UNDERSTANDING THE SPATIAL EMPLOYMENT DISTRIBUTION IN BARCELONA METROPOLITAN AREA

MSc GIMA Thesis

02/03/2023

Author:

Ferran Oró Arán (Id: 0706124)

f.oroaran@students.uu.nl

Supervisors:

Jesús Balado Frias (TUD)

Rafa Madariaga Sánchez (UVic-UCC)

Raymond Lagonigro Bertran (UVic-UCC)

GIMA Geographical Information Management and Applications

TU Delft Delft University of Technology

Universiteit Utrecht

UVIC UNIVERSITAT DE VIC UNIVERSITAT CENTRAL DE CATALUNYA

WAGENINGEN UNIVERSITY & RESEARCH

ITC UNIVERSITY OF TWENTE.

ACKNOWLEDGEMENTS

I would like to express my gratitude to the following individuals for their help with this research: Raymond Lagonigro and Rafa Madariaga for their guidance and mentorship throughout the project; Jesus Balado for his supervision during all the research process; and the GIMA MSc program for giving the opportunity of doing this Thesis and providing me the necessary tools for making it possible.

I would also like to thank my family and friends for their support during all the research and writing process. Also, to the Data Analysis and Modelling group and any other involved person for their help during the Thesis.

Finally, I would like to acknowledge that the author declares no conflicts of interest that could have influenced the findings or conclusions presented in this paper.

ABSTRACT

The spatial distribution of population, economic activity and income is important for public policies and the provision of public goods in cities. Understanding how employment is distributed in space has become a main factor when analysing the urban structure of a certain area/region. There has been a traditional lack of data on the spatial distribution of employment in Spain, and therefore, Catalonia. This project aimed to address this data gap by geocoding Catalan companies using different geocoders and techniques to obtain the first companies' geocoded data frame of Catalonia. It has revealed how geocoding process can be complex and the importance of having a clean and well-structured dataset before processing. Furthermore, to get a better understanding of the employment distribution in the BMA, the project has joined the employment data with socioeconomic data from census tracts. Global Moran's I and LISA analyses have been performed to comprehend the employment distribution with its most significant working clusters. While this research has been the initial step to accurately understand the employment distribution in the BMA, several tools have been provided that help to perform similar analysis and get significant and comparable results for more precise sectoral employment classifications. These calculations can be done alongside with the definition of the employment urban models: monocentric or polycentric.

TABLE OF CONTENTS

1. Introduction and Background Information

The spatial distribution of population, economic activity and income is important for public policies and the provision of public goods in cities. Understanding how employment is distributed in space has become a main factor when analysing the urban structure of a certain area/region. There has been a traditional lack of data on the spatial distribution of employment in Spain, and therefore, Catalonia. This project aims to improve the knowledge on the spatial structure of economic activity in Catalonia and help assessing employment distribution of Barcelona's metropolitan area (BMA).

The project is divided in two main parts, one which stands for creating a dataset with the geolocated addresses of all the employment centres in Catalonia (for the first time), and another which stands for studying the spatial distribution of employment in BMA and the role it plays in determining the urban economy.

The world in which we live nowadays is led by innovation, giving high importance in creating, exploiting and maintaining knowledge (Figueiredo & Pereira, 2017). Thus, it is important to be able to work with proper knowledge in order to keep this value creation for the good of the society. The acquired knowledge can be stored in datasets. Datasets help in giving people easier access to data and increases the possibility of analysing the knowledge that is stored in them. If researchers analyse data from datasets they can yield conclusions out of its information (Frankenfield, 2022). As part of this project, the desired outcomes that will be discussed later, largely rely on the dataset structure. Consequently, it is important to make the best effort in preparing a solid dataset which can be used to obtain trustworthy conclusions in this innovative study. The obtained dataset has specific locations based on addresses, which indicate the location by using predetermined directives. In order to make the correct data analysis cited before, it is crucial to convert the addresses into coordinates (Latitude and Longitude). This conversion process is known as geocoding (Küçük Matci & Avdan, 2018). As an example, Hellner (2021) shows how geocoding, in this case using Google Maps, can be used in order to obtain location points from addresses.

There are plenty of methods for geocoding addresses, the article of Prener (2021) provides a list of several geocoding services and performance comparisons between them. The global geocoders that appear in the article are ArcGIS (ESRI), Bing, Geocodio, Google, HERE, OpenCage, TomTom and OSM (Nominatim). Other geocoders like the user-friendly Batchgeo can be added to the list (Duncan et al., 2011). The most meaningful ones for the researchers in terms of availability and payment methods are described in more detail on the Methodology section.

To further analyse the data obtained by the geocoding processes with the best accuracy there is the need to clean and preprocess the dataset. This can be done via different programming languages such as R, MATLAB or Python. Python has increased its usage for data manipulation, providing powerful libraries that help to this task, like Pandas. Pandas is a library built on top of the NumPy package, and supplies tools for working with structured data in a easy, fast and expressive way. Moreover, NumPy (shorten form of Numerical Python) is a crucial package that can use functions to perform element-wise computations with arrays, read and write array-based data sets to disk, etc. (McKinney, 2013). R is a free and open-source software for statistical computing and graphics, and one of the most used worldwide. It provides numerous statistical analysis methods that can be self-created, copied or adapted. Lately, R developers have developed packages that include spatial analysis in R. Thus, it has been a key tool for GIS analysts. Packages like *sp*, *rgdal* and *sf* are the most used in R for this purpose (Bivand et al., 2008).

A crucial population change started around 1950 and continues in the present. The beginning of this transition took place when there was an increment in population in the developed and developing countries of the globe, which has been denominated as the demographic transition. In 1950 the estimated global population was about 2,5 thousand million, in 1990 5,3 thousand million and 7,7 thousand

million in 2019. The economists have associated this transformation to what is known as a demographic dividend. The demographic dividend stands for an increase in the working adult age groups which is connected to a faster economic growth (Abbafati et al., 2020). The economic growth has also led to what is known as deindustrialization. Industrialized rich economies have been losing their industrial nature and manufacture (Liboreiro et al., 2021). This economic change makes these cities, like Barcelona, to become centres of service economy. If we want to understand urban economics, it is crucial to study how the employment is distributed in cities with this new configuration.

This society development, with the economic and population growth/change as the central axis, has led to different possible urban expansions. In cities, two distinguished economic urban models can be found: monocentric and polycentric. The monocentric model stands for the presence of a Central Business District (CBD) in which the major part of the employment is found and a dominant part of the citizens commute. The polycentric model stands for a decentralized urban structure. In this model, the employment is distributed in several spots that act as multiurban poles and therefore they individually centralize the economic and population activity (Abozeid & AboElatta, 2021; Huai et al., 2021).

Figure 1 shows a map of Catalonia where 41 second rank cities (with more than 20.000 habitants) are located together with the Barcelona Metropolitan Area (BMA) and the boundaries of the Barcelona Metropolitan Region (BMR). A big percentage of the Catalan population (more than 40%) live in the BMA, a 30% of the population are located in the 41 second rank cities. The remaining 30% is found in the surplus towns and small cities (Madariaga et al., 2019). As stated earlier, there is a relation between the urban models and the economic growth, which relates it directly to the employment distribution. For this reason, the high percentage of Catalan population in BMA (composed by 36 municipalities) is of elevated significance and enhances the will to research about the employment distribution located in this area. There is some research in the employment distribution of the BMA. However, the research done in these studies is not well aligned with the one of this project; see Research Objective (section 2) and Methodology (section 4). For example, the articles from Coll-Martínez et al. (2017) & Maddah et al. (2021) only describe the employment distribution in cultural and creative industries. Other researches from Garcia-López & Muñiz (2010), García & Muñiz, (2005) & Marmolejo et al. (2010) are based on BMR and are based on larger geographical unit areas (municipalities).

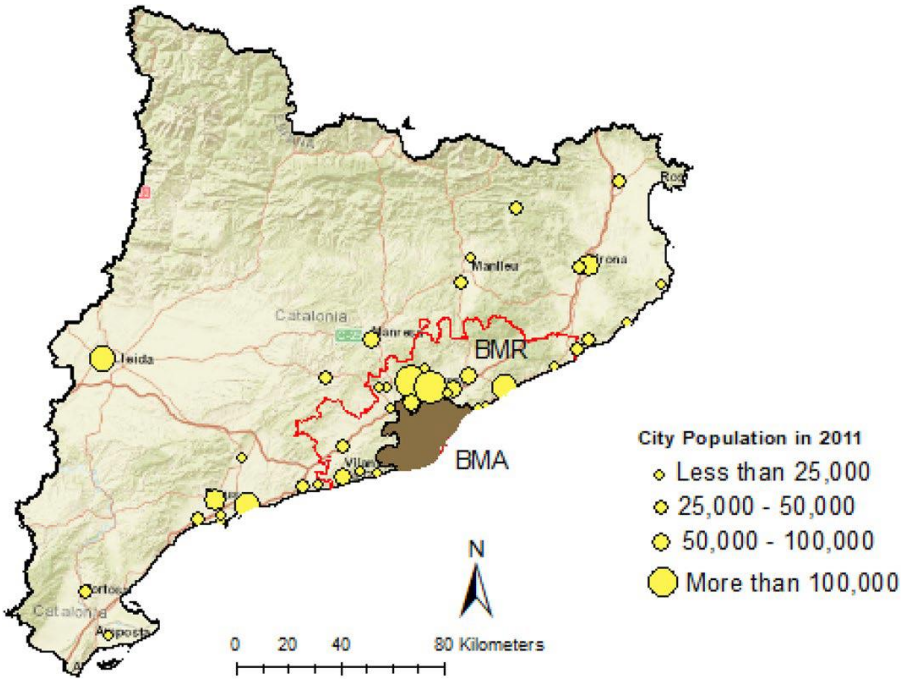

Figure 1 Location of the 41 second rank cities, BMA and BMR (Source: Madariaga et al., (2019))

To better identify unique employment centres, their locations are provided in point data. Many documents have shown the possible applications of point analysis. Bivand et al., (2008) explains how points can be analysed in two main possible ways. The first pertains to their distribution over a delimited space, while the second focuses on the possible interaction between those points. These two considerations give place to an exploratory analysis and to a more statistically rigorous one.

