Spatial disparities in incidence of COVID-19 in relation to economic and socio-demographic factors in the Autonomous Community of Madrid, Spain

Severino Escolano-Utrilla
Universidad de Zaragoza
severino@unizar.es

Andrés Roca-Medina
Diego Barrado-Timón
Universidad Autónoma de Madrid
andres.roca@uam.es
diego.barrado@uam.es

Received: December 2023
Accepted: June 2024
Published: July 2024

Abstract

This article models the relationship between the incidence of COVID-19 and several socio-economic factors during the second period of epidemic (22 June 2020 to 06 December 2020) in the Autonomous Community of Madrid, Spain. Data collected from Basic Health Zones (BHZs) is adjusted using the random forest method, which proves very appropriate for capturing non-linear relationships and obtaining accurate and robust predictions. The results show that the impact of the examined socio-economic variables on rates of incidence of COVID-19 was not uniform, and that levels of mean income by neighborhood exerted stronger influence than population density, proportion of the Spanish population, mean age of the population or average household size. A complex spatial pattern emerges from the combination of impacts, reflecting the relative weights of the different factors in terms of intensity of the pandemic. This information may be considered strategic for the effective future management of health resources.

Keywords: COVID-19; economic factors; socio-demographic factors; random forest method; Autonomous Community of Madrid

https://doi.org/10.5565/rev/dag.904 ISSN 0212-1573 (imprès), ISSN 2014-4512 (en línia)

Resum. *Desigualtats espacials de la incidència de la COVID-19 en relació amb factors econòmics i sociodemogràfics a la Comunitat Autònoma de Madrid (Espanya)*

En aquest article es modela la relació entre la incidència de la COVID-19 i diversos factors socioeconòmics durant el segon període epidèmic (22/06/2020 – 06/12/2020) a la Comunitat de Madrid (Espanya). Les dades obtingudes a zones bàsiques de salut (ZBS) s'ajusten mitjançant el mètode del bosc aleatori, molt apropiat per capturar relacions no lineals i obtenir prediccions més precises i robustes. Els resultats mostren que l'impacte de les variables socioeconòmiques en les taxes d'incidència de la COVID-19 no és uniforme i que la renda té una influència més forta que la densitat de població, la proporció de població espanyola, l'edat mitjana de la població i la mida mitjana de les llars. De la combinació dels impactes emergeix un patró espacial complex que reflecteix el pes relatiu dels diferents factors en la intensitat de la pandèmia. Aquesta informació és estratègica per a la gestió eficaç dels recursos sanitaris.

Paraules clau: COVID-19; factors econòmics; factors sociodemogràfics; bosc aleatori; Comunitat Autònoma de Madrid

Resumen. *Desigualdades espaciales de la incidencia de la COVID-19 en relación con factores económicos y sociodemográficos en la Comunidad Autónoma de Madrid (España)*

En este artículo se modela la relación entre la incidencia de la COVID-19 y varios factores socioeconómicos durante el segundo período epidémico (22/06/2020 – 06/12/2020) en la Comunidad de Madrid (España). Los datos tomados en zonas básicas de salud (ZBS) se ajustan mediante el método del bosque aleatorio, muy apropiado para capturar relaciones no lineales y obtener predicciones más precisas y robustas. Los resultados muestran que el impacto de las variables socioeconómicas en las tasas de incidencia de la COVID-19 no es uniforme y que la renta media tiene una influencia más fuerte que la densidad de población, la proporción de población española, la edad media de la población y el tamaño medio de los hogares. De la combinación de los impactos emerge un patrón espacial complejo que refleja el peso relativo de los diferentes factores en la intensidad de la pandemia. Esta información es estratégica para la gestión eficaz de los recursos sanitarios.

Palabras clave: COVID-19; factores económicos; factores sociodemográficos; bosque aleatorio; Comunidad Autónoma de Madrid

Résumé. *Inégalités spatiales de l'incidence de la COVID-19 en rapport avec les facteurs économiques et sociodémographiques dans la Communauté Autonome de Madrid (Espagne)*

Dans cet article, nous modélisons la relation entre l'incidence de la COVID-19 et plusieurs facteurs socio-économiques au cours de la deuxième période épidémique (22/06/2020 au 06/12/2020) dans la Communauté de Madrid, Espagne. Les données recueillies dans les zones de santé de base (ZSB) sont ajustées à l'aide de la méthode des forêts aléatoires, qui est très appropriée pour capturer les relations non linéaires et obtenir des prédictions plus précises et plus robustes. Les résultats montrent que l'impact des variables socio-économiques sur les taux d'incidence de la COVID-19 n'est pas uniforme et que le revenu a une influence plus forte que la densité de population, la proportion de la population espagnole, l'âge moyen de la population et la taille moyenne des ménages. Un schéma spatial complexe émerge de la combinaison des impacts, reflétant le poids relatif des différents facteurs dans l'intensité de la pandémie. Ces informations sont stratégiques pour une gestion efficace des ressources sanitaires.

Mots-clés : COVID-19 ; facteurs économiques ; facteurs sociodémographiques ; forêts aléatoires ; Communauté Autonome de Madrid

Summary

1. Introduction

Considerable inequalities observed in the incidence of COVID-19 (between individuals, social groups or proximate versus distant spaces) have led researchers to wonder at their possible causes. Studies have identified certain factors that can condition incidence, both in individuals and in populations, making it possible to model the spatio-temporal dynamics of the pandemic. Most research has focused on the relationship between one or more factors and the epidemiological situation prevalent in a specific place and time. The units of analysis (thematic, spatial, temporal) and the techniques employed in these studies have varied greatly. The explanatory variables most frequently used have measured some particular property of the target population's health status and/or the degree of socio-economic vulnerability of individuals or populations, or the environmental and urban conditions of the territory, or varying degrees of human mobility, or other cultural and ethnographic aspects.

This paper aims to contribute to greater understanding of the ways in which certain socioeconomic and geographical factors may have been related to the degree of incidence of the pandemic, taking two critical aspects of this relationship into account. First, it must be recognized that these factors do not operate independently but instead intersect and interact – for example, income levels are correlated with other variables, such as age (Mena et al., 2021) – and this association makes it difficult to estimate the impact of each factor separately (De Cos et al., 2022; Kolak et al., 2020; Maza and Hierro, 2021). Second, due to the interaction between factors, the influence of each on the incidence (and diffusion) of COVID-19 can vary across space and time. While the most common techniques used for analysis assume that these factors are constant (spatial stationarity), such an assumption may in fact limit the identification of spatial patterns of incidence. Tokey (2021) points out that studies on COVID-19 have not devoted attention to the effects of spatial stationarity, and Tchicaya et al. (2021) state that methods which consider spatial stationarity produce patterns distinct from those that do not.

To achieve the stated goals, determinations have been made about the data, methods and study area. To measure the incidence of the pandemic, public data have been collected on the number of cases diagnosed, separated by Basic Health Zones (BHZs). The five selected socio-economic and geographic variables have been made discrete, at various levels. For each level of each variable, the incidence of COVID-19 has been calculated from 10 March 2020 to 22 March 2022. From this analysis, it was found that the period between 22 June 2020 and 06 December 2020 – corresponding to the second wave or period of epidemic (SPE) (Instituto de Salud Carlos III, 2021) – is most appropriate for studying possible non-random associations between incidence of the pandemic and the explanatory factors considered. During that phase, and prior to the launch of vaccination campaigns, the pandemic had considerable impact, as non-pharmacological measures (mobility restrictions, distancing, use of masks) did not prevent notable inequalities in infection rates according to income levels. To model non-stationary spatial variability, a machine-learning method known as the *random forest* (RF) algorithm has been employed, with regressions carried out between incidence of the pandemic and the selected factors. Finally, a typology of combinations of impact values for the tested variables on the incidence rates of each BHZ was developed. The area of study is the Autonomous Community of Madrid (Spain), which features great territorial and social diversity, generating multiple patterns of influence of the variables on the incidence rates[1] of COVID-19, which were very high, especially in 2020.

The remainder of the article is organized as follows: The next section presents an overview of prior approaches to research on this topic and related results. Section three describes the data and methodological procedures employed here, along with general characteristics of the area of study. The fourth section presents the results obtained and their interpretation, followed by a proposal for future lines of research.

2. Bibliographic review

Studies of diverse factors determining the incidence of COVID-19 are numerous and have employed different units of analysis (spatial, thematic, temporal) and geographical scopes. Here we mention only the results of research deemed relevant to the objectives of this study.

One key aspect of research around the pandemic is the choice of model variables. Those most frequently used to measure the incidence of pandemic have been the number of confirmed positive cases, and the estimated mortality from COVID-19 (Franch-Pardo et al., 2020, 2021). This is mainly due to the reliability of available data. To measure socio-economic dimensions, the most commonly used indicators are those of income level or those (both simple and compound) representing the socio-economic status of persons, households or the specific spatial units under study. The age variable is represented by the mean age of the population in each spatial unit, or else grouped into intervals according to the objectives of a given work. Geographical and environmental conditions are often expressed by population density and by mean temperature, among other variables. Due to this diversity of variables, indicators, techniques and units of analysis, precise comparisons between different locations are difficult to draw.

1. The expression ‘incidence rate’ has been maintained out of fidelity to the data source, although strictly speaking, the value used to measure incidence is a ratio between the number of cases diagnosed with COVID-19 and the total population of each Basic Health Zone.

