Boletín de la Asociación de Geógrafos Españoles, (91)
eISSN: 2605-3322

How to cite this work: De Cos Guerra, O., Castillo Salcines, V., & Cantarero Prieto, D. (2021). Data mining and socio-spatial patterns of COVID-19: geo-prevention keys for tackling the pandemic. *Boletín de la Asociación de Geógrafos Españoles,* (91). https://doi.org/10.21138/bage.3145

Data mining and socio-spatial patterns of COVID-19: geo-prevention keys for tackling the pandemic

Minería de datos y patrones socio-espaciales de la COVID-19: claves de geoprevención para hacer frente a la pandemia

Olga De Cos Guerra

olga.decos@unican.es

Valentín Castillo Salcines

valentin.castillo@unican.es

David Cantarero Prieto

david.cantarero@unican.es

Research Group on Health Economics and Health Services Management (IDIVAL)

University of Cantabria (Spain)

Abstract

A geographic perspective is essential in tackling COVID-19. This research study is framed in the collaboration project set up by the University of Cantabria, the Valdecilla Hospital Research Institute (IDIVAL) and the Regional Government of Cantabria. The case study is the Santander functional urban area (FUA), which is considered from a multi-scale perspective. The main source is the daily records of micro-data on COVID-19 cases and the methodology is based on ESRI geo-technologies, and more specifically on a tool called SITAR (a Spanish acronym which stands for Fast-Action Territorial Information System). The main goal is to analyse and contribute to

Receipt: 10.05.2021 **Acceptance:** 02.10.2021 **Publication:** 01.12.2021

knowledge of the spatial patterns of COVID-19 at neighbourhood level from a space-time perspective. To that end the research is based on data mining methods (3D bins and emerging hot-spots) and exploratory geo-statistical analysis (Global Moran's Index, Nearest Neighbourhood and Ordinary Least Square analyses, among others). The study identifies space-time patterns that show significant hot-spots and demonstrates a high presence of the virus at building level in neighbourhoods where residential and economic uses are mixed. Knowing the spatial behaviour of the virus is strategically important for proposing geo-prevention keys, reducing spread and balancing trade-offs between potential health gains and economic burdens resulting from interventions to deal with the pandemic.

Key words: emerging hot-spots; geo-technologies; micro-data; social space; multi-scale.

Resumen

La perspectiva geográfica es esencial para afrontar la COVID-19. Este estudio se enmarca en el convenio de colaboración establecido por la Universidad de Cantabria, el Instituto de Investigación Sanitaria de Valdecilla (IDIVAL) y el Gobierno de Cantabria. El ámbito de estudio es el área urbana funcional de Santander y la investigación se desarrolla con perspectiva multiescalar. La principal fuente es el registro diario de microdatos de casos positivos COVID-19 y la metodología está basada en geo-tecnologías de ESRI, y más concretamente en la herramienta SITAR (Sistema de Información Territorial de Acción Rápida) implementada por el equipo investigador. El principal objetivo de este estudio es contribuir al conocimiento de los patrones espaciales de la COVID-19 a nivel de vecindario con perspectiva espacio-temporal. Para conseguir este objetivo la investigación incorpora métodos de minería de datos (cubos 3D y análisis de puntos calientes emergentes) así como análisis geo-estadísticos exploratorios (Índice de Moran global, vecino más cercano y mínimos cuadrados ordinarios). Con relación a los resultados, el estudio identifica patrones espacio-tiempo diferenciados con significación estadística como puntos calientes y demuestra la coincidencia de elevada presencia de casos a nivel de edificio con vecindarios donde la función residencial está combinada con actividades económicas. En definitiva, avanzar en el conocimiento del comportamiento espacial del virus es estratégico para proponer claves de geoprevención, reducir la propagación y equilibrar las compensaciones entre los posibles beneficios para la salud y las cargas económicas que surgen de las intervenciones pandémicas.

Palabras clave: puntos calientes emergentes; geo-tecnologías; microdatos; espacio social; multiescalar.

1 Introduction

More than a year later after the onset of the COVID-19 pandemic, health care workers and the scientific community continue to be on the front line of the battle against the virus. This research is taking place at a time when Spain is facing the fourth wave of infections, and the paper is framed in the contribution of the social sciences, especially geography, in bringing to light spatial patterns in the virus from a multi-scale perspective.

Many interesting contributions have been presented in the field of urban health research, especially focused on how COVID-19 is distributed, against an interesting background of knowledge of other respiratory diseases, such as influenza, analysed at neighbourhood level and using census data (Brizuela et al., 2021). Urban design has been put forward as key to analysing the spatial distribution of respiratory diseases, because it implies certain densities and roles of intra-urban areas depending on housing locations, activities and services. It is essential to explain the spread of influences because of transportation and mobility. Urban design enables space to be presented as a network for spreading with the concentration of activities, jobs and services and, indirectly, a large proportion of the population not straying outside small parts of urban areas (Brizuela et al., 2021).

Our hypothesis is based on a geographic approach, with urban design understood as the medium in which different content concerned with COVID-19 incidence can be analysed. That content is related to population concentration (residents), volume of activities and services (visiting population), population profile from a socio-economical perspective, household size, etc. One of the best documented variables in relation to the pandemic is population density, which seems to be assumed everywhere to be related to the spread of COVID-19. However, the importance of density in spreading COVID-19 depends on the scale and on the geographic units of reference. Hamidi et al. (2020) demonstrate it in their spatial study of COVID-19 severity in relation to population density at county level. In fact, the multi-scale behaviour of COVID-19 is seen as the main reason for the change of density relation with pandemic incidence. This is challenging society globally and locally, with global and local tensions associated with density, the spread of disease, interrelation, travel and transport (Salama, 2020).

Many studies focus on the spread of the virus (Fatima et al., 2021; Franch-Pardo et al., 2021), but our main goal is to analyse and contribute to knowledge of the spatial patterns of COVID-19, considering possible reasons for the "concentration" of infected individuals and looking not only at spread but also at spatial inequalities of incidence over time. The research is framed in the

collaboration project set up by the University of Cantabria, the Valdecilla Hospital Research Institute (IDIVAL) and the Department of Health of the Regional Government of Cantabria. This project focuses on the Regional Autonomous Community of Cantabria (Northern Spain) and is conducted using a multi-scale perspective. It is therefore essential to implement methodologies based on Geographic Information Systems (GIS). More specifically, we manage and analyse data using SITAR (a Spanish acronym which stands for Fast-Action Territorial Information System), a desktop and cloud GIS tool implemented by our research team using ESRI Technologies (De Cos et al., 2020). SITAR includes relevant data structured in thematic geo-databases as follows: health (health administrative boundaries, location and hierarchy of health centres, location and capacity of care homes and location of pharmacies), socio-demographic (gender and age groups, demographic structure indicators, incomes or household size, among others, at census section level) and, finally, the geo-database of buildings (detailed Cadastral Register data on use, conservation level and activities at individual building level). Taking into consideration the importance of the micro-scale in this contribution, the building geo-database is a very important context geo-database in this research. At the same time, we look at other types of data (dependent variables in the research) focused on COVID-19 distribution. In this sense, we must highlight that we use daily micro-data records for positive cases of COVID-19 in the Community of Cantabria. These key anonymised micro-data are provided by the health authorities of the Government of Cantabria (Spain) with the permission of the Medical Ethics Committee of Cantabria (CEIm, ID: 2020.238).

The research is based on two main stages (equivalent to two different scales). Firstly, using data mining tools, we analyse the spatiotemporal trend in COVID-19 cases in the most dynamic area of Cantabria —the functional urban area (FUA) of Santander—, where we analyse how cases are distributed and identify the main areas of incidence. Then, at a deeper scale, using building characteristics, we analyse the links between COVID-19 incidence and other variables.

Built environment and social context have a strong influence on patterns of health. As Huang et al. (2020) state in their geographic analysis to determine the relationship between built environment and COVID-19 incidence rate transmission in Hong Kong, the built environment is important not only directly in guaranteeing health conditions related to quality of build, but also indirectly in tackling disease transmission, because it involves other variables such as density, type of housing, expected mobility and intensity of social interactions. All these characteristics are important in designing and identifying safe areas to stop the transmission of the virus. In addition, promoting behaviour conducive to individual health (hand washing, social distancing and mask wearing,

among others) is essential in defining strategies to contain the virus spread (Pinter-Wollman et al., 2018). In this framework, our research seeks to contribute to knowledge of spatial patterns of COVID-19 and associated variables at detailed scales. This is very important in tackling the pandemic from a geo-prevention perspective. The adaptation of the concept of prevention to a geographic approach is framed in the field of environmental criminology, related to safe areas from the point of view of crime rates (Hernando, 2008). This approach can be adapted to other fields related to safety and wellbeing, so geo-prevention is also adaptable to health analysis and management because geo-prevention is helpful in designing local and regional public health measures for tackling the pandemic. In fact, studies of the difficulties of tackling COVID-19 consider that the pandemic has a global dimension, and can thus be approached globally, but actions must be designed locally. From this perspective, some authors propose a local custom approach as a sensible way of making decisions that balance the economy/health binomial better (Campagna, 2020). In this sense, geo-prevention is a strategic approach to the tackling of COVID-19.

Other authors look to disciplines related to architecture as a basis for arguing that even in the detailed-scale design of complementary decisions there are particularities that can help to reduce infection rates, such as the positioning of sanitizer dispensers (Pinter-Wollman et al., 2018). Our study refers to detailed scales from a geographic approach at building level and no deeper, but it is important to consider that small details, rules and behaviours can make all the difference between spreading and containing the pandemic. Individual or small-group decisions, preventive habits, social distancing and relationships in our living space are important keys for tackling COVID-19. Indeed, from the point of view of engineers specialised in environmental health, it is necessary to take into account that quarantine or staying at home is determinant in reducing spread outdoors but can entail other risks indoors because of the increase in air pollution and contact between occupants, as documented in the "sick building syndrome" (SBS) understood as a complication of the health status of occupants and building characteristics (Hosseini et al., 2020).

In the background there are different research lines involving buildings and health. On the one hand there is the link between built space and transmission and on the other the contribution of urban planning and buildings design to preventing spread, and the challenges arising from new urban dynamics during and after the pandemic, new proposals to secure distancing rules and new uses of living and working spaces from an architectural perspective (Salama, 2020).

In relation to the first approach, some studies before the COVID-19 pandemic focused on the contribution of building characteristics to the movement of air between flats and consequently to the spread of some viruses inside them. Li et al. (2004) study the severe acute respiratory syndrome (SARS) epidemic of 2002-2003 in Hong Kong, and demonstrate that infections in several housing blocks had a non-random spatial pattern. In relation to this, they design a multi-zone model for analysing aerosol distribution in different scenarios. Based on a similar approach, there are now studies of building conditions and COVID-19 spread by aerosol transmission at the micro-scale. For instance, there is the controversial hypothesis of Hwang et al. (2021), who from an architectural point of view, analyse ten positive cases (micro-data) in seven households in an apartment building in Seoul, where cases appeared along two vertical lines that connected poorly ventilated bathrooms. Deeper research is needed to obtain evidence about ventilation and air-condition systems not only in residential buildings but also public and work buildings (Chirico et al., 2020). This is an interesting point of view, and one may wonder whether it points to a hidden cause in buildings where there is a high concentration of positive cases in the same period of time. What is the contribution of indoor infection to pandemic spread? Related to this approach, a recent study published in *The Lancet* (Greenhalgh et al., 2021, April 15) includes a scientific statement arguing that there is airborne transmission of SARS-Cov-2 and referring to long-range transmission between people in adjacent rooms with their own bathroom inside and no balconies during quarantine in a hotel in New Zealand (Eichler et al., 2021).

In this sense, it is important to consider that studies of the spatial patterns of COVID-19 such as the one reported here were conducted at a time when there were many rules aimed at ensuring social distancing, so the distribution and spread of COVID-19 could be influenced by the effect of constraints on "new-normal" life. Containment rules can be considered at two levels: institutional (workplaces, schools, culture and economic activities) and individual, as people tend to reduce interaction by avoiding gatherings of family and friends and increasing the use of technology as a way of contacting others (Salama, 2020). This in turn is directly related to mobility, another important factor in the spread of the virus. As mentioned by Roy & Kar (2020), part of mobility is related to intrinsic factors (own motivations) but another part depends on extrinsic factors (for instance, rules from institutions to increase social distancing).

In short, the theoretical background on the spatial incidence of COVID-19 has many pillars that are important in our research. They concern variables and context characteristics linked with virus incidence. Moreover, many studies establish methodological advances in regard to the approach needed.

