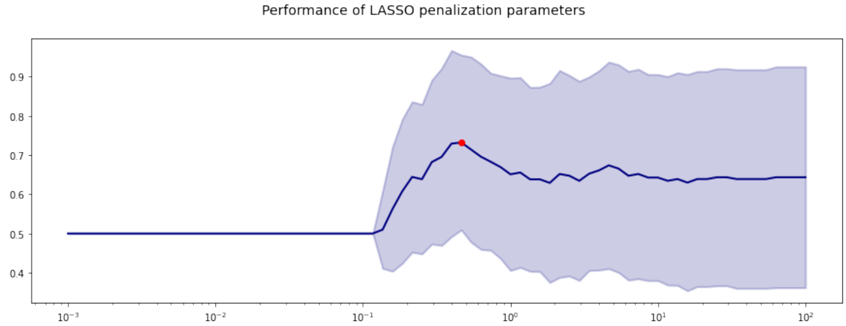


Figure A.22: 3-1: Chosen λ in grid search for wave 2.

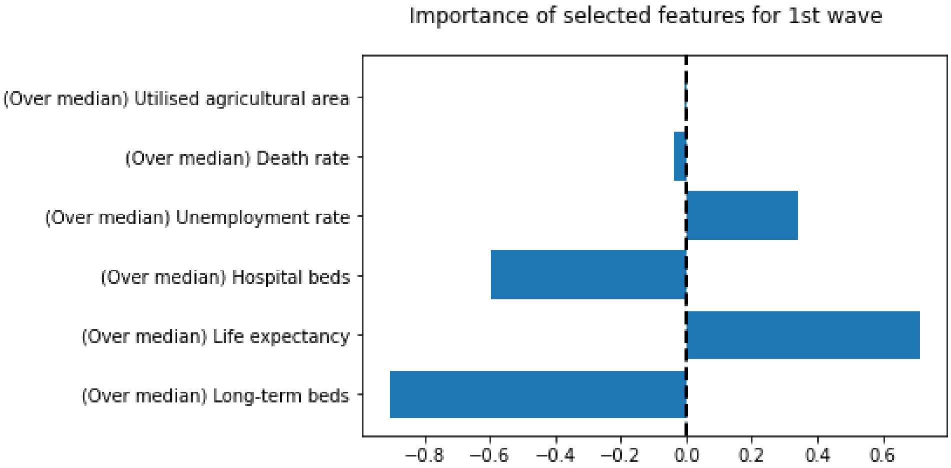


Figure A.23: 3-1: Weights of relevant features selected by LASSO in wave 1

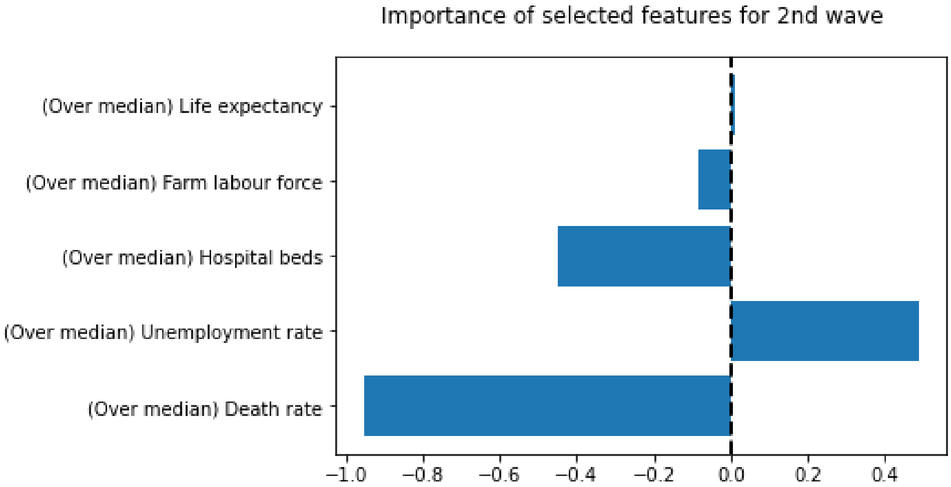


Figure A.24: 3-1: Weights of relevant features selected by LASSO in wave 2.