Many of the works consulted have included multiple socio-economic, demographic or environmental variables in their models in order to analyze joint effects on the incidence and spread of the pandemic. A negative association between wealth and the incidence of COVID-19 has been verified in many studies, at scales ranging from the national (Almendra et al., 2021; Benita and Gasca-Sanchez, 2021; Florida and Mellander, 2022; Gaynor and Wilson, 2020; Karaye and Horney, 2020; Raymundo et al., 2021) to the sub-national (Tepe, 2023) to the local, in cities and metropolitan areas of diverse economic base and located in both emerging countries (Cestari et al., 2021; Hierro and Maza, 2023; Maza and Hierro, 2021; You et al., 2020) and advanced economies (Amdaoud et al., 2020; Cordes and Castro, 2020). Several studies with a focus on Spain have verified this relationship at the metropolitan scale (Amengual-Moreno et al., 2020; Arauzo-Carod et al., 2021; Baena-Díez et al., 2020; González Pérez and Piñeira Mantiñán, 2020; Hierro and Maza, 2023; López-Gay et al., 2022; Ruiz-Pérez et al., 2023).

Although greater socio-economic vulnerability has been shown to have increased the incidence of COVID-19, the relationship between these two variables changes across time and space as mediated by the age structure of the population, the urban form and population density, and the relative extent of public health systems, among other factors. Thus in some cases a direct relationship between wealth and incidence has been detected – that is, where the highest incidence rates are associated with the highest income groups, especially in urban areas (De Mattos et al., 2020; Schellekens and Sourrouille, 2020).

The age of a population has been found to have influenced incidence as well as mortality. Reports from the Carlos III Health Institute (a Spanish Public Health and Research Institute located in Madrid) show that the incidence and lethality of the pandemic in Spain was higher in groups aged over 50 years old, but also that increases were fluctuating rather than linear (Instituto de Salud Carlos III (ISCIII), 2021). This relationship was elsewhere corroborated at an urban scale (Romero Starke et al., 2021; Santesmasses et al., 2020; Sasson, 2021). In addition to reasons intrinsic to age, unequal incidence and spatial patterns of spread between age strata have been attributed to differences in lifestyles, social relationships and mobility (Marí-Dell'Olmo et al., 2021). Household size has also been determined as relating positively to both increased vulnerability to COVID-19 and the speed of spread (De Cos et al., 2022; Sigler et al., 2020).

Another group of studies has investigated the relationship between COVID-19 and certain urban, environmental and mobility aspects. Population density has long been considered a health risk factor in contagious disease, given its relationship to physical distancing between people. The results reveal that the influence of density (and other factors) depends on local conditions as well as on general context. In some cases, high density is thought to have had a negative impact on the incidence of pandemic while favoring its spread (De Cos et al., 2022; Niu et al., 2020); in other cases, density was found to have scarcely affected the incidence and spread of COVID-19 (Hamidi et al., 2020), even constituting a possible advantage for controlling spread (Kling, 2020).

Table 1. Methods, variables, areas and units of selected studies on the factors of COVID-19 incidence

Method/Variables	Study area; spatial units; variable considered	Publication
Correlation and regression analysis; principal component analysis. Dependent variables: positive cases, mortality. Independent variables: twelve socio-economic variables (place-based factors) and three variables relating to diffusion factors	Sweden; all municipalities and neighborhoods of three cities (neighborhoods); positive cases and mortality from COVID-19	Florida and Mellander (2022)
Epidemic models; calculation of excess mortality; correlations between socio-economic status, excess mortality and tests performed	Santiago (Chile); municipalities; positive cases, estimated mortality and tests performed and estimated	Mena et al. (2021)
Least squares regression (OLS); spatial autoregression (SAR), conditional autoregressive model (CAR) and multiscale geographically weighted regression (MGWR) model	Brazil; municipalities; positive cases	Raymundo et al. (2021).
Multiscale geographically weighted regression (MGWR) model	U.S.; county; positive cases	Mollalo et al. (2020)
Multiple quantile regression analysis	Rio de Janeiro (Brazil); districts (neighborhoods); accumulated cases	De Mattos et al. (2020)
Moran's global index I; cluster space via kernel (Kulldorff); descriptive statistics, correlation	New York (U.S.); ZIP code; positive cases and number of tests	Cordes and Castro (2020)
Least squares regression (OLS); spatial autoregression (SAR), General Spatial Model (GSM); Moran's global index I	Florida; (U.S.); ZIP codes; confirmed COVID-19 cases	Tepe (2023)
Structural equation modeling to measure the direct and indirect impacts of density on COVID-19 incidence and mortality	U.S.; metropolitan counties	Hamidi et al. (2020)
Linear regression and random forests regression between positive cases (dependent variable) and climate variables, social distance and population, area and shape of counties (explanatory variables)	U.S.; counties	Zhang et al. (2022)

Source: Own elaboration.

Finally, to model the relationship between those variables that express the incidence of COVID-19 and those related to socio-economic, environmental and territorial aspects, regression methods have been preferred. In addition to ordinary linear regression, quantile regression or other types of regression, the use of geographically weighted regression (GWR) and of techniques addressing spatial and temporal dependence, such as autocorrelation, is also quite common. This group of local regression techniques explains the variability of incidence of the pandemic more adequately than ordinary linear regression (Mollalo et al., 2020). Only one of the works reviewed employed machine-learning methods – specifically random forest regression (Zhang et al., 2022) – to study transmission of the pandemic. The present analyses (see Table 1) have mainly been undertaken using Geographic Information Systems programs (Franch-Pardo et al., 2021).

3. Materials and methods

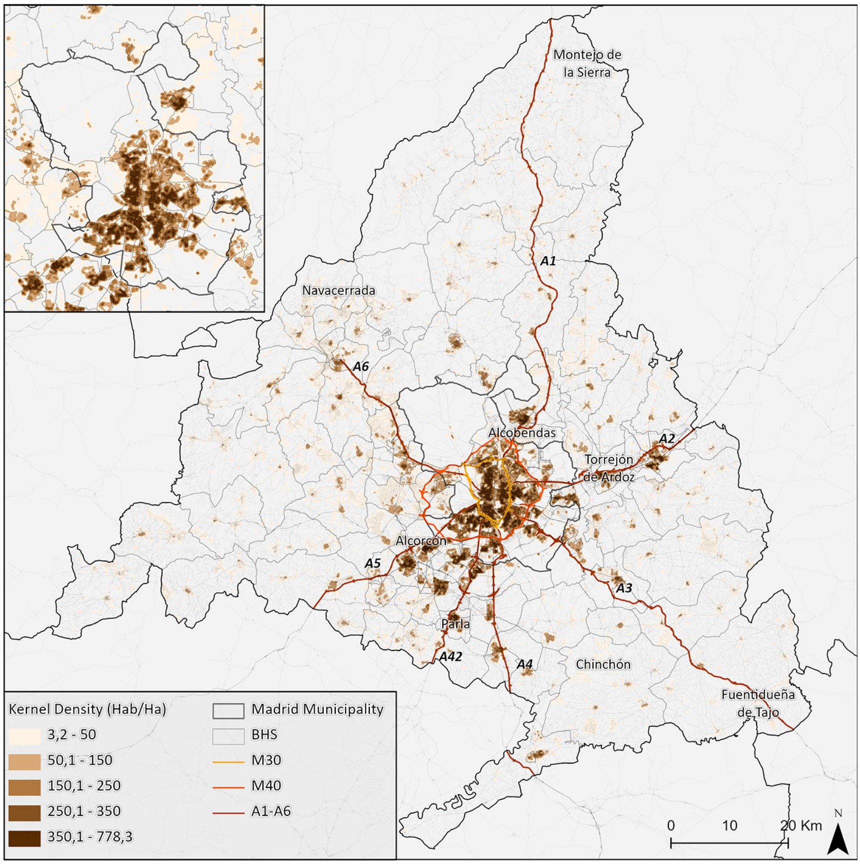
3.1. Presentation of the area of study

The Autonomous Community of Madrid presents a case of great interest to this research, due to several particularities. First, it is the largest metropolitan area in Spain and the second-largest in the European Union (after Paris) in both total population and GDP (Eurostat, 2023). At the same time, since the end of the 20th century, Madrid has been experiencing a gradual process of socio-spatial segregation that, although not unique in Europe, stands out in comparative studies in terms of severity (Musterd et al., 2017; Sorando and Leal, 2019). Among the main problems in this region are access to housing and health services, and social exclusion. Arroyo-Menéndez et al. (2022) have further noted a significant digital divide (although this study covers only the municipality of Madrid, and not the Autonomous Community as a whole).

At the same time, the impact of the pandemic in the Community of Madrid was among the most severe in Europe, particularly during 2020. The World Health Organization (WHO) placed the Community of Madrid in the 95th percentile of regions with the highest cumulative rate of incidence for 2020, and in the 85th percentile for 2021. In addition, mortality was very high, comparable to that of other communities in the European Union such as Castilla-La Mancha, Catalonia and Milan (Konstantinoudis et al., 2022; Rodríguez-Pose and Burlina, 2021). On the other hand, a study by Wang et al. (2022) estimating excess mortality around the world found that the Community of Madrid was not as badly affected as other regions in the world.

Management of the pandemic both politically and in health-care terms has also been a topic of debate. The government of the Community of Madrid argued for an open policy while the national government took a more cautious position; the media echoed this dichotomy, especially with regard to perimeter lockdowns. Such measures, which had already been applied by municipalities in other Autonomous Communities in Spain such as Catalonia, were based on three criteria: a cumulative incidence rate (CIR) above an established threshold (which ranged from 400 to 1,000 cases per 100,000 inhabitants); a stable or increasing trend in the CIR; and geographic contiguity. Nevertheless, according to a study by Fontán-Vela et al. (2021) the impact of these restrictions proved quite limited.

Figure 1. Distribution of population in the Community of Madrid (inhabitants/hectare) and BHZ (2021)



Source: IECM, Population data by buildings, based on the 2021 register. Own elaboration.

