pagepress

Geospatial Health 2022; volume 17(s1):1073

Spatial correlates of COVID-19 first wave across continental Portugal

Bruno Barbosa,[1] Melissa Silva,[2,3] César Capinha,[2,3] Ricardo A.C. Garcia,[2,3] Jorge Rocha[2,3]

[1]European Centre on Urban Risks (CERU), Lisbon; [2]Institute of Geography and Spatial Planning, University of Lisboa, Lisbon; [3]Associated Laboratory TERRA, Lisbon, Portugal

Abstract

The first case of COVID-19 in continental Portugal was documented on the 2nd of March 2020 and about seven months later more than 75 thousand infections had been reported. Although several factors correlate significantly with the spatial incidence of COVID-19 worldwide, the drivers of spatial incidence of this virus remain poorly known and need further exploration. In this study, we analyse the spatiotemporal patterns of COVID-19 incidence in the at the municipality level and test for significant relationships between these patterns and environmental, socioeconomic, demographic and human mobility factors to identify the mains drivers of COVID-19 incidence across time and space. We used a generalized liner mixed model, which accounts for zero inflated cases and spatial autocorrelation to identify significant relationships between the spatiotemporal incidence and the considered set of driving factors. Some of these relationships were particularly consistent across time, including the 'percentage of employment in services'; 'average time of commuting using individual transportation'; 'percentage of employment in the agricultural sector'; and 'average family size'. Comparing the preventive measures in Portugal (*e.g.*, restrictions on mobility and crowd around) with the model results clearly show that COVID-19 incidence fluctuates as those measures are imposed or relieved. This shows that our model can be a useful tool to help decision-makers in defining prevention and/or mitigation policies.

Correspondence: Jorge Rocha, Edifício IGOT, Rua Branca Edmée Marques, Cidade Universitária, 1600-276 Lisboa, Portugal.
Tel.: +351210443000.
E-mail: jorge.rocha@campus.ul.pt

Key words: COVID-19; drivers of transmission; socio-economic conditions; spatial incidence; Portugal.

Acknowledgements: this research was developed in the context of ABS-Covid - Anthropogenic Base factors of Spreading COVID project, CERU, Council of Europe, EUR-OPA Major Hazards Agreement and the CEG/IGOT Research Unit UIDB/00295/2020 and UIDP/00295/2020. Bruno Barbosa was granted by CERU/CE.

Received for publication: 16 January 2022.
Revision received: 26 April 2022.
Accepted for publication: 26 April 2022.

Geospatial Health 2022; 17(s1):1073
doi:10.4081/gh.2022.1073

Introduction

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is an extremely infectious coronavirus first identified in humans in December 2019 in the Chinese city of Wuhan and reported to the (WHO - World Health Organization, 2020). The disease caused by SARS-CoV-2 is known as coronavirus disease 2019 (COVID-19) has since become a worldwide public health problem (Gorbalenya *et al*., 2020; Ma *et al*., 2020). The virus quickly spread all over the world and as reported by the Ministry of Health in Portugal, *i.e*. Direção-Geral da Saúde (DGS) first detected in this country on the 2nd of March 2020 (DGS, 2020a). Following this, several measures were put in place aiming to contain its spread, such as strong mobility restrictions including home confinement and closure of many public services. These and other measures were successively alleviated or strengthened in response to perceived variation in the number of infection cases. However, despite these efforts the virus continued to spread and by the 30th of September, more than 75,000 infections had been reported, affecting all regions of the territory (DGS, 2020b).

Previous work has explored the association between the spatial patterns of occurrence or incidence of COVID-19 and characteristics of the underlying regions (Mohammad Ebrahimi *et al*., 2021; Mollalo *et al*., 2021). As a result, a number of environmental and anthropogenic factors were identified as potential drivers. Namely clinical-epidemiological conditions (Bai *et al*., 2020; Guan *et al*., 2020), climatic conditions (Chan *et al*., 2011; Paez *et al*., 2020; Quilodrán *et al*., 2020), socioeconomic and demographic structure (Laires and Nunes, 2020; Maroko *et al*., 2020; Marques, 2020; Murgante *et al*., 2020; Roy *et al*., 2020; Prazeres *et al*., 2021) and human mobility patterns (Chen *et al*., 2020; Melo *et al*., 2020; Orea and Álvarez, 2020; Tamagusko and Ferreira, 2020). These associations provide information on possible mechanisms by which the spread or impact of COVID-19 is hampered or magnified, in turn assisting the design of preventive or mitigation measures.

Although a few studies have analysed the spatial and temporal patterns of the disease across the country (Azevedo *et al*., 2020; Bai *et al*., 2020), a comprehensive analysis of how these patterns relate to the physical and human attributes of the territory at a higher temporal resolution (15 days) remains so far missing. To fill this gap, we here analyse how the number of new COVID-19 cases distributes across Portuguese municipalities over time and test if the observed patterns relate to any specific features of these areas.

Materials and methods

Study area

The study area corresponds to continental Portugal. The territories of the autonomous regions of Azores and Madeira were not possible to include due to the unavailability of data for several of the explanatory variables considered (see below). Notwithstanding the exclusion of these regions, the vast majority of recorded infections (140,470, 99.4%) occur in continental Portugal (DGS, 2020b) supporting the relevance of the spatial delimitation used. Our units of analysis correspond to the 278 municipal administrative units (Figure 1), which also represents the European Union (EU) hierarchical system for collecting, developing and harmonising European regional statistics named nomenclature of territorial units for statistics (NUTS). This is the highest spatial resolution for which data on COVID-19 infections can be obtained with reliability at a daily temporal resolution.

NUTS 2021 classification divides the territory into three levels according to socio-economic standing: NUTS 1 represents the major socio-economic regions, NUTS 2 the basic regions for the application of regional policies and NUTS 3 the smallest regions. Thus, the 278 municipalities in continental Portugal are now grouped into one NUTS-1, 5 NUTS-2 and 23 NUTS-3, the latter of which is important for the operation of our model.

Data on COVID-19 incidence and independent variables

Data on the daily number of new COVID-19 cases, per municipality, were collected from public reports of the Portuguese (DGS, 2020b). Data collection comprised the period from 1 April until 30 September 2020. We used the accumulated number of new daily cases over 15 days per municipality, which corresponds to the approximate time lag between exposure and first symptoms, *i.e.* incubation period of the disease (DGS, 2020c).

We did not include the month of March 2020 in the analyses because the disease was in an initial stage of dispersal, assuming that the location of the first cases were completely random and mainly imported. As one of the aims of this work is to analyse determinants of incidence of COVID-19 and not only its spatial spread, but this month was also not considered. In total, we obtained 12 maps representing the accumulated number of new COVID-19 cases by 10,000 inhabitants in each municipality for each 15 days period, going or the same period.

The selection of factors to consider in the selection of descriptor variables was based in prior scientific peer-reviewed articles, whose objectives were to explain spatial variability in COVID-19 incidence (Arashi *et al*., 2020; Coccia, 2020; Lakhani, 2020; Marques, 2020; Ricoca Peixoto *et al*., 2020; Sajadi *et al*., 2020; Wang *et al*., 2020). Thus, having as primary source the Statistics Portugal (https://www.ine.pt/), we used 33 variables. These variables describe each municipality in terms of long-term patterns in human population, demography, socioeconomic conditions, human housing characteristics, and human mobility and health services (Figure 2).

In addition, we also represented the effects of environmental conditions by including the variables 'average temperature' and 'total accumulated precipitation' and air pollution for each month. Unlike the other variables (which correspond to long-term descriptors), these climatic variables represent the conditions occurring in the month matching each analysed 15-day period. These climatic data were obtained from the E-OBS database (Cornes *et al*., 2018).

Data processing and modelling

A generalized linear mixed-effects model (GLMM) (McCulloch and Neuhaus, 2014) was used to test for statistically significant relationships between the independent variables and the incidence of COVID-19 cases in each 15-day period (Figure 3).

GLMM models are highly capable tools for investigating possible changes in the behaviour of COVID-19 cases distribution through changes in independent variables, as well as in the execution of planning actions and measures (Jamil *et al*., 2013).