	λ	**ROC AUC**	**Selected features**
Wave 1	0.4642	0.7417	6
Wave 2	0.4642	0.7318	5

Table A.29: 3-1: LASSO feature selection summary.

	Wave 1		Wave 2	
	Severity	Mildness	Severity	Mildness
Amount	1	2	0	2
Support	0.06383	0.3262 (min) 0.4184 (max)		0.2837
Confidence	0.6	0.9833 (min) 1.0 (max)		1.0
P-value	4.151e-07	1.804e-03 (min) 1.921e-03 (max)		4.862e-03
Factors	4 (all)	2 (all)		2 (all)

Table A.30: 3-1: AR mining results summary.

A.1.7. Case 3-2

Here we present the case of the target with the third choice of dichotomization (using the quantile 90) and the features with the second choice of dichotomization (using the inter-quantile range). As stated in section 2.1, the AR mining step considers the adding of complementary columns for the selected features. This way, all regions have always the same amount of items, where a single item can refer to the factor being outside the IQR or inside it.

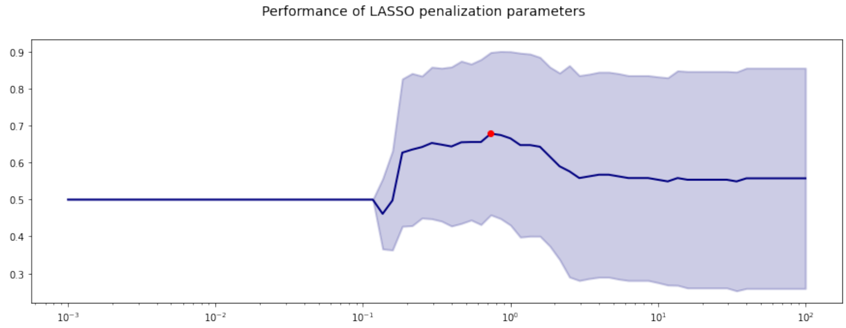


Figure A.25: 3-2: Chosen λ in grid search for wave 1.

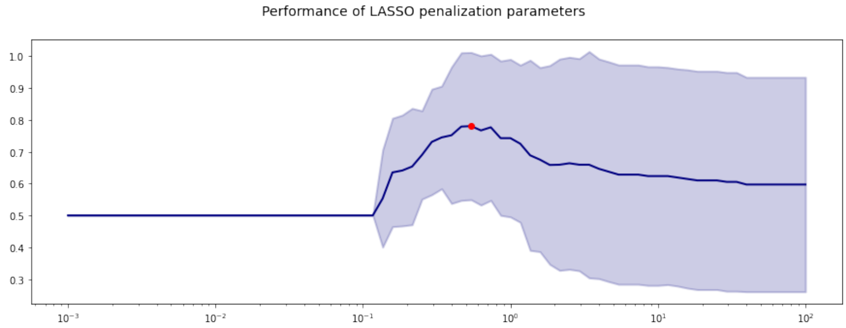


Figure A.26: 3-2: Chosen λ in grid search for wave 2.