In this framework, mobility and socio-demographic context are related to spatial patterns of COVID-19, as demonstrated by Roy & Kar (2020) in their study on the city of Los Angeles based on data at census block level and using machine learning methods. They define a vulnerability approach based on socioeconomic status (poverty, unemployment, schooling level and incomes), household composition and disability (age older 64, age under 17, disabilities and single parent), minorities (ethnic minority and English-language skills) and, finally, housing type. This last group of variables are related to the built environment (group quarters, multi-unit structures, mobile homes, crowding) and vehicle availability (Roy & Kar, 2020, p. 42). Moreover, some studies show that commonly accepted variables can lead to inappropriate results depending on the type of area. Huang et al. (2020), after geocoding and analysing positive tested cases data, find that the risk of COVID-19 transmission is underestimated in suburban areas (due to the presence of large open and green areas) if studies only consider the incidence rate. This brings us to the second pillar, related to methodology: many studies of COVID-19 and other previous virus at detailed scales reveal spatial non-stationarity (Brundson et al., 1996) in distribution patterns and links with environment variables (Mou et al., 2017; Huang et al., 2020). Therefore, it is not possible to establish an overall model or general knowledge of COVID-19 and context relationships, because it could hide the real spatial behaviour of the virus. This is the main reason why a more adaptable approach to space is needed, as we seek to provide with our multi-scale methodology.

2 Overview of the study area: focusing on urban areas

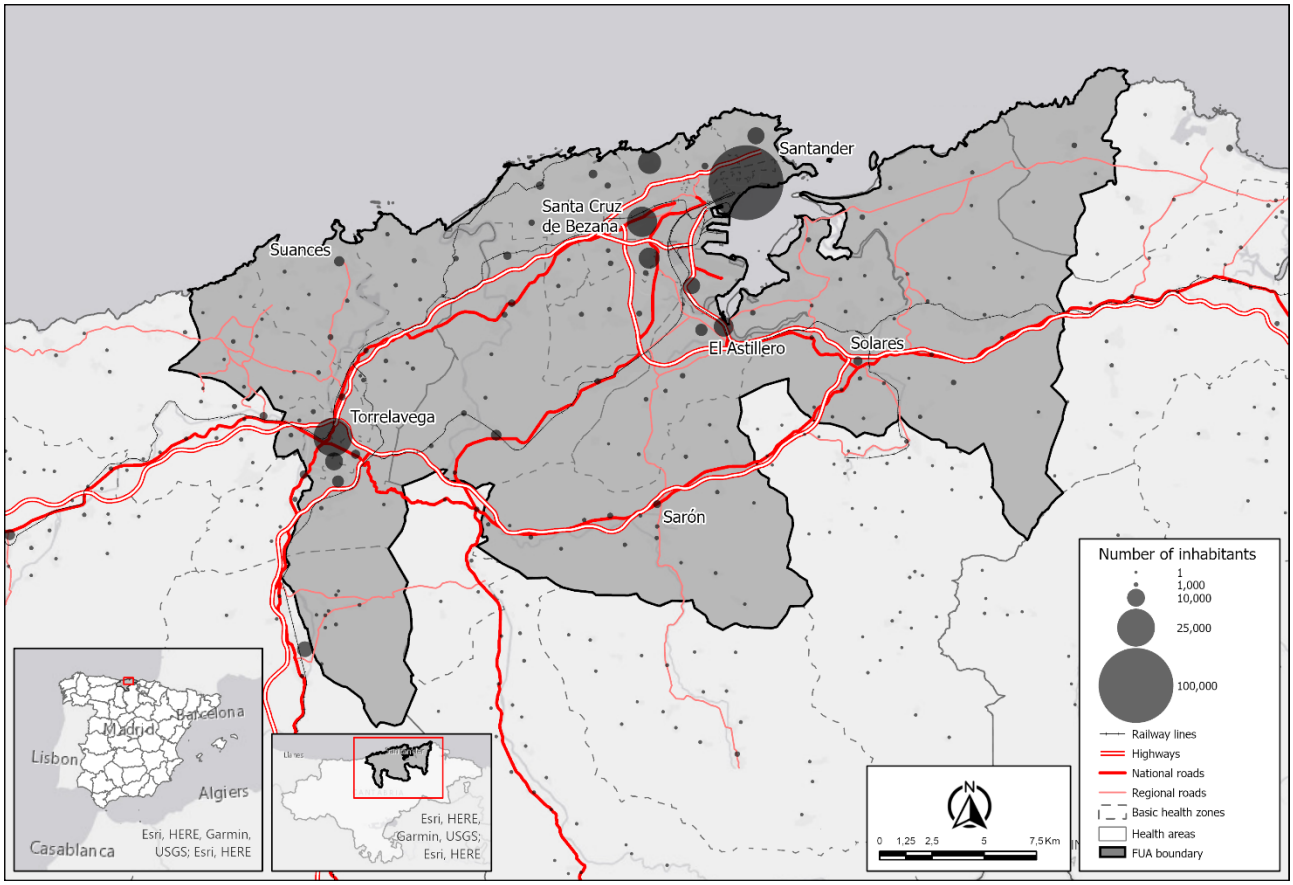
The study focuses on the Autonomous Community of Cantabria in northern Spain. This region has a population of almost 583,000 and a surface area of just over 5,300 km^2 (National Institute of Statistics. Register of residents, 2021). The average population density is thus close to 110 inhabitants per km^2. However, there are substantial internal differences in population distribution and density, with a sharp contrast first between the coastal area and the inland valleys, where densities are low or even critical except in the few towns where the main services are concentrated in rural areas.

In this framework, we analyse the case study of the functional urban area (FUA) of Santander. It is a dynamic unit identified at European level with the criteria of population concentration and intense mobility due to commuting, as proposed Batista & Poelman (2016). It has a surface area of 688 km^2, with 21 municipalities which between them are home to more than 380,000 people. It is an important area for analysing COVID-19 spatial patterns because of its urban and

metropolitan role in the context of Cantabria. The importance of urban areas in health research is worth highlighting here, in that cities, as complex systems, have many characteristics that impact the health of people (Brizuela et al., 2021). Moreover, some studies of virus spread highlight the role of urban settlements and a factor related to the "metropolitan city effect" considering aggregated data at the level of province in Italy (Gargiulo et al., 2020).

In this sense, the Santander FUA is the most interesting part of Cantabria for analysing COVID-19 patterns on detailed scales. It comprises a polycentric hinterland around the capital city, Santander, and Torrelavega, the second biggest city in the region (Figure 1). The Santander FUA extends across two different health areas (out of the four in the region) and intersects 25 basic health zones (out of 42 in the region).

Figure 1. The Santander FUA (Cantabria) study area:
settlements and main travel infrastructures



Source: authors' own elaboration based on ESRI (Administrative Base map), National Geographic Institute (National Cartographic Base 200 and Urban Atlas) and National Institute of Statistics (data from register of residents, 2021)

This area contains 65.6% of the population of Cantabria, in only 13% of its surface area (Table 1). Consequently, its population density is about five times the regional average. The difference is twice as great if the FUA density is compared with that of the rest of Cantabria, where rural areas are predominant.

Table 1. Main data on the study area in the context of Cantabria

ZONE	Municipalities		Inhabitants		Area (km^2)		Density
	Total	%	Total	%	Total	%	
Santander FUA	21	20.6	383,429	65.8	685.7	12.9	559
Rest of Cantabria	81	79.4	199,476	34.2	4,640.5	87.1	43
Total Cantabria	102	100.0	582,905	100.0	5,326.2	100.0	109

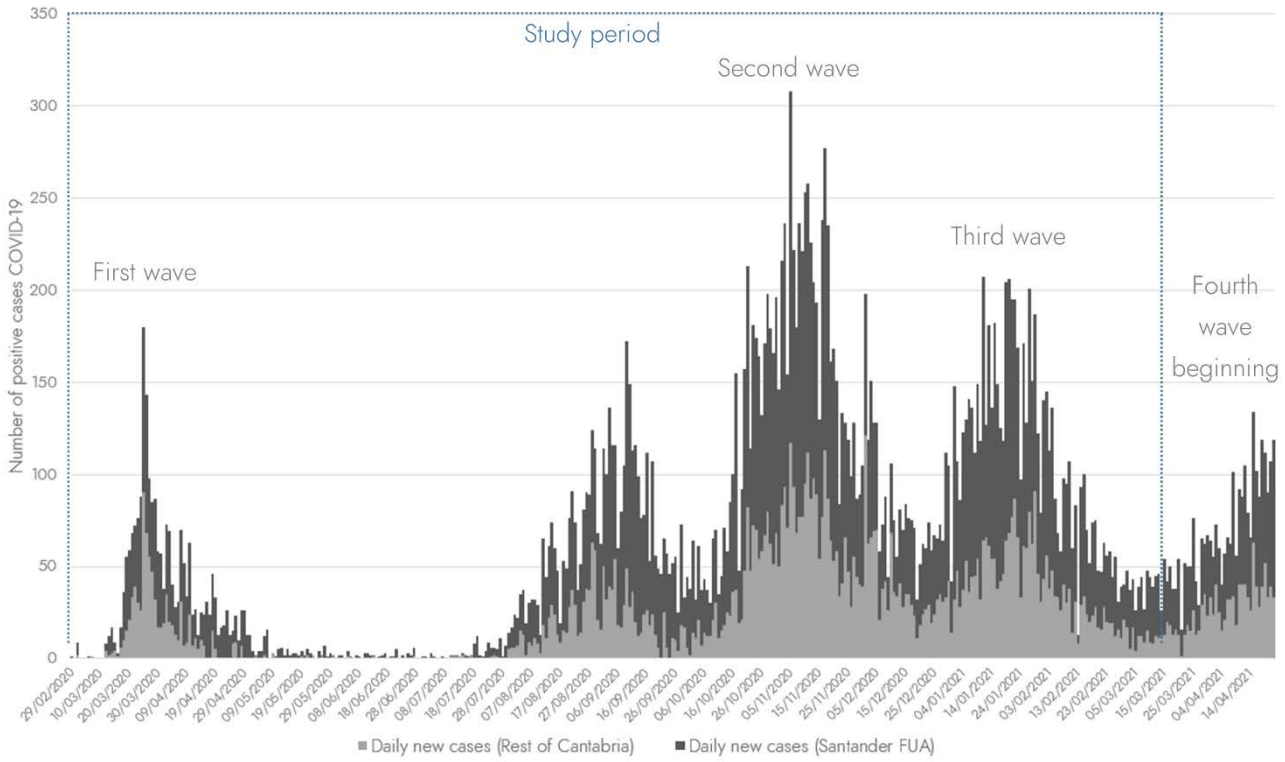
Key: density is expressed in inhabitants per square kilometre.

Source: authors' own elaboration based
on National Institute of Statistics (data from register of residents, 2021)

The Santander FUA is interesting as a case study not only for its density, but also for its concentration of activities and main transport infrastructures and the high number of daily commutes.

The study period covers practically the whole of the first year of the pandemic. It therefore includes the three waves from February 29, 2020 (the beginning of pandemic records) to March 15, 2021. During that time Cantabria suffered a total of 26,470 reported cases of COVID-19, 15,992 of them (around 60%) located in the Santander FUA and 10,478 (around 40%) in the rest of Cantabria. As Figure 2 shows, the trends during that period were similar in the Santander FUA and the rest of Cantabria, but in the case of the FUA a continuous presence of new cases can be observed even in inter-wave periods such as the "new normal" period from May to July after the easing of the strict lockdown in Spain. The daily dataset of new cases ends (at the time of this study) at the beginning of the fourth wave, so we do not have a full perspective for this wave. We therefore adjust our study period to the three complex waves documented using COVID-19 micro-data.

Figure 2. Daily trend in new cases in Cantabria distinguishing between the Santander FUA and the rest of Cantabria (February 2020–April 2021)



Source: authors' own elaboration based on records of the trend in COVID-19 in Cantabria (Open data from the Government of Cantabria)

3 Materials and methods

This section describes many important details regarding to the research methodology into spatial patterns of COVID-19 in the Autonomous Community of Cantabria. Firstly, datasets and their sources are presented as evidence that the research can be exported to other areas with similar data. Then, we explain the methods, in which Geographic Information Systems and geo-statistical and data mining tools are the two pillars that can be used to replicate the study elsewhere with the same or different scales.

3.1 An approach based on daily micro-data records of positive cases and the SITAR tool

Data are essential in tackling the pandemic. Indeed, research teams working on COVID-19 lines often have to overcome difficulties in analysing the influence of context determinants of health (social, economic or territorial) on smaller scales (for instance, neighbourhood level) because the resolution of epidemiological datasets is not appropriate. On the other hand, it seems accepted

that where people live is a key determinant for modelling vulnerability in relation to the incidence of the pandemic.