Data normalization

To improve the robustness of the statistical significance of the GLMM coefficients, we used logarithmic transforms of dependent and independent variables. Indeed, when a dataset does not behave as a normal (Gaussian) distribution, it would be imprudent to scale it using other traditional methods such as standard deviation (SD) normalization. As our data has a skewed (or kurtotic) behaviour the logarithmic function fitted well the purpose of implementing non-linear transformations in order to turn the distribution as close as possible to a Gaussian one. Thus, non-Gaussian distributions frequently have outliers that exceed the threshold of 3.29 SDs. By applying logarithmic transformations, these observations are moved closer to the mean, generating a distribution closer to normal (Mei-Ling and Lee, 2004).

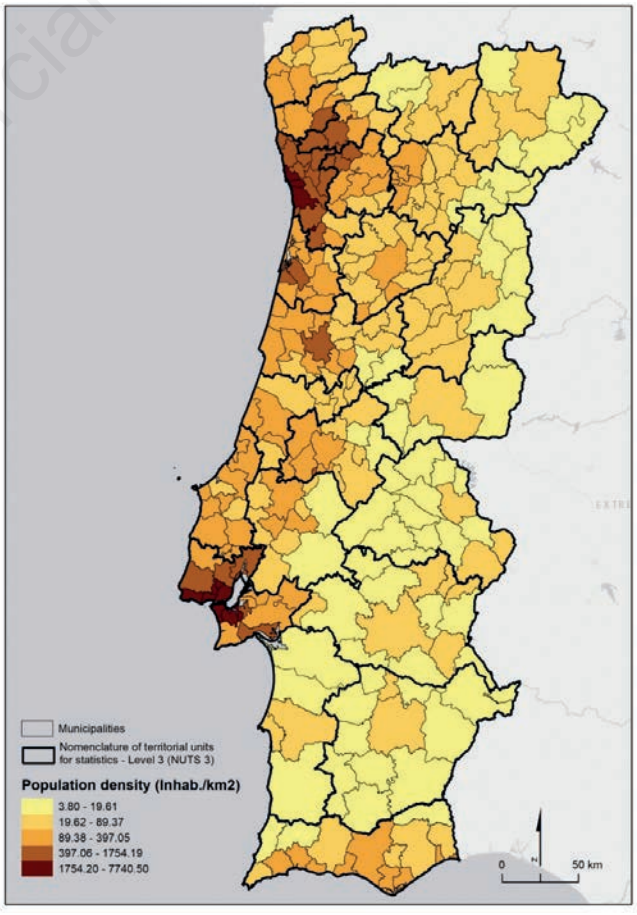


Figure 1. Delimitation of the study area and its municipal administrative units including population density.

The logarithm base can be 2, 10 or the natural logarithmic constant (e=2.7). There is no mathematical reason to prefer any one particular base, so the choice should be a matter of suitable explanation. A doubling, *i.e.* the reduction to 50%, is frequently seen as a biologically significant variation (Bate and Clark, 2014). The log2 scale converts to a scale between –1 and +1. This simple scale gives more flexibility to the modeller than using log10 (doubling using the log10 returns a modification of 0.3). In addition, modellers commonly use log2 for the binary representation of information. Nevertheless, one chose the natural because the subsequent coefficients are straightforwardly understandable as approximate proportional differences (Gelman and Hill, 2006), *i.e.* having a 0.05 coefficient means that a change of 1 in an independent variable implies around 5% change in the dependent variable. To allow the logarithmic transformation of variables where a value of zero is present, a constant value of one is added to all instances. The mean and SD of the converted variables must not be analysed in the traditional way. One should only interpret the P-value generated from the model.

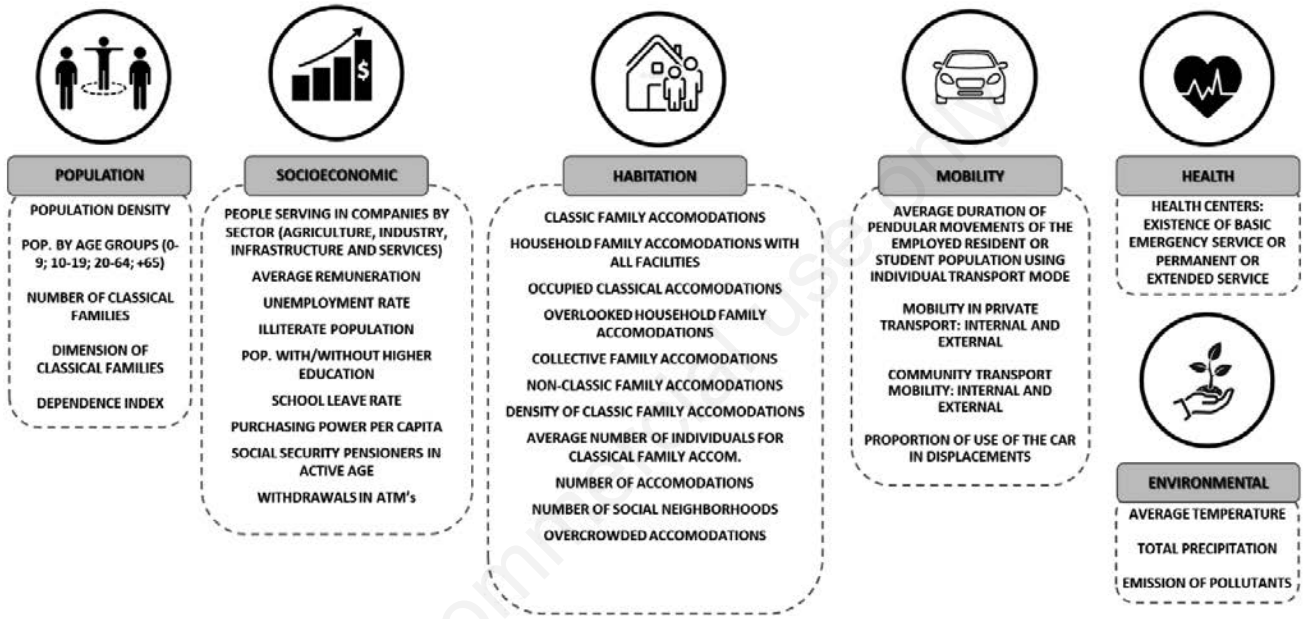


Figure 2. Selected explanatory variables based on literature review (Arashi *et al.*, 2020; Coccia, 2020; Lakhani, 2020; Marques, 2020; Ricoca Peixoto *et al.*, 2020; Sajadi *et al.*, 2020; Wang *et al.*, 2020).

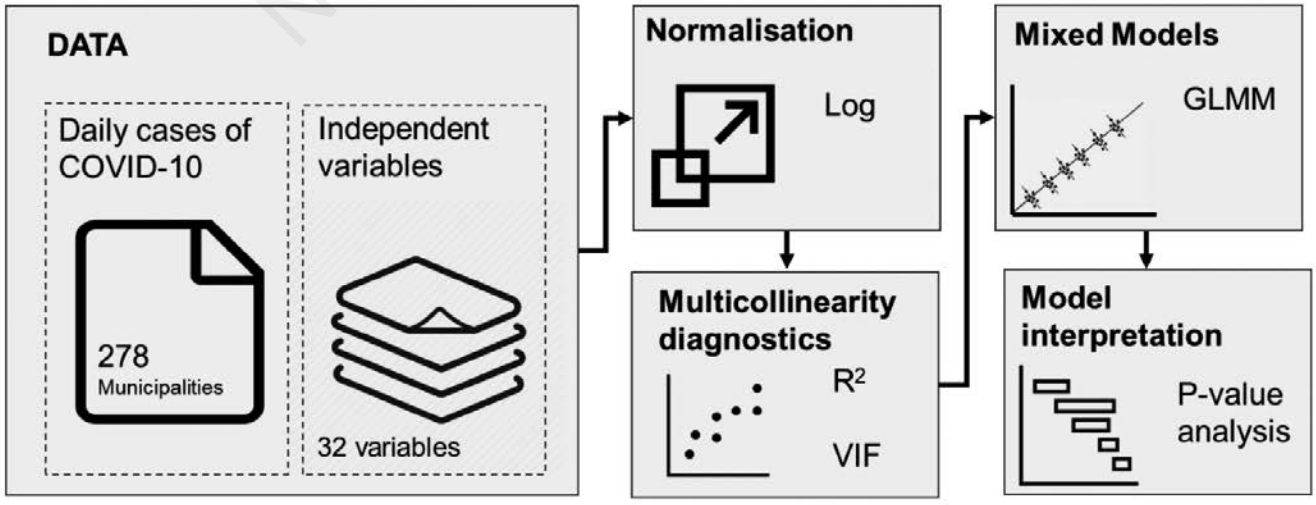


Figure 3. Schematics of the data analysis procedures.