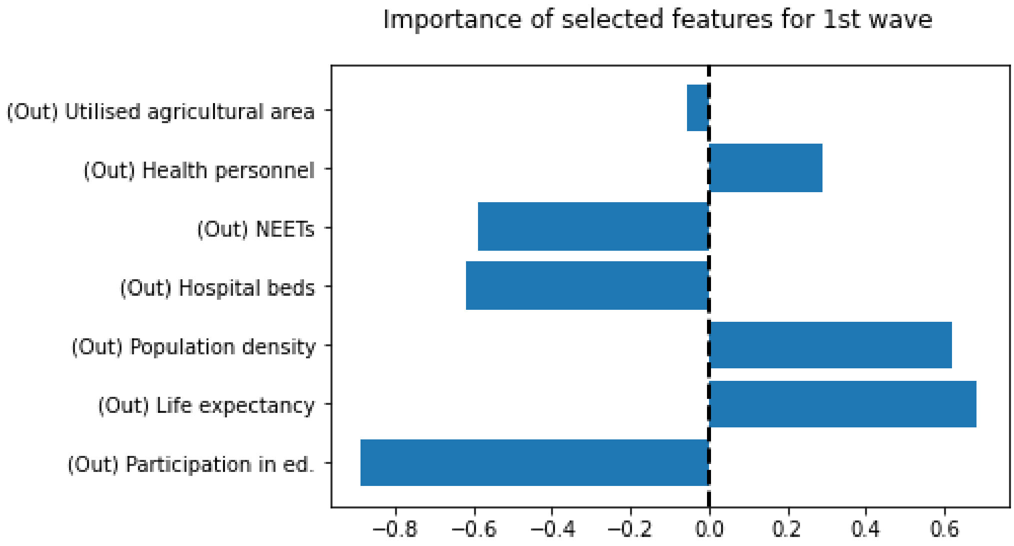


Figure A.27: 3-2: Weights of relevant features selected by LASSO in wave 1

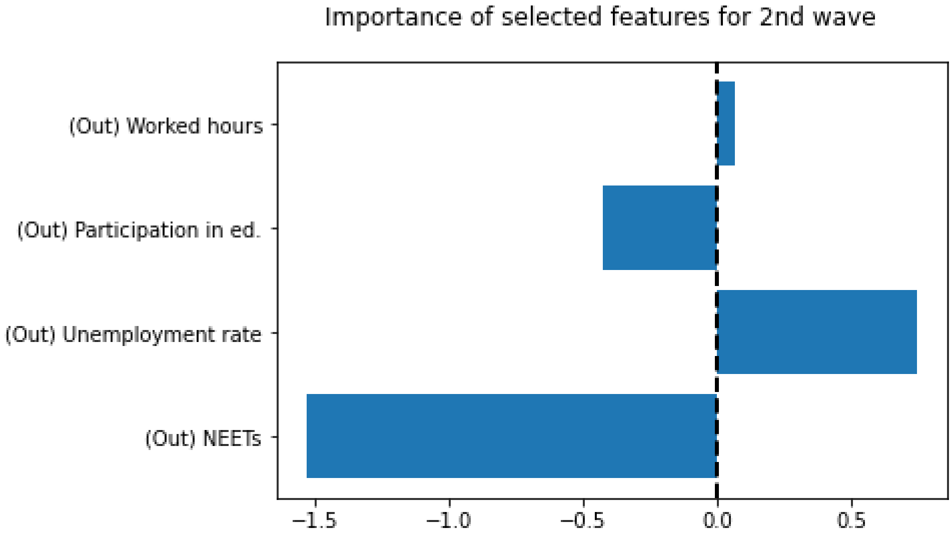


Figure A.28: 3-2: Weights of relevant features selected by LASSO in wave 2.

	λ	**ROC AUC**	**Selected features**
Wave 1	0.7356	0.6788	7
Wave 2	0.5412	0.7803	4

Table A.31: 3-2: LASSO feature selection summary.

	Wave 1		Wave 2	
	Severity	Mildness	Severity	Mildness
Amount	10	1	1	1
Support	0.02128 (min) 0.03546 (max)	0.3262	0.03546	0.2482
Confidence	0.7143 (min) 1.0 (max)	1.0	0.625	1.0
P-value	1.285e-04 (min) 3.720e-03 (max)	1.804e-03	3.215e-04	1.060e-02
Factors	4-6 (all)	2 (all)	4 (all)	2 (all)

Table A.32: 3-2: AR mining results summary.

	Wave 1		Wave 2	
	Severity	Mildness	Severity	Mildness
Precision	72.73%	100.00%	62.50%	100.00%
Recall	53.33%	36.51%	33.33%	27.78%
Accuracy	92.91%	43.26%	90.78%	35.46%
F1-measure	61.54%	53.49%	43.48%	43.48%

Table A.33: 3-2: Subject analysis summary.