At the time of study, the Community of Madrid had a population of 6,750,336, of whom almost half (3,280,782) resided within the municipality of Madrid. The remaining population was distributed mainly in proximate metropolitan rings on the margins of the main highways (A1 to A6): to the north, in Alcobendas, San Sebastián de los Reyes, Colmenar and Tres Cantos; to the east, in Coslada, San Fernando de Henares, Torrejón de Ardoz, and Alcalá de Henares; and to the south, in Móstoles, Leganés, Getafe, Fuenlabrada, and Parla (Figure 1).

3.2. Data and analytical procedures

This study models the relationship between the incidence of COVID-19 and five independent socio-demographic and geographic variables. The dataset used is the one that offers the highest spatial and temporal resolution, and which is also openly available and supported by an official body. This dataset contains cases reported by Basic Health Zones (BHZs) – the unit of analysis – subsequently reviewed and adjusted, based on health records. The incidence of COVID-19 is expressed in the number of diagnosed cases over periods of 14 days, aggregated in BHZs (Comunidad de Madrid, n.d.). This two-week-period approach modulates the daily variance of the disease and allows analysis of more stable aspects (Hierro and Maza, 2023) – which is the objective of this study.

To analyze the factors that influenced the incidence of COVID-19, five explanatory variables represent distinct dimensions that determine health – the social and environmental conditions under which a population lives and that influence its state of health (Artiga and Hinto, 2018). The biological variable is the *average age of the population* (in years), which, according to results from other studies, is expected to be positively related to incidence of the pandemic. The socio-economic variables are: the *average net income per person and year* (log; euros), reflecting the levels of wealth and economic stability and negatively associated with incidence of COVID-19; the *proportion of the population that is Spanish* (immigration), an indicator of social environment also negatively related to incidence; and the *average size of household*, an indicator of demographic structure that is positively associated with incidence. The geographical variable is *population density* (inhabitants/hectare), which measures the degree of agglomeration of the built space. The data for all variables derive from the Household Income Distribution Atlas (National Institute of Statistics, n.d.) and refer to 01 January 2020 and to census sections.

These variables were selected due to their significant relationship with incidence of the pandemic as well as availability of data, whether original or calculated, at BHZ resolution (Table 2). The maximum correlation between the variables is r = 0.52, observed between income and the proportion of the population that is Spanish.

The variables *population under 18 years of age* and *population over 65 years of age* have been excluded from the model due to their strong correlation with the *average age of the population* and *proportion of single-parent households*, as well as

Table 2. Parameters of the distributions of the variables

Variable	Average	Minimum	Maximum	Standard deviation
Population 2020	23,706	2,699	66,281	10,175
Population density (hab./ha)	139	0.1	568	137
Income 2020	14,887	7,173	30,694	4,643
Average age (years)	43	32	50	3
Household (persons)	3	2	3	0.3
Spanish population (%)	86	63	96	6

Source: Own elaboration.

their high correlation with the *average size of household*. Likewise, variables on land-coverage type (derived from Corine Land Cover and the Spanish Land Occupation Information System, or SIOSE) and building density have been excluded due to their weak relationship with the incidence rates of COVID-19 (Barros-Guerton and Ezquiaga-Domínguez, 2023).

The following transformations and analytical procedures have been applied to this basic information:

First, the rate of incidence of COVID-19 in the second period of epidemic (SPE) in each BHZ has been estimated using the spatial empirical Bayesian smoothing technique (Anselin, 2020) to reduce the instability of variance of the raw rates. The procedure combines spatial smoothing with Bayesian estimation.

The Bayesian empirical estimate (ESB) of the values of *a priori* distribution is obtained from the actual data. The smoothed rate is expressed as a weighted mean of the gross rate (r) and the *a priori* estimate (θ) (Anselin, 2020; Souris, 2019: 108). The risk estimates for a location i is expressed as:

$$\pi_i^{EB} = w_i r_i + (1 - w_i)\,\theta;\ w_i = \sigma^2 / (\sigma^2 + \mu/P_i)$$

where P_i is the base or at-risk population of area i, and μ and σ are the mean and variance. The value μ is the general mean rate of the area of study:

$$TFG = \sum_{i=1}^{i=n} O_i \Big/ \sum_{i=1}^{i=n} P_i ,$$

where O_i represents the observed cases of area i and P_i is the base population of area i. The variance is estimated as follows:

$$\sigma^2 = \frac{\sum_{i=1}^{i=n} P_i\,(r_i - \mu)^2}{\sum_{i=1}^{i=n} P_i} - \frac{\mu}{\sum_{i=1}^{i=n} \frac{P_i}{n}}$$

Spatial smoothing replaces the overall reference rate with specific rates calculated for each individual observation around a spatial window. The reference mean for each location i is calculated as follows:

$$\mu_i = {\sum_j w_{ij} O_j} \Big/ {\sum_j w_{ij} P_j}$$

where w_{ij} are spatial weights (0, 1), $w_{ij} = 1$, O_j: values observed in j

To estimate the variance, the above equation is used, but with rates and populations replaced by their local equivalents.

$$\sigma_i^2 = \frac{\sum_j w_{ij}\left[P_j (r_i - \mu_i)^2\right]}{\sum_j w_{ij} P_j} - \frac{\mu_i}{\sum_j w_{ij} P_i (k_i + +1)}$$

where k_i is the number of neighboring locations *i*. In this analysis, a 'neighborhood' has been delimited by means of the spatial contiguity criterion by borders and vertices ('queen contiguity'), so that a BHZ neighboring the target area identified by this method receives a weight of 1, and if not, of 0.

Afterwards, the weighted mean values of the socio-economic variables in the BHZ were calculated and the variables categorized according to natural intervals (for which the mean rates were calculated). The values for levels of income (measured in euros/person/year) were classified into five groups: up to 10,000; from 10,001 to 15,000; from 15,001 to 20,000; from 20,001 to 25,000; and more than 25,000. These income values were restated in logarithms to normalize their distribution. Also grouped into four intervals were: the population densities in inhabitants per hectare (less than 100, between 100 and 200, between 200 and 300, and more than 300); the mean age of the population (under 35 years of age, from 35 to 40, from 40 to 45, and over 45); and the proportion of Spanish population (less than 70%, between 70 and 80%, between 80 and 90%, and more than 90%). The variable for average household size was categorized into three levels: fewer than two people, between two and three people, and more than three people.

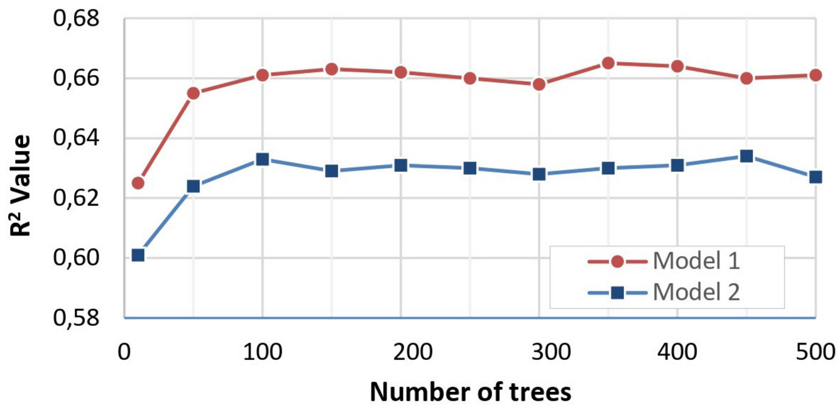
Finally, the relationship between the incidence rates (dependent variable) and the socio-economic variables was modeled using the random forest (RF) regression method, a relatively robust technique that reduces the overfit produced by correlation between variables. This method demands fewer requirements from the data than ordinary regression (Breiman, 2001) and has greater capacity to detect the complex and non-linear interactions of COVID-19 with environmental variables (Molnar, 2019; Zhang et al., 2022). The RF model has been applied in two versions:

— Model 1, with five predictor variables: *mean net income per person and year* (in euros), *population density* (inhabitants/hectare), *proportion of Spanish population*, *mean age of population,* and *average household size.*
— Model 2, with three predictor variables (the first three variables used in Model 1).

The sensitivity analysis was carried out with decision trees without cuts to their growth and with selection of the number of variables in each partition of √p (p: number of attributes). As Figure 2 shows, from 100 trees upwards, the value of the R^2 coefficient is very stable in both models, as is the mean square error (MSE), which remains constant at 0.006. In this analysis, the rates of incidence were estimated with 150 trees and five variables.

SHAP values (SHapley Additive exPlanations) are a cooperative game-theory-based technique used to increase the transparency and interpretability of machine-learning models. SHAP values measure the impact of each variable on the model's predictions, both globally and locally. In this research, we use SHAP values to estimate the influence of each variable on the estimated

Figure 2. Variability in response explained by the explanatory variables



Sources: Community of Madrid (n.d.). INE (n.d.). Own elaboration.

rates of COVID-19 in each BHZ. The SHAP values and their distribution permit the identification of the most relevant variables as well as the direction of their effect, both in the regression model as a whole and in each BHZ (Molnar, 2019). The SHAP values of each variable in each BHZ have been grouped using a hierarchical clustering procedure to identify sets of BHZs whose rates of incidence are explained by similar combinations of SHAP values.

The rates of incidence have been calculated with the *Geoda* program (Anselin, 2021); the regression model has been carried out with the *Orange* program (Demsar et al., 2013) The cartography has been prepared with the *ArcGISPro* (ESRI) program.

This study encountered limitations relating to the characteristics of the original data and its aggregation into spatial units. On one hand, the spatial resolution (Basic Health Zones) and temporal resolution (cumulative cases in 14-day periods during the Public Health Emergency) employed do not allow for an appreciation of the potential relationship between the incidence of COVID-19 and the average age and household size, a relationship that has been observed in other locations. Furthermore, the aggregation process results in the original information gathered from discrete and mobile individuals being transformed into static and continuous data within each spatial unit. This process reduces variability but also has advantages, such as the possibility of calculating per capita indices (rates) and densities. On the other hand, the variables included in the model explain close to 70% of the variation in incidence, and this value could increase if information were available that related living conditions and routines with modes of transmission of the pandemic (overcrowding, activities, others).

