COVID-19 spread over Europe

A statistical study to detect regional contextual factors

Francesca Anfossy

Supervisor: Francesca Ieva

Co-supervisors: Pietro Pinoli and Laura Savare'

Table of contents

1. Objectives
2. Data
3. Methods
4. Analysis results
5. Discussion
6. Conclusions

Objectives

Motivation

We study the COVID-19 spread at the level of EU regions, where we consider non-epidemiological data, regional social-demographic data of the COVID@Lombardy dataset.
We aim to detect the role of complementary predispositional characteristics that can shed light on preemptive policies for a further preparation towards a potential future outbreak.

Approaches

We study two approaches to understand the phenomenon:

- An association rule mining approach, that is efficient in obtaining results and easy to understand.
- A geo-statistical analysis approach, that provides richer analysis and considers the key aspect of geographical virus spreading.

Data

Francesca Anfossy

POLITECNICO MILANO 1863

Summary

- 144 regions from 19 countries
- Cases density of the two waves of 2020
- 23 factors: Education (5), Population (5), Healthcare (4), Economy (5), Mobility (2) and Primary sector (2).

Regional factors

Education	Demography	Healthcare
Early leavers rate	Life expectancy	Resp. discharges
Tert. ed. students	Pop. density	Long term care beds
Participation rate	Population	Health personnel
Pupils enrolled	Death rate	Hospital beds
	Total deaths	

Economy	Other sectors
Unemployment rate	Air passengers
Worked hours	Stock of vehicles
GVA growth rate	Farm labour force
Employees' compensation	Utilised agricultural area
GDP	

Table: Regional factors grouped in types related to economy, population, healthcare, economy, mobility and primary sector

Methods

Association rule mining approach

Pre-processing

Choose an adequate dichotomization of both the response and the covariates.

Association rule mining approach

Pre-processing

Choose an adequate dichotomization of both the response and the covariates.

LASSO selection

Perform Ordinary Logistic Regression with LASSO penalization to select significant factors for each wave.

Association rule mining approach

Pre-processing

Choose an adequate dichotomization of both the response and the covariates.

LASSO selection

Perform Ordinary Logistic Regression with LASSO penalization to select significant factors for each wave.

AR mining

Generate association rules with the selected features. This is done for each wave and each type of response.

Geo-statistics approach

Pre-processing

Approximate NAs and filter out incomplete columns and observations.

Geo-statistics approach

Pre-processing

Approximate NAs and filter out incomplete columns and observations.

LASSO selection

Perform linear regression with LASSO penalization to select significant factors for each wave.

Geo-statistics approach

Pre-processing

Approximate NAs and filter out incomplete columns and observations.

LASSO selection

Perform linear regression with LASSO penalization to select significant factors for each wave.

Variogram modelling

Fit a variogram model over the residuals to study spatial dependency.

Geo-statistics approach

Pre-processing

Approximate NAs and filter out incomplete columns and observations.

LASSO selection

Perform linear regression with LASSO penalization to select significant factors for each wave.

Variogram modelling

Fit a variogram model over the residuals to study spatial dependency.

LISA cluster analysis

Compare LISA clusters over the regions regarding the cases density, transformed response, and model residuals.

Analysis results

Francesca Anfossy

POLITECNICO MILANO 1863

AR results for wave 1

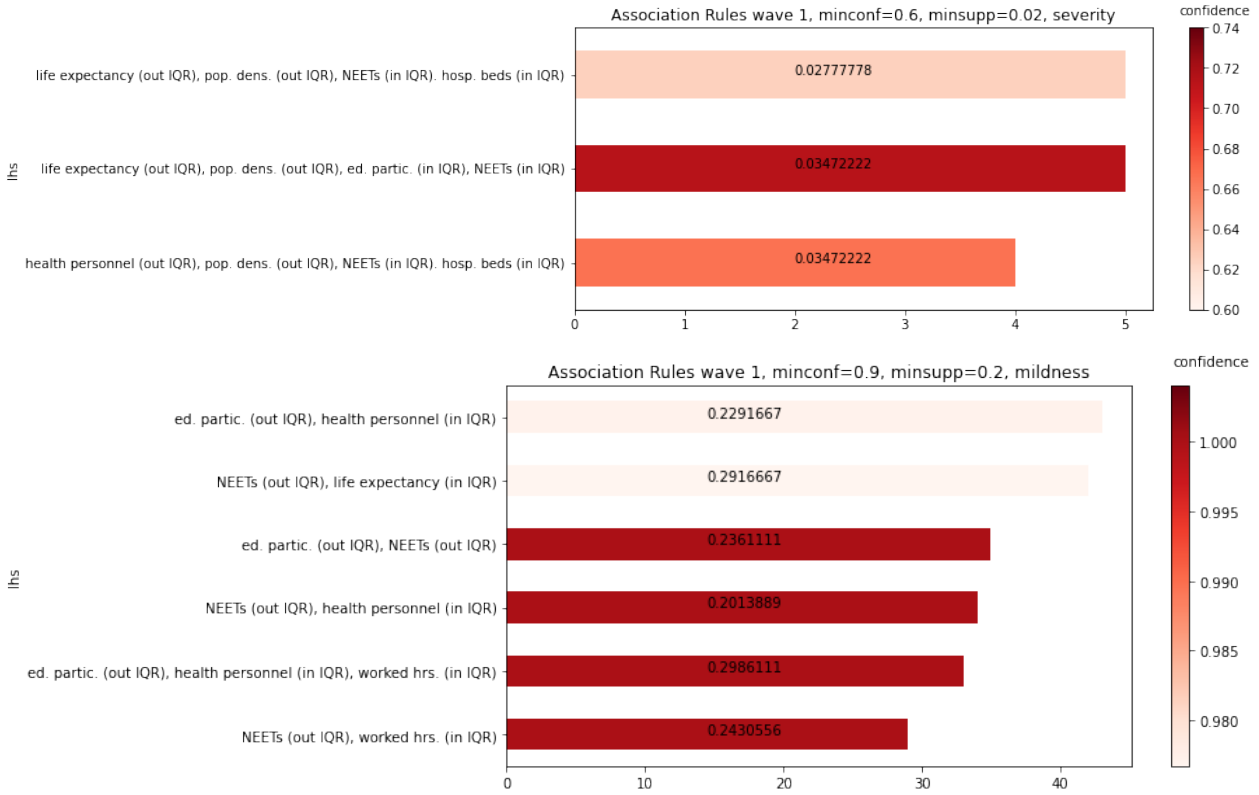


Figure: Description of association rules for the first wave.