	Mildness		Severity	
	1+ rules	0 rules	1+ rules	0 rules
High	0	15	8	7
Low	46	80	3	123

Table A.34: 3-2: Subject analysis for wave 1.

	Mildness		Severity	
	1+ rules	0 rules	1+ rules	0 rules
High	0	15	5	10
Low	35	91	3	123

Table A.35: 3-2: Subject analysis for wave 2.

	Wave 1		Wave 2	
	Severity	Mildness	Severity	Mildness
High	2.133	0.0	0.333	0.0
Low	0.048	0.365	0.024	0.278

Table A.36: 3-2: Average satisfied rules.

	Wave 1		Wave 2	
	Severity	Mildness	Severity	Mildness
Severity	35	13	4	0
Mildness	0	2	0	2

Table A.37: 3-2: Feature analysis counting all repetitions.

	Wave 1		Wave 2	
	Severity	Mildness	Severity	Mildness
Severity	6	3	4	0
Mildness	0	2	0	2

Table A.38: 3-2: Feature analysis counting unique appearances.

A.1.8. Case 3-3

Here we present the case of the target with the first choice of dichotomization (using the mean) and the features with the third choice of dichotomization (above and below the IQR). As stated in section 2.1, the AR mining step does not add complementary columns for the selected features. This way, regions have items only where they are above or below the IQR, and not when they are inside it.

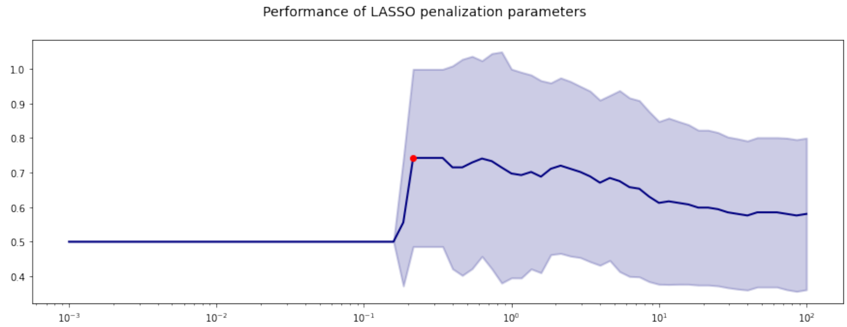


Figure A.29: 3-3: Chosen λ in grid search for wave 1.

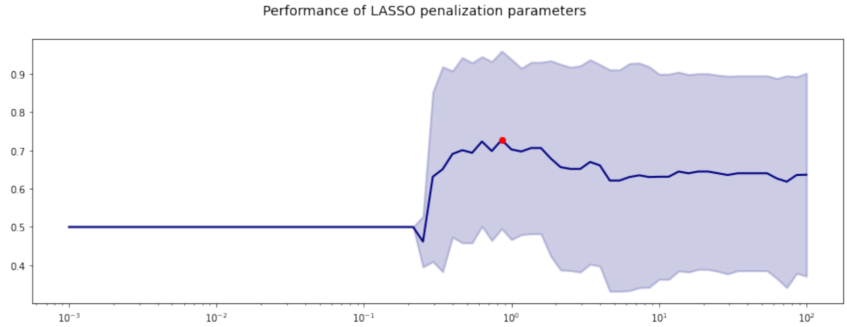


Figure A.30: 3-3: Chosen λ in grid search for wave 2.