> A person's place of residence can largely influence their role and vulnerability during an epidemic. In particular, the higher contact rates of people living near major activity hubs can give rise to predictable patterns in the spread of disease (Brizuela et al., 2021, pp. 1-2)

Taking into consideration the circumstances and references set out below, the most revealing source for our goals lies in the micro-data records of positive cases of COVID-19. This data is produced daily in Spain by the governments of the autonomous communities. Micro-data are essential for analysing the spatial patterns of the pandemic from a multi-scale approach. There are few research teams in Spain which can currently access such records (De Cos et al., 2020; Perles et al., 2021). The anonymised use of these data guarantees compliance with data protection rules not only at national level but also internationally (European Union Regulation 2016/679).

The research reported here is based on positive cases of COVID-19 reported daily by the Regional Government of Cantabria (Spain). Access to these data was permitted by the Medical Research Ethics Committee of Cantabria (CEIm) in June 2020 (ID: 2020.238) and the cumulative data series started at the beginning of records on the pandemic (29 February 2020). Micro-data on all individuals who have tested positive for COVID-19 in Cantabria are held initially in a tabular structure, but in the study these micro-data records are geo-coded using the multiple field geo-coding tool from ArcGIS Pro, that considers several location fields, such as address and other fields about polygonal administrative units (post code, town, municipality, and country). Geo-coding tool finds the position as a point connecting with the ArcGIS World Geocoding Service. We obtained the geo-codification for 97.8% of initial records, so it is an efficient tool. The missing records correspond to infected people without an address in the Autonomous Community of Cantabria or records without address matching in the ArcGIS World Geocoding Service. This geo-coding provides a point layer dataset that conserves other basic fields concerned with demographic structure (age and gender), time data (start and end dates), COVID-19 severity and status (hospitalisation, intensive care, status as positive —if the virus is active—, cured or deceased). Additionally, in relation to particularities of the incidence of the pandemic in Spain, the micro-data records of the Government of Cantabria include two important binary fields that enable us to filter if necessary according to whether an individual lives in a care

home and whether the positive case is related to an occupational category in health or care activities.

These micro-data are the main source for spatial research into COVID-19, but the methodology we design involves several sources for demographic, economic and residential context produced by relevant public institutions (National Institute of Statistics, National Geographic Institute and National Government of Spain) and by the private sector (ESRI Spain COVID-19 GIS Hub and ESRI ArcGIS Geo-Enrichment Service).

To manage and analyse these sources we implemented the tool called SITAR (the Spanish acronym of Fast-Action Territorial Information System). SITAR is based on ESRI Technologies accessed via the user license held by the University of Cantabria.

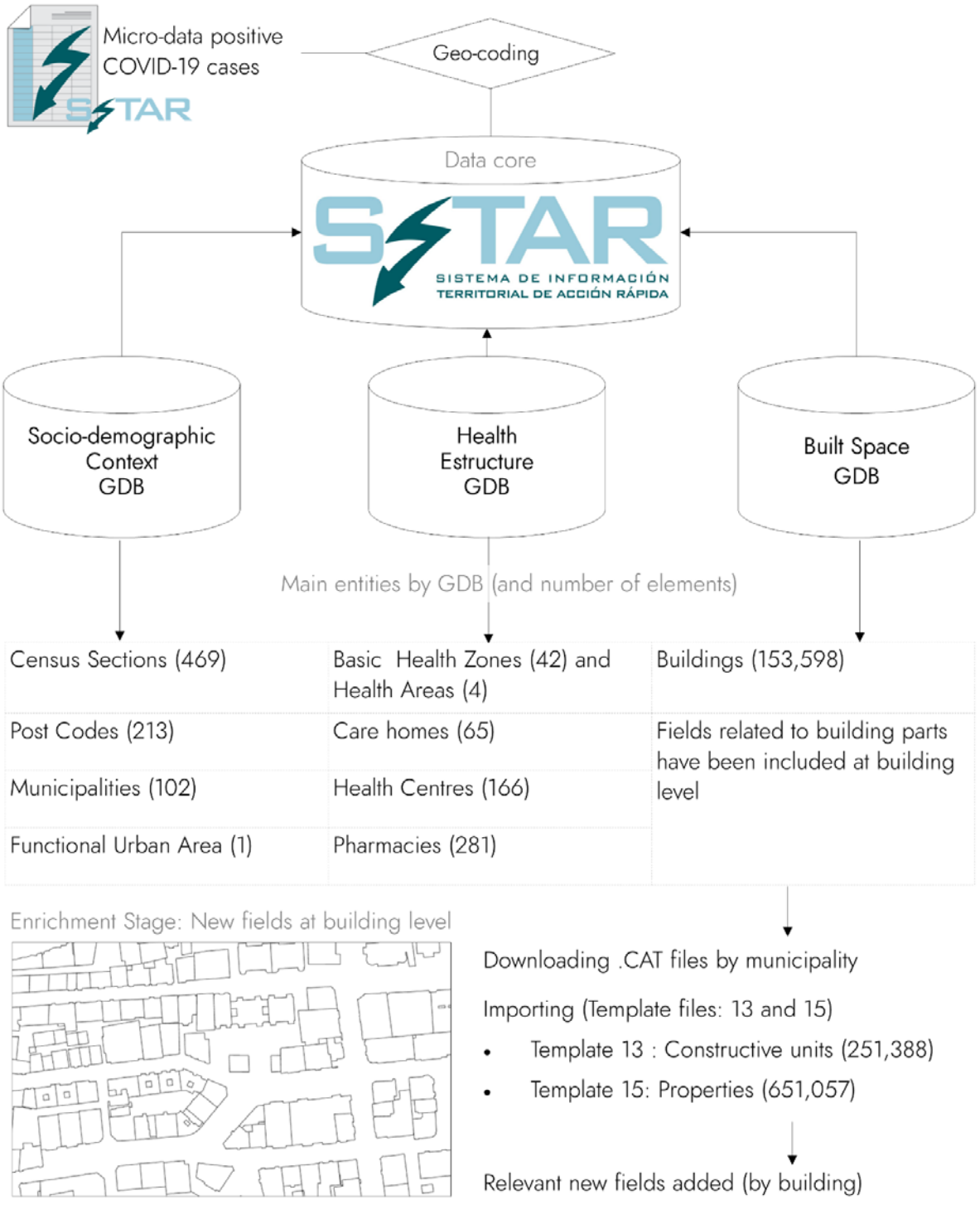
From the beginning of the study, SITAR core data were structured into three thematic geo-databases (GDB): health structure, socio-demographic context and built space (Figure 3). The different sources for each topic mean that different spatial units are used. The SITAR health GDB includes health areas and basic health zones, but socio-demographic data are organised at the level of census sections, municipalities or post codes. In this context, the more detailed polygon entities correspond to buildings from the Spanish Cadastral Register source. Building data enabled us to conduct a deeper analysis of spatial patterns of COVID-19 on detailed scales.

However, the basic information on buildings from the Cadastral Register's ATOM Service is not enough for our goals. It must be considered that the initial fields for buildings in the cadastral register only provide general data on current use (not details of activities in each building), date of entry in the Cadastral Register (not date built), conservation status and number of floors (available for building parts).

In the past few months, the SITAR GDB for Cadastral Register building data has been improved by incorporating more fields and entity counts at building level from additional Cadastral Register files (.CAT extension file). CAT files must be imported using templates. The Cadastral Register Service offers five initial templates, but in our research we use two of them (templates 13 and 15). We took into consideration the fields that we needed to incorporate into SITAR at building level. Template 13 is for construction units (more detailed than buildings) and it is important in our study because it includes fields for "year built" and "surface area occupied (in square meters)". Template 15 is for properties. Significantly, it includes information on potential economic activities per building, which is relevant to the spread of the virus on detailed scales, as mentioned in the Introduction. Those new fields detail the number of properties in each building, broken down by

uses: storage/parking, residential, industrial, offices, commercial, sports, shows, leisure and catering, health and welfare, cultural, religious, urbanisation work and gardens/undeveloped land, unique buildings, agricultural storage, agricultural industrial and, finally, agricultural.

Figure 3. SITAR data core and Cadastral Register enrichment stage



Source: authors' own elaboration

The use of .CAT files requires a laborious process. Firstly, the original source files must be downloaded from the Spanish Cadastral Register Service at municipal level (102 downloads in the case of the Autonomous Community of Cantabria). Secondly, the process of preparing data includes several steps. We import CAT files by municipalities using templates 13 and 15. It results 102 municipal spreadsheets by template where we obtain records of construction units (Template 13) and properties (Template 15). Then, municipal spreadsheets are imported in a database, using Microsoft Access as database management system. In the database it is necessary to manage data using action queries (aggregation to convert 102 tables of templates 13 and 15 into one new table by template). Resultant tables include the total number of records of Cantabria (251,388 records of construction units and 659,057 records that correspond to properties). Initially, our Cadastral layer represents buildings (153,598 records) so we have one-to-many relationships. Nevertheless, the preparation of Cadastral data continues in non-spatial framework (Microsoft Access). We use "make table queries" to summarize records by cadastral building register code (counting number of records, for instance, properties filtered by type, or adding numeric fields). Finally, we obtain tables at building level that include prepared fields from original templates 13 and 15. The process concludes in GIS framework, using ArcGIS Pro to join external tables to Cadastral building layer with one-to-one relationship.

3.2 Research workflow and GIS tools

The research workflow involves two stages, both framed in geo-statistical methods implemented by GIS, using the SITAR tool. The first comprises many exploratory analyses to reveal general spatial patterns of COVID-19 in the Santander FUA from different points of view: on the one hand the statistical significance of the distribution of COVID-19 cases and on the other the general pattern in relation to land use or coverage in the Santander FUA. The second stage uses a more in-depth method to identify risk areas from a multi-scale perspective: the whole FUA using 3D bins and emerging hot-spot analysis and building level by identifying the high-incidence buildings (over 30 COVID-19 cases), i.e. those buildings that accumulated more than ten times the average number of cases per building in the period (Figure 4).

The exploratory stage seeks to learn the likelihood of the distribution of cases being non-random. It is based on many geo-statistical methods applied to point datasets (geo-coded COVID-19 cases):

- The nearest-neighbour and Global Moran's Index methods are the main statistical tools used to calculate the probability of the pattern of case distribution being non-random.

Other interesting measures are also obtained, such as observed and expected distance. Expected distance serves as a parameter in the 3D bins dimension in later stages. Average Nearest-neighbour is calculated using the Equation 1 (Evans & Evans, 1954) and Global Moran's Index is calculated using the Equation 2 (Moran, 1948) as follows:

Equation 1

$$ANN = \frac{\bar{D}_O}{\bar{D}_E}$$

Where: $\bar{D}_O$ is the observed mean distance between each feature and its nearest neighbour and $\bar{D}_E$ is the expected mean distance for the features in a random distribution.
Note: features in our model correspond to individual points of COVID-19 cases.