Multicollinearity diagnosis

Statistically speaking, multicollinearity in a multiple regression model can make it difficult to evaluate the relationship between the independent variables and the dependent one. Therefore, we avoided multicollinearity in the information provided by the variables in each model, which can give rise to unreliable model estimates since small changes in the data can lead to large and erratic alterations in the predicted coefficients of the independent variables. One common measure of collinearity is R^2 (Velleman and Welsch, 1981). Having this in mind, we measured Pearson's correlation coefficient (r) between all pairs of variables and removed those showing an $|R|>0.7$, as this is an suitable threshold for when collinearity starts to harshly mislead model assessment and following prediction (Dormann *et al*., 2013). For cases where two variables were correlated above this threshold, we removed the one with the largest mean absolute correlation among all pair-wise measurements. This procedure was performed interactively, using the function 'find Correlation' of R package 'MuMIn' (Barton, 2020). However, R^2 was designed to compute models accuracy rather to check collinearity (Velleman and Welsch, 1981). One statistic often proposed to measure it, is the variance inflation factor (*VIF*) (Marquaridt, 1970) seen as:

$$VIF = (1 - R^2)^{-1} \quad (1)$$

VIF displays the variable variance intensification in result of the collinearity as matched with an perfect case of uncorrelated variables, *i.e*. how often collinearity inflates the variance of the regression coefficients (Ferré, 2009). The varies from one (uncorrelated coefficients) to infinity (perfect correlation). Thus, a *VIF* >1 shows that collinearity affects the variable and, though there is no exact threshold, it has been argued that *VIF* >10 is problematic (Vittinghoff *et al*., 2012), *VIF* >5 is problematic (Gareth *et al*., 2021) or *VIF* >2.5 indicates considerable collinearity (Johnston *et al*., 2018). We follow this last work and remove all variables with a *VIF* >2.5 leaving 16 dependent variables for the analysis.

Models fitting

To account for different population sizes in each municipality, the numbers of COVID-19 cases were converted into an incidence-weighted proportion, representing the number of cases per 10,000 residents. In the models we included NUT-3 regions as a random-effect term, to account for the non-independence in the observations of municipalities that are geographically close and that share similar rulings in the management of public services (Law n.75/2013 of 12 of September). We fitted the models using the template model builder glmmTMB package for programming language R (R Core Team, 2019), which can account for zero inflation (Brooks *et al*., 2017), a characteristic found in our response variable. All models that overfit zeros (ratio>1) at P=0.01 are zero inflated models (Table 1).

We used separate multivariate GLMMs to test the significance and type of association of the descriptor variables with the number of infections per 10,000 residents for each 15 days.

Modelling of infection cases is frequently complex since these datasets are typically right skewed, non-negative and have excess zeros for municipalities without registered occurrences (Saha *et al*., 2020). These characteristics inhibits the usage of linear models like the Gamma (Dagpunar, 2019) or Gaussian (Dasgupta and Wahed, 2014) distributions. A conventional way to address this problem is to use a Two-part or a Tobit model (Kurz, 2017). However, these models make the results interpretation more problematic. As such, one can test Gaussian (as reference), Poisson, Conway-Maxwell-Poisson (COM-Poisson) and Tweedie distributions.

Exponential, Gamma and Poisson distributions model distinct facets of the Poisson process. Exponential distribution models the time lag until the first occurrence. Gamma is applied to predict the n^{th} occurrence waiting period and Poisson to model the number of future occurrences, if COVID-19 cases frequency distribution has equal mean and variance, *i.e*. equidispersion. However, every so often one can observe overdispersion, *i.e*. the variance surpasses the mean suggesting that COVID-19 occurrences are clustered around certain dates. The COM–Poisson distribution was tested as an alternative to Poisson distribution that represents any dispersion characteristic (Mitchell and Camp, 2021). It can enhance the accuracy with which fortnight counts of COVID-19 cases are modelled. Finally, one can explore the potential of Tweedie distribution (Saha *et al*., 2020). This is a special case of exponential distribution models, that can model both the probability of zero outcome, *i.e*. a non-cases municipality and continuous infections occurrences (Kurz, 2017).

We tested four functions (Gaussian, Poisson, COM-Poisson and Tweedie) for each model and evaluated their goodness-of-fit

Table 1. Check for zero-inflation in models.

Model	Observed zeros (Oz)	Predicted zeros (Pz)	Ratio (Pz/Oz)
15 of April	94	99	1.05
30 of April	87	109	1.25
15 of May	125	154	1.23
31 of May	137	174	1.27
15 of June	129	177	1.37
30 of June	154	163	1.06
15 of July	132	156	1.18
31 of July	149	179	1.20
15 of August	145	177	1.22
31 of August	106	131	1.24
15 of September	84	106	1.26
30 of September	56	73	1.30

using the Akaike information criterion (AIC) (Akaike, 1974) and the corrected AIK (AICc) (Bozdogan, 1987; Hurvich and Tsai, 1989). Tweedie performed generally better except for 15 of June 30 of September. However, even in these models, it was the second best performing function (Table 2). Therefore, for the data analysis one assumes a Tweedie distribution of errors for all models.

Model validation

The essential step in any model is to evaluate its accuracy. The mean squared error (MSE) and the mean absolute error (MAE) are used to evaluate models performance in regression analysis (Kaliappan *et al*., 2021). MAE is calculated from the average of the absolute errors, that is, we used the module of each error to avoid underestimation, because outliers affect the result less. MAE measures the average of the residuals in the dataset:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|\overline{y_i} - \hat{y_i}| \quad (2)$$

where $\overline{y_i}$ is the mean value of y and $\hat{y_i}$ is its predicted value.

MSE, on the other hand, is commonly used to verify the accuracy of models and gives highest weight to the largest errors, since each error is squared individually when calculated, and the mean of these squared errors is calculated afterwards. It measures the variance of the residuals. Using the same error concept used in (5), we have the equation below:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(\overline{y_i} - \hat{y_i})^2 \quad (3)$$

Due to the squared exponent that the error assumes, MSE is quite sensitive to outliers and, if the data had many significant errors, this metric can be extrapolated. MSE penalizes the highest prediction errors *vis-a-vis* MAE. Thus, MSE is a differentiable function that turns easier to perform mathematical operations than in non-differentiable function like MAE.

Lower values of MAE and MSE indicate higher accuracy of the model. As we can see from Table 3 the values of all models in both measures are generally low, indicating a good model performance. Higher values of MSE between 15 of May and 15 of August indicate that more outliers occurred at these dates.

Table 2. Models goodness-of-fit according to AIC and AICC.

	Gaussian	Poisson	COM-Poisson	Tweedie
15 of April	722.3549 (726.3235)	741.2999 (744.9092)	720.3607 (724.3293)	769.4969 (773.8433)
30 of April	677.9044 (681.8730)	690.1893 (693.7986)	658.86 (662.8287)	754.7326 (759.0790)
15 of May	569.4477 (573.4163)	553.5544 (557.1638)	519.2992 (523.2678)	607.9225 (612.2690)
31 of May	442.6945 (446.6631)	471.1886 (474.7980)	NA (NA)	525.4794 (529.8258)
15 of June	386.8227 (390.4321)	466.3996 (469.6681)	720.3607 (724.3293)	493.3228 (497.2914)
30 of June	640.2762 (644.2448)	576.9462 (580.5555)	569.9709 (573.9395)	672.5513 (676.8977)
15 of July	640.2762 (644.2448)	576.9462 (580.5555)	569.9709 (573.9395)	672.5513 (676.8977)
31 of July	NA (NA)	487.0757 (490.6850)	463.2975 (467.2661)	571.3985 (575.7450)
15 of August	570.1512 (573.7606)	500.2459 (503.5144)	488.9160 (492.5254)	588.6866 (592.6552)
31 of August	663.8363 (667.4457)	637.6903 (640.9588)	613.5246 (617.1340)	718.1531 (722.1218)
15 of September	707.3059 (711.2745)	705.9059 (709.5153)	683.2107 (687.1794)	760.4494 (764.7959)
30 of September	737.6834 (741.6520)	788.9751 (792.5845)	738.9016 (742.8702)	741.0631 (745.4096)

Table 3. The mean squared error and the mean absolute error.