Moreover, it may be interesting for researchers to aggregate points inside polygons. This can be because some other data that aims to be worked with is not in point data but in polygon data. Thus, the *spatial join* aggregation method (either performed in ArcGIS or R) can be of use (*Spatial Joins by Feature Type—ArcMap | Documentation*, n.d.).

The easiest map data visualization is to simply plot the absolute values of a specific variable that are represented by each feature unit (point, line, or polygon). For the Exploratory Spatial Data Analysis (ESDA), clusters are a common way in which data can be visualized. Thus, they help to study the distribution and patterns of points/polygons over the space. Spatial clustering is used to find groups of features with similarities and to differentiate different groups or outlayers. The similarities will depend on the data that is being worked with. To find clusters over space Alidadi & Dadashpoor, (2018) suggests that Moran's I is the most used method. The authors also mention that this method has been used to identify employment peaks or sub-centres. It is used to identify possible autocorrelations over space and differentiate them between positive, negative or random.

On the other way, other techniques are used to study statistical data analysis. The interaction between the different points or polygons relies on the different variables that are attached to them. Geographically Weighted Regressions (GWR) can be used to understand how different variables affect a specific field in space. GWR move a weighted window over the data in space based on a specific kernel (Bivand et al., 2008). Several articles have used it for these purposes, some examples can be found in Diana et al., (2021) & Lagonigro et al., (2020) A GWR can also produce visual-spatial models that will help to understand and have better information about the possible relations (Saputra & Radam, 2022).

The article of Oller et al., (2017) study the relationship between the monocentricity and the directional heterogeneity that may appear in these cases. In the monocentric models, the distance to the CBD is a

key factor for the variables that are related to it or aimed to be studied. Directional heterogeneity happens when the weight/effect of the variables is not the same in all directions regarding the distance to the CBD. In the article different methods have been discussed, they conclude that if there are reasons of having directional heterogeneity the best model to be used is the LWR1, a locally weighted regression with the coefficients varying according to directions. If the CBD is not well located according to the right coordinates, using non-parametric models is the best option, in these only the geographic coordinates and the distance to the CBD are taken into account.

The reminding sections of the thesis are structured as follows. The second part provides the research objectives with their respective research questions. The third section describes the study area. The methodology that is applied for this research is found in the fourth section. The fifth section provides the results obtained following the methodology, and works as a basis for the sixth section, the discussion of these results. In the seventh section is compounded by the conclusions. Finally, the appendices that contain the summary of the code used are found after the bibliography.

2. Research Objectives

As previously mentioned, the existing research on employment in the Barcelona region does not completely align with the goals of this research. For this reason, the research objectives and questions will be addressed in understanding this specific innovative study.

The research questions that this project wants to answer are listed below. The main objective of this research project is to obtain the first dataset with the geographical information of all the employment centres from Catalonia. This will be the basis to provide a good dataset which will help to understand in the maximum extent possible the employment distribution in the BMA. A central research question is found below:

Is it possible to geolocate all the employment centres' addresses from Catalonia and understand the spatial employment distribution in Barcelona Metropolitan Area?

Following, several research questions related to the main research question are listed. The first set is based on the geocoding process for the maximum number of Catalan employment centres. The second set aims to study the socioeconomic distribution in the BMA based on the previously obtained results.

- Is there any method that can be applied to automate in an efficient way the conversion from addresses to geolocated points?
    - How can the process be automated in order to accelerate the process?
    - Is it possible to use open software for this process with meaningful results?
    - Which geocoding services have better performance? Which can geocode more addresses?

- Which is Barcelona's metropolitan area economic spatial distribution?
    - How is the BMA intraurban employment distribution? Which spatial employment pattern does it follow?
    - Are there any clusters per sectors in certain specific regions of the BMA?
    - How can maps be displayed in a way that are easy to understand and provide significant and accurate results about the analytical study?

3. Study Area

This research is focused on two important Spanish regions. The first one is the autonomous community of Catalonia. The second one, which is the main element of the study, is the Barcelona Metropolitan Area (BMA).

Catalonia is located in the north-east of the Iberian Peninsula (bottom-left corner in Figure 2). It has a total area of 32,113.86 square kilometres and it delimitates with France in the north. Catalonia population was of 7,763 million people in 2021 (based on the municipal register), being the 16.4% of the total Spanish population. In date of 13th of January 2023, in Catalonia there were find over 622,967 companies with their register office, being the 18.5% of the total number companies in Spain. Around 57.7% are companies without employees, and 42.3% with employees (Generalitat de Catalunya, 2019).

The BMA is composed by Barcelona and 36 other municipalities (Figure 2) As stated earlier, BMA has the 40% of the population of Catalonia: 3,222,117 inhabitants in 2011 (Martori et al., 2016). In 2010, produced the 52% of the Gross Domestic Product (GDP) of Catalonia, placing the BMA as a key area for the Catalan economy (*Àrea Metropolitana de Barcelona*, n.d.).

Be aware that data regarding to companies' registration, population and GDP fluctuates over time, however, the data provided can be used to get an estimation of the dimension of the Catalan and BMA economy and demographics. Also, consider that a company may have several employment centres.

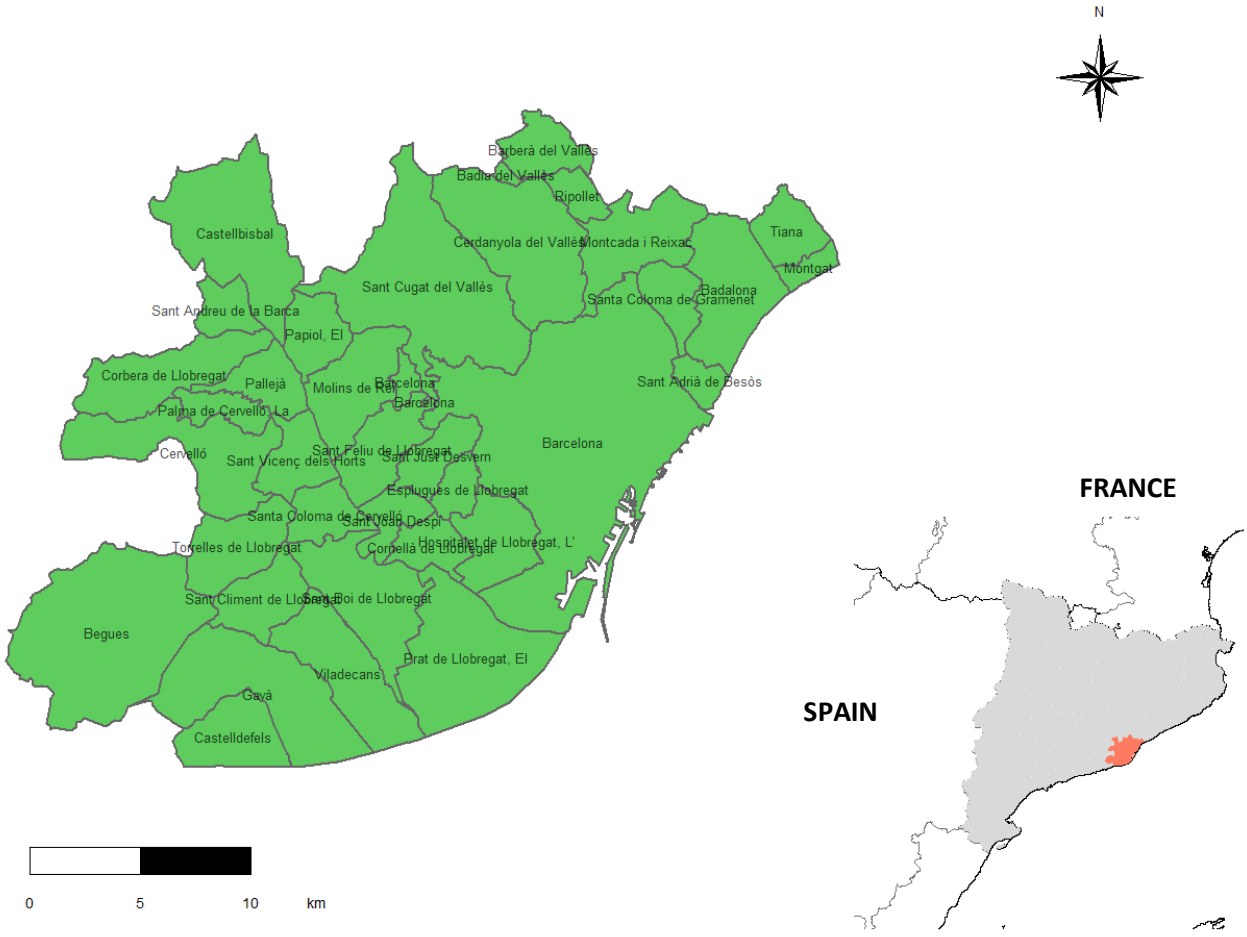

Figure 2 Representation of the Catalonia location (bottom-right) with the BMA highlighted. BMA municipalities in the top-left.