AR results for wave 2

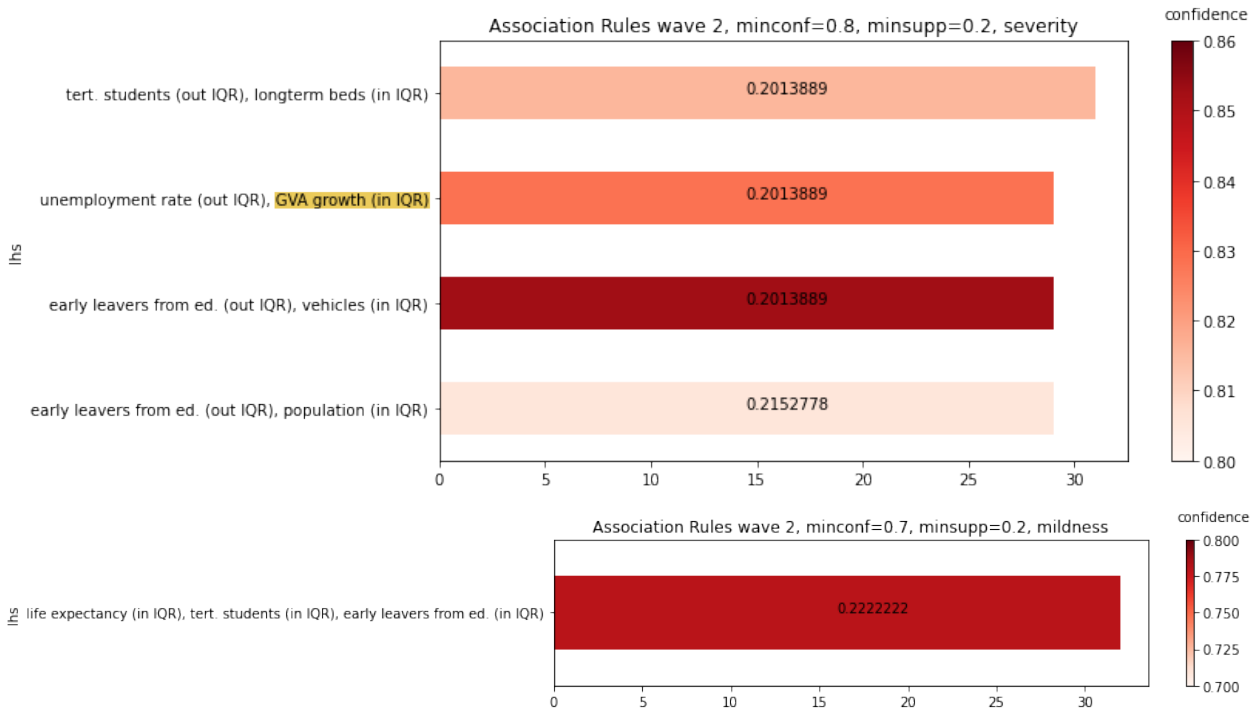


Figure: Description of association rules for the second wave.

GS results for wave 1

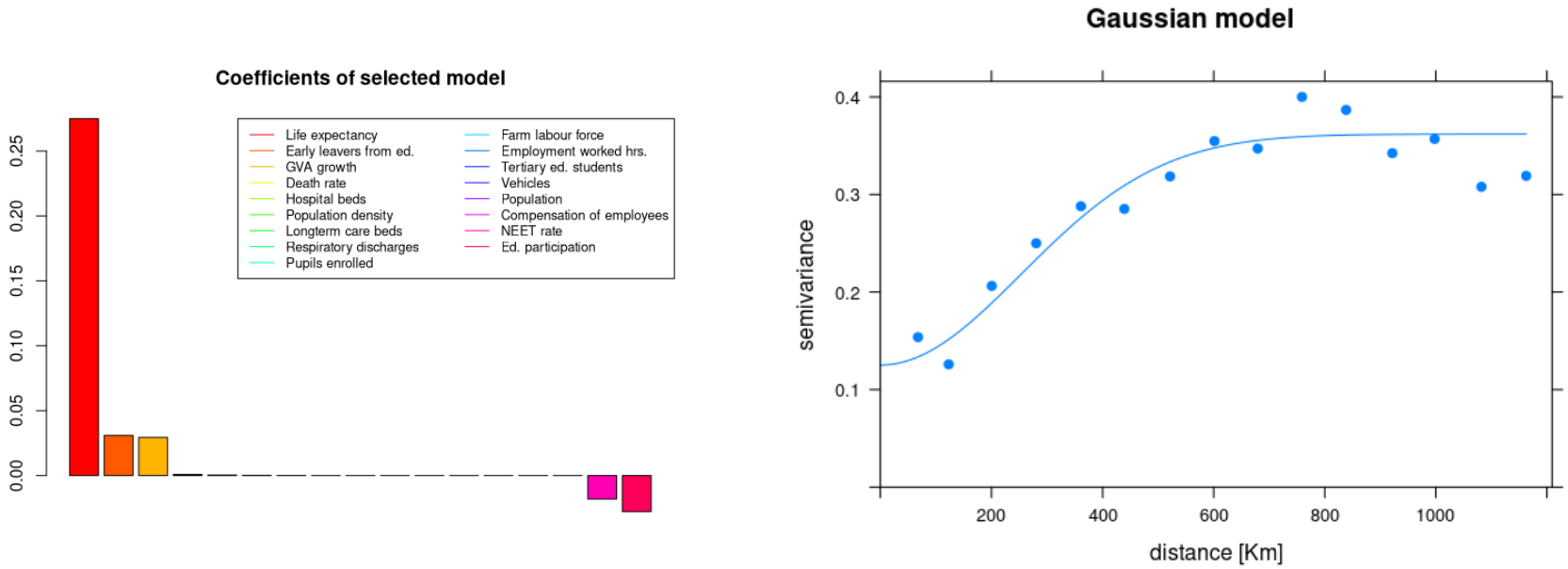
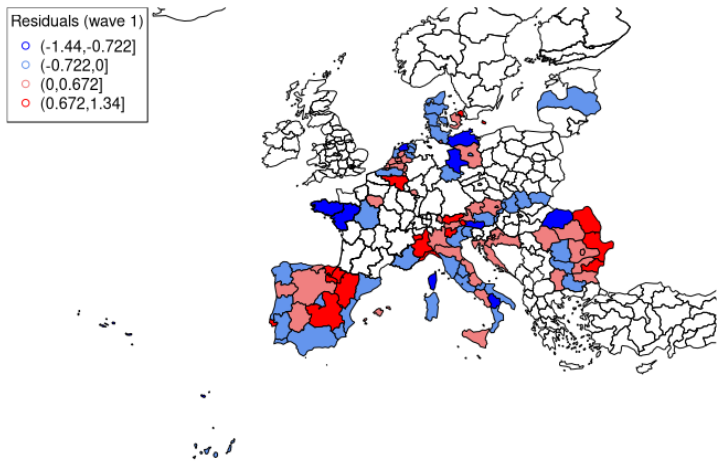


Figure: Coefficients of linear model and variogram plot of gaussian model.