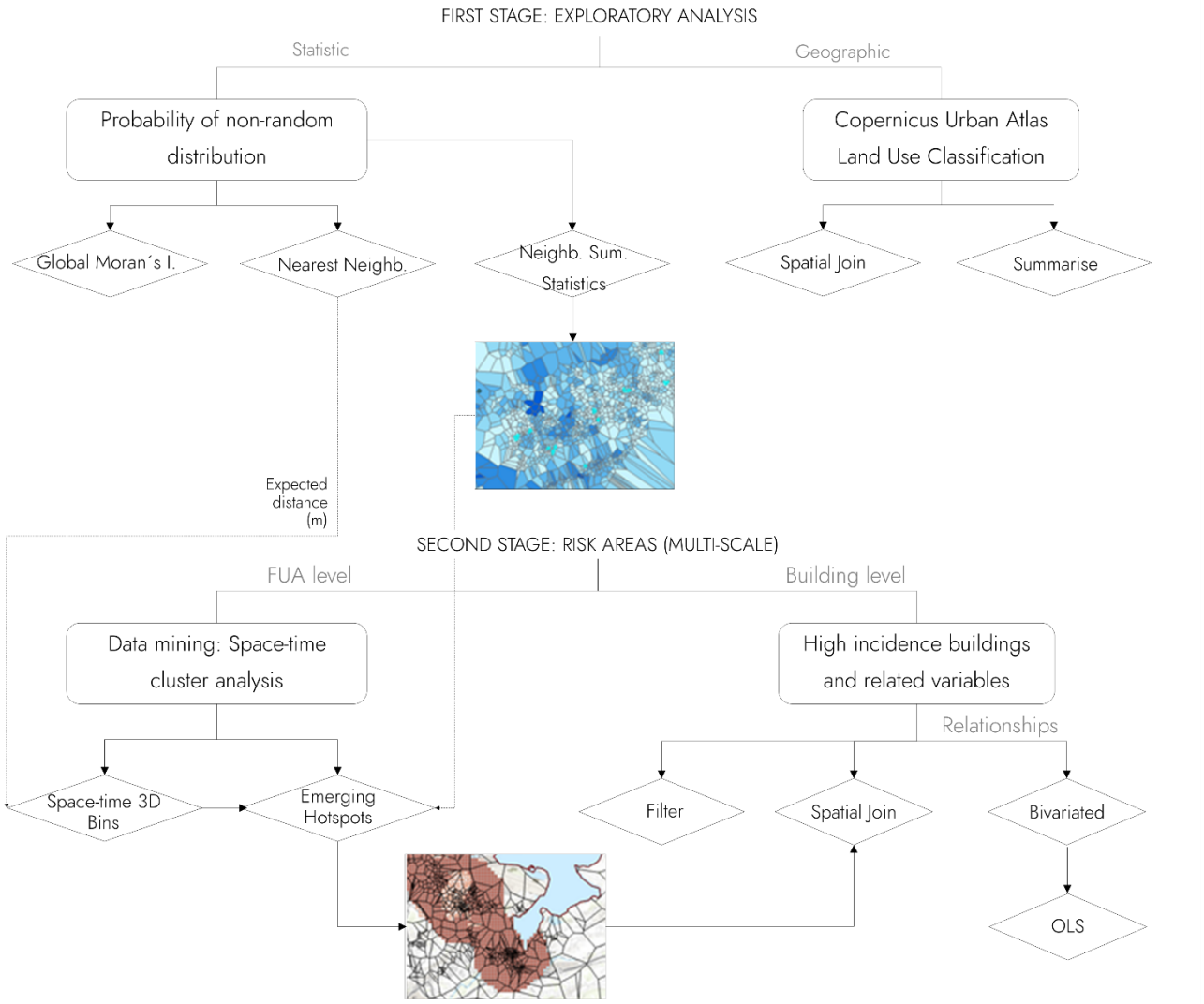
Equation 2

$$I = \frac{n}{S_0} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j} z_i z_j}{\sum_{i=1}^{n} z_i^2}$$

Where: z_i is the deviation of an attribute for feature i from its mean value, $w_{i,j}$ is the spatial matrix weight between features i and j, n is the total number of features, and S_0 is the normalisation factor (aggregated spatial weights).
Note: features in our model are point locations with COVID-19 cases and the attribute corresponds to the number of cases in each point.

- Neighbourhood summary statistics (Brundson et al., 2002) result in a further polygon layer based on Delaunay triangulation from the distribution of COVID-19 cases. These new areas provide interesting descriptive statistics on centrality, position and dispersion associated with each new polygon, including average distance between polygon cases and neighbours, which show internal disparities in distribution. Furthermore, the shape and area of each polygon is revealing from the point of view of accumulated cases (inversely proportional). This method contributes to the research not only through the exploratory results but also because it models a new polygon layer that is used to represent emerging hot-spots in later stages, overcoming possible constraints of administrative or management units such as basic health zones or census sections.
- The third exploratory analysis is not focused on statistical methods. From a geographical approach, it is useful to learn the general distribution of COVID-19 cases in line with the Copernicus Urban Atlas Land Use Classification (2018), through which density and intensity of occupation can be identified in an urban context. For that purpose, spatial joins and summaries per field (classification) are needed.

Figure 4. Research workflow based on two stages and a multi-scale approach



Source: author's own elaboration

The second stage, focused on determining risk areas, is based on a multi-scale perspective:

- The analysis at FUA level is based on a model of 3D bins and emerging hot-spots. This data-mining method has been researched in previous studies (waves 1 and 2 in Cantabria) which have shown its predictive potential in relation to spatial patterns of COVID-19 (De Cos et al., 2021). Previous research contrasted the importance of parameters in 3D bins analysis (bin size and temporal slides), so here we base the new 3D bins parameters of the FUA on the same relative criteria. The bin dimension is the expected distance obtained in the nearest neighbour analysis (127.11 m) and time is divided into 4-week intervals because (as shown in previous research) it thus covers 2 of the usual 2-week reference periods for accumulated incidence and meets the condition for the method of at least 10

points in time for the development of bins (De Cos et al., 2021). Using these parameters, new space-time 3D bins are created in a NetCDF (Network Common Data Form) layer, where COVID-19 cases are accumulated into a regular and constant structure with both spatial and temporal perspectives. These 3D bins are used for the emerging hot-spot analysis that clusters the space-time trend according to the nearby bins and distinguishes statistically significant patterns. The ArcGIS emerging hot-spots tool is based on Getis-Ord Gi* statistics (Getis, 1992) to identify hot-spots and Mann–Kendall statistics (Kendall & Stuart, 1976) to determine trends. The method is based on the key field count (COVID-19 aggregated cases) of each bin recorded over time (one year in our study divided into 4-week periods). The emerging analysis provides a maximum of 17 pattern types (1 no pattern, 8 coldspot and 8 hot-spot types). Emerging patterns are calculated in the framework of the polygon layer of neighbourhood summary statistics, which produces a model based on the units modelled by the distribution of cases itself. According to these patterns, and focusing on the risk, no pattern and cold-spots are not problematic areas, but hot-spots are related to the spread and a significant presence of the virus.

- At building level, the research focuses on high-incidence buildings by filtering. We summarise many variables to show differences between buildings, with three possible types being considered: high-incidence buildings, other buildings with COVID-19 cases and buildings with no cases. The research also includes some exploratory analysis to determine the main variables in relation to COVID-19 distribution at building level: initially we explore linear bivariate analysis and then Ordinary Least Square analysis (OLS) as a correct and expressive way of analysing the non-stationarity of COVID-19 distribution (Zhou, 2017).

4 Results from a multi-scale approach

According to the above methodology, we present the results of the two main stages of analysis. The spatial analysis of the 15,374 geo-coded cases out of care homes in the Santander FUA account for 63.8% of all cases in the Autonomous Community of Cantabria from the beginning of the recording of COVID-19 micro-data (February 29, 2020) to the third wave in continuous daily records, which end for the purposes of analysis on March 15, 2021.

Before we present our empirical results, it must be highlighted that the number of cases in the Santander FUA as a proportion of the total for the whole region is not particularly high: the location coefficient (which expresses the number of COVID-19 cases in relation to the number of

inhabitants) is 0.92, with 1.0 indicating perfect correspondence between the number of positive cases and the population size (Table 2). But the Santander FUA is still an interesting case study and the results are significant because the area has shown a continuous presence of COVID-19 cases throughout the first year of the pandemic, with an uneven distribution and reiterative damage in certain specific parts of the territory, as outlined below.

Table 2. Location Coefficient of COVID-19 cases in the Santander FUA

ZONE	Number of inhabitants	COVID-19 cases	Ratio of inhabitants (FUA respect to Cantabria)	Ratio of COVID-19 cases (FUA respect to Cantabria)	Location Coefficient
Santander FUA	383,429	15,992	0.6578	0.6042	0.9185
Total Cantabria	582,905	26,470	-	-	-

Key: The location coefficient is calculated using positive cases reported by health authorities per municipality (independently of the geo-coding process, so cases with no data for correcting geo-coding are also included).

The coefficient should be interpreted as follows: a figure of 1.0 denotes areas with exactly the number of COVID-19 cases expected for their population size, figures of over 1.0 denote areas with more COVID-19 cases than expected for their population size, and finally figures of less than 1.0 denote areas with fewer COVID-19 cases than expected for their population size.

Source: authors' own elaboration based on National Institute of Statistics (Data from Register of Residents, 2021) and COVID-19 cases reported by the Regional Government of Cantabria

A preliminary stage analysis at FUA level shows that the nearest neighbour distance and Global Moran's Index confirm a non-random spatial pattern for COVID-19 with a confidence level of more than 99% in all three areas considered, coinciding with clustered distributions. Moreover, the preliminary analysis also shows interesting values in relation to distance, and it can be related to the density of cases (Table 3). Indeed, in the nearest neighbour analysis the average observed distance between cases is 11.56 metres in the Santander FUA, compared to 37.40 in the rest of Cantabria, so there are three times more observed distance for the same period as in the rest of the region. The Z score (standard deviation) is -201.58 (under -2.58), so the spatial pattern is clustered and non-random, again with a confidence level of more than 99%.

Table 3. Preliminary analysis results in relation to distance of COVID-19 cases in the Santander FUA and the rest of Cantabria

ZONE	Observed average distance (m) between cases	Expected average distance (m)	Z Score under Nearest Neighbour	Critical Value Z Score under Moran´s Index	P Value
Santander FUA	11.56	127.11	-215.66	6.61	<0.01
Rest of Cantabria	37.40	510.47	-165.57	4.45	<0.01
Total Cantabria	20.84	307.10	-276.82	7.39	<0.01

Key: The Z Score under Nearest Neighbour can be interpreted as follows: <-2.58 means that the distribution is non-random and clustered. The critical value for the Z Score under Moran's Index can be interpreted as follows: >2.58 means that the distribution is non-random and clustered. The P Value can be interpreted as follows: <0.01 means a confidence level in regard to non-randomness of more than 99%.

Source: authors' own elaboration based on COVID-19 micro-data in daily records from the health authorities (Government of Cantabria)

Moreover, the location of cases is analysed under the Copernicus Urban Atlas Land Use Classification (2018). As Table 4 shows, 61.0% of the COVID-19 cases in the Santander FUA are in the more densely occupied areas, but the location coefficient brings the significance of this figure in 0.87 (under 1.0, so the number of cases in more densely populated areas is lower than expected for the number of inhabitants). Secondly, 17.2% of cases are found in discontinuous dense peri-urban areas. This figure is higher than theoretically expected based on population size. In fact, the discontinuous urban fabric has more cases in general than expected for the number of people who live in such areas. These areas are very characteristic of the Santander FUA given the metropolitan dynamic of the main cities of Santander and Torrelavega in a polynuclear metropolitan system.

An explanation is necessary in relation to the high location coefficient in areas of "Pastures and Forest". It corresponds to large areas (67.2% of the FUA), with isolated buildings (only 3.5% of buildings of the FUA), close to urban and peri-urban areas in a context of high commuting. Although in percentage terms it is not high (5.2% of COVID-19 cases) the high Location

Coefficient is conditioned by the low volume of residents (3,128 inhabitants, i.e., 0.8% of the FUA residents). This value could be related to the effect that high commuting has in the virus spread, even in isolated buildings close to the most intense occupied urban areas.

Table 4. Location Coefficient of COVID-19 cases in the Santander FUA under the Copernicus Land Use Classification

COPERNICUS LAND USE	Number of cases	Percentage of cases	Location Coefficient
Continuous urban fabric (80% occupancy)*	9,373	61.0	0.87
High-density discontinuous urban fabric (50% - 80% occupancy)	2,646	17.2	1.11
Medium-density discontinuous urban fabric (30% - 50% occupancy)	1,211	7.9	1.24
Pastures and Forest**	794	5.2	6.31
Low-density discontinuous urban fabric (10% - 30% occupancy)	733	4.8	1.71
Industrial, commercial, public, military and private units	497	3.2	1.04
Very low-density discontinuous urban fabric (<10% occupancy)	123	0.8	0.51

Key: The Location Coefficient is calculated using the micro-data records reported by the health authorities (Government of Cantabria).

The coefficient can be interpreted as follows: a figure of 1.0 denotes areas with exactly the number of COVID-19 cases expected for their population size; > 1.0 means more COVID-19 cases than expected for the population size; and < 1.0 means fewer cases than expected for the population size.

*Continuous urban fabric category includes green urban areas.