4. Methodology

As stated in the introduction, this project is divided into two parts. Both parts aim to achieve the research objectives mentioned in section 2. The methodological processes followed in this research are presented in Figure 3. The detailed steps are described later.

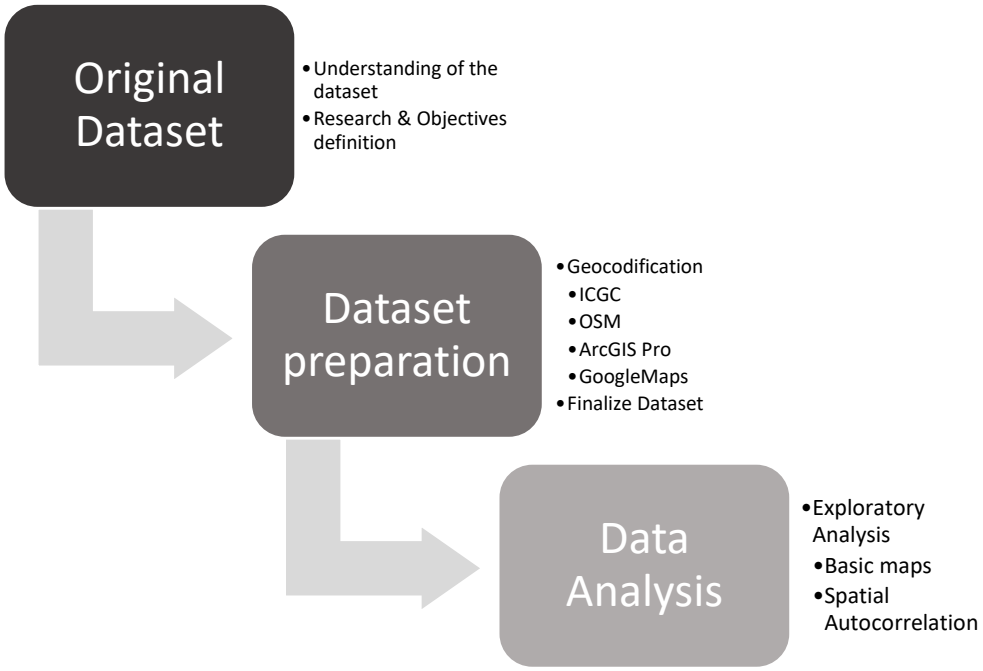

Figure 3 Diagram of the methodological process.

In the background information section, following the articles from Frankenfield (2022) and Küçük Matci & Avdan (2018) it has already been explained how datasets are key for data analysis. Moreover, they highlight the significance of geocoding the addresses given by each employment centre for research. Thus, different approaches are used to obtain the major number of geocoded addresses possible and in the best quality. 60% of the total addresses have already been geocoded (which consists on getting the latitude and longitude values through addresses) using Open Street Map (OSM) and Nominatim, and the Geocoder from the Institut Cartogràfic i Geològic de Catalunya (ICGC). However, numerous addresses are not well located and the remaining 40% still needs to be geocoded.

The provided dataset consists of a list of all the employment center's names from Catalonia, the addresses, some geocoded points, the economic activity based on the standardized National Classification of Economic Activities (CNAE) and the number of employees for each employment centre at the time the data was provided (Table 1). This data set has been obtained from the *Departament de treball, afers socials i família* who collects the results of trade union elections from the years 2014 – 2020. The registered employment centres may have at least 6 employees although some exceptions may happen through company agreement. Also, is important to remind that not all employment centres with 6 to 10 employees decide to do trade union elections as it is not mandatory. However, its is used as an accurate and significant representation of Catalan employment centres.

To further study the employment distribution in a socioeconomic context, an extra dataset about the census tracts from the BMA has been obtained. It contains information related to family income, attainment, poverty, and immigrant population percentages per each census tract. A total number of 2,136 census tracts are found in the BMA for the Catalan general elections of 2019.

Table 1 Description of the columns that conform the first employment centres' dataset.

Column name	Description
Codi	Code to identify each employment centre
Employment centre name	Employment centre name
Address	Location of the employment centre
City	City where the employment centre is settled
Latitude	Coordinates (not all are filled)
Longitude	Coordinates (not all are filled)
CNAE ID	ID of the activity sector regarding a standardized classification of the economic activity
CNAE Name	Standardized description of the employment centre economic activity
Number of employees	Number of employees in the employment centre

4.1 Geocoding

Before the geocoding, some errors corresponding to the original dataset must be corrected. The process of developing better data quality can be known as data preprocessing and is recognized as being a fundamental process for solid data analysis (Fan et al., 2021). Several errors have been found during the data exploration and checking process, and therefore, changes are applied. Some of them are described in this section but others are found in the following sections where they can be better justified.

First, each record in the dataset must be uniquely identified. This allows to use an identifier for margining newly obtained/created data without errors. For this, the *CODI* (Code in English) column has been used. However, some employment centres did not have a code and a new one had to be created. For them, a numeric code which consists of *9999000001* and the up following numbers has been created. This codification ensures no duplicate values. Some other employment centres appeared more than once in the dataset, with duplicated addresses and codes. In these cases, the duplicated records are removed and only a single significant observation (row) is kept. Datasets were first divided considering the existence of a code value to facilitate the dataset set-up process. Finally, both datasets have been merged.

In the rest of the section we define how the geocoding process has been done, which validation methods have been used in order to assess the performance of the geocoding, introduces some other data preprocessing methods and finally describes how all these processes previously mentioned have led to the construction of the first dataset with geolocated employment centres of Catalonia.

Below, a simplified diagram of the geocoding process and its validation methods is created (Figure 4). This diagram must be used while reading all the methodological section, the rows on the top make reference to the 4 geocoders used.

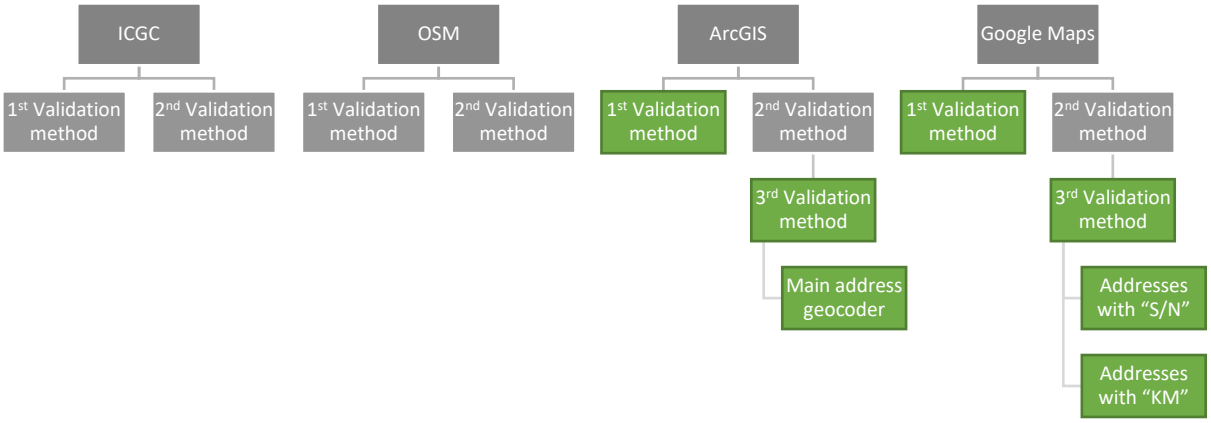

Figure 4 Simplified diagram of the geocoding and validation processes. The green cells correspond to the best approaches for all the process.

4.1.1 Geocoding Process

Multiple geocoders have been discussed in the introduction section regarding the work from Prener (2021). However, just three of them are considered for the analysis, plus an extra one from a specific cartographic institution in the country. Those are ArcGIS, Google Maps and OSM (Nominatim), as well as the ICGC geocoder. ArcGIS and Google Maps have been chosen as they are widely known geocoders and have been applied in many articles (Duncan et al., 2011; Lemke et al., 2015; Panasyuk et al., 2019; Prener, 2021; Schootman et al., 2006; Zhan et al., 2006). Moreover, the author has the use License from ArcGIS, which is the paid methodology. Furthermore, OSM is chosen because is a completely free licensed software. The ICGC geocoder choice is based on it being a local geocoder only made for Catalonia. Although one of the main aims of this research is to get the largest number of addresses geocoded, it also wants to compare different geocoders. Not all possibilities can be chosen due to scope limits.

OSM is a crowdsourced project that collects volunteered geographical information intending to create an open geodatabase that covers the whole world (*OpenStreetMap*, 2022). Nominatim is a geocoding software that uses the data from OSM to find locations all over the globe using names and addresses, it can also be used in the reverse functionality (*Nominatim*, 2022). The Geocoder from the ICGC is a geocoding service that works through other software, which aim matches the one from Nominatim, to obtain coordinates from a specific address, municipality, crossroad, street, or place name. This last approach only works for Catalonia (ICGC, 2020). Both of the methods are free to use.

The geocoded addresses using the OSM and Nominatim have been performed firstly using a R script (APPENDIX I) to automatize the process and geocode the maximum number of addresses possible. At the same time, the Geocoder from the ICGC has been used via the Windows Console (CMD), where the comma-separated values (csv) document is read with an eventual data processing (geocoding) that is finally copied and saved to a new csv document. Finally, a Macro via Visual Basic in Excel is programmed. A query is created that takes the addresses from a specific column on the dataset and uses the Nominatim software to obtain the specific location in longitude and latitude format. This last approach has been used to perform single geocoding operations of the addresses that have not yet been geocoded (APPENDIX II). However, not all the addresses have coordinates and new operations need to be done.