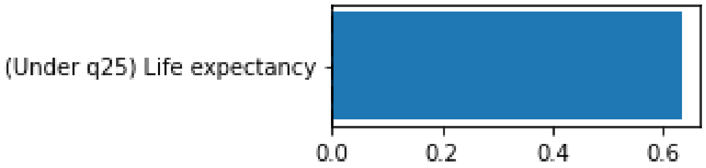


Figure A.31: 3-3: Weights of relevant features selected by LASSO in wave 1

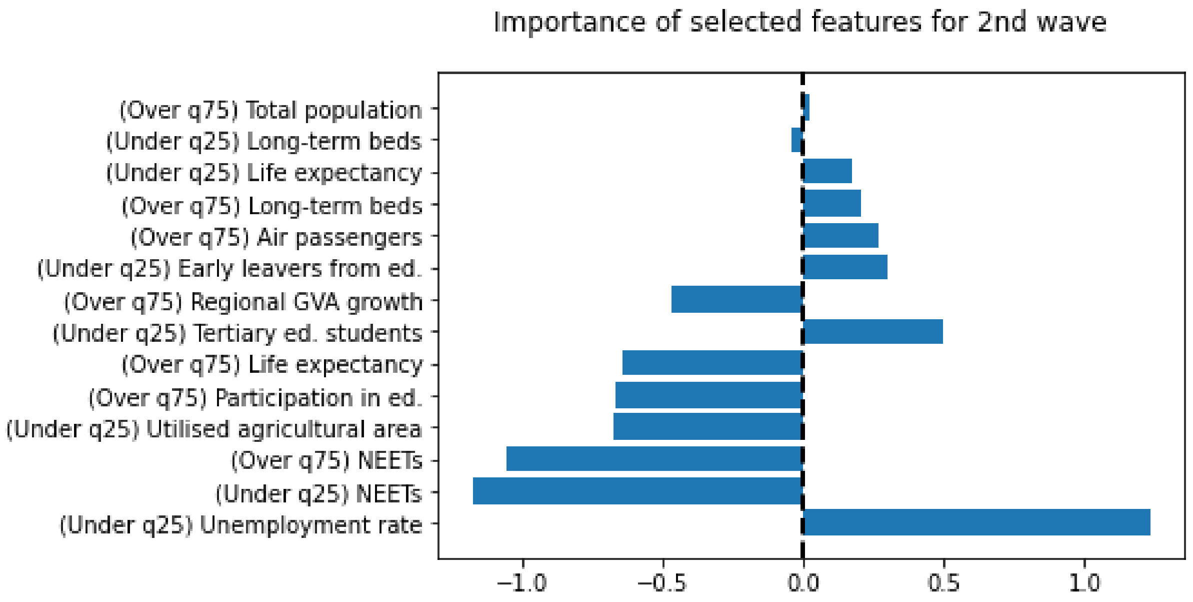


Figure A.32: 3-3: Weights of relevant features selected by LASSO in wave 2.

	λ	**ROC AUC**	**Selected features**
Wave 1	0.2154	0.742	1
Wave 2	0.8577	0.7273	14

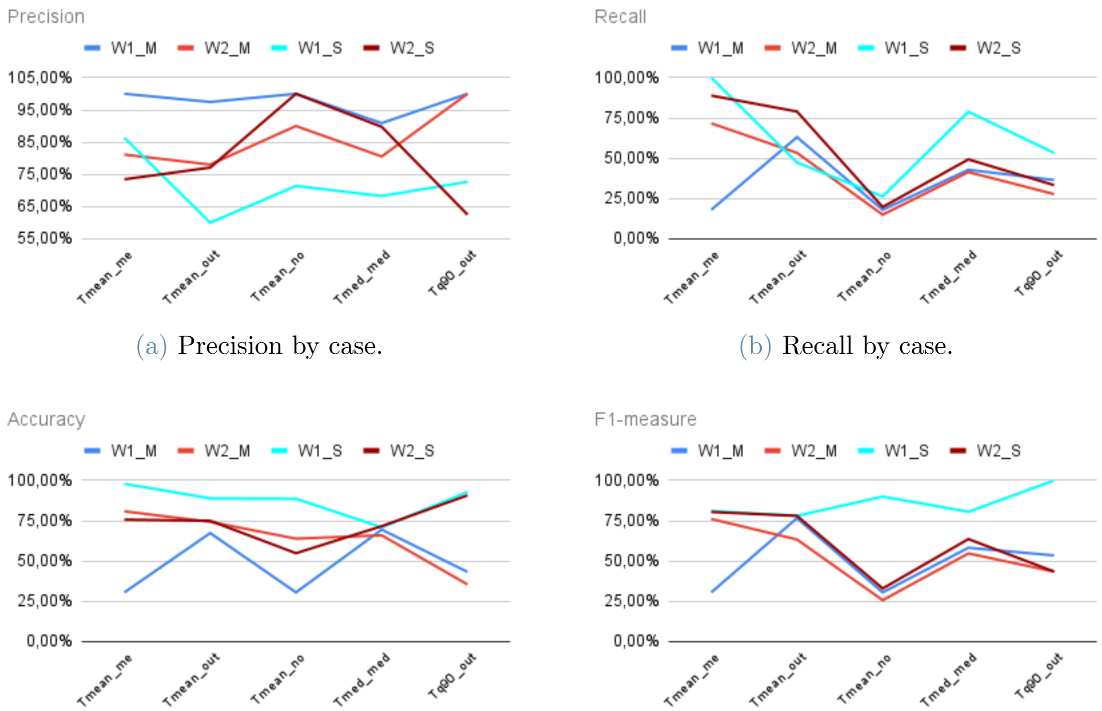
Table A.39: 3-3: LASSO feature selection summary.