** Pastures and Forest category includes two original land uses that correspond to the 67.2% of the area.

Source: authors' own elaboration based on COVID-19 micro-data in daily records from the health authorities (Regional Government of Cantabria), National Institute of Statistics (data from Register of Residents, 2021) and European Union Urban Atlas (2018)

4.1 Risk areas in the Santander FUA from a space-time perspective

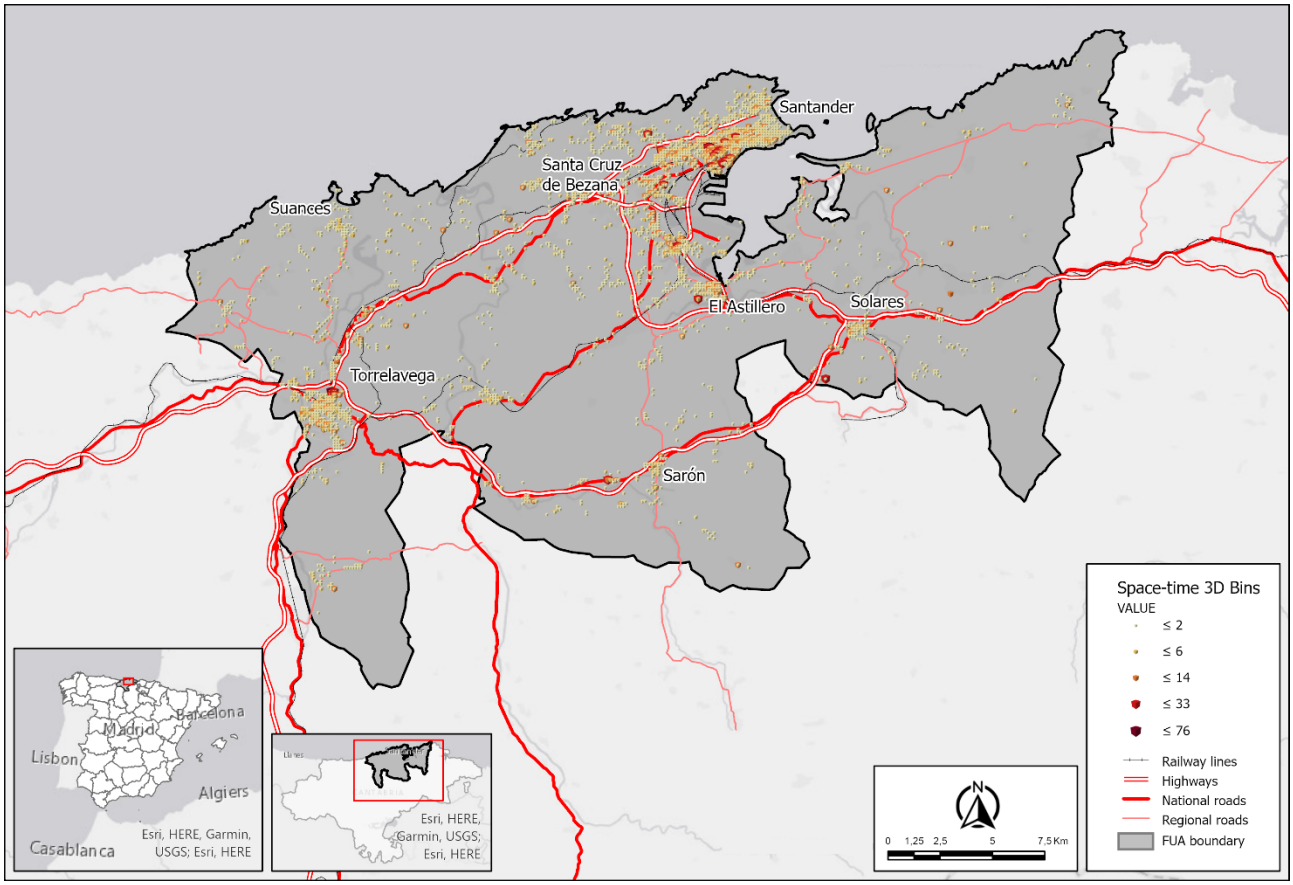
The results of 3D bins and emerging hot-spots analysis are revealing, in that they enable a distinction to be drawn between risky areas and non-significant patterns in relation to the space-time trend of cases at the level of homogeneous units (3D bins with expected distance sized).

As Figure 5 shows, the distribution of 3D bins is noteworthy in relation to the metropolitan dynamic, mobility and the main residential areas. In fact, the spatial pattern of the virus highlights the urban centres of Santander and Torrelavega and their peri-urban areas around the western side of the Santander Bay and the municipalities adjacent to Torrelavega, respectively. Thirdly, there are significant 3D bins near main transport routes, especially on the Santander-Torrelavega corridor.

This model is supplemented by a diagnostic model of emerging hot-spots (Figure 6). First of all, the contribution of this model is useful in distinguishing significant risk areas from the rest. Indeed, it is possible to clearly identify areas with specific emerging hot-spot patterns. Thus, significant emerging types are found in Santander and its periphery, Torrelavega and its periphery and secondary locations (smaller in size) in medium sized population centres, with a progressive demographic trend due to the metropolitan dynamic (such as Renedo, Santa María de Cayón and Solares to the south). The same pattern emerges in the north, on the coast, in the case of Suances, which differs from the above in that it comprises over 40% second homes and holiday residences.

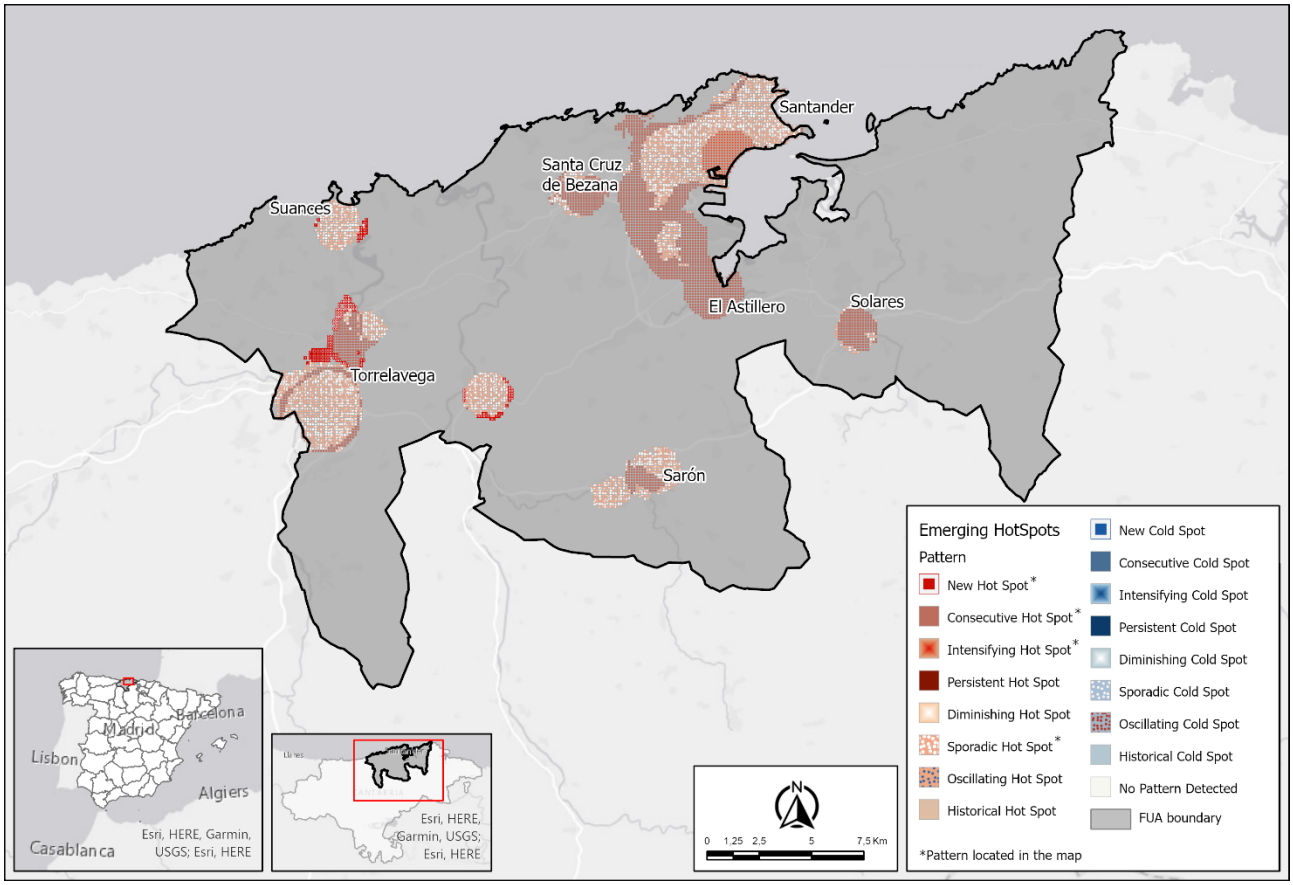
Other interesting findings are the distribution of consecutive hot-spots in areas with high commuting and the location of new hot-spots near previous significant areas.

Figure 5. 3D bins for COVID-19 in the Santander FUA.
An analysis of the three waves from March 2020 to March 2021



Source: authors' own elaboration based on COVID-19 micro-data in daily records from the health authorities (Regional Government of Cantabria) and European Union Urban Atlas (2018)

Figure 6. Emerging hot-spots for COVID-19 in the Santander FUA.
An analysis of the three waves from March 2020 to March 2021



Source: authors' own elaboration based on COVID-19 micro-data in daily records from the health authorities (Regional Government of Cantabria) and European Union Urban Atlas (2018)

The most significant emerging hot-spot pattern in terms of the number of COVID-19 cases is the sporadic pattern (Table 5), with 128.72 cases per square kilometre. These areas had recurrent periods with and without cases during all three waves and are associated with continuous urban areas of Santander and Torrelavega, plus some peri-urban areas. The second biggest emerging pattern is found in intensifying areas of COVID-19 in the centre of Santander, with 3,643 cases in the period analysed. This means that this part of the city is a statistically significant hot-spot in 90% of the time slides throughout the three waves, including the last period (near March 2021).

Table 5. Emerging hot-spots for COVID-19 cases in the Santander FUA

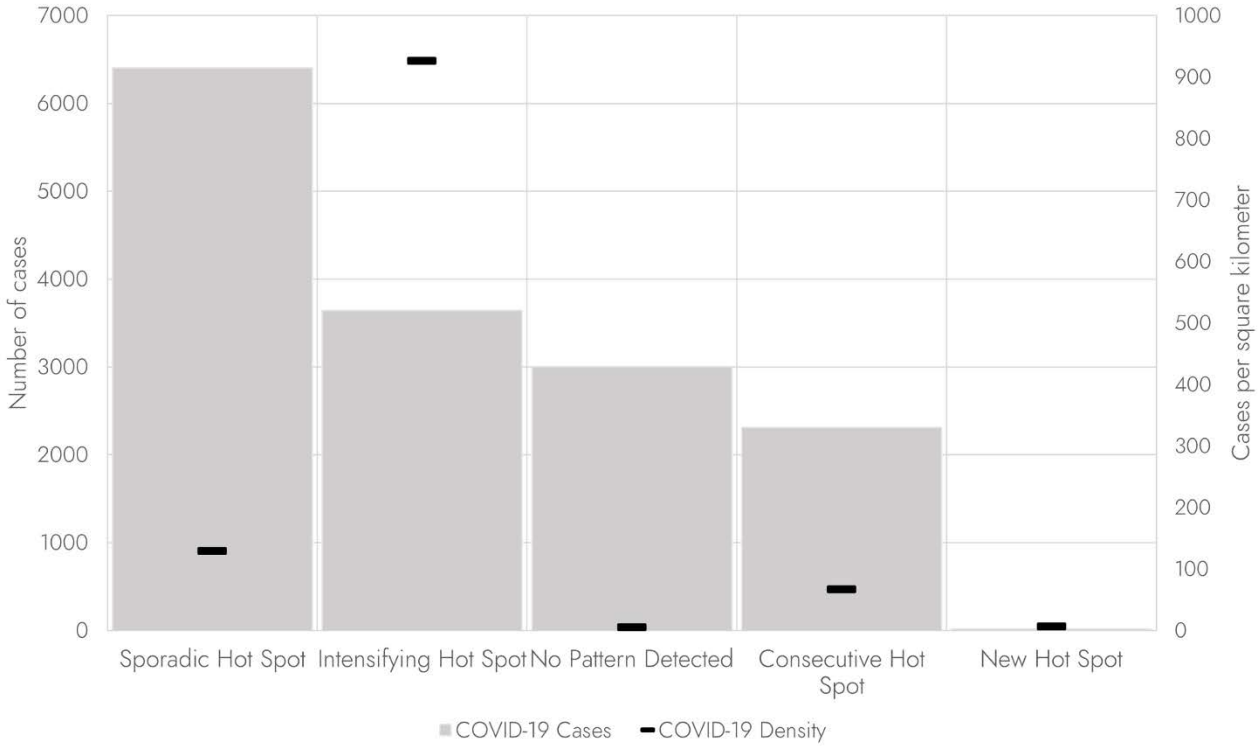
EMERGING PATTERN	Number of COVID-19 cases	Number of bins	Area in square kilometres	Mean Z-value	Mean P-value	Cumulative density of cases per km^2
Sporadic hot-spot	6,405	3,085	49.76	2.7353	0.0109	128,72
Intensifying hot-spot	3,643	244	3.94	2.6037	0.0096	925,68
No pattern detected	3,002	36,818	593.84	1.0841	0.2025	5,06
Consecutive hot-spot	2,311	2,160	34.84	2.8835	0.0067	66,33
New hot-spot	16	159	2.56	3.0658	0.0027	6,24
Total emerging	15,377	42,466	684.93	2.4745	0.0465	22,45

Key: The P-value shows the probability of a random pattern. Low values as the table presented can be interpreted as non-random distribution of COVID-19 cases. The Z-value measures the trend based on standard deviation and extreme values refer to the edges of normal distribution. Between them a low P-value and a high Z-value denote a non-random pattern. Indeed, all significant patterns —except the "No pattern detected" category (with a higher P-value and lower Z-value)— are non-random.