4. Results

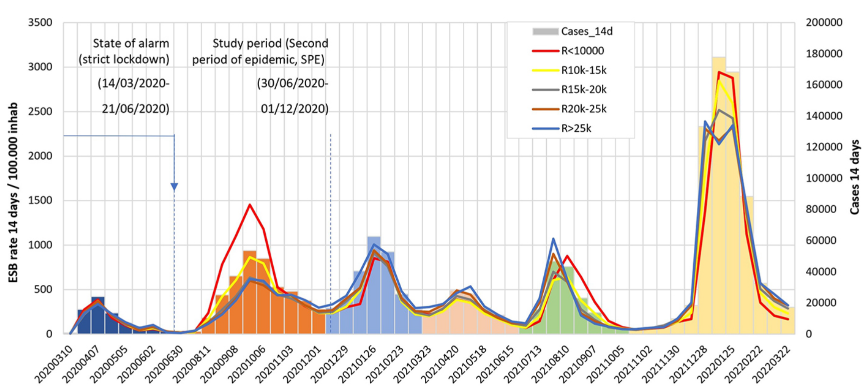
4.1. Significant inequalities in rates of incidence between groups for each variable in the second period of epidemic

During the second period of epidemic (SPE) of COVID-19, significant differences were observed in the rates of incidence between the groups defined by each variable analyzed. In contrast, in other periods of the pandemic, the rates were similar between different groups.

In the case of income, during the SPE the rates of incidence were significantly higher for the lowest income group. Differences are also found at the start of the fifth 14-day period and at the peak of the sixth, but not during other phases (Figure 3). The variables *proportion of Spanish population*, *population density* and *mean age of the population* all show a similar pattern and no significant differences outside the SPE. In the variable *household size*, those households with fewer than two people were found to have higher rates than the others during the fifth period of epidemic.

As a general feature, it should be noted that the number of positive cases and the mean rates of incidence of COVID-19 were not higher in the SPE than in other epidemic periods (except the fourth), but that differences in rates between levels of the variables (especially income) are more pronounced.

Figure 3. Diagnosed cases of COVID-19 and rates of incidence by 14-day periods and by income groups in the Community of Madrid (March 2020 to March 2022)



R: income; k: thousands of euros; colors: epidemic periods

Sources: Community of Madrid (n.d.). INE (n.d.). Own elaboration.

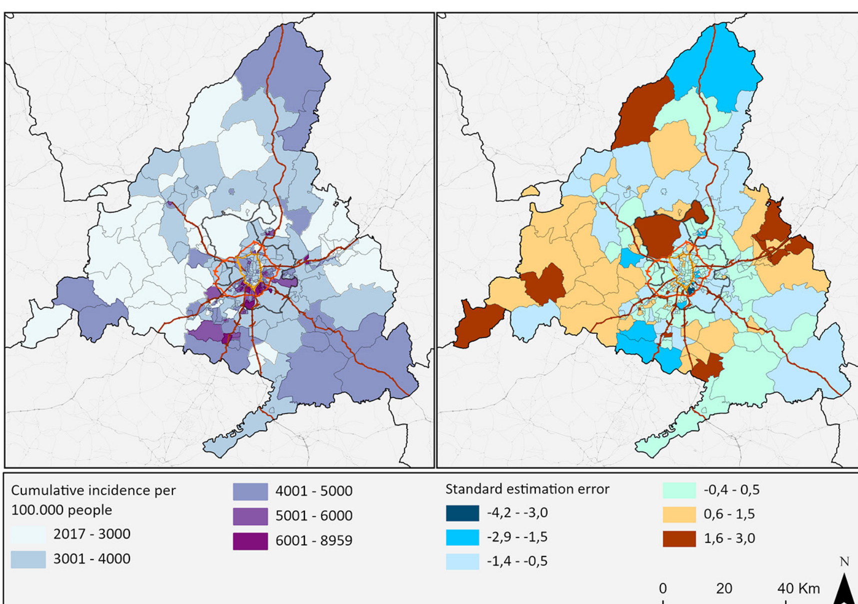
4.2. Neighborhood mean income as the main explanatory variable for the variation in rates of incidence

The applied regression model includes five predictor variables of the rates of incidence during the SPE. The explanatory capacity of each predictor variable has been estimated by adding their (absolute) SHAP values.

During the SPE, the highest rates of incidence (between 6,000 and 9,000 confirmed cases per 100,000 inhabitants) were located in a strip to the south of the city of Madrid between the M30 and M40 beltways on both sides of the N4 highway (in San Diego, San Cristóbal, Pozo del Tío Raimundo, Peña Prieta). The next highest values were scattered throughout the Autonomous Community (in Parla, parts of Alcorcón, Alcobendas, Torrejón de Ardoz). The next group of high rates (between 5,000 and 6,000 cases) were located in the northern triangle (Montejo de la Sierra, Buitrago de Lozoya, Navacerrada) and in various zones in the south of the Community (San Martín de Valdeiglesias, Pelayos de la Presa, Villa del Prado, Pinto, Torrejón de la Calzada, Chinchón, Colmenar de Oreja, Fuentidueña del Tajo). The average rates and lowest rates (fewer than 3,000 confirmed cases per 100,000 inhabitants) formed various groupings, the largest of which was in the southeast of the Community of Madrid (between San Lorenzo del Escorial, Navas del Rey and Navalcarnero) (Figure 4a).

The results show that *variation in mean income* is the main explanatory variable for the values representing the rate of incidence, followed at great distance by *population density*, the *proportion of Spanish population*, and (with considerably less weight) the *mean age of the population* and *average size of household* (Table 1).

Figure 4. *a*) left: accumulated incidence rates in the SPE by Basic Health Zones (per 100,000 inhabitants); *b*) right: estimated incidence rates in the SPE by Basic Health Zones (per 100,000 inhabitants)



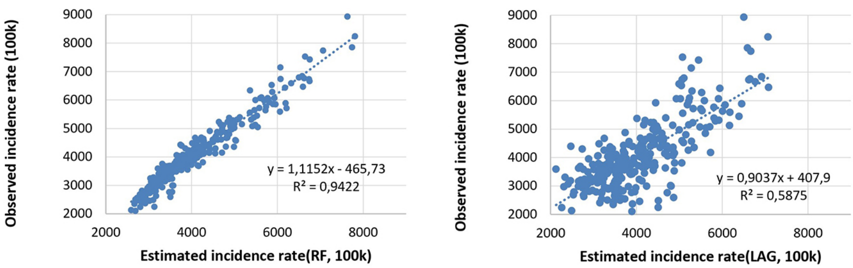
Sources: Community of Madrid (n.d.). INE (n.d.). Own elaboration.

Table 3. Most significant variables of the RF model

Predictor variable	SHAP values (absolute)
Mean net income per person per year (€)	0.89
Population density (inhabitants/hectare)	0.40
Proportion of Spanish population	0.34
Mean age of the population	0.14
Average household size	0.10

Sources: Community of Madrid (n.d.). INE (n.d.). Own elaboration.

Figure 5. Correlation between the observed rates of incidence and those estimated by the random forest (RF) method (left) and by spatial regression (LAG) (right)



Sources: Community of Madrid (n.d.). INE (n.d.). Own elaboration.

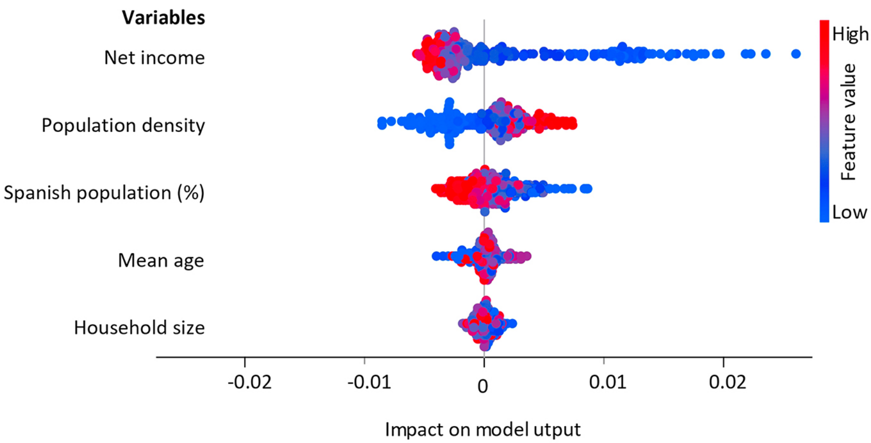
As mentioned, the correlation between the observed and estimated rates is very high (R2 = 0.94) while the standard error of the estimate is low (0.003100, or 310 per 100,000 inhabitants) and its distribution is normal. The standard error of the estimated rates tends to be larger when the rates are further from the mean, meaning that the predictions are less accurate for BHZs with extreme rates of incidence, whether due to underestimation or overestimation. In contrast, predictions are more accurate for a BHZ with incidence rates close to the mean (Figure 5, left). This result surpasses that obtained by other methods, such as multiple linear regression or spatial regression (Figure 5, right), as in the study by Zhang et al. (2022).

4.3. The unequal contribution of variables to rates of incidence in BHZs

The SHAP values measure the influence of each variable on the estimated rates of incidence of COVID-19 in each BHZ. Positive SHAP values mean that the variable increases rates of incidence, while negative SHAP values indicate a reduction in rates.