		MAE		MSE	
Month	Day	Value	Scaled	Value	Scaled
Apr	15	1.040	1.000	1.510	1.000
	30	0.939	0.902	1.320	0.872
May	15	1.500	1.443	2.790	1.846
	30	1.630	1.561	3.120	2.065
Jun	15	1.550	1.490	2.790	1.843
	30	1.410	1.357	2.540	1.678
Jul	15	1.330	1.273	2.260	1.491
	30	1.540	1.478	2.840	1.879
Aug	15	1.580	1.520	3.080	2.036
	30	0.831	0.798	1.110	0.732
Sep	15	0.915	0.879	1.270	0.839
	30	1.110	1.064	1.640	1.081

To give more support to our validation we further accessed the regression error characteristic (REC) and the regression receiver operating characteristic (RROC). REC curves assisted to interpret the performance of the models, while RROC also shows model asymmetry.

REC is the regression counterpart of the receiver operating characteristic (ROC) in classification models. It plots the error tolerance against the percentage of cases predicted within the tolerance, resulting in a curve that estimates the error cumulative distribution function. The REC area over the curve (AOC) works as a biased estimate of the expected error (Figure 4).

The RROC is a plot of total over-estimation *versus* total under-estimation. In this analyse one used a shift similar to the ROC threshold. For every occurrence, we computed a new prediction $\hat{y}' = \hat{y} + s$ where s represents the shift (Hernández-Orallo, 2013). Hence, one gets differe nt errors for each shift $e_i = \hat{y}_i' - y_i$. The under-estimation is given by $\Sigma(e_i | e_i < 0)$ and

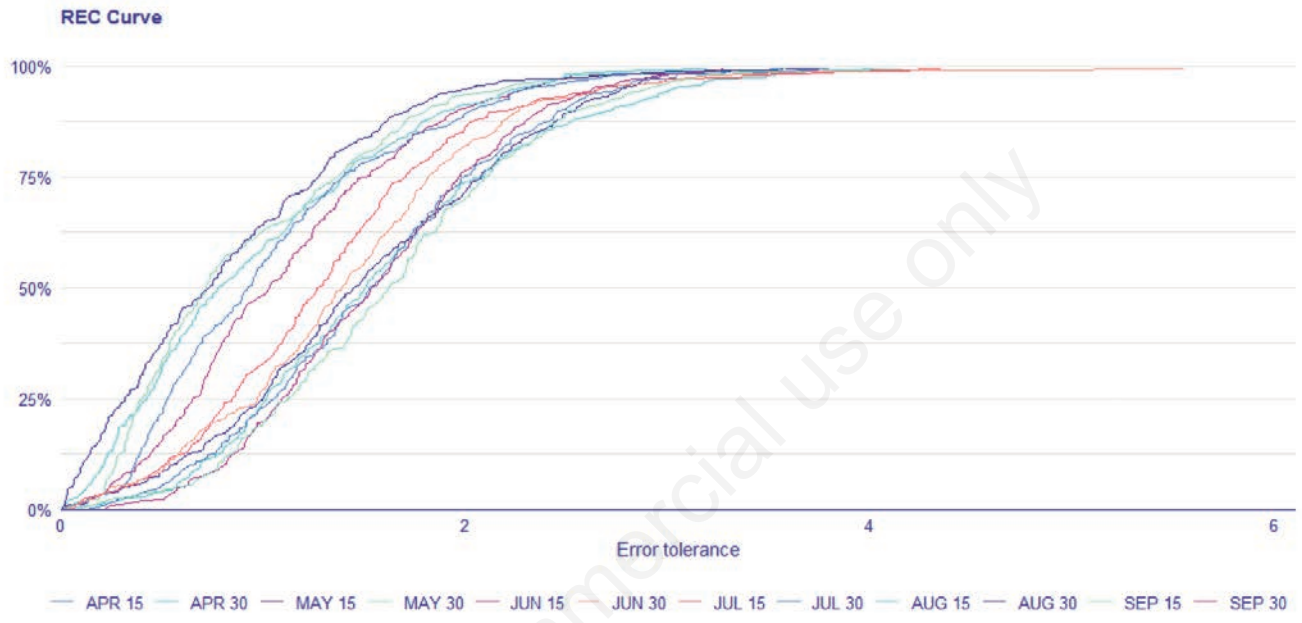


Figure 4. Regression error characteristic (REC) of all models.

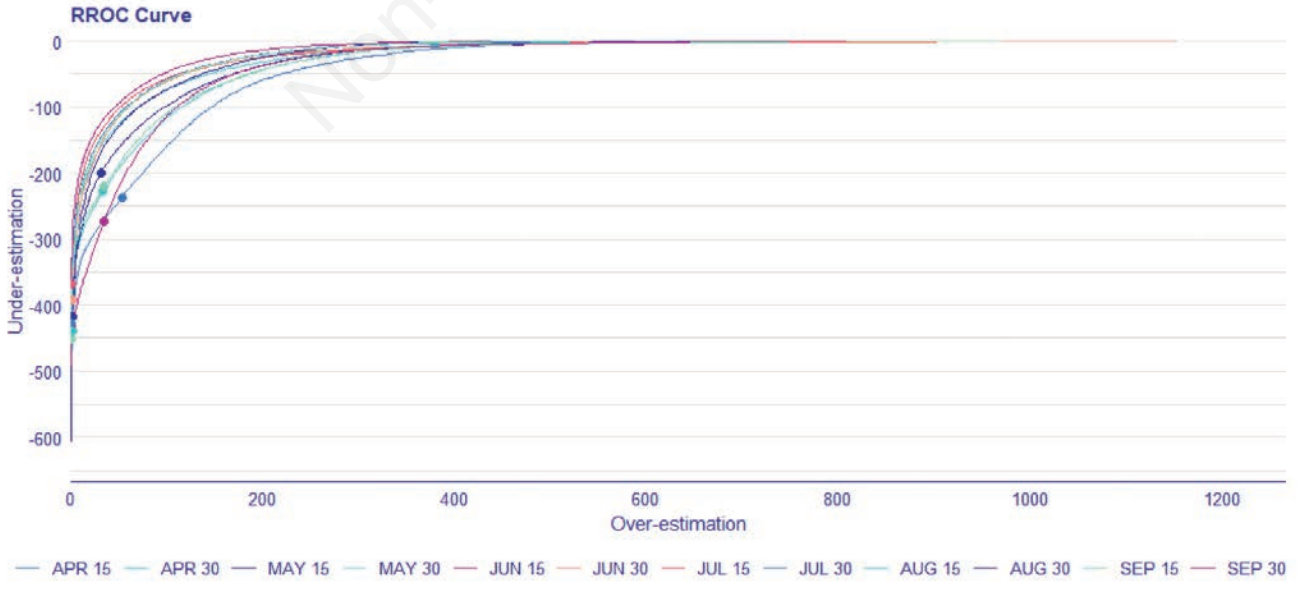


Figure 5. Regression receiver operating characteristic (RROC) for all models.

over-estimation by $\Sigma(e_i|e_i > 0)$. The AOC and the equal error rate (EER) are two performance measures usually applied in RROC analysis. The EER is the point where the false positive rate and the false negative rate are equal (*i.e.* the shift equals 0), and a good model have to keep this value as small as possible. Figure 5 gives a good idea of the model performance by keeping all values of both AOC and EER low.

Results

The number of new cases of COVID-19 showed a consistent trend for decrease and then stabilization from the beginning of April to the beginning of August. From mid-August onwards, the number of new cases increased continuously until the end of September (Figure 6). It is also possible to observe that in the first half of April (approximately one and a half-month since its first detection), the disease had already dispersed widely, with cases being reported from all major regions of the territory.

The results show that the huge dispersion in inner Portugal are more relevant from July onwards. In fact, these cases should in general be related to the end of lockdown and when circulation again increases. However, holiday influence should not be overlooked, since many might opt for rural tourism in less populated areas and the proximity to the border, where tourism and other activities may lead to the increasing contacts between populations with differentiated restrictions and mitigation measures.