Source: authors' own elaboration based on COVID-19 micro-data in daily records from the health authorities (Regional Government of Cantabria)

The results show the difference in the density of cases in the different patterns. Many contrasts appear in relation to the number of cases and their spatial distribution (Figure 7). The two main patterns have contrasting dimensions in their numbers of cases and densities. Indeed, sporadic hot-spots show a low cumulative density in comparison to the high volume of cases (the area is large) while intensifying hot-spots —in small areas with many cases— are interesting due to their relatively large number compared to the number of cases per square kilometre.

Figure 7. COVID-19 total cases and densities
by emerging hot-spots patterns in the Santander FUA



Source: authors' own elaboration based on COVID-19 micro-data in daily records from the health authorities (Regional Government of Cantabria)

4.2 Evidence of the spatial patterns of COVID-19 on an intra-urban scale. An analysis based on building characteristics

This section focuses on revealing evidence on an intra-urban scale. To that end it is essential to consider the results obtained from the building variables analysis and their link to COVID-19 cases. This section is thus framed in the field of urban health research, as stated in the Introduction. From a geographical perspective it seeks to analyse the link between built space and virus incidence.

The Santander FUA shows 53,584 buildings from the Cadastral Register source in SITAR, but our study is based on those buildings where there is at least one residential dwelling. Our results are therefore based on 39,787 buildings totally or partially for residential use. As Table 6 shows, in the period considered 4,786 buildings presented at least one case, and the cumulative average was 3.21 cases per building. However, there are substantial differences in COVID-19 incidence at building level (standard deviation 4.42): the number of cases ranges from 89 in the building

with the highest cumulative total for the year considered to just 1 case. Indeed, 2,207 (42.35%) of the buildings in the Santander FUA affected by COVID-19 only present one case.

Table 6. Initial data on buildings and COVID-19 in the Santander FUA

TOTAL BUILDINGS	Residential buildings	Buildings with COVID-19 cases	Average cumulative cases per building	Standard deviation in cases per building
53,584	39,787	4,786	3.21	4.42

Source: authors' own elaboration based on COVID-19 micro-data in daily records from the health authorities (Regional Government of Cantabria) and the Spanish Cadastral Register Service

Having in consideration the contrast in incidence at building level, we focus our research on these buildings that were hit hardest by the pandemic and identifying variables related to COVID-19 incidence.

An analysis of many linear bivariate correlations revealed that there is little or no association between the number of COVID-19 cases as a dependent variable and other independent or explanatory variables, such as number of inhabitants, ratio of inhabitants per dwelling, properties used for economic activities (offices, retail, etc.), surface area of buildings, year built, properties of residences, square meters of useful area per dwelling, incomes from the ESRI ArcGIS Geo-Enrichment Service, etc. We suspect that many variables behave in a non-stationary manner in the territory, which means that bivariate linear correlation is often not expressive or even misleading.

Consequently, as indicated in the Methodology section, the analysis includes other perspectives which are more advanced than common linear bivariate coefficients. Two main lines of results emerge, as outlined below.

Firstly, in regard to high-incidence buildings and characteristics possibly associated with the presence of more cases, our analysis identifies 20 buildings with at least 30 cases counted in the first year. At various times these buildings accumulated a total of 848 cases, which means that nearly 6% of the total cases for the Santander FUA occurred in 0.4% of the buildings with COVID-19 cases. Table 7 shows the difference in residential and functional contexts in terms of building characteristics in the areas with and without COVID-19. It is even more expressive when the average pattern is considered for high-incidence buildings (at least 30 cases). The main result is that a large number of virus cases coincide with neighbourhoods where residential and

economic uses are found in the same building. Indeed, an average of 19.73% of the properties in buildings with a high COVID-19 incidence are used for economic activities, compared to just 1.72% of those in buildings with no COVID-19 cases. High-incidence buildings have an average of 6.45 commercial premises compared to practically none in buildings with no cases. Similarly, the number of storage facilities and car parks in high-incidence buildings is four times greater than in other buildings with COVID-19 cases. Similar patterns are obtained in relation to offices, leisure, catering, cultural and industrial properties, among others.

It is also noteworthy that the number of dwellings per building is much higher in high-incidence buildings (81.45) than in other buildings with COVID-19 cases (18.16) and in buildings with no cases (3.64). However, the results presented here do not seek to establish a bivariate correlation or a causal link between certain activities and virus incidence. They are an overall approximation of the idea of building context in places where infected people live and work. Therefore, we highlight the higher incidence in buildings located in neighbourhoods where people can buy, work, do official business and enjoy free time, while areas with only residential use may occasionally have cases but do not fall under the high incidence pattern.

The specific locations and addresses of high-incidence buildings are not published here to maintain the confidentiality commitment required under the permission given by the Medical Ethics Committee of Cantabria (CEIm, ID: 2020.238). The results reported here show only general patterns and type case studies (referring to specific buildings) without revealing identifying data. We therefore refer to previous profiles as emerging hot-spot patterns where high-incidence buildings are located. Specifically, 19 of the 20 buildings in question are in significant emerging hot-spot patterns. The predominant pattern is sporadic hot-spots: this heading accounts for 50% of high-incidence buildings and 51.5% of cases in high-incidence buildings (Figure 8). It must be remembered, as shown in Table 5, that sporadic hot-spots make up the main pattern in terms of the number of cases in the whole Santander FUA and the second biggest in terms of surface area at 49.76 km^2 (behind no pattern areas). Secondly, five buildings with a total of 218 cases are in intensifying areas, and thirdly four high-incidence buildings are in consecutive hot-spots (Figure 8). It is worth highlighting that 19 of the 20 high-incidence buildings are in significant emerging pattern areas; only one building, with 37 cases, is in an area where no significant pattern was detected.

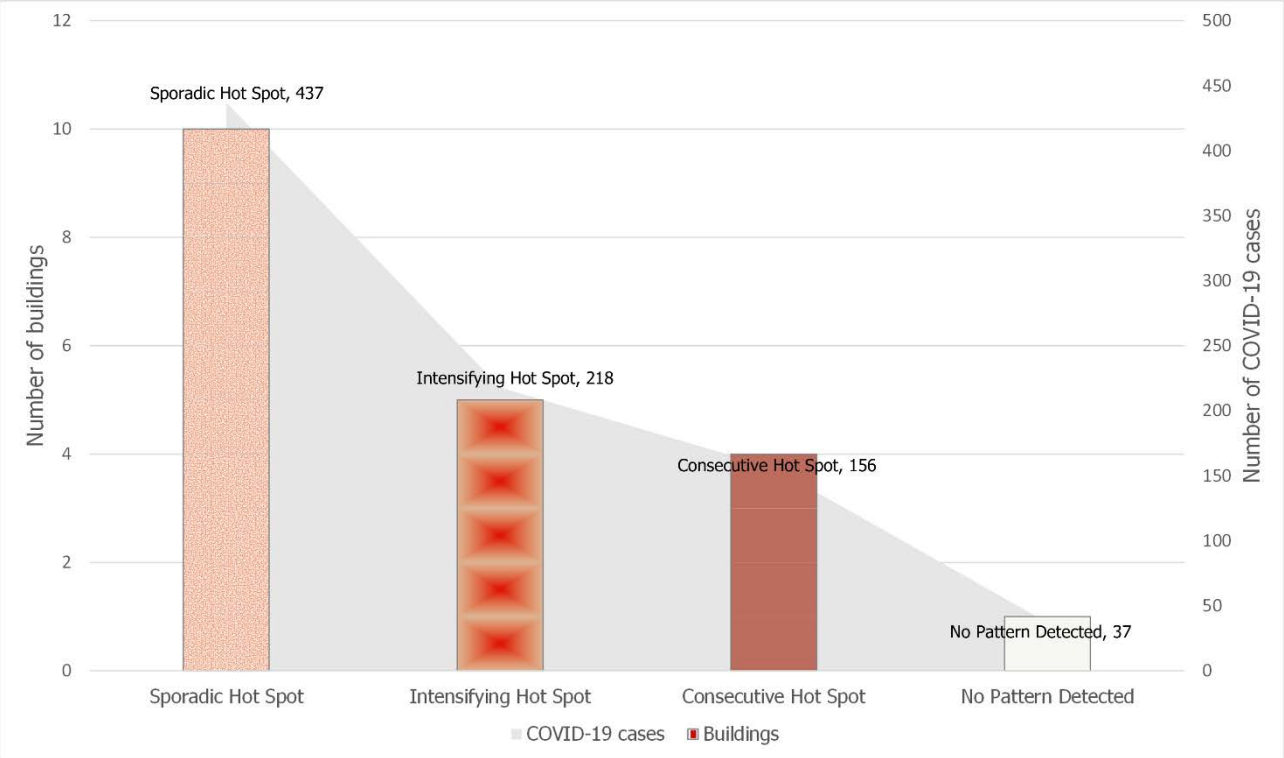
Table 7. Characteristics of residential buildings in the Santander FUA in relation to the incidence of COVID-19

VARIABLE	High-incidence buildings	Rest of buildings with cases	Buildings with no cases
Average inhabitants per dwelling	2.00	1.94	1.93
Percentage of properties that are not dwellings	19.73	16.01	1.72
Average number of dwellings per building	81.45	18.16	3.64
Average square meters per dwelling	113.20	166.27	228.04
Average number of commercial properties per building	6.45	1.36	0.20
Average number of storage facilities and car parks per building	46.55	10.82	1.51
Average number of office properties per building	1.90	0.26	0.03
Average number of leisure and catering properties per building	0.05	0.01	0.00
Average number of healthcare and charity properties per building	0.10	0.02	0.00
Average number of cultural properties per building	0.30	0.03	0.00
Average number of industrial properties per building	0.10	0.05	0.03

Key: High-incidence buildings are filtered as cases with approximately ten times the average number of cases per building (as shown in Table 6: 3.21 cases), which means that buildings with 30 cases or more are included.

Source: authors' own elaboration based on COVID-19 micro-data in daily records from the health authorities (Regional Government of Cantabria) and Spanish Cadastral Register Service.