The geolocations from the new geocoding processes are stored as new columns in a new version of the dataset. The first geocoding process made by the author has been obtained using the ArcGIS Pro software, applying the *Geocode addresses* tool. This tool is used to geocode a table of addresses. It uses

a basis data set that stores the addresses to geocode and an address locator. Then it matches them with the main dataset to finally create a new vector layer with all the geolocations. However, the ArcGIS World Geocoding Service can be used instead of the basis data set. This last approach will consume credits from the ArcGIS account (*Geocode Addresses (Geocoding)—ArcGIS Pro | Documentation*, n.d.). For 15,091 employment centres, approximately 600 credits have been used. For this reason, the ArcGIS *Geocode addresses* tool has been categorized in this article as a paid method. This geocoding process has been divided in 4 parts: 2 sets of datasets containing around 5,000 employment centres each, and 2 more of around 2,500 employment centres. It has been done this way to ease the detection of possible problems/errors that can occur during the process and to make adequate corrections. After the process, the datasets are merged into one. The *Geocode addresses* tool separates the data according to street type, street address, street name, subregion, region, city, etc…. Thus, it uses the whole table that contains the location specifications in different columns. Finally, the results obtained (latitude and longitude values) have been merged to the main, newest dataset version, where the OSM and ICGC outputs are present.

On the other hand, R software is used to obtain extra addresses. For this process the *ggmap* library package is used, followed by a registration to the Google Maps Platform and Google Cloud Platform using an API KEY. This package allows the use of sources from Google Maps and Stamen Maps and has tools that can be applied for the geocoding function. The API of Google (*Geocoding API*) is used for the desired process. It has been known that around 2017 there was a limit for its usage of 2,500 queries per day (Bajak, 2017; Kahle et al., 2019; Tran, 2018). However, policy changes inside the Google Platforms have been applied lately. The changes allow users to have a monthly credit of 200$ for the Google Maps Platform API usage, and a starting credit of 315$ to use the Google Cloud Platform (*Google Cloud Platform*, 2022). These changes make the Google Maps *Geocoding API* a semi-paid method as users can get free budgeting, but if they exceed the threshold they need to pay. This credit given by Google has been used to obtain geocoded results for the 15,091 employment centres. As an estimation, a total of 75$ is used for geocoding this number of addresses. This process has been performed 3 times. In all of them a *for* loop in R was used along with the *geocode* function (see code in APPENDIX III). The first two processes just considered either the address (street name + number) (GM1) or the name of the employment centre (GM2). This is because only one column can be implemented for the process, many wrong geocoding operations have appeared in both processes (see results section).

For the last application, a *for* loop has been used along with the *paste* tool for merging the different columns and adding the country variable (Spain) (APPENDIX IV). Thus, a new variable is created (Complete Address) which contains the street name and number, city and country. Then, the same process has been applied to this new variable and it has resulted in more accurate address locations for the *Geocoding API* (GM3). The different types of licenses used for each method are summarised in Table 2.

Table 2 Summary of the different types of Geocoding methods used.

Method	Type of license
OSM (R)	Free license
ICGC	Free license
ArcGIS Pro	Paid license
Google Maps (R)	Semi-paid license

4.1.2 Geocoding Validation

To answer the question regarding the performance of the different geocoding services and their capability to geolocate a higher number of addresses, three main geocoding analyses are performed.

The first one consists of checking if the geocoded points are inside their corresponding municipality, the second one calculates the distance to the municipality centroid, and the third one consists of checking manually the results between two good performing geocoding methodologies.

The first one considers if each geolocated point corresponding to the employment centre is located inside the polygon equivalent to the municipality given in the address field in the dataset. For this, a polygon layer containing all the municipalities is used. R software is again used for this process, the function *is_covered_by* is applied in this case (refer to APPENDIX V). On it, a *for* loop with an *if* conditional is employed to obtain a dummy variable where value 0 indicates a mistake in the geocoding according to the municipality and the value 1 indicates a correct geocoding by each specific method. A column is created next to each pair of coordinate columns for each geocoding method.

It may happen that the geocoders locate to the municipality centroid what they are not able to geocode in address level (street name and number). Thus, the distance from the municipality centroid is computed. With the function *st_centroid* the centroids for each municipality are created. The *st_*distance function is used to calculate the distance between two points. The code used for it appears in APPENDIX VI. In this process, a new variable for each geocoding process is created. The loop iterates through the dataset and calculates the distance as a straight line, in meters, to the corresponding municipality centroid as a new variable. For these processes is crucial to have the same coordinate reference system in all layers, in the study case, WGS84 (EPSG: 4326) has been used. Afterwards, a 50 metres threshold is set to differentiate the values which are close to centroid or not. This detects possible errors in the geocoding location around the centroid point.

The third approach regarding the validation of the results aims to determine which method, ArcGIS or GM3 (Google Maps), perform better. In this approach, a set of 95 random employment centres is selected, and the distance between the geocoding locations of both methods is computed using the *st_distance* and the code variable as a link. The locations are then checked manually with the Google Maps web application and a topographic map to assess if they are well located or not. A number 1 is given to well-located and 0 to not well-located. Finally, some evaluation comments are written for each employment centre to look for patterns in the errors obtained.

As some pattern errors are found, some last methods are used to obtain the most suitable results in the geocoding section. Two error patterns are found, the first one is with addresses including "KM" (indicating a specific km in a road) and the second one is with addresses containing "S/N" (buildings or addresses without a number), which are isolated and treated apart. The total number of employment centres with these specific characteristics is 994 for "KM" and 1,907 for "S/N". In those, a new address column is created with the name of the employment centre, town, and country (Spain). Again, the Geocoding API from Google Maps with the R scripts applied in the first geocoding process is used. This is based on the widely known capability of Google Maps to detect employment centre names and locate them on the map. The results are compared manually (based on the 3rd approach) and it is decided if the geocoding results from the best approach or the ones from this last method are selected. A total of 30 cases for "S/N" and "KM" addresses are chosen. The number of significant improvements compared to the most suitable approach is computed per each of the two error patterns. An example of addresses containing these two different errors is shown in Table 3.

Table 3 Examples of addresses containing kilometric points ("KM") and addresses without number ("S/N").

"KM" Example	**"S/N" Example**
Carretera NACIONAL II KM 9-A, Llers, SPAIN	AFORES S/N, Bruc, el, SPAIN
CarreteraN-240,KM 38,1, Montblanc, SPAIN	Polígon industrial CAN COMELLES S/N, Esparraguera, SPAIN

The best approach from the 3rd validation method is used as the main driver for the final Lat and Lon values (only applied to the "KM" addresses) together with the best general geocoding choice for all the other employment centres.

After all these processes, a checking according to the municipality checking (1st validation approach) has been performed again. This last inspection showed how many of the employment centres that are not located inside the municipality, are correctly located in the GM3 processes. Consequently, the latitude and longitude values from the GM3 have been assigned to the ones with the geocoding outside the municipality. A final check has been performed from which the employment centres with the location inside the municipality have been used to create the final dataset. The remaining employment centres (with no correct geolocation) have been kept apart from this data set. From the 15,091 initial employment centres, only 204 (1.35%) have been deleted.

4.1.3 Data Improvement

This section describes other data preprocessing processes needed to create the final dataset. These consists of correcting municipality names and assigning new CNAE ID code for some employment centres.

Python coding has also been used in the data preparation process for the validation process between the municipality layer and the geolocated points to standardize the municipality names. An R code is used to check if all the municipalities from the two layers have the same name. Then *NumPy* and *Pandas* Python libraries have been applied for this standardization process. The scripts used are in APPENDIX VII. The Spyder environment has been selected for this process due to the capacity and easy environment to work with datasets. Some municipality names have an article word (*el, la, els…*) at the end of the text. The *np.where* and *str.replace* functions are used to find and place them at the beginning in the correct format. Moreover, some more specific changes to few municipality names have been done. All these processes have made possible the data comparison between the two data tables. Some examples are shown in Table 4:

Table 4 Example of data standardization using Python.

Old format	New format
Prat de Llobregat, el	el Prat de Llobregat
Garriga, la	la Garriga
Hospitalet de Llobregat, l'	l'Hospitalet de Llobregat
El Guiametss	els Guiamets
Vimbodí I Poblet	Vimbodí

The CNAE ID is crucial for the analytical part. The Spanish Government approved the National Classification of Economic Activities (CNAE) in the Royal Decree 475/2007, which consists of a set of alphanumeric classifications. Each employment centre is listed in this classification depending on its professional specialization. The most general classification is the Section, followed by the Division, Group and Class (working like the taxonomy classifications in biology) (Ministerio de Economía y Hacienda, 2007). Thus, it is key to have all the employment centres with their corresponding CNAE ID. Some of the records in the dataset do not contain any value for this field. Therefore, CNAE ID identifiers (only for the BMA) have been generated based on similar employment centres that already had the CNAE ID value, or, by searching them on web portals which provide employment centres' information (the one used in this research is (*Información de Empresas Españolas | EInforma*, n.d.)).

After this process, different blocks have been done based on the different section groupings: Class, Group, Division and Section. This was done to get the employment centre frequencies within each group in the BMA.