Figure 6 shows the distribution of the SHAP values for each explanatory variable; variables are ordered from highest to lowest contribution to the prediction, and the horizontal axis represents the variation of the SHAP values in each BHZ. The BHZs are symbolized by points colored according to the value

Figure 6. Impacts of the predictor variables on the rates of incidence of COVID-19 in the BHZs (SHAP values x 1,000

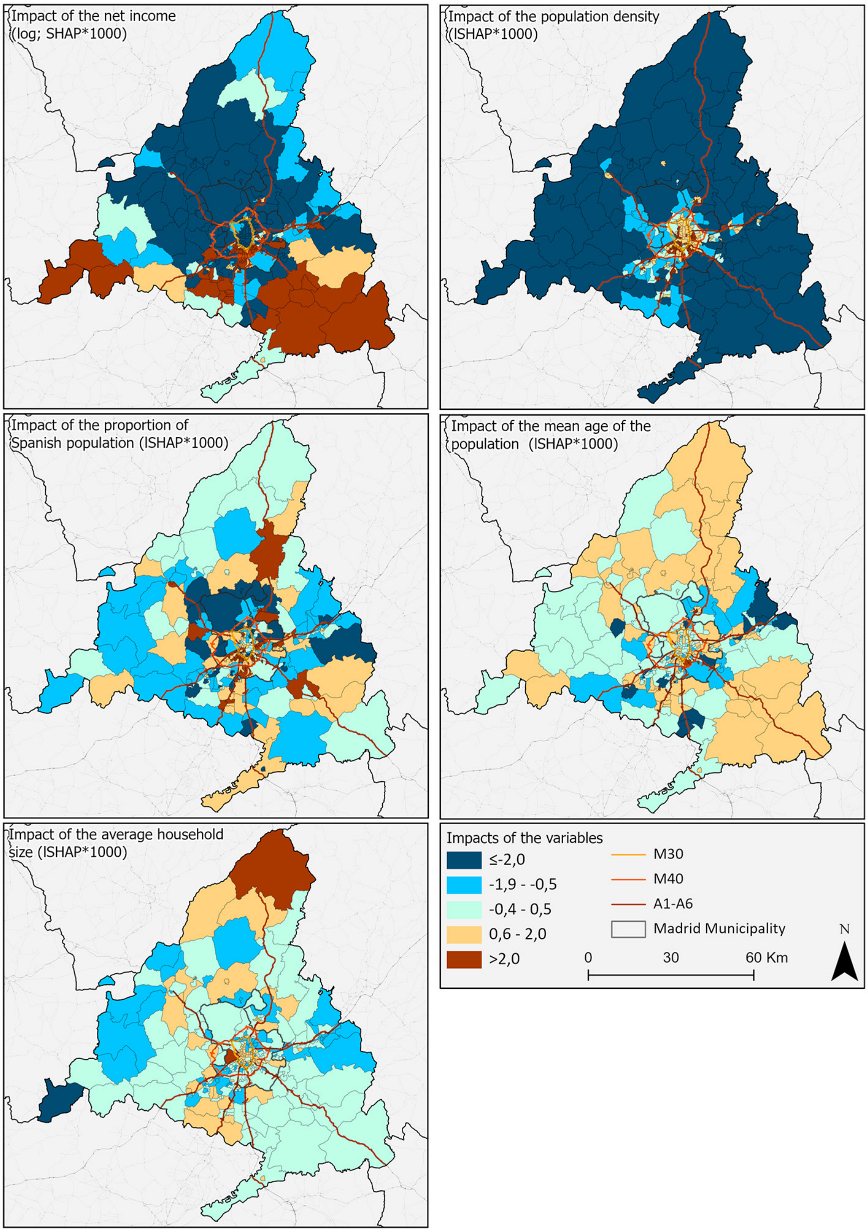


Sources: Community of Madrid (n.d.). INE (n.d.). Own elaboration.

of the variable. *Income* proves to be the variable that most influenced rates of incidence, but in unequal ways, depending on the level of increase or decrease: that is, BHZs with low income levels (such as Martínez de la Riva, San Diego, Pozo del Tío Raimundo, San Cristóbal and Entrevías) showed much higher rates of incidence than others, while BHZs with high incomes had slightly lower rates of incidence (similar to one another, as indicated by the clustering of red dots). *Population density* had a smaller effect than income level, and is symmetrically distributed around the value SHAP=0 (neutral influence): high densities increased rates of incidence while low densities reduced them with the same intensity. The *proportion of Spanish population* had a positive impact on the rates of some BHZs (Peña Prieta, Las Calesas and Martínez de la Riva). The *mean age of the population* had a limited, ambivalent influence on rates of incidence, with high and low values showing both positive and negative effects. Finally, the *size of household* showed scant relevance in the predictions, with very small influence on the BHZs.

The five maps in Figure 7 represent the impacts of each variable on the BHZs (with SHAP values multiplied by 1,000 to facilitate visualization). The variables for *income* and *population density* here present more defined spatial patterns than the others. The strongest impacts by neighborhood mean income on the increase in rates of incidence corresponded to several zones concentrated between the M30 and M40 beltways to the south of the city of Madrid (Abrantes, Entrevías, Alcocer) as well as a few located to the south of the Community of Madrid (Villa del Prado, Humanes, San Blas, Las Américas, Colmenar de Oreja). The zones in which higher incomes contributed to reductions in rates of incidence are near the center (most of the BHZs inside the M30 beltway: Pozuelo de Alarcón, Las Rozas, La Moraleja), to the north-northeast of the city

Figure 7. Location of the impact of the predictor variables on incidence rates of COVID-19 in the Basic Health Zones (SHAP values x 1000)



Sources: Community of Madrid (n.d.). INE (n.d.). Own elaboration.

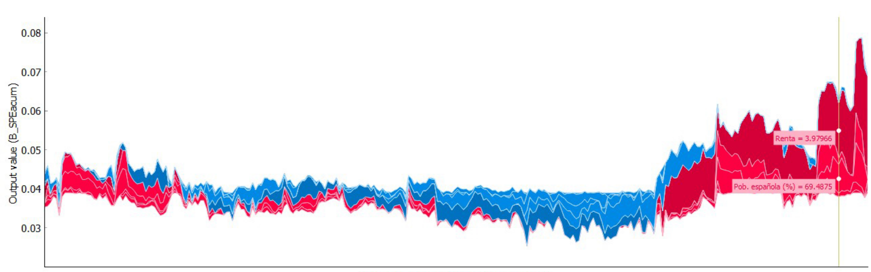
(Alameda de Osuna), and scattered throughout the Community (Soto del Real, El Escorial, Virgen del Val). The spatial pattern of impact according to density on the rates of incidence of COVID-19 shows a clear center/periphery configuration: the greatest increases in rates of incidence caused by density were located in the center of the city of Madrid; these rates decreased towards the city outskirts, to the point of changing their sign of influence. Other variables influencing the rates of incidence show a less regular, more dispersed spatial distribution, except the *mean age of the population*, which reveals a certain grouping of areas with similar impacts in the north, center and southeast of the Community.

4.4. Typologies of the combined impact of variables on rates of incidence

SHAP values are combined differently in each BHZ. Figure 8 represents the combination of each BHZ by means of a column on the *x*-axis, while the *y*-axis symbolizes the stacked contribution of each variable to the rate estimation: values that increase the prediction (rates of incidence) are in red; values that decrease it are in blue; red and blue and their lighter degrees indicate variable income. The group on the right stands out, with values of the variables that increase their incidence rates of COVID-19.

To identify BHZs exhibiting similar structures of SHAP values, we used a hierarchical method of grouping that produced six groups (distinct sets of BHZs whose rates of incidence are explained by similar combinations of SHAP values). The spatial pattern presents a center/periphery configuration with three differentiated zones: *a*) a nucleus at the center of the city of Madrid, where rates of incidence (whether high or low) depend mainly on income levels; *b*) a mosaic of groupings around that nucleus and following transport corridors, where rates of incidence are explained by combinations of variables; and *c*) an outer extreme where the most relevant explanatory variable is population density (Figure 9).

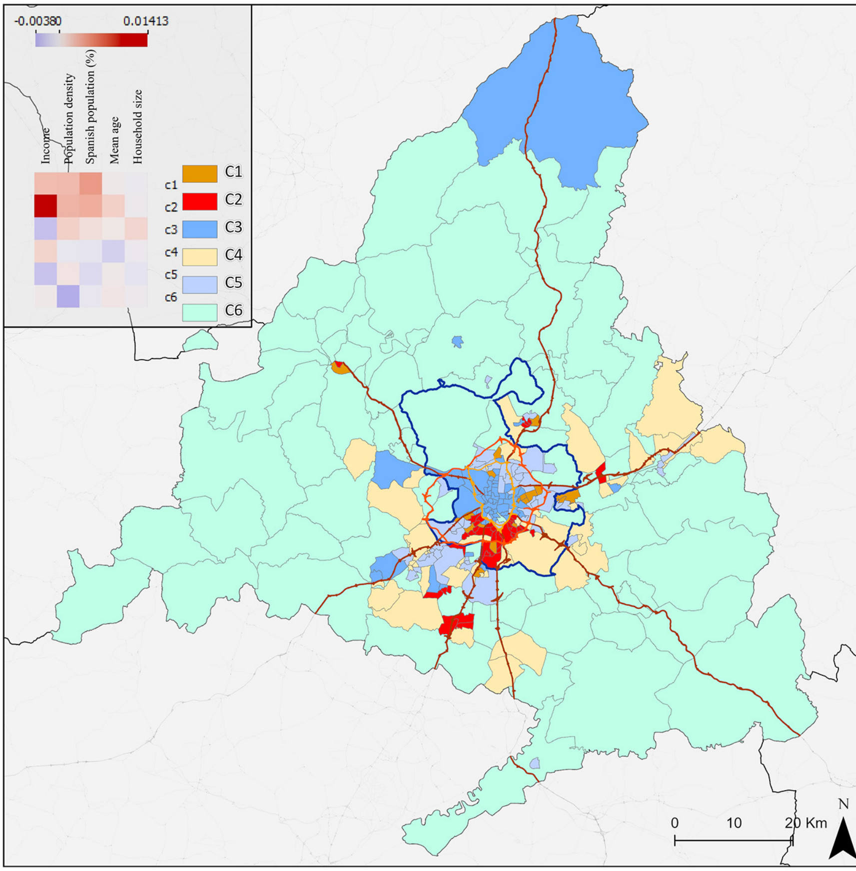
Figure 8. Structure of SHAP values in the BHZs, ordered according to the similarity of factors that explain their rates of incidence (x-axis)



Red: push prediction probabilities towards 1; blue: push probabilities against the prediction

Sources: Community of Madrid (n.d.). INE (n.d.). Own elaboration.