We find an $R^2 = 0.578$, while the Shapiro-Wilk test for the residuals has a p-value of 0.135.

GS results for wave 1



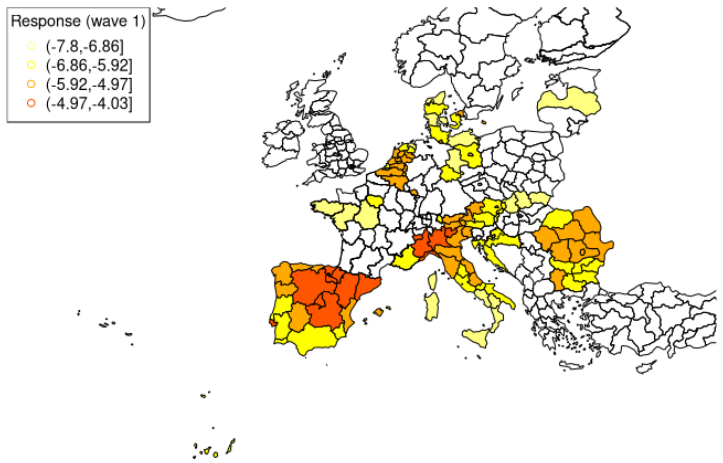


Figure: Residual and response values for each region over the european map.

LISA clusters for wave 1

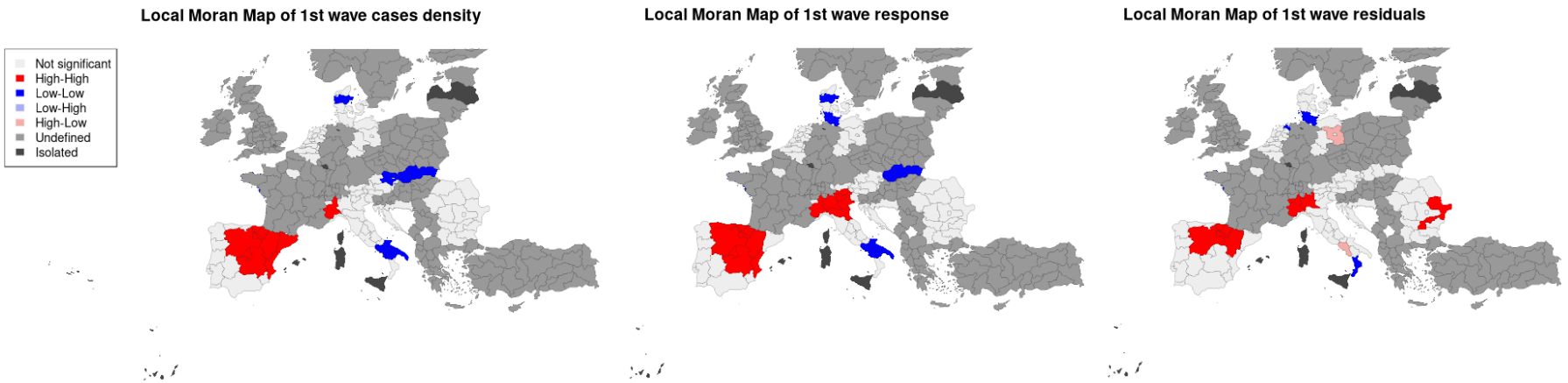


Figure: LISA clusters for cases density, transformed response, and model residuals.

GS results for wave 2

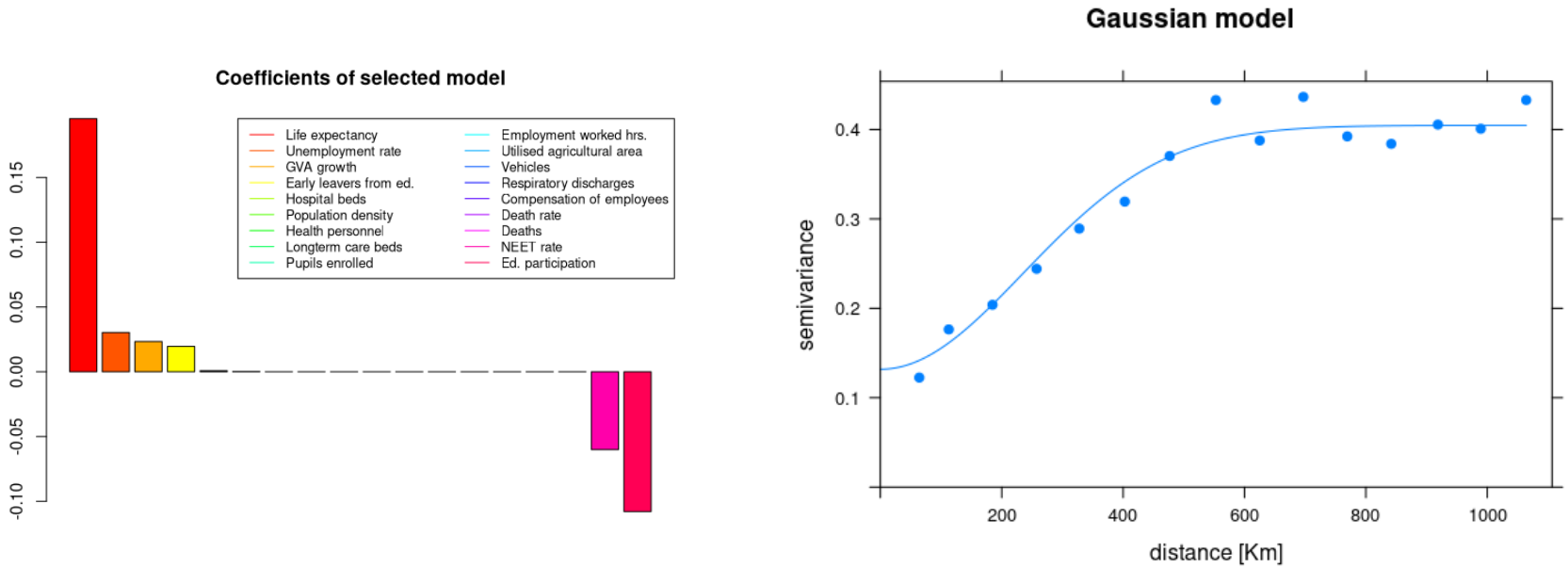
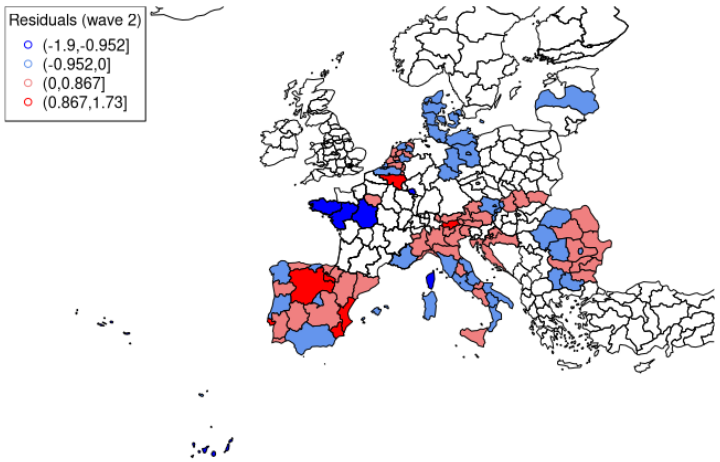