4.1.4 Final Datasets

After all the processes explained previously, a dataset containing all the employment centres in Catalonia has been generated. Subsequently, a second dataset has been obtained subtracting the employment centres located within the BMA. Both have the same variables; the only difference is the number of records in each one.

Table 5 show the variable's results of the final dataset. All geocoding groups have their validations (municipality and centroid) in their neighbour columns for easy comparison. An example of 5 rows from the dataset is shown in Table 8 in the results part. The Lat and Lon fields correspond to the best geocoding choice. These fields will be used for the analytical part of the project.

Table 5 Description of the columns that conform the final employment centres' dataset.

Column name	Description
Code	Code to identify each employment centre
Employment centre name	Employment centre name
CNAE ID	ID of the employment centre regarding a standardized classification of the economic activity
CNAE Name	Standardized description of the employment centre economic activity
Number of employees	Number of workers in the employment centre
Employment centre type	Employment centre type
Lat	Final latitude value
Lon	Final longitude value
Lat ICGC	Latitude regarding ICGC geocoding
Lon ICGC	Longitude regarding ICGC geocoding
Valid.Mun.ICGC	Dummy variable to know if the coordinates from ICGC are in the corresponding municipality. 1 is correct, and 0 incorrect
Dist.Cent.ICGC	Distance from the coordinates from ICGC to the centroid of the corresponding municipality. Units in meters
Lat OSM	Latitude regarding OSM geocoding
Lon OSM	Longitude regarding OSM geocoding
Valid.Mun.OSM	Dummy variable to know if the coordinates from OSM are in the corresponding municipality. 1 is correct, and 0 incorrect
Dist.Cent.OSM	Distance from the coordinates from OSM to the centroid of the corresponding municipality. Units in meters
Lat ARC	Latitude regarding ArcGIS geocoding
Lon ARC	Longitude regarding ArcGIS geocoding
Valid.Mun.ARC	Dummy variable to know if the coordinates from ArcGIS are in the corresponding municipality. 1 is correct, and 0 incorrect
Dist.Cent.ARC	Distance from the coordinates from ArcGIS to the centroid of the corresponding municipality. Units in meters
Lat GM1	Latitude regarding Google Maps (GM1) geocoding
Lon GM1	Longitude regarding Google Maps (GM1) geocoding
Valid.Mun.GM1	Dummy variable to know if the coordinates from GM1 are in the corresponding municipality. 1 is correct, and 0 incorrect
Dist.Cent.GM1	Distance from the coordinates from GM1 to the centroid of the corresponding municipality. Units in meters
Lat GM2	Latitude regarding Google Maps (GM2) geocoding
Lon GM2	Longitude regarding Google Maps (GM2) geocoding
Valid.Mun.GM2	Dummy variable to know if the coordinates from GM2 are in the corresponding municipality. 1 is correct, and 0 incorrect
Dist.Cent.GM2	Distance from the coordinates from GM2 to the centroid of the corresponding municipality. Units in meters
Lat GM3	Latitude regarding Google Maps (GM3) geocoding
Lon GM3	Longitude regarding Google Maps (GM3) geocoding
Valid.Mun.GM3	Dummy variable to know if the coordinates from GM3 are in the corresponding municipality. 1 is correct, and 0 incorrect
Dist.Cent.GM3	Distance from the coordinates from GM3 to the centroid of the corresponding municipality. Units in meters
Full address	Complete address of the employment centre. Street name, number, city and country
Municipality name	Municipality name standardized
Comarca name	Comarca name
Province name	Province name
Distance from GM	Distance from the ARC geocoding to the GM3 geocoding

4.2 Data Analysis

This methodology section has been only developed using the R software (R Core Team, 2022). It is necessary to match the employment centre's point data with the census tracts polygon data from the BMA. This allows to get a deep understanding of the employment centres' location in a socioeconomic context and provide a more meaningful dataset. Moreover, it gets a good employment centres' representation per unit area and at the same time ensure that employment centres' data privacy/confidentiality is protected. To join the point data with the census tract data a pipe with the *st_join*, *group_by* and *summarize* functions has been used (APPENDIX VIII). Below, there is a simplified representation of the process carried out in this section (Figure 5). To understand better the process list read the methodological part.

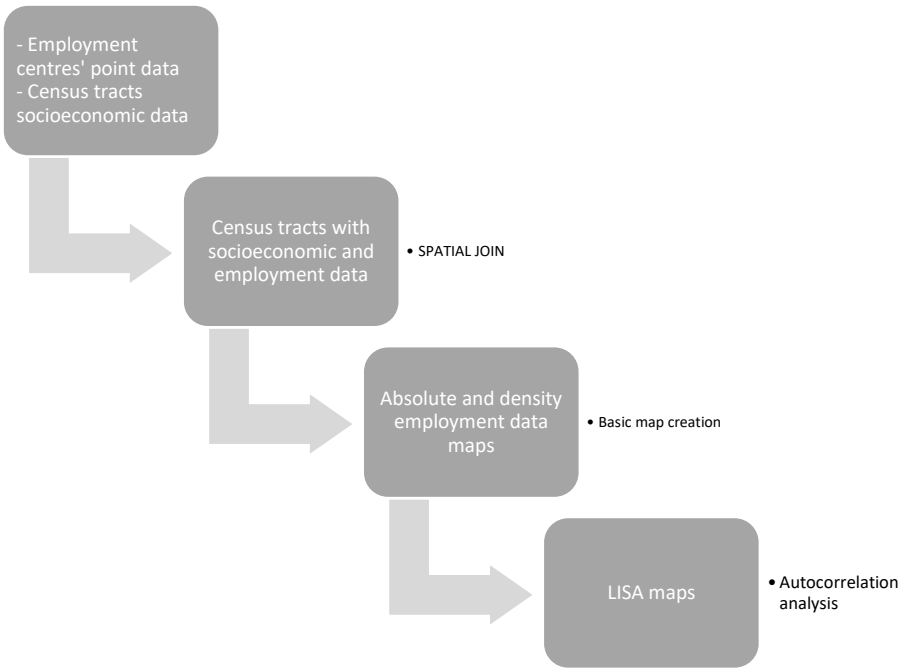

Figure 5 Simplified process list for the Data Analysis part. This process may be used while reading the methodology part.

The United Nations Development Group makes focus on Big Data handling for the achievement of the 2030 agenda and the Sustainable Development Goals (United Nations Development Group, 2017). They are concerned about data privacy and confidentiality and for the safe and responsible usage of Big Data. For instance, Kounadi & Leitner (2015), outline the significance of ensuring data privacy for the individuals (in our study case: Employment centres) involved in the datasets. For this reason, following the article from Armstrong et al. (1999) and the idea of areal aggregation, the use of Census Tracts can be helpful to ensure employment centres' data confidentiality.

As it was previously mentioned, the CNAE ID field is of high importance for the data analysis (section 4.1.3). Its relevance is because is key if we don't want to lose employment centres' information when joining the employment centres' dataset and the Census Tract data. Frequencies of employment centres and employees per each census tract on each CNAE division group have been calculated. An adapted classification focused on the BMA has been created considering the employment similarities and the frequencies of each class. The results obtained from these can be found in APPENDIX IX with their attached description. To ease the process a function has been created that counts the number of employment centres and employees per each census tract and each grouping classification (APPENDIX X).

The approach described above is inspired by a mix of several articles which aim to work with census tracts data for similar types of analysis. These articles are Lagonigro et al., (2018), Li & Monzur (2018) & Madariaga et al. (2014).

The data table structure of the census tracts with the employment centres' information is found below (Table 6). The last two rows refer to a specific adapted classification. A total of 58 adapted classifications are found in the dataset.

Table 6 Census tracts structure. It is represented the name of each variable and its description.

Column name	Description
CUSEC	Code from each census tract
Municipality Name	Municipality where the census tract belongs
Area	Census tract's total area
TotalPopulation	Census tract's total population
TotalForeigners	Census tract's total number of foreign/immigrant people
Economic Immigration	Immigration people for economic reasons
Density	Census tract's population density. Inhabitants per square kilometer
Analphabets Percentage	Census tract's total analphabet population. They don't know how to read nor write
Without Studies Percentage	Census tract's total population without studies of any kind
P1erg	Census tract's total population with studies of first grade
P2og	Census tract's total population with studies of second grade
P3erg	Census tract's total population with studies of third grade
IngrDebajo40	Percentage of population with an income lower than 40% of the median
Medium Rent	Census tract's mean annual family income
Rumania Immigrants	Census tract's total number of immigrants from Rumania
Maghreb Immigrants	Census tract's total number of immigrants from Maghreb region
Sub-Sahara Immigrants	Census tract's total number of immigrants from Sub-Sahara region
South America Immigrants	Census tract's total number of immigrants from South America
Total employment centers number	Census tract's total number of employment centers
Total employees number	Census tract's total number of employees
Employment centers density	Census tract's employment centers' density
Employees density	Census tract's employees' density
Total Agricultural employment centers' number	Census tract's specific employment centers' number
Total Agricultural employees' number	Census tract's specific employees' number

The first understanding of the data above described has been obtained using the scripts from APPENDIX XI, which allows getting the univariate statistics of a dataset. The table has been created using the R package *skim*, which allows the creation of the output file directly to Word using the *RMarkdown* (Allaire et al., 2023). The output is found in the results sections.

4.2.1 Exploratory Analysis

This section is based on studying the employment distribution in total and density values for each census tract, and calculating the autocorrelation values for each of the census tracts: either for total employment and for the adapted CNAE classifications.