	Wave 1		**Wave 2**	
	Severity	Mildness	Severity	Mildness
Amount	0	0	2	0
Support	0.02128			
Confidence	0.75			
P-value	3.720e-03			
Factors	3 (all)			

Table A.40: 3-3: AR mining results summary.

A.2. Comparison between cases

Given that cases 2-2, 2-3, 3-1 and 3-3 do not find significant rules of all types for all waves, they are discarded from the comparison to consider the pre-processing that presents the most satisfying results. With the remaining cases, we plot for each metric the performance of each pre-processing case, in order to find the best set of performances by measure. These are showed in Figure A.33.



(a) Precision by case.

(b) Recall by case.

(c) Accuracy by case.

(d) F1-measure by case.

Figure A.33: Subject analysis performance by pre-processing case.

We can notice that case 1-2 presents the best fitting set in recall, accuracy and F1-measure. The only measure where it is outperformed is precision, where case 1-3 has the best performing set, but performs far worse regarding specially accuracy and F1-measure. For this reason, we choose case 1-2 as the best case for analysis, since its resulting rules associate better to the proposed response.

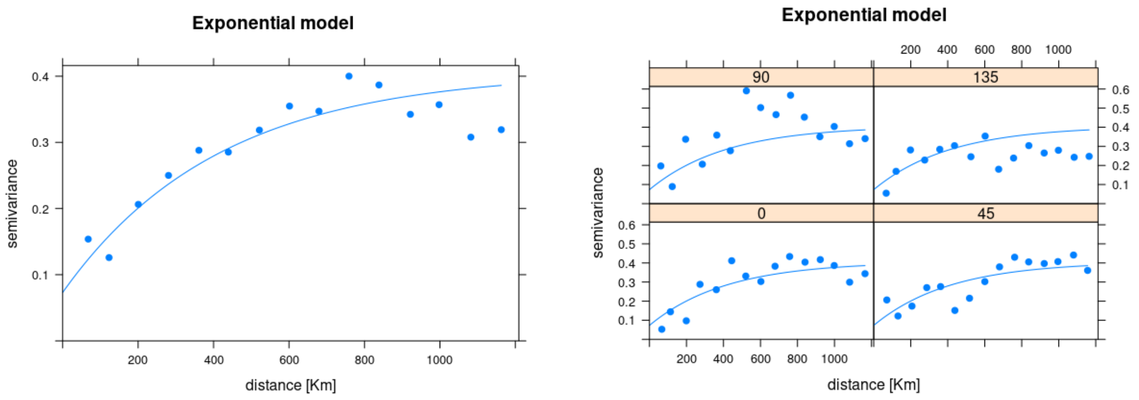
B | Appendix B

B.1. First wave variogram model selection

We fit the exponential, spherical and gaussian models, all added to a nugget model, in order to find the best approximation considering the main characteristics of the shape of such models. We do this by fitting the main residual variogram, visualizing their behavior over the directional variograms, and measuring their approximation errors. We can assess visually the fit of the discarded models (see Fig. B.1 and B.2), while Table B.1 shows the results for all models, where we see that the lowest errors correspond to the gaussian model.