Figure 8. Emerging pattern of COVID-19 cases
in high-incidence buildings in the Santander FUA

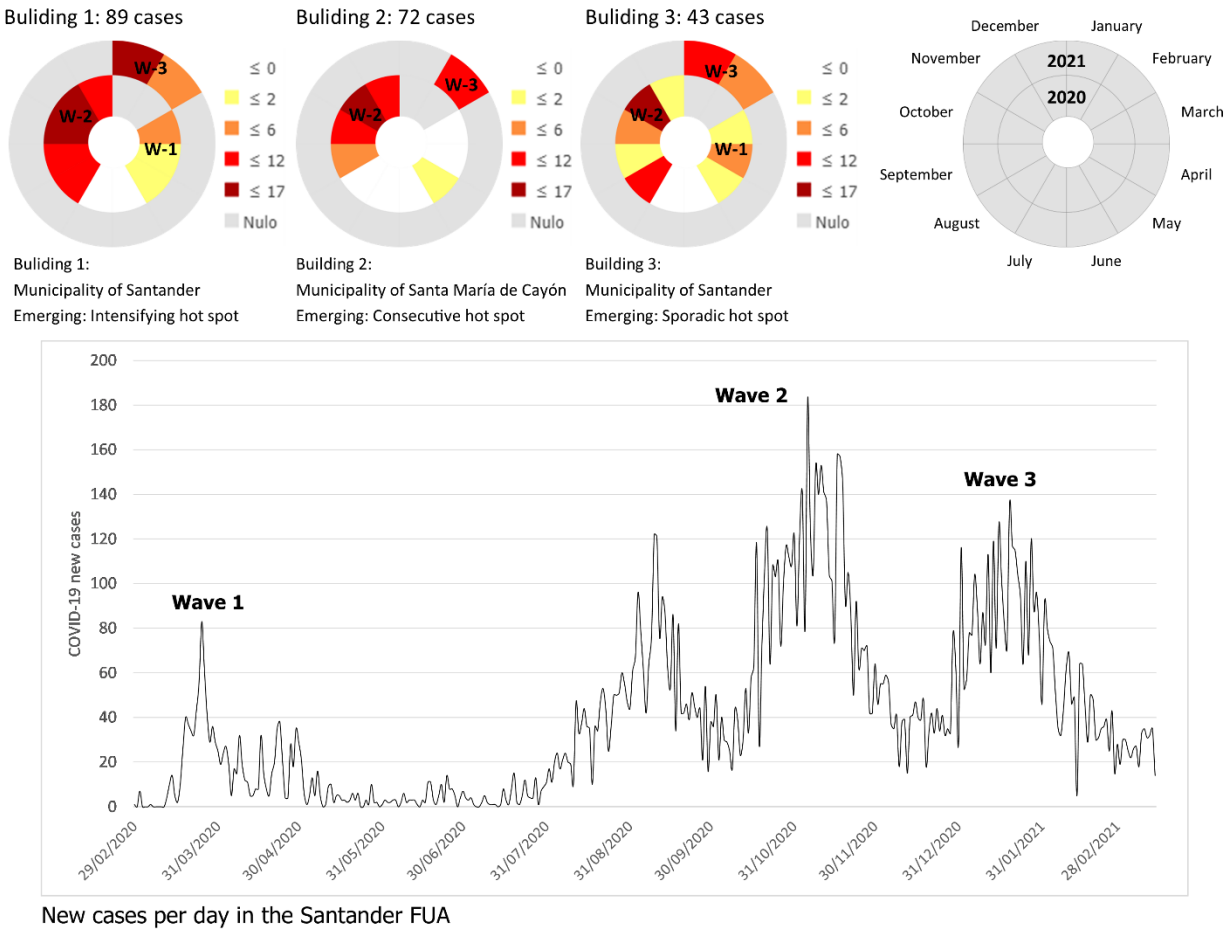


Source: authors' own elaboration based on COVID-19 micro-data in daily records from the health authorities (Regional Government of Cantabria) and Spanish Cadastral Register Service

Secondly, high-incidence buildings show expressive patterns in timing that can be interpreted as internal temporal patterns at building level. In fact, except for the period closest to the "new normal" stage that followed the strict lockdown in Spain (June and part of July) it is possible to identify a pattern per month in relation to the presence of positive cases in these buildings (Figure 9). The general trend in the FUA can also be perceived at building level, especially in case of high-incidence buildings. The examples shown in the following figure have in common an absence of cases in June and July and the reiterative presence of cases over several months. They all report new positive cases each month after summer 2020, with a large number of cases coinciding with the second wave (in the last interval of the legend) and in the third wave.

This micro-scale detail helps to show how the general trend in the Santander FUA waves affects certain buildings at micro-scale level.

Figure 9. Temporal trend in new COVID-19 cases in the Santander FUA. Example of three high-incidence buildings month by month from March 2020 to March 2021



Source: authors' own elaboration based on COVID-19 micro-data in daily records from the health authorities (Regional Government of Cantabria) and Spanish Cadastral Register Service

At building level, considering the disparities between buildings in the number of cases and the absence of significant links with other variables accepted on other scales, we check for non-stationarity of COVID-19 cases with other explanatory variables. We thus consider many variables related to each building (residential and functional characteristics) and conduct an exploratory analysis based on Ordinary Least Square analysis (OLS), as explained in the Methodology section. The model shows expected values between 0 and 1, more precisely 0.172 and 0.169 in multiple R-squared and adjusted R-squared, respectively. In the model explored, the independent variables thus explain about 17% of spatial variation in COVID-19.

In this regard, the statistical significance of the Koenker Index ($p<0.01$) must be considered, which implies that the links between variables are not consistent because of non-stationarity. Therefore, the behaviour and closeness of the link between the independent building variables

and the number of COVID-19 cases changes depending on the spatial framework. However, this does not prevent the model from showing overall significance, according to the Wald Index ($p<0.01$).

For the exploratory assessment of each variable, we base our selection on three main statistics: coefficient, robust probability (we dismiss only the probability parameter due to non-stationarity) and the variance inflation factor (VIF). Dispersion diagrams are also analysed to complete the information on the links. Looking at economic activities, we obtain a coefficient that is acceptable in the presence of cultural venues (0.821), healthcare and charity premises (0.41) and, to a lesser extent, sports facilities (0.23) and religious properties (0.23). The robust probability test gives significant results in the percentage of residential and non-residential properties. It seems that the mix of residential and business properties is important in the spread and incidence of the virus. Finally, the exploratory OLS analysis identifies several variables as redundant, with figures of more than 7.5: commercial properties (32.20), population (25.94) and useful area in square meters per dwelling (17.66), among others.

Estimating a predictive model at building level goes beyond the goals of this research, but the OLS results make some interesting contributions in terms of analysing the spatial behaviour of the virus on a micro-scale and confirm non-stationarity as an important characteristic in the analysis of the pandemic from a geographic perspective on an intra-urban scale.

5 Discussion

In line with the results presented here, we first argue that it is important to analyse residential areas for positive cases. The micro-data geo-coding is related to the address of each person who has tested positive for COVID-19, although the contagion may have originated elsewhere. We admit as a limitation that we have no information on the precise location of outbreaks, but we argue that the spatial pattern and spread of the virus differ depending on the characteristics of the areas where people live. It is also important to consider where people are in the study period, with the constraints on movement and social distancing rules imposed by the health authorities. In this regard, a mobility report (Google, 2021) confirms that people in Cantabria spent more time than in the pre-pandemic reference period in residential areas (+4%), parks (+15%) —often near their home neighbourhoods— and on essential purchases (+11% in supermarkets and pharmacies). These data confirm the presence of new urban behaviour patterns and a return to a proximity-based city with journeys under 15 minutes, now linked to the pandemic, with a simpler relationship between people and urban spaces (Marin & Palomares,

2020). Shorter distances and the distinction between essential and non-essential activities are basic points. Thus, leisure and shopping areas, transport stations and workplaces (with the new tele-working framework) have fewer people than before the pandemic. Consequently, the patterns obtained are for a period when there is more use of residential areas and living spaces, and shorter journeys. This makes the analysis of data of places where infected people live and those nearby particularly interesting.

In regard to the role of residential areas or neighbourhoods in the pandemic, some authors have made sound contributions based on the hypothesis of neighbourhood contagion, constituting a cluster focus that needs to be considered as a relevant unit for diagnosis (Perles et al., 2021). If people tend to spend more time in residential areas, the first key point towards contributing from a geo-prevention approach is to consider the areas where COVID-19 patients live. So, residential areas are relevant for research over waves. In this regard, our study confirms the importance of additional variables other than population density in relation to COVID-19 cases. Population density seems relevant on global and national scales, but its influence is lower and fuzzier on more detailed scales.

In fact, in the Santander FUA we show that the discontinuous urban fabric has more cases than theoretically expected for the number of people who live there. Other authors agree on the fuzzy role of population densities at intra-urban level and develop other approaches using, for instance, the venue density in terms of venues in buildings visited by people confirmed as having the virus (Huang et al., 2020) or the presence of urban vegetation associated with lower densities as ways to reduce virus spread (You & Pan, 2020). It is necessary to point out that Huang et al. (2020) had tracking contact data to calculate venue density. This source enables them to make a clear distinction in their analysis depending whether cases are imported or local. They find a close relationship between the built environment and COVID-19 risk. This is a very interesting approach, but in our case we do not have access to contact tracking data. Although our study at building level is linked to these research lines, we take an indirect approach due to source limitations, so our analysis is based on the presence of venues in residential buildings, even though we cannot check whether Covid-positive individuals have visited them. As mentioned before, other relevant studies have endorsed a similar hypothesis to ours in relation to the importance of residential areas in COVID-19 cases and nearby activities (Huang et al., 2020; Perles et al., 2021).

Empirical analysis in intra-urban research into COVID-19 includes many variables. Depending on availability and access permissions it is possible to obtain a wide range of research datasets to analyse the distribution of the virus. Some authors analyse the distribution of COVID-19 on intra-urban scales in relation to other co-morbidity-related variables (Zúñiga et al., 2020; Mansour et al., 2021). This is an interesting approach, but in our case we do not have access to comorbidity data on our scale of analysis. Other authors consider imprecise topics such as urban green areas with a twin role in relation to virus incidence and sprawl. Some studies identify a beneficial effect of urban green areas in decreasing spread with lower population densities (You & Pan, 2020) but other authors use contact tracking data to show that green areas tend to attract more visitors more frequently, which is risky from the point of view of possible contagion (Huang et al., 2020).

In regard to results for emerging hot-spots, we consider our analysis to be revealing from a space-time perspective and a major contribution from the point of view of geo-prevention. The model that we present shows the different patterns of risk and reveals areas with recurrent patterns in the period considered. This is an strategic contribution to help policy-makers to design future rules for coexisting with the virus, because the study reveals locations with a significant presence of cases, areas with an increasing trend in the last period (new and consecutive) and areas with recurrent presence of cases (sporadic pattern). Most prospective studies seek to model future trends using geographically weighted regression (GWR) to analyse the link between COVID-19 and space (Rahman et al., 2020). We, however, argue for 3D bins and emerging hot-spots as a necessary first stage in modelling the pandemic because this method does not influence the result by selecting variables; it directly identifies problematic areas by combining space and time. GWR is widely used in spatial patterns for healthcare, which we intend to consider in further research. In any case, regression methods need to include geographic adaptation because, as demonstrated in our study with the Koenker Index ($p<0.01$), non-stationarity implies variable behaviour of COVID-19 with contextual variables depending on places, as found by other authors in health-related spatial studies (Mou et al., 2017; Rahman et al., 2020; Mansour et al., 2021). Here, we obtain another important result in relation to geo-prevention keys. If each variable related to COVID-19 distribution presents different links on intra-urban scales, it does not seem appropriate to design rules based on topics that are implemented in different territories in the same conditions. Adapting rules to specific characteristics of the areas where people live, work and relate to others is very important in tackling the pandemic with detailed measures that cater for not only health but also economic activities.

The building level and the context of areas near those with COVID-19 are the focus of original research which considers new approaches, including some related to urban landscape, as Nguyen et al. (2020) state in their study using Google Street View images. The authors conclude that indicators of physical disorder such as dilapidated buildings and visible utility wires are associated with more cases, perhaps as an indirect measure of social and economic vulnerabilities. In any event, studies at building level are an original approach, as presented here, to focus on high-incidence buildings and analyse the characteristics of the urban context. In that regard, our most revealing result is that there are more cases in buildings where residential functions and economic activities share the same space. This coincides with the conclusions of Huang et al. (2020), who analyse the close links between environmental built variables and COVID-19 risk in Hong Kong and find that land-use mix (diversity) and accessibility are positively linked to virus incidence.

One limitation of our research in relation to building level analysis is that in the case of the Santander FUA we cannot check on what happens inside buildings or analyse the hypothesis of internal contagion between neighbours. So, our SITAR tool cannot look in more depth at the controversial hypothesis of Hwang et al. (2021) because we do not have floor number data; our micro-data records have only a field for geo-location (street, number). Even so, our geo-statistical analysis can help health authorities to inspect suspicious buildings and study these cases in depth. In that regard, epidemiological advances have now provided society with a great deal of knowledge of the short-range transmission of COVID-19, but it is also necessary to consider previous and present research into indoor transmission and cross infection between different rooms, and between different apartments due to the dispersion routes of airborne pathogens (Xu et al., 2020). The selection of high-incidence buildings and corresponding calendar diagrams focus the attention of policy-makers on specific buildings as a key for geo-prevention at building level. It is true that considering many variables linked to urban areas could make the analysis less representative in low-density locations or peri-urban areas, as stated by certain authors who consider in their models that the risk of COVID-19 transmission is underestimated in suburban areas due to the significant presence of open and green areas (Huang et al., 2020). This same effect could be attributed to our research, because the major spatial patterns are in urban and peri-urban areas with high commuting and urban land uses, while rural areas remain in the background. Nevertheless, this urban-centric approach is widespread and common in most methods and results cited in this paper.