To gain better understanding of the spatial distribution of employment in the BMA an exploratory analysis using the dataset previously created (Table 6) is necessary. This analysis involves creating maps that display the absolute number of employees and employment centres per census tract, with the maps represented by quartiles and deciles. In addition, more exploratory maps based on employment centre and employees' densities per square kilometres are created. The *tmap* package has been used to create the maps and the corresponding scripts can be found in APPENDIX XII.

Understanding and dealing with spatial correlation has become a backbone in the field of spatial econometrics. Testing methods are becoming widely used and have been proven to have many advantages. To support what has been written in the Data Analysis section, the articles from Indriyani & Widaningrum (2021), Li & Monzur (2018) & Madariaga et al. (2014) are inspiring. These authors have already used Moran's indicators, as it is one of the most common ways to test for spatial autocorrelation. This type of autocorrelation is used to determine the class of correlation that may exist within variables across space. Moran's can be differentiated between global or local. Global autocorrelation analysis (Moran's I in this case) uses the entire map for the analysis of the values for the variable under study over the area and provides a first knowledge of the autocorrelation along the space. However, it does not give a local spatial indication of where the autocorrelation occurs. On the other hand, local autocorrelation tests the clustering values of the interested observation and its neighbours and gives the spatial location of where autocorrelation is happening (Getis, 2007; Li & Monzur, 2018). The usage of the Moran's I index in this research is divided into both scales: global (Global Moran's I) and local (Local Moran's I or Local Indicator of Spatial Autocorrelation (LISA)). A description of how both statistics work is described below.

The Univariate Global Moran's I result is calculated as follows:

$$Moran's\ I = \frac{n\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

where x_i is the observed variable in the census tract i, n is the number of census tracts, and w_{ij} indicates the elements in the spatial weight matrix. The result values lie between -1 and 1, with values close to 1 meaning the presence of positive spatial autocorrelation, the ones close to 0 result in no autocorrelation and, if they are close to -1 suggests negative autocorrelation. The presence of positive autocorrelation means that similar values cluster together on a map, while negative means dissimilar values cluster together.

The Univariate Local Moran's I for the census tract i is calculated as follows:

$$I_i = \frac{(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2/(n-1)}\sum_{j=1}^{n} w_{ij}(x_j - \bar{x})$$

where x_i is the variable value in the census tract i and x_j is the variable value in adjacent areas.

Additionally, plotting the LISA covers 4 types of association if autocorrelation occurs: clusters with high values surrounded by high values (HH), clusters of low values surrounded by low values (LL), clusters of high values surrounded by low values (HL) and the other way around (LH). Finally, the areas with no significant distribution , i.e. the p-value is higher than 0.05, will appear as NS (Shi et al., 2022). Besides from understanding the patterns in the space, this analysis can be helpful for understanding and discussing the statistical analysis results with higher accuracy.

The scripts used to perform this part of the analysis are found in the APPENDIX XIII, where a main function is created by the author. This allows the calculation of the Moran's I Index (Global) and to plot a LISA map for its interpretation over space for each variable of interest. Another script is designed to

apply the function to all the variables of interest (Employees per each classification) at once using the *lapply* function compacting the resulting code. The principal R package used for this calculation is the *spdep*. For a better understanding of this process, a description of this function is below:

- The function computes the neighbours' system of each area using the Rook's case contiguity and calculates a weight matrix for each neighbour sample. From this, a Global Moran's I Index is obtained based on the selected study variable. The function also displays a plot highlighting the relation between a specific census tract and its neighbours (based on the weighted matrix previously created) and provides a first overview of how data is clustered in a logarithmic transformation (HH, LL, LH, HL). The logarithmic transformation is made for better data visualisation. Then a new column with the classifications is created with the LISA values. This last column's values are based on the raw values of the variable, the estimated neighbours (lagged values) and an assigned significance level. Finally, it plots a specific map (using the *tmap* library) with the area of the BMA and the LISA corresponding values.

A first study on the economic spatial distribution is made by using the main employees' variable in the function. Moreover, due to the high number of classifications, only the sectors with the highest representation on the BMA (higher values of employment) are used for the ESDA.

Maps are key to present spatial information, and creative skills are important in this process of communication. One can think that creating maps is a simple process. However, information obtained in the research process must be displayed in an understandable way. For the map making process the author has used the (Lovelace et al., 2019) book ideas, mainly focused on the *tmap* library, and looked closely other maps of his liking. There is not a straight and delimited process to follow, so constantly editing the maps has been an important part of the process. The created maps belong to the choropleth maps class. Map data needs to be divided in classes for its spatial evaluation. There are many ways in which data classes can be defined (Brewer, 2006). Decile representation has been chosen as the best breaking style because it is the method which represent the data distribution best. In some cases, quartiles might have to be used depending on the data of the adapted classification sectors. Apart from the breaks; colours and political boundaries have been fundamental for data visualization. The same colour ramp (reds) has been used together with a varying intensity of the red colours that range from less employment (low intensity) to high employment (high intensity) to allow the reader a rapid understanding of the map and locate the most relevant commuting places. Moreover, municipality boundaries have been highlighted (with wider line width) to help readers locate each of the census tract values with the municipalities map from the Study Area section.

5. Results

This section presents the results obtained from the detailed methodology exposed in the previous chapter. It is divided in two main parts: one that emphasizes the first set of questions about the dataset and geocoding topics, and the second which draw some ideas about possible analysis of the results for the employment distribution in the BMA.

5.1 Geocoded Dataset

The validation results obtained from the three geocodification methods can be observed in Table 7. The values from the two first rows are over the total of 15.091 employment centres' of the dataset. The last, is based on a 95 random sample selection.

It can be differentiated two different results regarding the number of coordinates inside the corresponding municipality: low values in cases of ICGC, OSM, GM1 and GM2; high values in ArcGIS and GM3 methodologies (Table 7). On the other hand, low numbers in the coordinates close to the geographical centroid of the corresponding municipality are found. compared to the other geocoding methodologies GM3 presents a higher number of coordinates close to the geographical centroid, with more the double. Finally, the results from the 3rd validation show a slightly higher number on the ArcGIS (68,5%) compared to the GM3 (63%).

Table 7 Validation's results obtained by the different Geocoding methods

Geocoding method	**ICGC**	**OSM**	**ArcGIS**	**GM1**	**GM2**	**GM3**
1st Coordinates inside municipality	6148	6998	14295	5138	669	13991
2nd Number of points at less than 50m from centroid	0	5	7	2	6	16
3rd Deep validation of geocoding results. Correct results out of 95	–	–	65	–	–	60

A graphical representation of the percentage of values matching the coordinates inside municipality is shown in Figure 6.

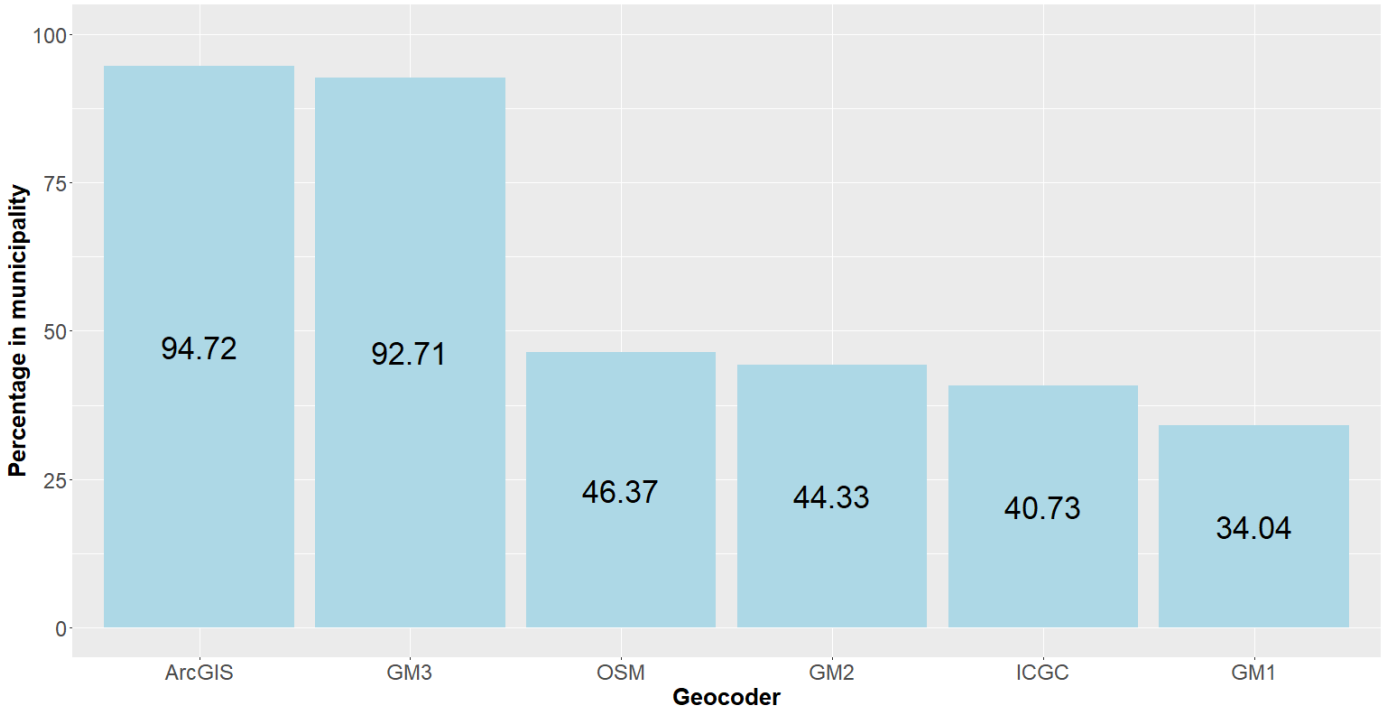

Figure 6 Representation of the percentage of Coordinates inside municipality for each Geocoding method.