Figure 9. Typology of the combined impacts of the variables on rates of incidence



Sources: Community of Madrid (n.d.). INE (n.d.). Own elaboration.

— *Group 1*. In this group of BHZs, the factors that contributed to raised rates of incidence were a lower proportion of Spaniards in a population, a lower level of income, and a high population density. These are found around the center of the city of Madrid (Campamento, Lucero, Los Ángeles, Alegría) as well as scattered through other areas (Collado Villalba Estación; Alcobendas V Centenario).

— *Group 2*. The BHZs in this group showed the strongest increase in rates of incidence associated with low income levels, while a low proportion of Spanish population and high population density operated in the same direction. These zones are concentrated between the M30 and M40 beltways south of the city of Madrid (Fuenlabrada, Parla).

— *Group 3*. In contrast to the previous groups, this group was distinguished by the negative impact of mean neighborhood income on rates of incidence (stronger than in the other groups). The BHZs of this group form a compact area in the center of Madrid extending to the northwest (along the A-6 axis) (Aravaca, Majadahonda, Las Rozas) and scattered elsewhere (Colmenar Viejo Norte, El Soto).
— *Group 4*. This group includes BHZs in which all the variables showed a mild influence on rates of incidence, with *mean neighborhood income* slightly negative and *mean age* slightly positive. These are grouped in sectors on the outskirts of Madrid in the vicinity of the M50 (Boadilla del Monte, Miguel Servet, Parque Oeste, Butarque, Ensanche-Vallecas, Valderribas) and along the Henares corridor (Meco, Miguel de Cervantes).
— *Group 5*. In these BHZs, neighborhood mean income exerted a moderately negative impact on rates of incidence, while the other variables showed minimal influence. These are found in diverse clusters on the northern and northeastern periphery of the city of Madrid (Barajas, Mar Báltico, Mirasierra, Fuentelarreina, Peñagrande) as well as to the southwest (Alcorcón, Leganés, Móstoles).
— *Group 6*. This group consists of BHZs in which density exerted the most prominent negative effect on rates of incidence, while other variables showed no significant effects. These occupy most of the Autonomous Community, forming a border around the city of Madrid, along with a BHZ at the extreme north (Buitrago de Lozoya).

5. Discussion and conclusions

This paper examines the relationship between rates of incidence of COVID-19 and several socio-economic variables in the Autonomous Community of Madrid. It has been found that income levels and other variables were significantly associated with variations in rates of incidence during the second period of epidemic (SPE); this association was not detected in the other periods considered.

One possible cause for the differentiated behavior observed during the SPE may have been the scarcity of testing conducted during the previous (first) wave of the pandemic, as tests were at that time reserved for the most serious cases and at-risk populations. In June 2020, in a seroprevalence study by the Carlos III Health Institute, it was found that over 10% of the population of the overall Autonomous Community of Madrid showed signs of having suffered a COVID-19 infection (ISCIII, 2020); at the same time, the official number of positive tests according to the Community of Madrid was close to 70,000 (≈1% of the population). This coincides with the results of Florida et al. (2021), in which no association was found between the level of income by district and the incidence of COVID-19 on 10 June 2020 in the city of Madrid, unlike in other cities analyzed, such as New York. However, this situation changed during the second period of the epidemic (from 22 June 2020 to 06 December 2020), following the state of emergency, when a negative correlation was observed between income and the incidence of COVID-19.

In addition, the general lockdown of the population may have masked differences in incidence of COVID-19 between social groups according to income levels or other factors. In the stages following the SPE, vaccinations and seroprevalence testing for a large portion of the population were able to modify the conditions of spread and reduce the impact of socio-economic factors (Salmon et al., 2021). It has further been found that the factors affecting the spread of the pandemic were subject to changes in significance, even within a single phase of epidemic (Maza and Hierro, 2021).

The results obtained here generally coincide with those of other studies, but they also reveal new aspects of the non-uniformity of the effect of socio-economic factors on the incidence of COVID-19 during the SPE.

On the one hand, this study confirms *income* as having been the factor with the greatest influence on incidence of the pandemic, followed by *population density* and the *proportion of Spanish population*. During the SPE, the pandemic affected the lowest-income areas of the Community of Madrid most, as it did in other territories (see Section 2. Bibliographic review, above). Nevertheless, inequalities between income levels were not as marked and persistent as in other, more segregated, metropolitan areas such as Santiago, Chile (Mena et al., 2021). In line with the results of studies by De Cos et al. (2022) and Niu et al. (2020), it is further observed that the incidence of COVID-19 was lower in BHZs with low population density, probably because greater distances between people reduces the risk of infection, as demonstrated in other studies mentioned in the bibliography. In the particular case of the Community of Madrid, low density is linked to lower economic/work and leisure activity (particularly around cultural activities and hospitality), which contributes to this mitigating effect. Likewise, a negative relationship is found between areas with a lower percentage of Spanish population – used as an indicator of foreign immigration – and incidence of the disease, and this might be ascribed to cultural differences in living habits and forms of socialization (Hierro and Maza, 2023). Finally, the model shows that the *mean age of the population* and the *average size of household* had insignificant effects on our predictions, being appreciable only in a few BHZs; moreover, the direction of this effect is less than clear, varying across BHZs with higher or lower average ages and household sizes.

Our analysis further reveals that not only did those factors under study carry differing weights in terms of incidence, but also that their impacts were not uniform, thus generating a complex spatial pattern. It was found that income had unequal influence on incidence of the disease: while low-income areas showed the highest rates of incidence, rates were not significantly reduced in high-income BHZs. This suggests that, once infection begins in an area, high incomes do not contain the spread, but rather low incomes exacerbate it. The distinct living conditions and routines relating to income levels (such as overcrowding, working conditions and mobility) can explain these differences. Population density is the most determining factor in limiting the incidence of the disease, and so in BHZs with low densities, rates were reduced more intensely than in high-income areas, but high densities drove rises in incidence to a lesser degree than low incomes. The effect of the proportion of the Spanish population was reflected mainly in the rise in incidence of COVID-19 in certain BHZs with the highest percentages of immigrant populations.

Between September 2020 and May 2021, the Community of Madrid applied selective measures of perimeter lockdown by BHZ (and in some cases by municipality) to frustrate the spread of COVID-19. According to multiple studies, those measures had no effect on control of the pandemic (De Miguel Arribas et al., 2023; García-García et al., 2022; Fontán-Vela et al., 2021); the less pronounced inequalities in incidence observed throughout the Community of Madrid might be explained by high seroprevalence, the public health system, and generalized non-pharmacological measures (restriction of capacity and activities, use of masks, follow-up of positive cases, etc.). It should be noted that in other territories such as Santiago de Chile, perimeter lockdowns did indeed have positive effects on pandemic control, depending on the circumstances of application (Li et al., 2022).

Additionally, the various factors examined were found to have been related to one another, generating a complex territorial configuration. The (positive or negative) effects of income in certain BHZs were more or less intense depending on the state of other factors. For example, when the proportion of the Spanish population was greater than 90%, the effect of income on the decrease in incidence was negligible, but when that percentage was less than 70%, the incidence decreased significantly with an increase in income.

Taken together, our results reveal that the pandemic had a greater impact in areas with lower incomes and higher population densities; and yet the results also reveal a non-uniformity of influence. High income is found to have reduced incidence, but not as much as low density or a low proportion of non-Spanish populations. Therefore, it can be said that measures for control of future epidemics should focus on the most vulnerable as well as the most population-dense areas.

Until now, the random forest regression method has not been widely used to model the incidence of COVID-19 and its relationship with other factors. Indeed, we have found no other research that applies this method to the study of territories in Spain. The present study shows that this method can offer more accurate predictions than the usual approaches, including those that consider spatial autocorrelation. Moreover, this method allows evaluation of the direction and impact of different variables in each particular location; Therefore, we believe it advisable that more studies be undertaken that apply this method, in order to further explore the relationships between socio-economic and environmental factors and the incidence of COVID-19.