Figure: Coefficients of linear model and variogram plot of gaussian model.

We find an $R^2 = 0.619$, while the Shapiro-Wilk test for the residuals has a p-value of 0.673.

GS results for wave 2



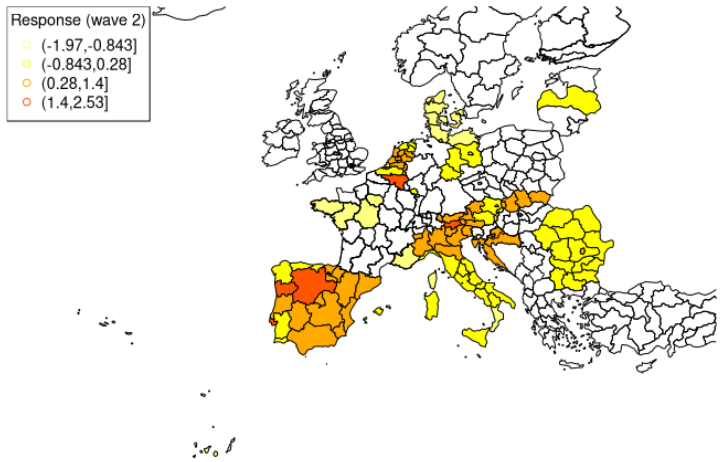


Figure: Residual and response values for each region over the european map.

LISA clusters for wave 2

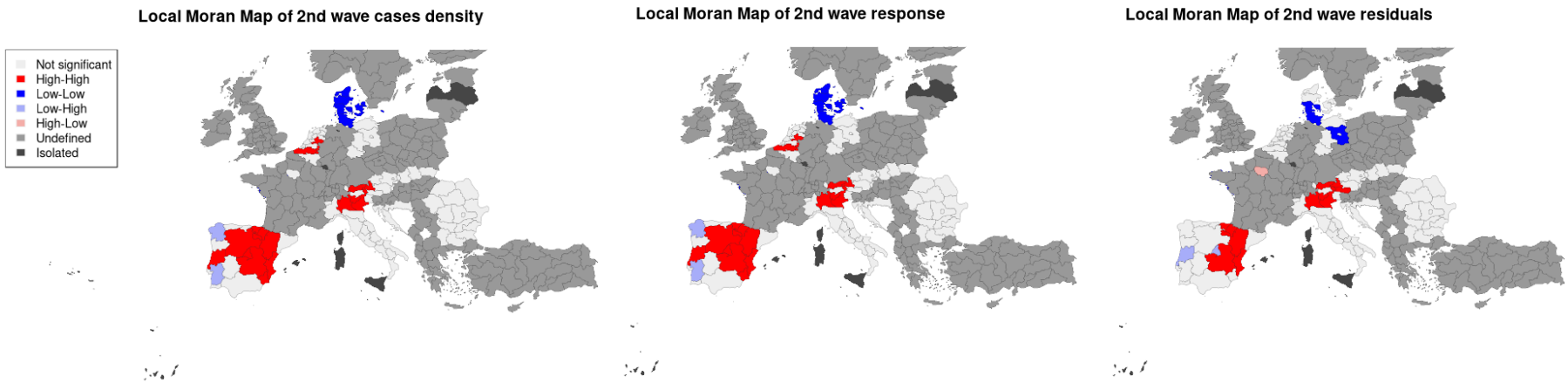


Figure: LISA clusters for cases density, transformed response, and model residuals.

Discussion

AR factors for wave 1

1. Population density [1] [11] [14]
2. Available hospital beds and health personnel [2] [5]
3. Life expectancy [2] [9] [11]
4. Worked hours [9]

AR factors for wave 2

1. Population [4]
2. Long term care beds, available hospital beds, and health personnel [2] [5]
3. Life expectancy [2] [9] [11]
4. Regional GVA growth rate [9]
5. Unemployment rate [9]

Discussion

There are different interactions between types of factors that can motivate further research.

- Wave 1
  - Risk: Population and education (and healthcare)
  - Mildness: Economy, population and healthcare
- Wave 2
  - Risk: Education and healthcare/population/economy
  - Risk: Economy factors alone
  - Mildness: Education and population

GS factors for both waves

For both waves, the following factors were most influential:

- Risk factors:
  - Life expectancy [2] [9] [11]
  - Early leavers from education*
  - GVA growth rate [9]
- Mildness factors:
  - Education participation*
  - NEET rate*

For the second wave, another risk factor was the unemployment rate, which can be associated with social fragility and vulnerability to contagion [7].

*Lower education levels are positively correlated with cumulative case rates [8]

Conclusions

Comparison of approaches

	AR approach	**GS approach**
Strength	Simple Interpretable Factors' interaction Robust Efficient [12]	Key spatial effects [6] Works with continuous data
Weakness	Needs dichotomic data Descriptive [13]	Sensible to approximations [10] Sensible to spatial distribution [10] Expensive and complex

Table: Points of strength and weakness for each approach.