The 3rd validation method considers the kilometric points and the addresses without building number as patterns of errors in the geocoding process. For the "KM" approach the significant improvements correspond to 63.5%, while for the "S/N" is 0%. The Google Maps Geocoding API has been used and show a significant percentage of improvements compared to the ArcGIS approach. "KM" and "S/N" are described in the methodology part (4.1.3).

Note that the results only show the cases in which the Google Maps method is used and which results are better than the ones from the ArcGIS Pro approach. Thus, all the other cases' results can be equally good, or worse than the ArcGIS method. If the results are compared, the Kilometric approach ("KM") is better than the ArcGIS, whereas the "S/N" is not. Therefore, the results will keep the ArcGIS approach for this later case. Thus, the final dataset Lat and Lon values first consider the "KM" Google Maps Geocoded points and then uses the ArcGIS method for the other employment centres. Finally, only in the cases where the location is wrong for the ArcGIS and correct for the GM3 approach, the location values from GM3 are used.

Below, a summary of the final table is given (Table 8), the total number of employment centres in the dataset is 14.887, which corresponds to the 98.65% of the initial employment centres. This table shows 5 example employment centres with their corresponding variables from the geocoding process. There are noticeable differences for the geocoding methodologies. The employment centres' information is not shown to not violate their privacy.

Table 8 Summary of the final table with 5 examples of employment centres' information. Employment centres' information is not shown to keep their privacy.

Code	**21**	**61**	**128**	**366**	**381**
Lat	41.56615	41.62016	41.35700	41.92941	41.28683
Lon	2.003733	2.666189	2.126855	2.257990	1.096822
Lat.ICGC	41.56397	41.61809	0	0	0
Lon.ICGC	2.00363	2.667786	0	0	0
Valid Mun ICGC	1	1	0	0	0
Dist Cent ICGC	1317.24	1852.86	4603693	4666967	4523424
Lat OSM	0	41.61627	41.35665	41.92927	40.71327
Lon OSM	0	2.669226	2.125649	2.257762	0.581099
Valid Mun OSM	0	1	1	1	1
Dist Cent OSM	4627385	2034.038	1147.292	1939.605	6422.107
Lat ARC	41.56615	41.62016	41.35701	41.92941	41.28683
Lon ARC	2.003733	2.666189	2.126855	2.25799	1.096823
Valid Mun ARC	1	1	1	1	0
Dist Cent ARC	1083.241	1666.649	1244.266	1963.084	78040.16
Lat GM1	41.41174	0	41.35811	41.92954	39.47091
Lon GM1	2.164653	0	2.123502	2.258579	-0.37707
Valid Mun GM1	0	0	1	1	0
Dist Cent GM1	22390.86	4635927	962.4965	2013.617	159467.8
Lat GM2	41.56616	41.61343	41.39837	41.92939	40.7134
Lon GM2	2.003703	2.653555	2.141006	2.257812	0.580997
Valid Mun GM2	1	1	0	1	1
Dist Cent GM2	1082.96	1223.295	5125.828	1948.655	6437.814
Lat GM3	41.56634	41.62022	41.35811	41.92954	40.71342
Lon GM3	2.00359	2.66623	2.123502	2.258579	0.581025
Valid Mun GM3	1	1	1	1	1
Dist Cent GM3	1067.384	1668.777	962.4965	2013.617	6437.295
Full address	PI I MARGALL, 201, Terrassa, SPAIN	BALMES,152, Calella, SPAIN	PLAÇA D'EUROPA 22-24 1er, Hospitalet de Llobregat, l', SPAIN	CARRER DE DUES SOLES, 7, CASA CLARIANA, Vic, SPAIN	Plaça de l'Ajuntament, 3-4, Amposta, SPAIN
Municipality name	Terrassa	Calella	l'Hospitalet de Llobregat	Vic	Amposta
Comarca name	Vallès Occidental	Maresme	Barcelonès	Osona	Montsià
Province name	Barcelona	Barcelona	Barcelona	Barcelona	Tarragona
Distance from GM	24.13975867	7.177268997	305.4541296	50.87296048	77065.20361

5.2 Spatial Analysis

Table 9 represent the univariate statistics from the most relevant variables used in the ESDA. The two first variables represent the geometry characteristics of the census tracts in kilometres. It is followed by 15 socioeconomic variables (not shown in the table) and ends with two employment variables obtained from the previous merged dataset. Notice that there are big differences in the census tracts, ranging from small census tracts (0.008 sq.km) to big ones (22 sq.km). Almost all of them (75%) are found below 0.15 sq.km. The same pattern can be appreciated in the employment data, there are low values according to employment centres and employees' number (1 and 3 respectively) and high values 269 and 39,525 respectively. In all cases, the high numbers appear to be in the 4th quartile of the data.

Table 9 Univariate Statistics from the most relevant variables of the final census tracts dataset.8

Variable	**mean**	**sd**	**min**	**Q1**	**median**	**Q3**	**max**
Shape_Leng (km)	2.142040	3.240970	0.3797200	0.8131200	1.091460	1.917010	43.58387
Shape_Area (sq.km)	0.4166472	1.548955	0.008261630	0.03443951	0.05492837	**0.1502757**	**22.82661**
Employment Centers	5.17	12.68	1.00	1.00	2.00	**4.00**	**269.00**
Employees	649.80	2274.66	3.00	35.00	111.00	**380.25**	**39525.00**

Next, Table 10 represents how the data is found in each census tract. The table presents a set of 6 census tract examples to provide an overview of the data set. The CUSEC is used to identify each census tract. The following rows summarize some socioeconomic data, which has not been used for this study but could lead to further analyse the relation between jobs and socioeconomic characteristics of the population living in each tract. The last 7 rows contain all the different employment data variables (except the area) used. The *QuimFarma* variables (two last rows) are just examples of the frequencies for this specific sector. However, in the real dataset there are up to 58 adapted CNAE sectors, and each of them is composed by the same pair of variables.

As an example, the major employment sector in the first census tract is within the QuimFarma sector, as the 94% of the total employment is based on it. In contrast, the second and third census tracts show that QuimFarma sector is just a portion of their total employment, while others have no representation of it. This example highlights the employment sector variability in the BMA, which are not homogenously distributed. Density of Employees and Employment centres, highly varies in census tract areas.

Table 10 Dataset sample of 6 census tracts in the BMA.

CUSEC	**820505007**	**820004001**	**801910080**	**801901031**	**801902103**	**810106001**
CLAU2	8205	8200	8019	8019	8019	8101
NMUN	Sant Cugat del Vallès	Sant Boi de Llobregat	Barcelona	Barcelona	Barcelona	Hospitalet de Llobregat, L'
Shape_Leng	4702.047	22279.567	3843.225	1114.809	1098.228	7542.662
Shape_Area (sq m)	1272925.26	10754066.57	464653.19	46112.02	78134.60	2134855.68
TotalPopulation	2894	2089	2295	1558	1628	1979
TotalForeigners	352	156	891	857	401	296
EconomicImmigration	114	126	280	432	217	254
Density	2.2735035	0.1942521	4.9391677	33.7872890	20.8358398	0.9269947
AnalphabetsPercentage	14	24	0	0	0	14
WithoutStudiesPercentage	43.5	93.0	25.0	65.0	42.0	59.0
P1erg	139.5	146.0	70.0	73.0	89.0	149.0
P2og	1072	1147	927	486	677	585
P3erg	1517	308	1626	344	746	21
IngrBelow40	3.4	4.1	13.6	21.0	9.2	6.7
MediumRent	14620	10926	22330	11015	20298	7895
Rumania Immigrants	2	5	15	7	4	17
Maghreb Immigrants	5	40	17	26	9	27
Sub-Sahara Immigrants	0	0	1	4	0	1
SouthAmerica Immigrants	78	39	95	131	120	106
Total Employment Centres	**3**	**43**	**30**	**3**	**9**	**19**
Total Employees	**2023**	**3465**	**3516**	**3577**	**6961**	**11681**
Area (sq km)	**1.26658875**	**10.70071682**	**0.46221385**	**0.04587309**	**0.07773354**	**2.12408897**
Employment Centres Density	**2.368567**	**4.018422**	**64.905022**	**65.397821**	**115.780133**	**8.945011**
Employees Density	**1597.2035**	**323.8101**	**7606.8685**	**77976.0024**	**89549.5009**	**5499.2988**
Total QuimFarma EmploymentCenters	1	1	1	0	0	0
Total QuimFarma Employees	**1913**	**449**	**684**	**0**	**0**	**0**

The arranged number of workers per 6 adapted CNAE classifications is in Table 11. These groups are the ones with the highest number of employees, they sum up a total of 52% of the employees' number from the BMA. The remaining 48% is divided in 52 other classes.

Table 11 Representation of the 6 adapted CNAE classifications with highest employees number in BMA.

CNAE group	**Employees number**
AAPP	105713
Comercio	73272
ACTSanitarias	72758
Educacion	63218
FinanzasSeguros	51567
LimpJardin	32238
Percentage in BMA	52%

Next, there is a list of different maps and plots illustrating the most significant results obtained for this research. Each of the map and plot explanations are described below.