Bibliographical references

Almendra, Ricardo; Santana, Paula and Costa, Claudia (2021). “Spatial inequalities of COVID-19 incidence and associated socioeconomic risk factors in Portugal”. *Boletín de la Asociación de Geógrafos Españoles*, 91. <https://doi.org/10.21138/bage.3160>

Amdaoud, Mounir; Arcuri, Giuseppe; Levratto, Nadine; Succurro, Marianna and Constanzo, Damiana (2020). “Geography of COVID-19 outbreak and first policy answers in European regions and cities”. *Halshs-03046489*. Retrieved from <https://halshs.archives-ouvertes.fr/halshs-03046489>

Amengual-Moreno, Miquel; Calafat-Caules, Marina; Carot, Aina; Rosa Correia, Ana Rita; Río-Bergé, Claudia; Rovira Plujà, Jana; Valenzuela Pascual, Clàudia and Ventura-Gabarró, Cèlia (2020). “Social determinants of the incidence of Covid-19 in Barcelona: a preliminary ecological study using public data”. *Revista española de salud pública*, 94. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/32935664>

Anselin, Luc (2020). *Documentation/GeoDa on Github/GeoDa Workbook*. Retrieved from <https://geodacenter.github.io/documentation.html>

— (2021). *GeoDa (Tm)* (1.20.). Retrieved from <https://geodacenter.github.io/>

Arauzo-Carod, Josep-Maria; Domènech, Antoni and Gutiérrez, Aaron (2021). “Do local characteristics act in a similar way for the first two waves of COVID-19? Analysis at intraurban level in Barcelona”. *Journal of Public Health*, 43 (3), 455-461. <https://doi.org/10.1093/pubmed/fdaa238>

Arroyo-Menéndez, Millán; Barañano-Cid, Margarita and Uceda-Navas, Pedro (2022). “¿Desiguales en la smart city? Segregación espacial y desigualdades digitales en Madrid/Unequal in the Smart City? Spatial Segregation and Digital Inequalities in Madrid”. *Revista Española de Investigaciones Sociológicas*. <https://doi.org/10.5477/cis/reis.180.19>

Artiga, Samantha and Hinto, Elizabeth (2018). *Beyond health care: the role of social determinants in promoting health and health equity*. KFF.org. Retrieved from <https://www.kff.org/racial-equity-and-health-policy/issue-brief/beyond-health-care-the-role-of-social-determinants-in-promoting-health-and-health-equity/>

Baena-Díez, Jose Miguel; Barroso, María; Cordeiro-Coelho, Sara Isabel; Díaz, Jorge Luis and Grau, María (2020). “Impact of COVID-19 outbreak by income: hitting hardest the most deprived”. *Journal of Public Health*, 42 (4), 698-703. <https://doi.org/10.1093/pubmed/fdaa136>

Barros-Guerton, Javier and Ezquiaga-Domínguez, José María (2023). “Caer y levantarse: como la ciudad de Madrid ha afrontado el COVID19 de acuerdo con los datos de libre reutilización”. *Ciudad y Territorio Estudios Territoriales*, 55 (218), 1033-1054. <https://doi.org/10.37230/CyTET.2023.218.3>

Benita, Francisco and Gasca-Sanchez, Francisco (2021). “The main factors influencing COVID-19 spread and deaths in Mexico: A comparison between phases I and II”. *Applied Geography*, 134. <https://doi.org/10.1016/J.APGEOG.2021.102523>

Breiman, Leo (2001). “Random Forests”. *Machine Learning*, 45, 5-32. <https://doi.org/10.1023/A:1010933404324>

Cestari, Virna Ribeiro Feitosa; Florêncio Sampaio, Raquel; Bezerra Sousa, George Jó; Santos Garces, Tiago; Araújo Maranhão, Thatiana; Ribeiro Castro, Révia; Luana Cordeiro, Ibiapina; Ventura Damasceno, Lara Lídia; Vera

Lucia, Mendes de Paula Pessoa; Duarte Pereira, Maria Lúcia and Magalhães Moreira, Thereza Maria (2021). "Social vulnerability and COVID-19 incidence in a Brazilian metropolis". *Ciencia & saude coletiva*, 26 (3), 1023-1033. <https://doi.org/10.1590/1413-81232021263.42372020>

Comunidad de Madrid (n. d.). *Covid 19 -TIA Zonas Básicas de Salud - Conjuntos de datos - Datos Abiertos Comunidad de Madrid*. Retrieved from <https://datos.comunidad.madrid/catalogo/dataset/covid19_tia_zonas_basicas_salud> [Accessed: 07 February 2022]

Cordes, Jack and Castro, Maria (2020). "Spatial analysis of COVID-19 clusters and contextual factors in New York City". *Spatial and Spatio-temporal Epidemiology*, 34, 100355. <https://doi.org/10.1016/j.sste.2020.100355>

De Cos, Olga; Castillo-Salcines, Valentín and Cantarero-Prieto, David (2022). "A geographical information system model to define COVID-19 problem areas with an analysis in the socio-economic context at the regional scale in the North of Spain". *Geospatial Health*, 17. <https://doi.org/10.4081/gh.2022.1067>

De Mattos, Ricardo; Russo, Rafael; Neto, Mercedes; Gomes-Depret, Davi; Costa Gil, Adriana; Silva Fonseca, Mary Hellem and Souza-Santos, Reinaldo (2020). "Effect of income on the cumulative incidence of COVID-19: an ecological study". *Revista Latino-Americana de Enfermagem*, 28. <https://doi.org/10.1590/1518-8345.4475.3344>

De Miguel Arribas, Alfonso; Aleta, Alberto and Moreno, Yamir (2023). "Assessing the effectiveness of perimeter lockdowns as a response to epidemics at the urban scale". *Scientific Reports*, 13 (1), 4474. <https://doi.org/10.1038/s41598-023-31614-8>

Demsar, Janez; Curk, Tomaž; Erjavez, Aleš; Group, Črt; Hocevar, Tomaž; Milutinovič, Mitar; Možina, Martin; Polajnar, Matija; Toplak, Marko; Starič, Anže; Štajdohar, Miha; Umek, Lan; Žagar, Lan; Žbontar, Jure; Žitnik, Marinka and Zupan, Blaž (2013). "Orange: Data Mining Toolbox in Python". *Journal of Machine Learning Research*, 14 (Aug.), 2349-2353. Retrieved from <http://jmlr.org/papers/v14/demsar13a.html>

Eurostat (n. d.). *Eurostat Metropolitan Region Database*. Retrieved from <https://ec.europa.eu/eurostat/web/metropolitan-regions/database> [Accessed: 15 November 2023]

Florida, Richard and Mellander, Charlotta (2022). "The geography of COVID-19 in Sweden". *The Annals of Regional Science*, 68, 125-150. <https://doi.org/10.1007/s00168-021-01071-0>

Florida, Richard; Rodríguez-Pose, Andrés and Storper, Michael (2021). "Cities in a post-COVID world". *Urban Studies*. <https://doi.org/10.1177/00420980211018072>

Fontán-Vela, Mario; Gullón, Pedro and Padilla-Bernáldez, Javier (2021). "Selective perimeter lockdowns in Madrid: a way to bend the COVID-19 curve?". *European Journal of Public Health*, 31 (5), 1102-1104. <https://doi.org/10.1093/eurpub/ckab061>

Franch-Pardo, Ivan; Desjardins, Michael; Barea-Navarro, Isabel and Cerdà, Artemi. (2021). "A review of GIS methodologies to analyze the dynamics of COVID-19 in the second half of 2020". *Transactions in GIS*, 25 (5), 2191-2239. <https://doi.org/10.1111/tgis.12792>

Franch-Pardo, Ivan; Napoletano, Brian; Rosete-Verges, Fernando and Billa, Lawal (2020). "Spatial analysis and GIS in the study of COVID-19. A review". *Science of The Total Environment*, 739, 140033. <https://doi.org/10.1016/j.scitotenv.2020.140033>

García-García, David; Herranz-Hernández, Rafael; Rojas-Benedicto, Ayelen; León-Gómez, Inmaculada; Larrauri, Amparo; Peñuelas, Marina; Guerrero-Vadillo, María; Ramis, Rebeca and Gómez-Barroso, Diana (2022). "Perimeter confinements of basic health zones and COVID-19 incidence in Madrid, Spain". *BMC Public Health*, 22 (1), 216. <https://doi.org/10.1186/s12889-022-12626-x>

Gaynor, Tia Sherèe and Wilson, Meghan (2020). "Social Vulnerability and Equity: The Disproportionate Impact of COVID". *Public Administration Review*, 80 (5), 832-838. <https://doi.org/10.1111/puar.13264>

González Pérez, Jesús and Piñeira Mantiñán, María José (2020). "La ciudad desigual en Palma (Mallorca): geografía del confinamiento durante la pandemia de la COVID-19". *Boletín de la Asociación de Geógrafos Españoles*, 87. <https://doi.org/10.21138/bage.2998>

Hamidi, Shima; Sabouri, Sadegh and Ewing, Reid (2020). "Does Density Aggravate the COVID-19 Pandemic?" *Journal of the American Planning Association*, 86 (4), 495-509. <https://doi.org/10.1080/01944363.2020.1777891>

Hierro, María and Maza, Alfonso (2023). "Spatial contagion during the first wave of the COVID-19 pandemic: Some lessons from the case of Madrid, Spain". *Regional Science Policy & Practice*, 15 (3), 474-492. <https://doi.org/10.1111/rsp3.12522>

Instituto de Salud Carlos III (ISCIII). *Informe n.º 173. Situación de COVID-19 en España*. Retrieved from <https://www.isciii.es/QueHacemos/Servicios/VigilanciaSaludPublicaRENAVE/EnfermedadesTransmisibles/Paginas/-COVID-19.-Informes-previos.aspx> [Accessed: 03 April 2023]

Instituto Nacional de Estadística (2021). *Estadística experimental. Atlas de la distribución de renta de los hogares*. Retrieved from <https://www.ine.es/experimental/atlas/experimental_atlas.htm> [Accessed: 24 July 2021]

Karaye, Ibraheem and Horney, Jennifer (2020). "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), 317-325. <https://doi.org/10.1016/j.amepre.2020.06.006>

Kling, Samuel (2020). "Is the City Itself the Problem? The Long History of Demonizing Urban Density". *Bloomberg City Lab*. Retrieved from <https://www.bloomberg.com/news/articles/2020-04-20/the-long-history-of-demonizing-urban-density>

Kolak, Marinya; Bhatt, Jay; Hong Park, Yoon; Padrón, Norma and Molefe, Ayrin (2020). "Quantification of Neighborhood-Level Social Determinants of Health in the Continental United States". *JAMA Network Open*, 3 (1), 1919928. <https://doi.org/10.1001/jamanetworkopen.2019.19928>