Comparison of results

We verify the most important factors selected by each, which are:

- AR approach: Factors in the significant rules.
- GS approach: Factors selected by LASSO with an absolute weight $\geq e^{-4}$.

	Mildness	**Severity**
Wave 1	Ed. participation NEET rate	Life expectancy Available hospital beds Population density
Wave 2		Early leavers from ed. GVA growth rate Life expectancy Unemployment rate

Table: Important features for both approaches.

Comparison of results

	Mildness	**Severity**
Wave 1	Worked hours (AR)	Health personnel (AR) Early leavers from ed. (GS) GVA growth rate (GS) Death rate (GS)
Wave 2	Students in tert. ed. (AR) Ed. participation (GS) NEET rate (GS)	Long term care beds (AR) Population (AR) Stock of vehicles (AR) Available hospital beds (GS) Population density (GS)

Table: Important features for only one approach.

Conclusions

Both methods should be considered, given that

- The results can be used as complementary while having mutual validation
- We gain interpretable but also rich insights
- We can understand how the covariates behave and also predict the spatial spread of COVID-19

References I

[1] Kawther Aabed and Maha M A Lashin. "An analytical study of the factors that influence COVID-19 spread". In: *Saudi Journal of Biological Sciences* 28.2 (2021), pp. 1177–1195.

[2] Kasim Allel, Thamara Tapia-Muñoz, and Walter Morris. "Country-level factors associated with the early spread of COVID-19 cases at 5, 10 and 15 days since the onset". In: *Global Public Health* 15.11 (2020), pp. 1589–1602.

[3] Luc Anselin. "Geographical Analysis". In: 1995. Chap. 27, pp. 93–115. URL: https://dces.webhosting.cals.wisc.edu/wp-content/uploads/sites/128/2013/08/W4_Anselin1995.pdf.

[4] Marco Ciotti et al. "The COVID-19 pandemic". In: *Critical reviews in clinical laboratory sciences* 57.6 (2020), pp. 365–388.

[5] Aleksandr Farseev et al. "Understanding economic and health factors impacting the spread of COVID-19 disease". In: *medRxiv* (2020).

[6] Munazza Fatima et al. "Geospatial Analysis of COVID-19: A Scoping Review". In: *International journal of environmental research and public health* 18.5 (2021). URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7956835/.

[7] Ibraheem M. Karaye and Jennifer A. Horney. "The Impact of Social Vulnerability on COVID-19 in the U.S.: An Analysis of Spatially Varying Relationships". In: *American Journal of Preventive Medicine* 59.3 (2020). URL: https://www.ajpmonline.org/article/S0749-3797(20)30259-2/fulltext.

[8] Ibraheem M. Karaye and Jennifer A. Horney. "The Impact of Social Vulnerability on COVID-19 in the U.S.: An Analysis of Spatially Varying Relationships". In: *American Journal of Preventive Medicine* 59.3 (2020). URL: https://www.ajpmonline.org/article/S0749-3797(20)30259-2/fulltext.

References II

[9] Rajesh Kumar et al. "Study of social and geographical factors affecting the spread of COVID-19 in Ethiopia". In: *Journal of Statistics and Management Systems* 24.1 (2021), pp. 99–113.

[10] C.R. Paramasivam. "Merits and Demerits of GIS and Geostatistical Techniques". In: *GIS and Geostatistical Techniques for Groundwater Science* (2019), pp. 17–21. URL: https://doi.org/10.1016/B978-0-12-815413-7.00002-X.

[11] Satyaki Roy and Preetam Ghosh. "Factors affecting COVID-19 infected and death rates inform lockdown-related policymaking". In: *PloS one* 15.10 (2020).

[12] Domenico Talia, Paolo Trunfio, and Fabrizio Marozzo. "Chapter 1 - Introduction to Data Mining". In: *Data Analysis in the Cloud* (2016), pp. 1–25. URL: https://www.sciencedirect.com/science/article/pii/B9780128028810000019.

[13] Choh Man Teng. "Data, Data, Everywhere: Statistical Issues in Data Mining". In: *Philosophy of Statistics 7* (2011), pp. 1099–1117. URL: https://www.sciencedirect.com/science/article/pii/B9780444518620500344.

[14] Thirumalaisamy P Velavan and Christian G Meyer. "The COVID-19 epidemic". In: *Tropical medicine & international health : TM & IH* 25.3 (2020), pp. 278–280.