The results from the regression analysis show that the spatial patterns of COVID-19 incidence per 10,000 residents are significantly associated with some of the descriptive variables considered (Figure 7). Of these relationships, a few are relatively frequent through time. A significant positive relationship with 'percentage of population employed in services' is the most frequent (found for eight of the twelve 15-day periods analysed).

Three variables show significant relationships in the five 15-day periods: 'average time of commuting using individual transportation', 'percentage of residents employed in the agricultural sector' and 'average family size'. The first variable shows a positive relationship with COVID-19 incidence and the second a negative relationship, while the latter shows a dominantly positive relationship (four out of five 15-day periods).

A positive significant relationship with average air pollution also emerges for four 15-day periods. A significant association with total precipitation is also found for four 15- day periods but the form of the relationships varied between a positive association, *e.g.*, in one 15-day period, and negative for the remainder three. A few other variables also show significant relationships, but their temporal consistency is even more reduced (*i.e.* in three of the periods or less) and these relationships are not discussed further.

One can also find higher COVID-19 cases in municipalities with a strong percentage of workers in the tertiary sector. The types of activities this sector involves are usually more common in urban areas, a fact which is in accordance with the obtained results for

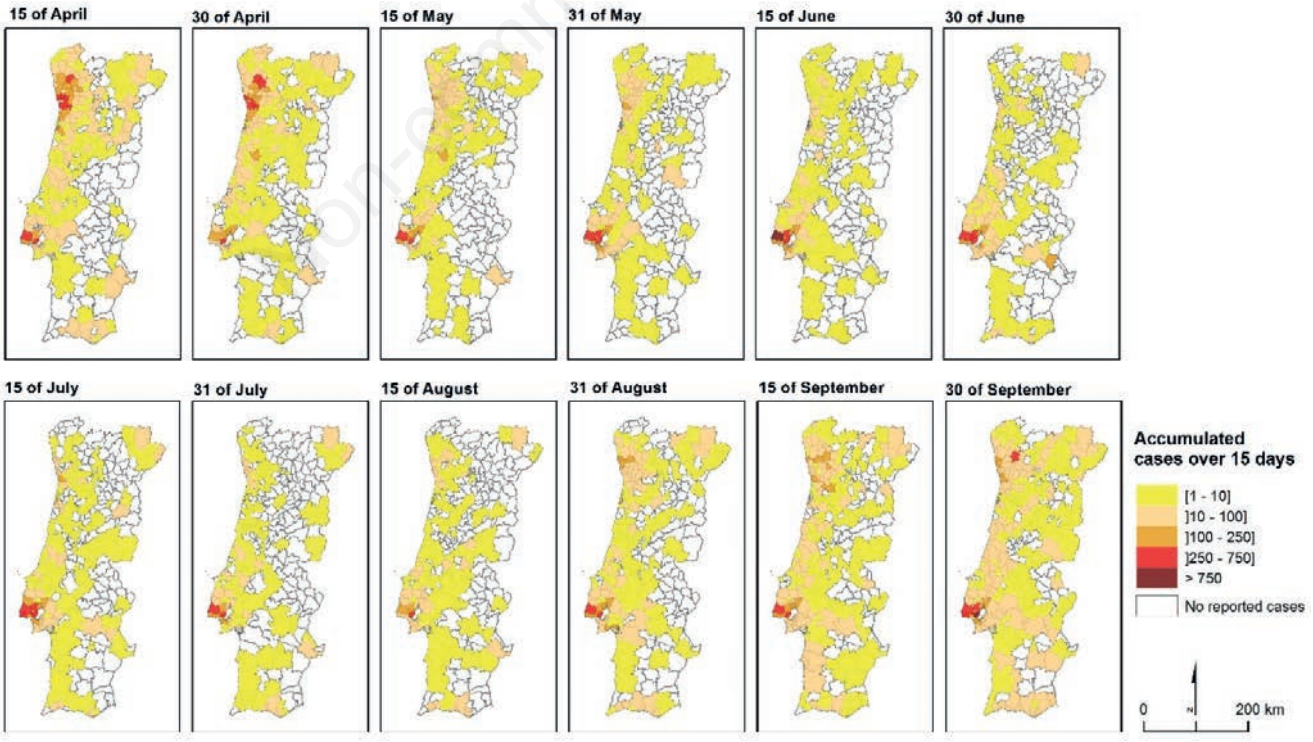


Figure 6. Total number of new COVID-19 cases per 15 days from the 1 April to 30 September, 2020. Source: DGS, 2020b.

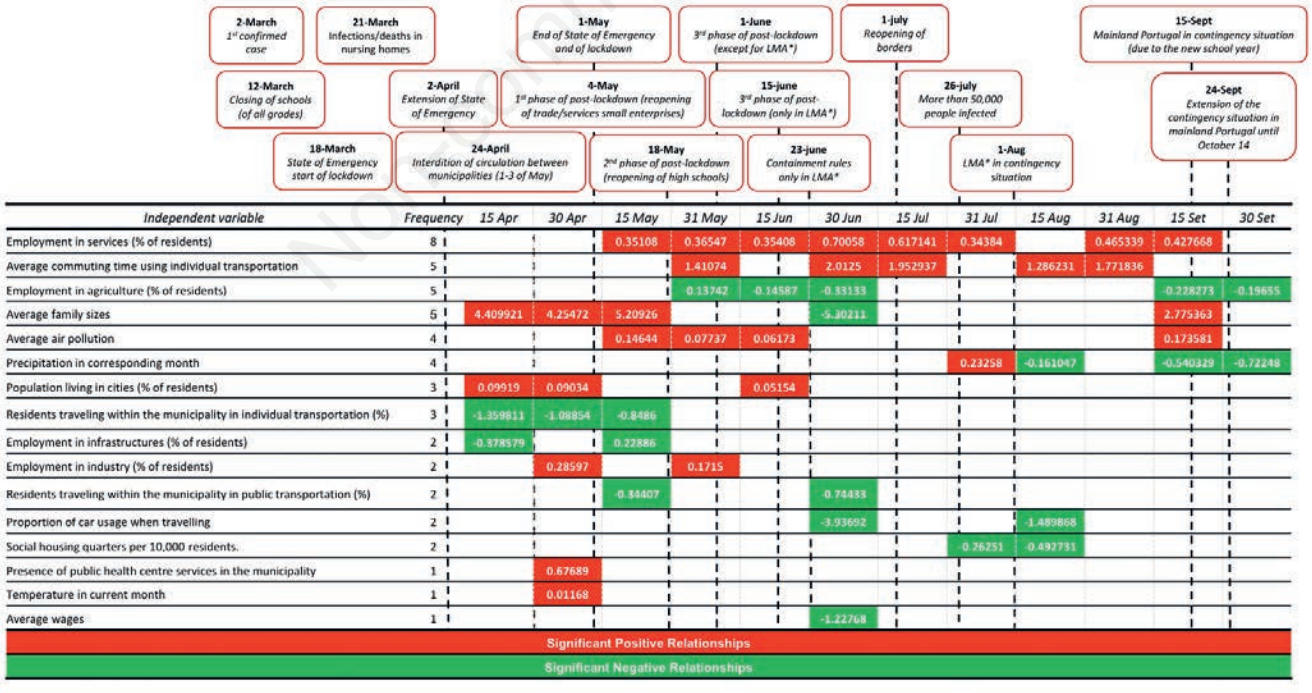
the first 100 COVID-19 days, which identify high urban density areas as the first spot of incidences. However, the temporal fluctuation in the significance of this relationship is interesting, showing that the relationship emerges immediately after the reopening of trade and services that followed the first lockdown (on 4 May 2020) and that it remained significant until the first half of August, the period when the number of disease cases was the lowest. Following this period, the relationship emerged again, and was only non-significant for the later period analysed here (16 to 30 September). Overall, these dynamics suggest that the role of the service sector as a disease transmission agent is strongly determined by the preventive measures that were put in practice by the national authorities.