Konstantinoudis, Garyfallos; Cameletti, Michela; Gómez-Rubio, Virgilio; León Gómez, Inmaculada; Pirani, Mónica; Baio, Gianluca; Larrauri, Amparo; Riou, Julien; Egger, Matthias; Vineis, Paolo and Blangiardo, Marta (2022). "Regional excess mortality during the 2020 COVID-19 pandemic in five European countries". *Nature Communications*, 13 (1), 482. <https://doi.org/10.1038/s41467-022-28157-3>

Li, Yige; Undurraga, Eduardo and Zubizarreta, José (2022). "Effectiveness of Localized Lockdowns in the COVID-19 Pandemic". *American Journal of Epidemiology*, 191 (5), 812-824.
<https://doi.org/10.1093/aje/kwac008>

López-Gay, Antonio; Spijker, Jeroen; Cole, Helen; Marques, Antonio; Triguero-Mas, Margarita; Anguelovski, Isabelle; Marí-Dell'Olmo, Marc; Módenes, Juan; Álamo-Junquera, Dolores; López-Gallego, Fernando and Borrell, Carme (2022). "Sociodemographic determinants of intraurban variations in COVID-19 incidence: the case of Barcelona". *Journal of Epidemiology and Community Health*, 76 (1), 1-7.
<https://doi.org/10.1136/jech-2020-216325>

Marí-Dell'Olmo, Marc; Gotsens, Mercè; Pasarín, María Isabel; Rodríguez-Sanz, Maica; Artazcoz, Lucía; Garcia de Olalla, Patricia; Rius, Cristina and Borrell, Carme (2021). "Socioeconomic Inequalities in COVID-19 in a European Urban Area: Two Waves, Two Patterns". *International Journal of Environmental Research and Public Health*, 18 (3), 1256.
<https://doi.org/10.3390/ijerph18031256>

Maza, Adolfo and Hierro, María (2021). "Modelling changing patterns in the COVID-19 geographical distribution: Madrid's case". *Geographical Research*, 60 (2), 218-231.
<https://doi.org/10.1111/1745-5871.12521>

Mena, Gonzalo; Martinez, Pamela; Mahmud, Ayesha; Marquet, Pablo; Buckee, Caroline and Santillana, Mauricio (2021). "Socioeconomic status determines COVID-19 incidence and related mortality in Santiago, Chile". *Science*, 372 (6545).
<https://doi.org/10.1126/science.abg5298>

Mollalo, Abolfazl; Vahedi, Behzad and Rivera, Kiara M. (2020). "GIS-based spatial modeling of COVID-19 incidence rate in the continental United States". *Science of The Total Environment*, 728, (138884).
<https://doi.org/10.1016/j.scitotenv.2020.138884>

Molnar, Christoph (2019). *Interpretable machine learning. A Guide for Making Black Box Models Explainable*. Retrieved from <https://christophm.github.io/interpretable-ml-book/>

Musterd, Sako; Marcinczak, Szymon; van Ham, Maarten and Tammaru, Tiit (2017). "Socioeconomic segregation in European capital cities. Increasing separation between poor and rich". *Urban Geography*, 38 (7), 1062-1083.
<https://doi.org/10.1080/02723638.2016.1228371>

Niu, Xinyi; Yue, Yufeng; Zhou, Xingang and Zhang, Xiaohu (2020). "How Urban Factors Affect the Spatiotemporal Distribution of Infectious Diseases in Addition to Intercity Population Movement in China". *International Journal of Geo-Information*, 9 (11), 615.
<https://doi.org/10.3390/ijgi9110615>

OMS (n.d.). *WHO Coronavirus (COVID-19) Dashboard*. Retrieved from <https://covid19.who.int/table>

Raymundo, Carlos Eduardo; Cini Oliveira, Marcella; de Araújo Eleuterio, Tatiana; André, Suzana Rosa; Gonçalves da Silva, Marcele; da Silva Queiroz, Eny Regina and de Andrade Medronho, Roberto (2021). "Spatial analysis of COVID-19 incidence and the sociodemographic context in Brazil". *PLOS ONE*, 16 (3), e0247794.
<https://doi.org/10.1371/journal.pone.0247794>

Rodríguez-Pose, Andrés and Burlina, Chiara (2021). "Institutions and the uneven geography of the first wave of the COVID-19 pandemic". *Journal of Regional Science*, 61 (4), 728-752. <https://doi.org/10.1111/jors.12541>

Romero Starke, Karla; Reissig, David; Petereit-Haack, Gabriela; Schmauder, Stefanie; Nienhaus, Albert and Seidler, Andreas (2021). "The isolated effect of age on the risk of COVID-19 severe outcomes: a systematic review with meta-analysis". *BMJ Global Health*, 6 (12), e006434. <https://doi.org/10.1136/bmjgh-2021-006434>

Ruiz-Pérez, Maurici; Moragues, Alexandre; Seguí-Pons, Joana Maria; Muncunill, Josep; Pou Goyanes, Albert and Colom Fernández, Antoni (2023). "Geographical Distribution and Social Justice of the COVID-19 Pandemic: The Case of Palma (Balearic Islands)". *GeoHealth*, 7 (2). <https://doi.org/10.1029/2022GH000733>

Salmon, Charleen; Flanagan, Jordyn; Farkas, Brenlea; Masthikina, Liza; Egunsola, Oluwaseun; Saxinger, Lynora; Skidmore, Becky and Clement, Fiona (2021). "Transmissibility of COVID-19 among vaccinated individuals". Retrieved from <https://sporevidencealliance.ca/wp-content/uploads/2021/10/Transmissibility-of-COVID-Vaccinated-Individuals_Final-Report_2021.09.24.pdf>

Santesmasses, Didac; Castro, Jose Pedro; Zenin, Alexandr; Shindyapina, Anastasia; Gerashchenko, Maxim; Zhang, Bohan; Kerepesi, Csaba; Yim, Sun Hee; Fedichev, Peter and Gladyshev, Vadim (2020). "COVID-19 is an emergent disease of aging". *Aging Cell*, 19 (10). <https://doi.org/10.1111/acel.13230>

Sasson, Isaac (2021). "Age and COVID-19 mortality: A comparison of Gompertz doubling time across countries and causes of death". *Demographic Research*, 44, 379-396. <https://doi.org/10.4054/DemRes.2021.44.16>

Schellekens, Philip and Sourrouille, Diego (2020). *The unreal dichotomy in COVID-19 mortality between high-income and developing countries*. Retrieved from <https://www.brookings.edu/blog/future-development/2020/05/05/the-unreal-dichotomy-in-covid-19-mortality-between-high-income-and-developing-countries/> [Accessed: 02/05/2023]

Sigler, Thomas; Mahmuda, Sirat; Kimpton, Anthony; Loginova, Julia; Wohland-Jakhar, Pia; Charles-Edwards, Elin and Corcoran, Jonathan (2020). "The Socio-Spatial Determinants of COVID-19 Diffusion: The Impact of Globalisation, Settlement Characteristics and Population". *Research Square, preprint.* <https://doi.org/10.21203/rs.3.rs-33615/v1>

Sorando, Daniel and Leal, Jesús (2019). "Distantes y desiguales: el declive de la mezcla social en Barcelona y Madrid/Distant and Unequal: The Decline of Social Mixing in Barcelona and Madrid". *Revista Española de Investigaciones Sociológicas*, 167, 125-148. <https://doi.org/10.5477/cis/reis.167.125>

Souris, Marc (2019). *Épidémiologie et géographie, principes, méthodes et outils de l´analyse spatiales*. London: ISTE Editions Ltd.

Tchicaya, Anastase; Lorentz, Nathalie; Leduc, Kristell and de Lanchy, Gaetan (2021). "COVID-19 mortality with regard to healthcare services availability, health risks, and socio-spatial factors at department level in France: A spatial cross-sectional analysis". *PloS One*, 16 (9), e0256857. <https://doi.org/10.1371/journal.pone.0256857>

Tepe, Emre (2023). "The impact of built and socio-economic environment factors on Covid-19 transmission at the ZIP-code level in Florida". *Journal of Environmental Management*, 326, 116806.
<https://doi.org/10.1016/j.jenvman.2022.116806>

Tokey, Ahmad Ilderim (2021). "Spatial association of mobility and COVID-19 infection rate in the USA: A county-level study using mobile phone location data". *Journal of Transport & Health*, 22, 101135.
<https://doi.org/10.1016/j.jth.2021.101135>

Wang, Haidong; Paulson, Katherin; Pease, Spencer; Watson, Stefanie; Comfort, Haley; Zheng, Peng; Aravkin, Aleksandr; Bisignano, Catherin; Barber, Ryan; Alam, Tahiya; Fuller, John; May, Erin; Jones, Darwin Phan; Frisch, Meghan; Abbafati, Cristiana; Adolph, Christopher; Allorant, Adrien; Amlag, Joanne; Bang-Jensen, Bree; … Murray, Christopher (2022). "Estimating excess mortality due to the COVID-19 pandemic: a systematic analysis of COVID-19-related mortality, 2020–21". *The Lancet*, 399 (10334), 1513-1536.
<https://doi.org/10.1016/S0140-6736(21)02796-3>

You, Heyuan; Wu, Xin and Guo, Xuxu (2020). "Distribution of COVID-19 Morbidity Rate in Association with Social and Economic Factors in Wuhan, China: Implications for Urban Development". *International Journal of Environmental Research and Public Health*, 17 (10).
<https://doi.org/10.3390/ijerph17103417>

Zhang, Xinxuan; Maggioni, Viviana; Houser, Paul; Xue, Yuan and Mei, Yiwen (2022). "The impact of weather condition and social activity on COVID-19 transmission in the United States". *Journal of Environmental Management*, 302, 114085.
<https://doi.org/10.1016/j.jenvman.2021.114085>