Annex

Data exploration

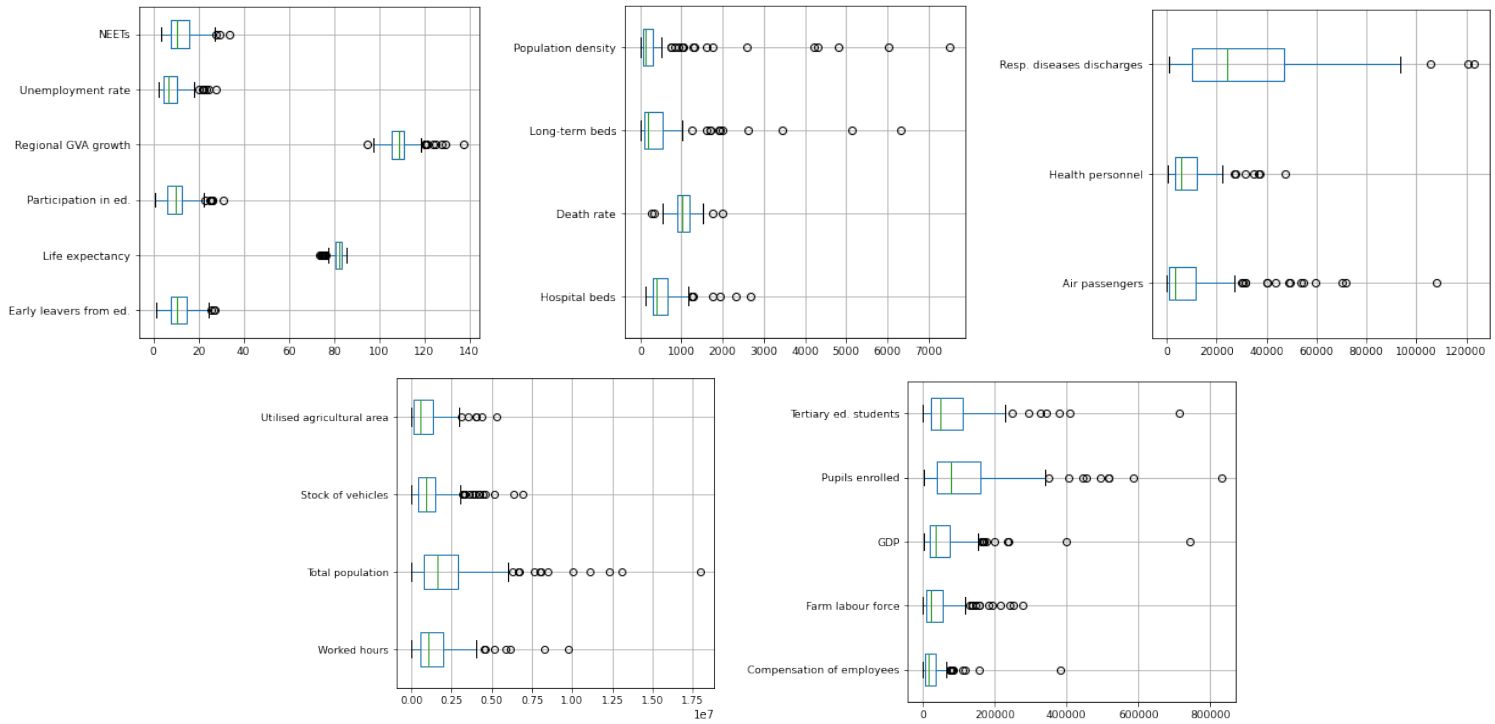


Figure: Box plots of features grouped by magnitude.

AR approach: Dichotomization method

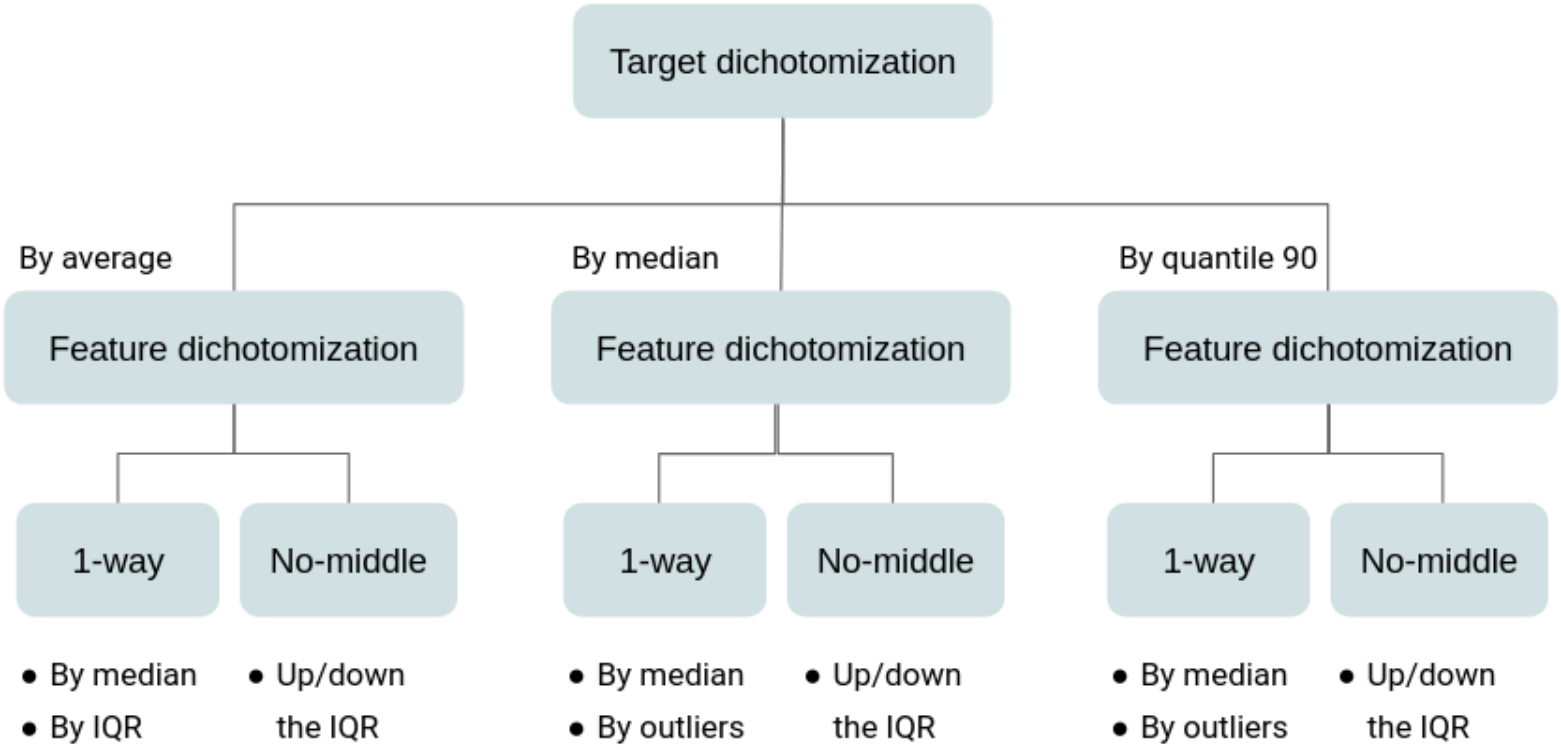


Figure: Cases diagram for the data pre-processing proposals.

AR approach: LASSO selection method

Logistic regression models the probability distribution of $y^{(i)}$ as in eq. 1.

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

The Maximum a Posteriori estimate is obtained by the optimization problem in 2.

$$\min_{\Theta \in \mathbb{R}^n} \sum_{i=1}^{m} -log(p(y^{(i)}|x^{(i)}; \Theta)) + \lambda|\Theta|_{l1} \tag{2}$$

To find the optimal penalization parameter, we tune it providing a grid of 76 equally spaced values between 10^{-3} and 10^2 and evaluate the performance using the average ROC-AUC score obtained through 10-fold cross validation. We select the factors with absolute weights bigger than 0.000001.