The variable 'average travel time using individual transportation' was deemed positively significant for five out of the twelve 15-day periods analysed, in which the obtained results clearly show the importance of the first lockdown, which strongly limited the circulation. Thus, this relationship suggests a higher incidence of the disease in municipalities where residents use their car for longer time of travel to and from their workplace. These situations are likely to refer to municipalities surrounding major cities, such as Lisbon and Porto, which attract workers from distant municipalities. A possibility is that people who routinely move to large cities increase their chances of becoming infected because human interactions in these places tend to be higher - as caused by being employed in, or making use of, service-related activities. To some extent, this is supported by the apparent temporal coincidence of the relationship found for this variable and of those for the variable 'percentage of residents working in services'.

Regarding the negative relationship found for the variable 'percentage of workers in the primary sector (agriculture)', this may reflect the higher isolation of people in predominantly rural municipalities. Supporting this interpretation is the fact that this relationship emerged only for periods of time when the preventive measures were lessened, allowing the signal of a generally higher isolation of people in these areas to emerge.

Municipalities where the average size of households is higher (*i.e.* with more individuals) were found to have a significantly higher number of cases for three 15-day periods and a significantly lower number of cases for one period. While the former relationship is unsurprising, given the higher availability of available human hosts for the virus in each household, the negative association suggests that the expected effect of this variable cannot be interpreted so lightly. Because the negative relationship shows up after and temporally close to a consistent phase of positive relationships, one possibility could be that the previous higher incidence of cases in the more populated households led to a phase of immunity to the virus, resulting in significantly lower number of COVID-19 cases afterwards. However, that cannot be supported due to lack of available data.

Air pollution has a positive correlation with COVID-19 cases, especially in post-lockdown times. The high importance of air pollution is always hand-in-hand with the percentage of people working in the tertiary sector. The relationships identified for total monthly precipitation are also intriguing. These relationships emerge only in the latter half of the period analysed, and they start with a positive relationship with COVID-19 incidence and then change into a more temporally consistent negative relationship.



Independent variable	Frequency	15 Apr	30 Apr	15 May	31 May	15 Jun	30 Jun	15 Jul	31 Jul	15 Aug	31 Aug	15 Set	30 Set
Employment in services (% of residents)	8			0.35108	0.36547	0.35408	0.70058	0.617141	0.34384		0.465339	0.427668	
Average commuting time using individual transportation	5				1.41074		2.0125	1.952937		1.286231	1.771836		
Employment in agriculture (% of residents)	5				0.13742	-0.14587	-0.33133					-0.228273	-0.19655
Average family sizes	5	4.409921	4.25472	5.20926			-5.30211					2.775363	
Average air pollution	4			0.14644	0.07737	0.06173						0.173581	
Precipitation in corresponding month	4								0.23258	-0.161047		-0.540329	-0.72248
Population living in cities (% of residents)	3	0.09919	0.09034			0.05154							
Residents traveling within the municipality in individual transportation (%)	3	-1.359811	-1.08854	-0.8486									
Employment in infrastructures (% of residents)	2	-0.378579		0.22886									
Employment in industry (% of residents)	2		0.28597		0.1715								
Residents traveling within the municipality in public transportation (%)	2			-0.84407			0.74433						
Proportion of car usage when travelling	2						-3.93692			-1.489868			
Social housing quarters per 10,000 residents.	2								0.76251	-0.492731			
Presence of public health centre services in the municipality	1		0.67689										
Temperature in current month	1		0.01168										
Average wages	1						-1.22768						

Figure 7. Statistically associations between descriptive variables and number of new COVID-19 cases per 15-day period. Coloured cells represent significant relationships (P<0.05). Variables in red (green) are positively (negatively) related to COVID-19 incidence. The descriptive variables are ranked according to the total number of 15-day periods for which an association was statistically significant. Timeline of key events is matched with the periods analysed to facilitate the interpretation of results.

Lastly, we analysed the models residuals (Figure 8). Residuals indicated the natural variation of the data, a random factor (or not) that the model did not capture. If the model's assumptions are violated, the analysis would lead to dubious and unreliable results for the inference. These model flaws with regard to the assumptions can come from several factors such as non-linearity, non-normality, heteroscedasticity, non-independence and this can be caused by atypical points (outliers), which may or may not influence the model fit.

With this analysis, it is possible to compare the observed values with the expected values. When the result is negative demonstrates that the number of COVID cases in a certain municipality was below the expected and when positive, it indicates that the values were higher than expected. Municipalities with high positive residuals represent the places of higher risks, where more resources should be allocated in terms of raising awareness.

Discussion

This research tries to build a comprehensive set of descriptive variables, representing many putative human-related (*e.g.*, number of inhabitants, population density; main economic activities, education levels, type of transportation; pendular movements, access to health care, *etc.*) and environmental drivers (*e.g.*, mean temperature, total precipitation and air pollution). A relation among the number of COVID-19 cases and the descriptor variables was done using a multivariate regression model, accounting for the spatial non-independence of the study units, and report both statistically significant relationships as well as the form of relationship formed (*i.e.* positive, or negative). Our results could serve as the basis to the adoption or refinement of future containment or prevention measures.

The consistently higher incidence of the disease in municipalities with a higher percentage of people working in the tertiary sector strongly suggests an involvement of this type of activities in the transmission of the disease, a hypothesis that has already received relevant support, concerning, for instance, restaurants and bars (Ahmed *et al*., 2018; Fisher *et al*., 2020; Xie *et al*., 2020).

The positive relationship identified between COVID-19 incidence and air pollution is interesting. A positive relationship between air pollution and transmissibility of SARS-CoV-2 was also identified previously for other regions (Frontera *et al*., 2020), namely in more densely populated areas (Sajadi *et al*., 2020; Xie and Zhu, 2020). For these cases, the authors suggest that the set of conditions leading to the concentration of higher pollution levels (*e.g.*, reduced air circulation) may promote a longer permanence of the viral particles in the air, increasing its transmissibility. This may be the reason behind the relationships identified here, but because the most polluted areas are largely coincident with the larger urban centres of the country, we cannot exclude the possibility of other, urban-related, factors driving this relationship.

Nevertheless, some contradictory results occur in particular locations such as in Oslo (Menebo, 2020) and Singapore (Pani *et al*., 2020) where a positive correlation was found and in several Chinese provinces where the two situations coexist (Shahzad *et al*., 2020). Additionally, if it seems that air pollution plays a determinative role in COVID-19 outbreak and mortality - the number of

Figure 8. Models residuals.

confirmed cases was extremely higher in cities with more than 100 days of air pollution than cities with cleaner air (Coccia, 2020), probably a substantial relationship could also be found between urban air pollution and the transmission dynamics of COVID-19, which is why pollutant emission was considered.

The initial expectation for total monthly precipitation was that of a negative relationship owing to rainfall discouraging people from leaving home (Menebo, 2020). While this expectation is somewhat supported by the dominance of a negative relationship identified for this variable, the positive relationship also identified for one 15 days period is hard to explain.

Previous worldwide work has found a significant correlation between temperature and COVID-19 spread (Huang *et al*., 2020; Mandal and Panwar, 2020; Ozyigit, 2020; Wu *et al*., 2020; Notari, 2021). However, this tends to be negative correlation as observed in several Latin American cities (Bolaño-Ortiz *et al*., 2020; Prata *et al*., 2020), China (Li *et al*., 2020; Shi *et al*., 2020), USA (Li *et al*., 2020) and in Japan (Ujiie *et al*., 2020).

Notwithstanding other work using spatial analysis (Ribeiro and Santos, 2020), several limitations are present and should be acknowledged to avoid over-interpretation of our results. The lack of updated data and low spatial resolution must be highlighted in this regard. Indeed, data used to characterize the non-environmental attributes of municipalities that refer to long-term patterns verified in previous years and not to the status of each municipality in each of the periods analysed. While these long-term descriptors should allow a characterization of the relative differences between municipalities to some extent, this characterization should be more accurate for the periods when no major restrictions to human activity were imposed (*i.e*. when human activity was allowed to proceed 'as usual').

The representation of patterns of human activity in periods where strong restrictions were imposed would be best represented using temporally matching data, which to our knowledge, is not available for most of the factors considered here. In addition, the spatial detail of our data cannot shed light on the precise mechanisms hindering or promoting the transmission of the disease. Local level, high-detail studies will be required to shed further light on the mechanisms behind the main relationships we found (*e.g*., regarding the role of services-sector, average household size or air pollution). Additionally, complete (spatial and temporal) data, regarding the patient characteristics are indispensable for a detailed study of mortality patterns.