Model	Nugget	Sill	Range	SSErr	Med err	Mean err
Exponential	0.07314	0.33311	414.4195	1.520e-05	0.1377	0.1285
Spherical	0.09069	0.27088	759.1032	1.206e-05	0.1004	0.1037
Gaussian	0.12522	0.23679	358.5561	9.883e-06	0.10259	0.09868

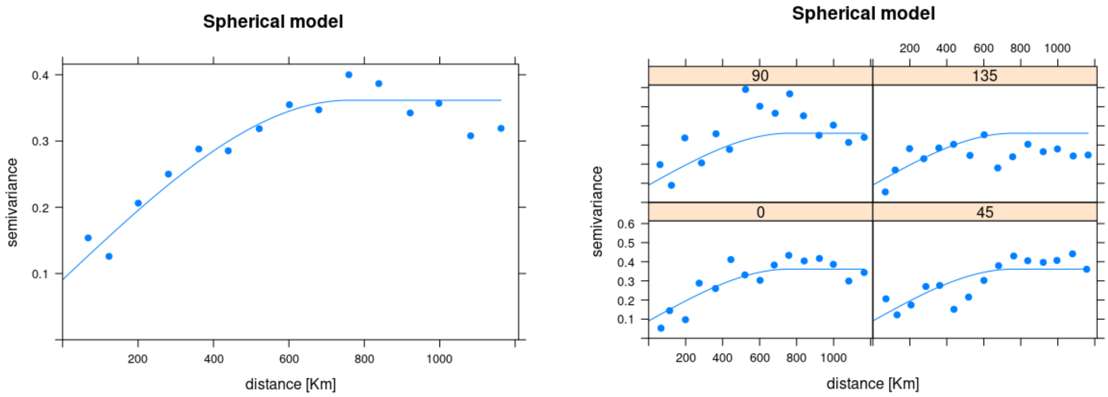
Table B.1: Summary fitted variogram models for wave 1.



(a) Residual variogram.

(b) Directional variogram.

Figure B.1: Exponential model of residuals from a linear trend depending on longitude.



(a) Residual variogram.

(b) Directional variogram.

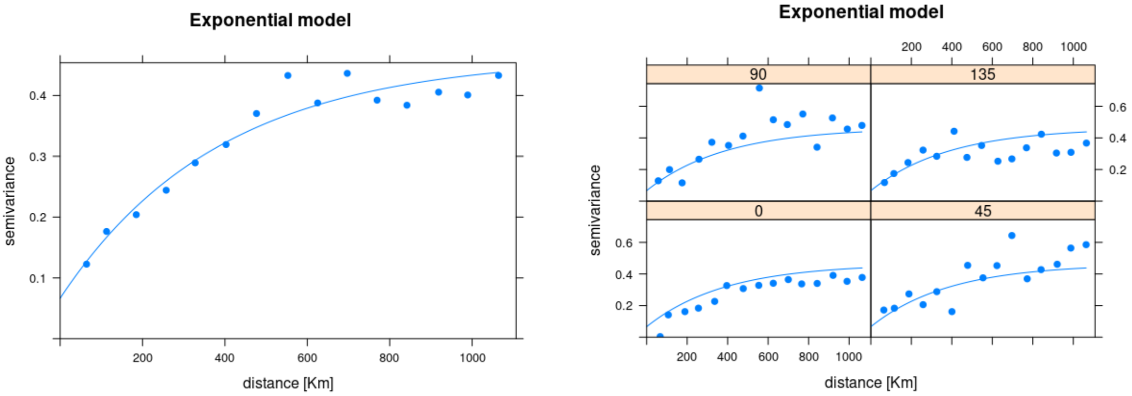
Figure B.2: Spherical model of residuals from a linear trend depending on longitude.