AR approach: LASSO selection results

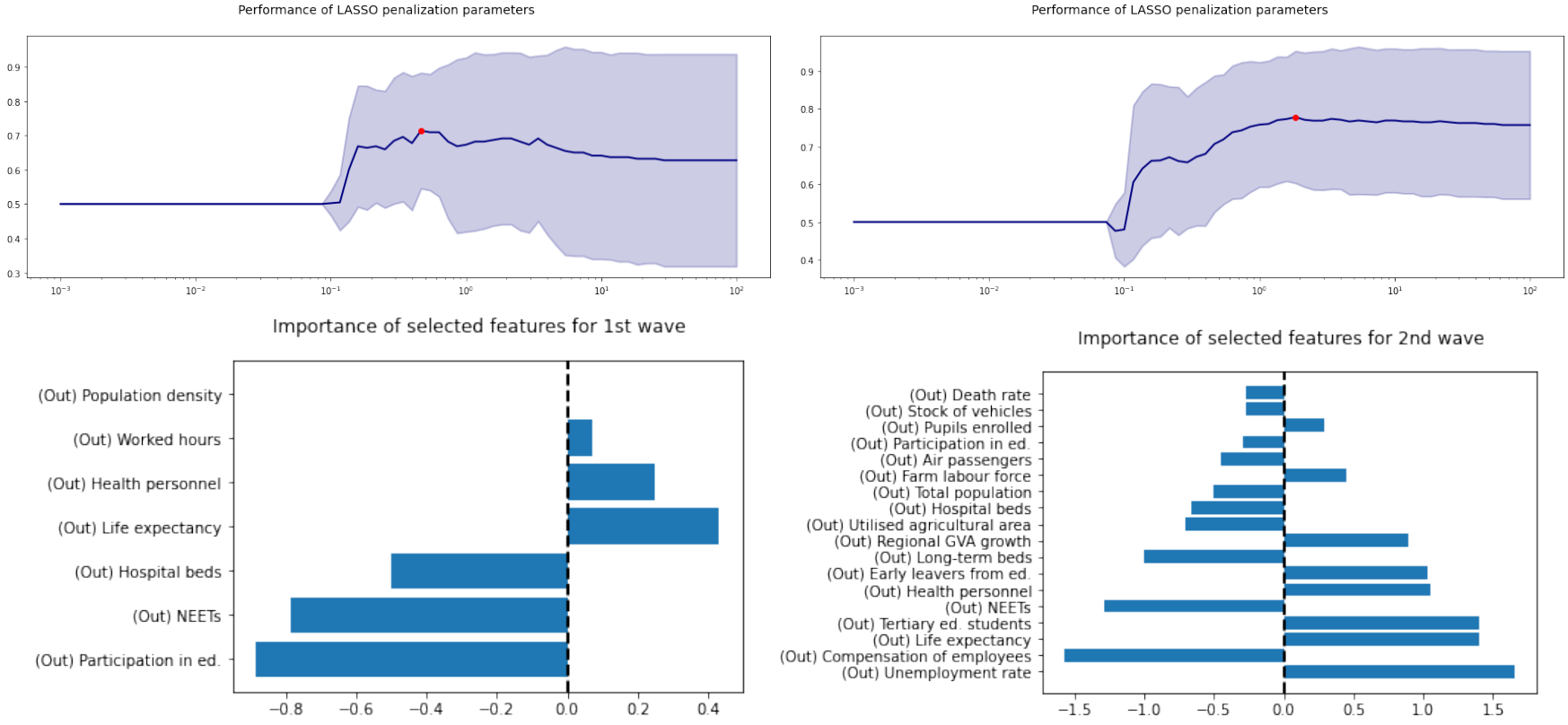


Figure: (Top) LASSO parameter tuning and (bottom) weights of selected features for wave 1 (left) and wave 2 (right).

AR approach: Association rule mining

In order to find significant 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: 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.

AR approach: AR results

	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: AR mining results summary.

AR approach: Output evaluation

	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: Subject analysis summary.

	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: Average satisfied rules.

Geo-statistics approach: LASSO methodology

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.

$$\min_{\beta \in \mathbb{R}^n} \sum_{i=1}^{m} (y^{(i)} - \beta^T x^{(i)})^2 + \lambda |\beta|_{l1} \tag{3}$$

First, in order to find the optimal penalization parameter, we tune it providing a grid of 100 equally spaced values between 10^{-3} and 10^2 and evaluate the performance of the feature selection using the average MSE (mean squared error) obtained through 10-fold cross validation.
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.