Conclusions

The obtained results show that the adopted methodological approach can identify significant relations among COVID-19 and several environmental, socioeconomic, demographic and human mobility factors. The number of employees in services activities was the variable identified as the most relevant. However, the use of a 15-day temporal resolution shows that the importance of each factor could change with time. These results are important to better understand the influence of the adopted safety and restriction measures but are even more relevant to support both spatially and temporally adjusted measures, allowing different 'realities' in the country to be worked out differently. Despite the limitations acknowledged, our work highlighted a few consistent relationships between COVID-19 incidence and attributes of municipalities that could be of interest for the future prevention of the disease and to further understand the factors that mediate the transmissibility of the SARS-CoV-2 virus.

References

Ahmed F, Zviedrite N, Uzicanin A, 2018. Effectiveness of workplace social distancing measures in reducing influenza transmission: a systematic review. BMC Public Health 18:518.

Akaike H, 1974. A new look at the statistical model identification. IEEE Trans Automat Contr 19:716-23.

Arashi M, Bekker A, Salehi M, Millard S, Erasmus B, Cronje T, Golpaygani M, 2020. Spatial analysis and prediction of COVID-19 spread in South Africa after lockdown. arXiv [preprint].

Azevedo L, Pereira MJ, Ribeiro MC, Soares A, 2020. Geostatistical COVID-19 infection risk maps for Portugal. Int J Health Geogr 19:25.

Bai Y, Yao L, Wei T, Tian F, Jin D-Y, Chen L, Wang M, 2020. Presumed asymptomatic carrier transmission of COVID-19. JAMA 323:1406-7.

Barton K, 2020. MuMIn: multi-model inference. R package version 1.43.17. Available from: https://cran.r-project.org/package=MuMIn Accessed: 15 February 2021).

Bate ST, Clark RA, 2014. The design and statistical analysis of animal experiments. Cambridge University Press, Cambridge, UK.

Bolaño-Ortiz TR, Camargo-Caicedo Y, Puliafito SE, Ruggeri MF, Bolaño-Diaz S, Pascual-Flores R, Saturno J, Ibarra-Espinosa S, Mayol-Bracero OL, Torres-Delgado E, Cereceda-Balic F, 2020. Spread of SARS-CoV-2 through Latin America and the Caribbean region: A look from its economic conditions, climate and air pollution indicators. Environ Res 191:109938.

Bozdogan H, 1987. Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions. Psychometrika 52:345-70.

Brooks ME, Kristensen K, van Benthem KJ, Magnusson A, Berg CW, Nielsen A, Skaug HJ, Mächler M, Bolker BM, 2017. glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. R J 9:378-400.

Chan KH, Peiris JSM, Lam SY, Poon LLM, Yuen KY, Seto WH, 2011. The effects of temperature and relative humidity on the viability of the SARS coronavirus. Adv Virol 2011:734690.

Chen Z-L, Zhang Q, Lu Y, Guo Z-M, Zhang X, Zhang W-J, Guo C, Liao C-H, Li Q-L, Han X-H, Lu J-H, 2020. Distribution of the COVID-19 epidemic and correlation with population emigration from Wuhan, China. Chin Med J (Engl) [Epub ahead of print].

Coccia M, 2020. Factors determining the diffusion of COVID-19 and suggested strategy to prevent future accelerated viral infectivity similar to COVID. Sci Total Environ 729:138474.

Cornes RC, van der Schrier G, van den Besselaar EJM, Jones PD, 2018. An ensemble version of the E-OBS temperature and precipitation data sets. J Geophys Res Atmos 123:9391-9409.

Dagpunar J, 2019. The gamma distribution. Significance 16:10-1.

Dasgupta A., Wahed, A., 2014. Laboratory statistics and quality control. In: Dasgupta A.W. (Ed.), Immunology and laboratory quality control - Clinical chemistry, immunology and laboratory quality control. Elsevier, San Diego, CA, USA, pp. 47-66.

DGS, 2020a. Casos de Infeção por Novo Coronavírus (COVID-19). Available from: https://covid19.min-saude.pt/comunica-

dos/ Accessed: 15 February 2021.

DGS, 2020b. COVID-19 - Relatórios de situação [National Health Delegation - COVID19 daily reports]. Available from: https://covid19.min-saude.pt/relatorio-de-situacao/ Accessed: 15 February 2021.

DGS, 2020c. COVID-19. Available from: https://covid19.min-saude.pt/category/perguntas-frequentes/ Accessed: 15 February 2021.

Dormann CF, Elith J, Bacher S, Buchmann C, Carl G, Carré G, Marquéz JRG, Gruber B, Lafourcade B, Leitão PJ, Münkemüller T, McClean C, Osborne PE, Reineking B, Schröder B, Skidmore AK, Zurell D, Lautenbach S, 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography (Cop.) 36:27-46.

Ferré J, 2009. Regression diagnostics. In: Brown SD, Tauler R, Walczak BBT-CC (Eds.), Comprehensive chemometrics chemical and biochemical data analysis. Elsevier, Oxford, pp. 33-89.

Fisher KA, Tenforde MW, Feldstein LR, Lindsell CJ, Shapiro NI, Files DC, Gibbs KW, Erickson HL, Prekker ME, Steingrub JS, 2020. Community and close contact exposures associated with COVID-19 among symptomatic adults ≥18 years in 11 outpatient health care facilities - United States, July 2020. Morb Mortal Wkly Rep 69:1258.

Frontera A, Martin C, Vlachos K, Sgubin G, 2020. Regional air pollution persistence links to covid19 infection zoning. J Infect 81:318-56.

Gareth J, Daniela W, Trevor H, Robert T, 2021. An introduction to statistical learning: with applications in R, 2nd ed. Springer, New York, NY, USA.

Gelman A, Hill J, 2006. Data analysis using regression and multilevel/hierarchical models, analytical methods for social research. Cambridge University Press, Cambridge, UK.

Gorbalenya AE, Baker SC, Baric RS, de Groot RJ, Drosten C, Gulyaeva AA, Haagmans BL, Lauber C, Leontovich AM, Neuman BW, Penzar D, Perlman S, Poon LLM, Samborskiy D, Sidorov IA, Sola I, Ziebuhr J, 2020. Severe acute respiratory syndrome-related coronavirus: The species and its viruses - a statement of the Coronavirus Study Group. bioRxiv 2020.02.07.937862.

Guan W, Ni Z, Hu Yu, Liang W, Ou C, He J, Liu L, Shan H, Lei C, Hui DSC, Du B, Li L, Zeng G, Yuen K-Y, Chen R, Tang C, Wang T, Chen P, Xiang J, Li S, Wang J-L, Liang Z, Peng Y, Wei L, Liu Y, Hu Y-H, Peng P, Wang J-M, Liu J, Chen Z, Li G, Zheng Z, Qiu S, Luo J, Ye C, Zhu S, Zhong N, 2020. Clinical characteristics of coronavirus disease 2019 in China. N Engl J Med 382:1708-20.

Hernández-Orallo J, 2013. ROC curves for regression. Pattern Recognit 46:3395-411.

Huang Z, Huang J, Gu Q, Du P, Liang H, Dong Q, 2020. Optimal temperature zone for the dispersal of COVID-19. Sci Total Environ 736:139487.

Hurvich CM, Tsai C-L, 1989. Regression and time series model selection in small samples. Biometrika 76:297-307.

Jamil T, Ozinga WA, Kleyer M, ter Braak CJF, 2013. Selecting traits that explain species-environment relationships: a generalized linear mixed model approach. J Veg Sci 24:988-1000.

Johnston R, Jones K, Manley D, 2018. Confounding and collinearity in regression analysis: a cautionary tale and an alternative procedure, illustrated by studies of British voting behaviour. Qual Quant 52:1957-76.

Kaliappan J, Srinivasan K, Mian Qaisar S, Sundararajan K, Chang C-Y, 2021. Performance evaluation of regression models for the prediction of the COVID-19 reproduction rate. Front. Public Heal 9:1319.

Kurz CF, 2017. Tweedie distributions for fitting semicontinuous health care utilization cost data. BMC Med Res Methodol 17:171.