6 Conclusions

Non-random distributions of COVID-19 cases with cluster patterns are found repeatedly in exploratory spatial analysis from the first to the third wave of infection throughout the Regional Autonomous Community of Cantabria and the Santander FUA with a multi-scale perspective. It supports our empirical findings using geotechnologies.

This study contributes to multi-scale knowledge of the virus. Taking into consideration the multi-scale behaviour of the virus and its spread and the strong influence of the social framework, we argue that COVID-19 needs to be analysed based on micro-data geo-coding in relation to characteristics of the building where positive tested individuals live and nearby locations.

The contribution of our study to knowledge of the spatial patterns of the virus is first of all prospective, given the predictive potential of the data mining methods used (3D bins and emerging hot-spot analysis). Previous research based on 3D bins and emerging hot-spots (De Cos et al., 2021) demonstrates that more than 80% of new cases were in statistically significant previous emerging hot-spots. Therefore, we understand our findings as a prelude of problems that are re-emerging in areas in the future. In this regard, we expect new concentration of COVID-19 cases where our maps show new, consecutive, intensifying, and sporadic significant hot-spots.

Secondly, we identify high-incidence buildings and reveal the pattern of mixed-use as a characteristic that contributes to risk. Both these results are interesting from the point of view of geo-prevention and pandemic management. The emerging hot-spots model is a strategic model for identifying possible areas at risk in future waves, in line with statistical significance and past and recent trends. However, studying high-incidence buildings and their context enables us to identify buildings where authorities could make major decisions related to investigating the contagion process, intensifying cleaning work and monitoring or even, in extraordinary circumstances, planning vaccination campaigns based on a spatial diagnosis of COVID-19 risk areas.

We consider that geo-prevention is an essential approach to cope with the multi-scale behaviour of the pandemic, in terms of thinking globally but acting locally (Salama, 2020) with rules coordinated at regional level and supported by municipalities.

We admit that it is not easy to make decisions about parts of a territory that do not coincide with administrative limits. On the other hand, some regional governments have had to make similar decisions to establish perimeter lockdowns and other limitations.

Acknowledgements: This research was funded by the Government of Cantabria (Spain) under grant number UC: 10.3834.64001 "Assistance in adapting Cantabria to the plan for the transition to a new normal in times of Covid-19: socioeconomic contributions" from the University of Cantabria-IDIVAL Valdecilla-Government of Cantabria. The APC was funded by the research project "Stress or resistance test in the Health System of Cantabria, development of innovative digital technologies to model scenarios of greater health utilization and socioeconomic and human impact solutions against COVID-19" INNVAL20/03 (IDIVAL).

Authorship statement: The authors declare no conflict of interest. The participation of the authors in the article is as follows. Conceptualisation, methodology and spatial analysis, Olga De Cos; visualisation and publishing (graphs and maps), Valentín Castillo & Olga De Cos; writing and writing-review Olga De Cos & David Cantarero; supervision, Olga De Cos; securing of funding, David Cantarero.

References

Batista, F., & Poelman, H. (2016). *Mapping Population Density in Functional Urban Areas. A Method to Downscale Population Statistics to Urban Atlas Polygons* (JRC Technical Reports). European Commission. https://doi.org/10.2791/06831

Brizuela, N. G., García-Chan, N., Gutiérrez, H., & Chowell, G. (2021). Understanding the role of urban design in disease spreading. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, 476,* 20200524. https://doi.org/10.1098/rspa.2020.0524

Brundson, C. A., Fotheringham, A. S., Charlton, M. (1996). Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical Analysis, 28*(4). 281-298. https://doi.org/10.1111/j.1538-4632.1996.tb00936.x

Brunsdon, C., Fotheringham, A. S., Charlton, M. (2002). Geographically weighted summary statistics — a framework for localised exploratory data analysis. *Computers, Environment and Urban Systems, 26*(6), 501-524. https://doi.org/10.1016/S0198-9715(01)00009-6

Campagna, M. (2020). Geographic Information and Covid-19 outbreak. Does the spatial dimension matter? *Tema. Journal of Land Use, Mobility and Environment, special issue,* 31-44. https://doi.org/10.6092/1970-9870/6850

Chirico, F., Sacco, A., Bragazzi, N. L., & Magnavita, N. (2020). Can air-conditioning systems contribute to the spread of SARS/MERS/COVID-19 infection? Insights from a rapid review of literature. *International Journal of Environmental Research and Public Health, 17*(17), 6052. https://doi.org/10.3390/ijerph17176052

De Cos, O., Castillo, V., & Cantarero, D. (2020). Facing a Second Wave from a Regional View: Spatial Patterns of COVID-19 as a Key Determinant for Public Health and Geoprevention Plans. *International Journal of Environmental Research and Public Health, 17*(22), 8468. https://doi.org/10.3390/ijerph17228468

De Cos, O., Castillo, V., & Cantarero, D. (2021). Differencing the risk of reiterative spatial incidence of COVID-19 using space-time 3D bins of geocoded daily cases. *International Journal of Geo-Information, 10,* 261. https://doi.org/10.3390/ijgi10040261

Eichler, N., Thornley, C., Swadi, T., Devine, T. Mackelnay, C., Sherwood, J., Brunton, C., Williamson, F., Freeman, J., Berger, S., Ren, X., Storey, M., de Ligt, J., & Geoghegan, J.L. (2021). Transmission of severe acute respiratory syndrome Coronavirus 2 during border

quarantine and air travel, New Zealand (Aotearoa). *Emerging infectious diseases, 27*(5). https://doi.org/10.3201/eid2705.210514

Evans, P. J., & Evans, F. C. (1954). Distance to nearest neighbour as a measure of spatial relations in populations. *Ecology, 35,* 445-453.

Fatima, M., O'Keefe, K. J., Wei, W., Arshad, S., & Gruebner, O. (2021). Geospatial analysis of COVID-19: A scoping review. *International Journal of Environmental Research and Public Health, 18*(2336). https://doi.org/10.3390/ijerph18052336

Franch-Pardo, I., Desjardins, M., Barea-Navarro, I., & Cerdà, A. (2021). A review of GIS methodologies to analyze the dynamics of COVD-19 in the second half of 2020. *Transactions in GIS, 00,* 1-49. https://doi.org/10.1111/tgis.12792

Greenhalgh, T., Jimenez, J. L., Prather, K.A., Tufekci, Z., Fisman, D., & Schooley, R. (2021, April 15). Ten scientific reasons in support or airborne transmission of SARS-Cov-2. *The Lancet.* https://doi.org/10.1016/S0140-6736(21)00869-2

Hamidi, S., Sabouri, S., & Ewing, R. (2020). Does Density Aggravate the COVID-19 Pandemic? Early Findings and Lessons for Planners. *Journal of the American Planning Association, 86*(4), 495-509. https://doi.org/10.1080/01944363.2020.1777891

Hernando, F. (2008). La seguridad en las ciudades: el nuevo enfoque de la geoprevención. *Scripta Nova, 12,* 270(14). http://www.ub.edu/geocrit/sn/sn-270/sn-270-14.htm

Huang, J., Kwan, M-P., Kan, Z., Wong, M. S., Tung Kwok, C. Y., & Yu, X. (2020). Investigating the Relationship between the Built Environment and Relative Risk of COVID-19 in Hong Kong. *ISPRS International Journal of Geo-Information, 9,* 624. https://doi.org/10.3390/ijgi9110624

Hwang, S. E., Chang, J. H., Oh, B., & Heo, J. (2021). Possible aerosol transmission of COVID-19 associated with an outbreak in an apartment in Seoul, South Korea, 2020. *International Journal of Infectious Diseases, 104,* 73-76. https://doi.org/10.1016/j.ijid.2020.12.035

Gargiulo, C., Gaglione, F., Guida, C., Papa, R., Zucaro, F., & Carpentieri, G. (2020). The role of the urban settlement system in the spread of Covid-19 pandemic. The Italian case. *Tema. Journal of Land Use, Mobility and Environment,* 189-212. https://doi.org/10.6092/1970-9870/6864

Getis, A. (1992). The Analysis of Spatial Association by Use of Distance Statistics. *Geographical Analysis, 24*(3).

Google (2021). Google Mobility Report 2021-04-29 Cantabria (Spain). https://www.gstatic.com/covid19/mobility/2021-04-29_ES_Cantabria_Mobility_Report_es.pdf

Hosseini, M.R., Fouladi-Fard, R., & Aali, R. (2020). COVID-19 pandemic and sick building syndrome. *Indoor and Built Environment, 29*(8), 1181-1183. https://doi.org/10.1177/1420326X20935644

Kendall, M. G., & Stuart, A. (1976). *The Advanced Theory of Statistics. Distribution Theory, 1.* Griffin.

Li, Y., Duan, S., Yu, I. T. S., & Wong, T.W. (2004). Multi-zone modelling of probable SARS virus transmission by airflow between flats in Block E, Amoy Gardens. *Indoor Air, 15*, 96-111. https://doi.org/10.1111/j.1600-0668.2004.00318.x

Mansour, S., Al Kindi, A., Al-Said, Al., Al-Said, Ad., & Atkinson, P. (2021). Sociodemographic determinants of COVID-19 incidence rates in Oman: Geospatial modelling using multiscale geographically weighted regression (MGWR). *Sustainable Cities and Society, 65*, 102627. https://doi.org/10.1016/j.scs.2020.102627

Marín-Cots, P., & Palomares-Pastor, M. (2020). En un entorno de 15 minutos. Hacia la ciudad de proximidad, y su relación con el Covid-19 y la crisis climática: El caso de Málaga. Ciudad y Territorio. *Estudios Territoriales, LII, 205*, 685-700. https://doi.org/10.37230/CyTET.2020.205.13.3

Moran, P. (1948) The Interpretation of Statistical Maps. *Journal of the Royal Statistical Society, 10*, 243-251.

Mou, Y., He, Q., & Zhou, B. (2017). Detecting the spatially non-stationary relationships between housing price and its determinants in China: Guide for housing market sustainability. *Sustainability, 9*(10), 1826. https://doi.org/10.3390/su9101826

Nguyen, Q. C., Huang, Y., Kumar, A., Duan, H., Keralis, J. M., Dwivedi, P., Meng, H-W., Brunisholz, K. D., Jay, J., Javanmardi, M., & Tasdizen, T. (2020). Using 164 million google street view images to derive built environment predictors of COVID-19 cases. *International Journal of Environmental Research and Public Health, 17*(17), 6359. https://doi.org/10.3390/ijerph17176359

Perles, M.-J., Sortino, J. F., & Mérida, M. F. (2021). The neighborhood contagion focus as spatial unit for diagnosis and epidemiological action against COVID-19 contagion in urban spaces: a

methodological proposal for its detection and delimitation. *International Journal of Environmental Research and Public Health, 18*, 3145. https://doi.org/10.3390/ijerph18063145

Pinter-Wollman, N., Jelic, A., & Wells, N. M. (2018). The impact of the built environment on health behaviours and disease transmission in social systems. *Philosophical Transactions R. Soc. B., 373*, 20170245. http://dx.doi.org/10.1098/rstb.2017.0245

Rahman, M. H., Zafri, N. M., Ashik, F. R., & Waliullah, M. (2020). GIS-based spatial modelling to identify factors affecting COVID-19 incidence rates in Bangladesh. *Medrxiv. The preprint server for health sciences*. https://doi.org/10.1101/2020.08.16.20175976

Roy, A., & Kar, B. (2020). Characterizing the Spread of COVID-19 from Human Mobility Patterns and Sociodemographic Indicators. In *3rd ACM SIGSPATIAL Workshop on Advances in Resilient and Intelligent Cities (ARIC'20)*, November 3–6, Seattle, WA, USA. ACM, New York, NY, USA. https://doi.org/10.1145/3423455.3430303

Salama, A. M. (2020). Coronavirus questions that will not go away: interrogating urban and socio-spatial implications of COVID-19 measures. *Emerald Open Research, 2*(14). https://doi.org/10.35241/emeraldopenres.13561.1

Xu, C., Luo, X., Yu, C., & Cao, S. J. (2020). The 2019-nCoV epidemic control strategies and future challenges of building healthy smart cities. *Indoor and Built Environment, 29*(5), 639-644. https://doi.org/10.1177/1420326X20910408

You, Y., & Pan, S. (2020). Urban vegetation slows down the spread of coronavirus disease (COVID-19) in United States. *Geophysical Research Letters, 47*. https://doi.org/10.1029/2020GL089286

Zúñiga, M., Pueyo, A., & Postigo, R. (2020). Herramientas espaciales para la mejora de la gestión de la información en alerta sanitaria por COVID-19. *Geographicalia, 72*, 141-145. https://doi.org/10.26754/ojs_geoph/geoph.2020725005