Geo-statistics approach: Linear model results

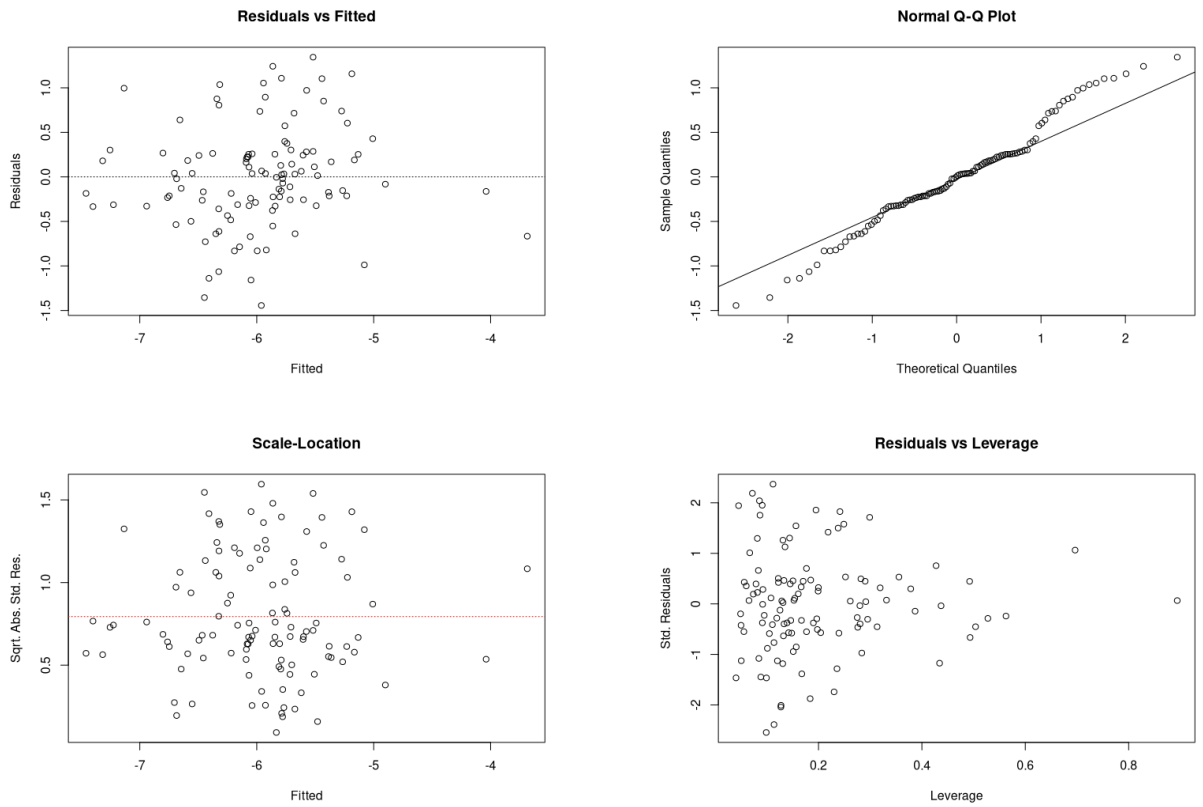


Figure: Residuals visualization and diagnostics for the model of the first wave.

Geo-statistics approach: Linear model results

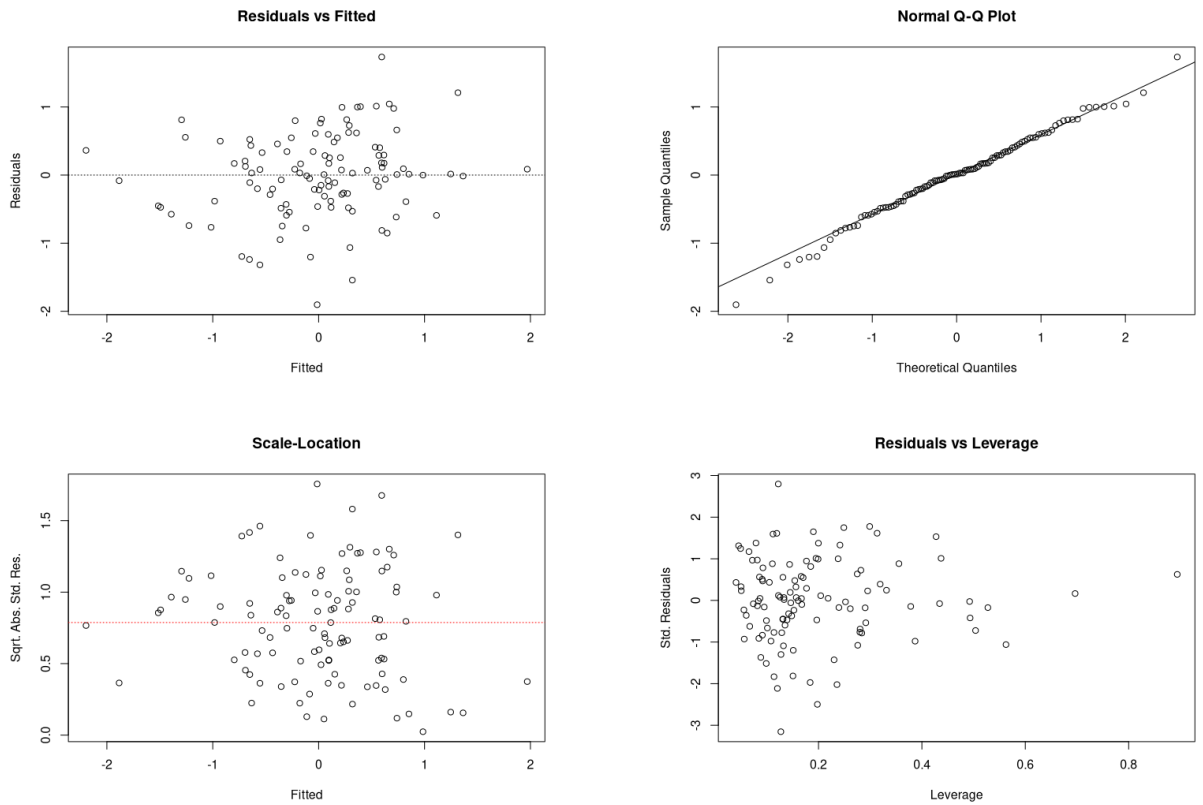


Figure: Residuals visualization and diagnostics for the model of the second wave.

Geo-statistics approach: Variogram models

- 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.

$$\gamma(h) = \begin{cases} \tau^2 & h > 0 \\ 0 & h = 0 \end{cases} \tag{4}$$

- 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.

$$\gamma(h) = \begin{cases} \sigma^2(1 - e^{-h/a}) & h > 0 \\ 0 & h = 0 \end{cases} \tag{5}$$

Geo-statistics approach: Variogram models

- 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.

$$\gamma(h) = \begin{cases} \sigma^2 & h \geq a \\ \sigma^2 \left[\frac{3}{2}\frac{h}{a} - \frac{1}{2}\left(\frac{h}{a}\right)^3 \right] & a > h > 0 \\ 0 & h = 0 \end{cases} \tag{6}$$

- 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.

$$\gamma(h) = \begin{cases} \sigma^2(1 - e^{-(h/a)^2}) & h > 0 \\ 0 & h = 0 \end{cases} \tag{7}$$

Geo-stats. approach: Variogram modelling methodology

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 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.

$$\sum_{k=1}^{K} \frac{1}{w_k} (\hat{\gamma}(h_k) - \gamma(h_k, \theta))^2 \tag{8}$$

After finding the best parameters of each model, we look into three error indicators, which are the residual 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.

Geo-statistics approach: Variogram modelling results

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

Table: Summary fitted variogram models for wave 1.

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

Table: Summary fitted variogram models for wave 2.

Geo-statistics approach: LISA clustering methodology

Moran's I statistic is a cross-product statistic between a variable and its spatial lag, with the variable expressed in deviations from its mean. We formulate the deviations from the mean z_i, the spatial weights w_{ij} such that $n = \sum_i \sum_j w_{ij}$, so Moran's I statistic in this case is stated in 9, where we can see the decomposition of the global statistic into local statistics (or LISAs [3]) and simplify the expression using constant $k = (\sum_{i=1}^{n} z_i^2)^{-1}$.

$$I = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} z_i^T z_j}{\sum_{i=1}^{n} z_i^2} = \sum_{i=1}^{n} I_i = \sum_{i=1}^{n} k \cdot z_i^T \sum_{j=1}^{n} w_{ij} z_j \tag{9}$$

We consider the queen contiguity criteria that, 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 for matrix W.