Laires PA, Nunes C, 2020. Population-based estimates for high risk of severe COVID-19 disease due to age and underlying health conditions. Acta Med Port 33:720-5.

Lakhani A, 2020. Which Melbourne metropolitan areas are vulnerable to COVID-19 based on age, disability, and access to health services? Using spatial analysis to identify service gaps and Inform Delivery J Pain Symptom Manage 60:e41-4.

Li AY, Hannah TC, Durbin JR, Dreher N, McAuley FM, Marayati NF, Spiera Z, Ali M, Gometz A, Kostman JT, Choudhri TF, 2020. Multivariate analysis of black race and environmental temperature on COVID-19 in the US. Am J Med Sci 360:348-56.

Li H, Xu X-L, Dai D-W, Huang Z-Y, Ma Z, Guan Y-J, 2020. Air pollution and temperature are associated with increased COVID-19 incidence: a time series study. Int J Infect Dis 97:278-82.

Ma Y, Zhao Y, Liu J, He X, Wang B, Fu S, Yan J, Niu J, Zhou J, Luo B, 2020. Effects of temperature variation and humidity on the death of COVID-19 in Wuhan, China. Sci Total Environ 724:138226.

Mandal CC, Panwar MS, 2020. Can the summer temperatures reduce COVID-19 cases? Public Health 185:72-9.

Maroko AR, Nash D, Pavilonis BT, 2020. COVID-19 and inequity: a comparative spatial analysis of New York City and Chicago Hot Spots. J Urban Heal 97:461-70.

Marquaridt DW, 1970. Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. Technometrics 12:591-612.

Marques TS, 2020. O mosaico territorial do risco ao contágio e à mortalidade por covid-19 em portugal continental. Finisterra 19-26.

McCulloch CE, Neuhaus JM, 2014. Generalized linear mixed models. Wiley Stats Ref Stat Ref Online, doi:10.1002/9781118445112.stat07540.

Mei-Ling LT, 2004. Transformation and normalization BT. In: Lee M.-L.T. (Ed.), Analysis of microarray gene expression data. Springer US, Boston, MA, USA, pp. 67-84.

Melo HPM, Henriques J, Carvalho R, Verma T, da Cruz JP, Araujo NAM, 2020. Heterogeneous impact of a lockdown on inter-municipality mobility. [Epub ahead of print].

Menebo MM, 2020. Temperature and precipitation associate with Covid-19 new daily cases: A correlation study between weather and Covid-19 pandemic in Oslo, Norway. Sci Total Environ 737:139659.

Mitchell TD, Camp J, 2021. The use of the Conway-Maxwell-Poisson in the seasonal forecasting of tropical cyclones. Weather Forecast 36:929-39.

MohammadEbrahimi S, Mohammadi A, Bergquist R, Dolatkhah F, Olia M, Tavakolian A, Pishgar E, Kiani B., 2021. Epidemiological characteristics and initial spatiotemporal visualisation of COVID-19 in a major city in the Middle East. BMC Public Health 21:1-18.

Mollalo A, Mohammadi A, Mavaddati S, Kiani B, 2021. Spatial

analysis of COVID-19 vaccination: a scoping review. Int J Environ Res Public Heal [Epub ahead of print].
Murgante B, Borruso G, Balletto G, Castiglia P, Dettori M, 2020. Why Italy first? Health, geographical and planning aspects of the COVID-19 outbreak. Sustain. [Epub ahead of print].
Notari A, 2021. Temperature dependence of COVID-19 transmission. Sci Total Environ 763:144390.
Orea L, Álvarez IC, 2020. How effective has been the Spanish lockdown to battle COVID-19? A spatial analysis of the coronavirus propagation across provinces. Health Econ 1-20.
Ozyigit A, 2020. Understanding Covid-19 transmission: the effect of temperature and health behavior on transmission rates. Infect Dis Heal 25:233-8.
Paez A, Lopez FA, Menezes T, Cavalcanti R, Pitta MG da R, 2020. A Spatio-temporal analysis of the environmental correlates of COVID-19 incidence in Spain. Geogr Anal [Epub ahead of print].
Pani SK, Lin N-H, RavindraBabu S, 2020. Association of COVID-19 pandemic with meteorological parameters over Singapore. Sci Total Environ 740:140112.
Prata DN, Rodrigues W, Bermejo PH, 2020. Temperature significantly changes COVID-19 transmission in (sub)tropical cities of Brazil. Sci Total Environ 729:138862.
Prazeres F, Passos L, Simões JA, Simões P, Martins C, Teixeira A, 2021. COVID-19-related fear and anxiety: spiritual-religious coping in healthcare workers in Portugal. Int J Environ Res Public Heal [Epub ahead of print].
Quilodrán CS, Currat M, Montoya-Burgos JI, 2020. Climatic factors influence COVID-19 outbreak as revealed by worldwide mortality. medRxiv 2020.04.20.20072934.
R Core Team, 2019. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Available from: https://www.r-project.org/ Accessed: 15 February 2021).
Ribeiro AI, Santos CJ, 2020. The importance of spatial analysis of COVID-19 pandemic for health geography: challenges and perspectives. Finisterra LV, 37-42.
Ricoca Peixoto V, Vieira A, Aguiar P, Carvalho C, Rhys Thomas D, Abrantes A, 2020. Initial assessment of the impact of the emergency state lockdown measures on the 1st wave of the COVID-19 epidemic in Portugal. Acta Med Port 33:733-41.
Roy S, Bhunia GS, Shit PK, 2020. Spatial prediction of COVID-19 epidemic using ARIMA techniques in India. Model Earth Syst Environ [Epub ahead of print].
Saha D, Alluri P, Dumbaugh E, Gan A, 2020. Application of the Poisson-Tweedie distribution in analyzing crash frequency data. Accid Anal Prev 137:105456.
Sajadi MM, Habibzadeh P, Vintzileos A, Shokouhi S, Miralles-Wilhelm F, Amoroso A, 2020. Temperature, humidity, and latitude analysis to estimate potential spread and seasonality of coronavirus disease 2019 (COVID-19). JAMA Netw Open 3:e2011834-e2011834.
Shahzad F, Shahzad U, Fareed Z, Iqbal N, Hashmi SH, Ahmad F, 2020. Asymmetric nexus between temperature and COVID-19 in the top ten affected provinces of China: a current application of quantile-on-quantile approach. Sci Total Environ 736:139115.
Shi P, Dong Y, Yan H, Zhao C, Li X, Liu W, He M, Tang S, Xi S, 2020. Impact of temperature on the dynamics of the COVID-19 outbreak in China. Sci Total Environ 728:138890.
Tamagusko T, Ferreira A, 2020. Data-Driven approach to understand the mobility patterns of the Portuguese population during the COVID-19 pandemic. Sustain. [Epub ahead of print].
Ujiie M, Tsuzuki S, Ohmagari N, 2020. Effect of temperature on the infectivity of COVID-19. Int J Infect Dis 95:301-3. [Epub ahead of print].
Velleman PF, Welsch RE, 1981. Efficient computing of regression diagnostics. Am Stat 35:234-42.
Vittinghoff E, Glidden DV, Shiboski SC, McCulloch CE, 2012. Logistic regression. In: Regression Methods in Biostatistics. Springer, Boston, MA, pp. 139-202.
Wang J, Tang K, Feng K, Lv W, 2020. High temperature and high humidity reduce the transmission of COVID-19. Available from: https://ssrn.com/abstract=3551767 2020. Accessed: 15 February 2021.
WHO - World Health Organization., 2020. Coronavirus disease 2019 (COVID-19) (Situation Report – 94). Available from: https://apps.who.int/iris/handle/10665/331865 Accessed: 15 February 2021.
Wu Y, Jing W, Liu J, Ma Q, Yuan J, Wang Y, Du M, Liu M, 2020. Effects of temperature and humidity on the daily new cases and new deaths of COVID-19 in 166 countries. Sci Total Environ 729:139051.
Xie J, Zhu Y, 2020. Association between ambient temperature and COVID-19 infection in 122 cities from China. Sci Total Environ 724:138201.
Xie K, Liang B, Dulebenets MA, Mei Y, 2020. The impact of risk perception on social distancing during the COVID-19 Pandemic in China. Int J Environ Res Public Heal. [Epub ahead of print].