B.2. Second wave variogram model selection

We fit the exponential, spherical and gaussian models, all added to a nugget model, in order to find the best approximation considering the main characteristics of the shape of such models. We do this by fitting the main residual variogram, visualizing their behavior over the directional variograms, and measuring their approximation errors. We can assess visually the fit of the discarded models (see Fig. B.3 and B.4), while Table B.2 shows the results for all models, where we see that the lowest errors correspond to the gaussian model, except for the residual sum of squares that, however has a neglectable difference with the other models.

Model	Nugget	Sill	Range	SSErr	Med err	Mean err
Exponential	0.06645	0.39698	386.9507	5.328e-06	0.2444	0.2240
Spherical	0.09043	0.31983	720.8476	3.617e-06	0.1932	0.1646
Gaussian	0.13183	0.27286	332.7346	5.387e-06	0.1918	0.1605

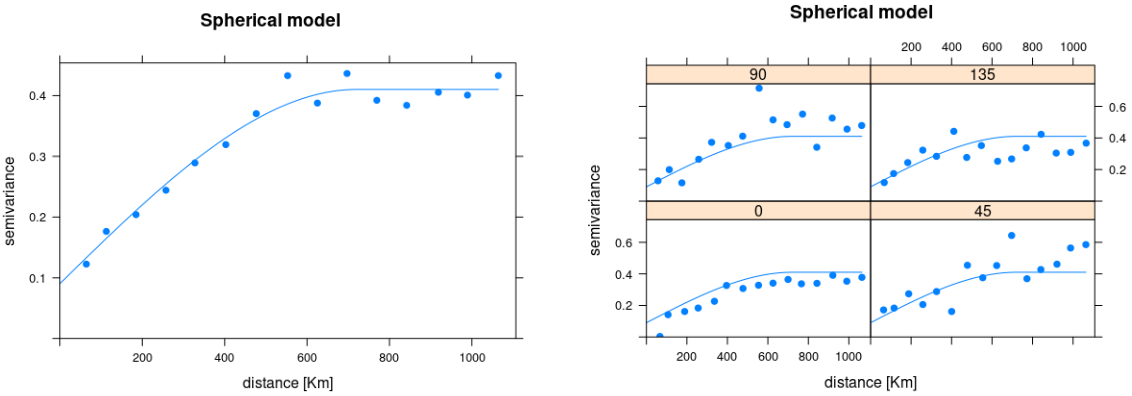
Table B.2: Summary fitted variogram models for wave 2.



(a) Residual variogram.

(b) Directional variogram.

Figure B.3: Exponential model of residuals from a linear trend depending on latitude.



(a) Residual variogram.

(b) Directional variogram.

Figure B.4: Spherical model of residuals from a linear trend depending on latitude.

List of Figures

List of Tables

Acknowledgements

I want to acknowledge the incredible support my family has given me; Giovanni, Paola, Giancarlo, Rocío, Katia, I will return home grateful that even the greatest distance will not keep us apart. I also want to acknowledge my family in Huentelauquen and Champa, to which I thank that kept rooting for me. I want to thank my family members that may not be physically here, but I carry deep in my heart; specially Raúl, Rorro, Carlos, Mimí, Margarita, I hope I can make you proud as well. I want to also acknowledge my second family, that I met during this time; Francesco specially, that helped me get through most hardships during this experience, and gave me inspiration and support.

I want to acknowledge many people that have been important in this path, that allowed me to discover many wonderful and important things, without them I would not have been able to arrive so far. Haru (and the Nakamas), Nati, Cami, Nattu, my teachers at the Liceo Carmela Carvajal de Prat, specially the ones that made me grow my enthusiast for mathematics, Víctor Corominas and Sergio "Joe" Santelices. My friends, colleagues and superiors from university, specially the greatest student project groups I could have ever asked for, at GoIng, SAI, SIAM, CEIINS, thank you for all the experience and space you gave me to learn and grow.

I want to acknowledge the special people that taught me and supported me; Anselmo, Coca, Ganter, Alonso, Pancho, Fran, and my dear community Hogwarts Chile. And last but not least, the group of people that I met here at Politecnico that led me to this thesis work and motivated me further into this field, my Applied Statistics and Bayesian Statistics project groups, thank you for our great work together.