Figure 7 shows a map with the number of employees per census tract represented in deciles. BMA municipalities are differentiated in thin borders while the Barcelona city boundary is highlighted with a bigger line width. The map legend provides information about how data is distributed: 80% of the census tracts have less than 516 employees, while the remaining 20% ranges from 516 to 39525. The census tracts with a high number of employees are found in the centre of Barcelona, Barcelona maritime port, el Prat de Llobregat area, Sant Boi de Llobregat, Gavà, Viladecans and finally in some surrounding municipalities of Barcelona corresponding to Sant Cugat del Vallès, Cerdanyola del Vallès, Barberà del Vallès, Montacada i Reixach, Badalona and Castellbisbal (see Figure 2 on page 7).

The map with the same characteristics but representing the density of employees per square kilometres appears in Figure 8. The high values are located inside Barcelona boundaries where census tracts are of smaller areas than the ones from the surrounding tracts and municipalities.

The Moran's I statistic is of 0.1957, suggesting a positive autocorrelation at global scale. Although this positive autocorrelation is slightly low, it denotes that the areas with high number of employees tend to cluster together. The p-value of the test is < 2.2e-16, indicating this clustering is not a result of a random spatial process.

The Moran's plot for the LISA analysis is shown in Figure 9, it represents the global employee variable (x-axis) and its corresponding lagged variable (y-axis). The lagged variable in the Local Moran's I is a calculation of the neighbours' variable values based on a weighted matrix previously created. The plot is divided into 4 quadrants: The top-right one represents the HH values (observations with a value greater than the mean value of all observations, and its neighbours have also values greater than the mean value); the bottom-right locates the HL values (observations with a value greater than the mean value of all observations, and its neighbours have values lower than the mean value); the top-left situates the LH values (observations with a value lower than the mean value of all observations, and its neighbours have values greater than the mean value); and the bottom-left places the LL values (observations with a value lower than the mean value of all observations, and its neighbours have also values lower than the mean value). The tendency line has a positive slope. To have a better visualization of the data, the lagged variable has been logarithmically transformed.

Figure 10 presents the results of the LISA map for the total number of employees in the BMA. Some HH and LH cluster values appear to be significant and are coloured according to their LISA category. Overall, most of the HH values are found inside the municipalities of Barcelona (centre and maritime areas) and el Prat de Llobregat. With less extension they are found in the municipalities of Sant Boi de Llobregat, l'Hospitalet de Llobregat, Sant Cugat del Vallès, Cerdanyola del Vallès, Barberà del Vallès and Cornellà de Llobregat. A total of 86 census tracts showed significant results compared to the total number of census tracts (2136) in the BMA. 47 of them are in the HH category. The remaining 39 significant values are part of the LH category. This map gives a representation of how generic employment clusters are distributed over space in the BMA.

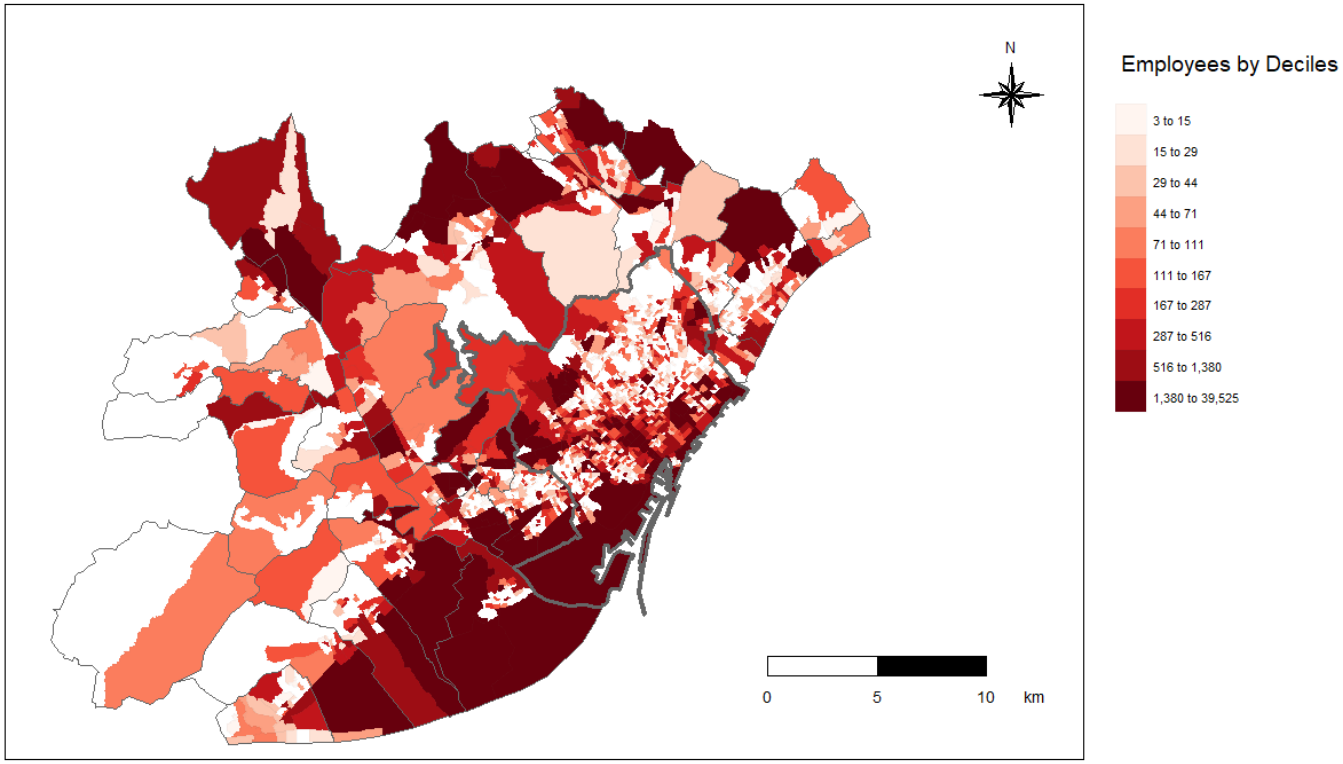

Figure 7 Absolute total employment values per census tract in the BMA. The values are divided by deciles.

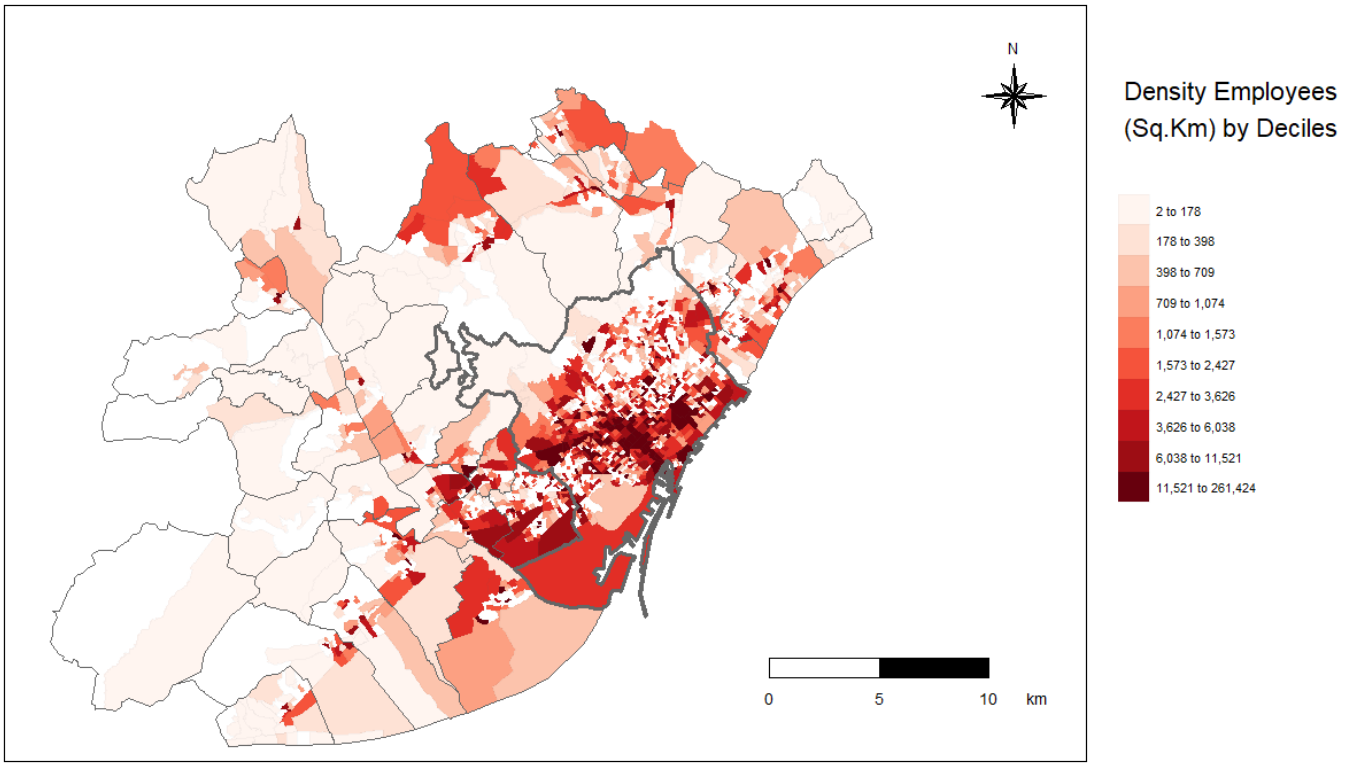

Figure 8 Density of the total employment per census tract in the BMA. The values are divided by deciles.

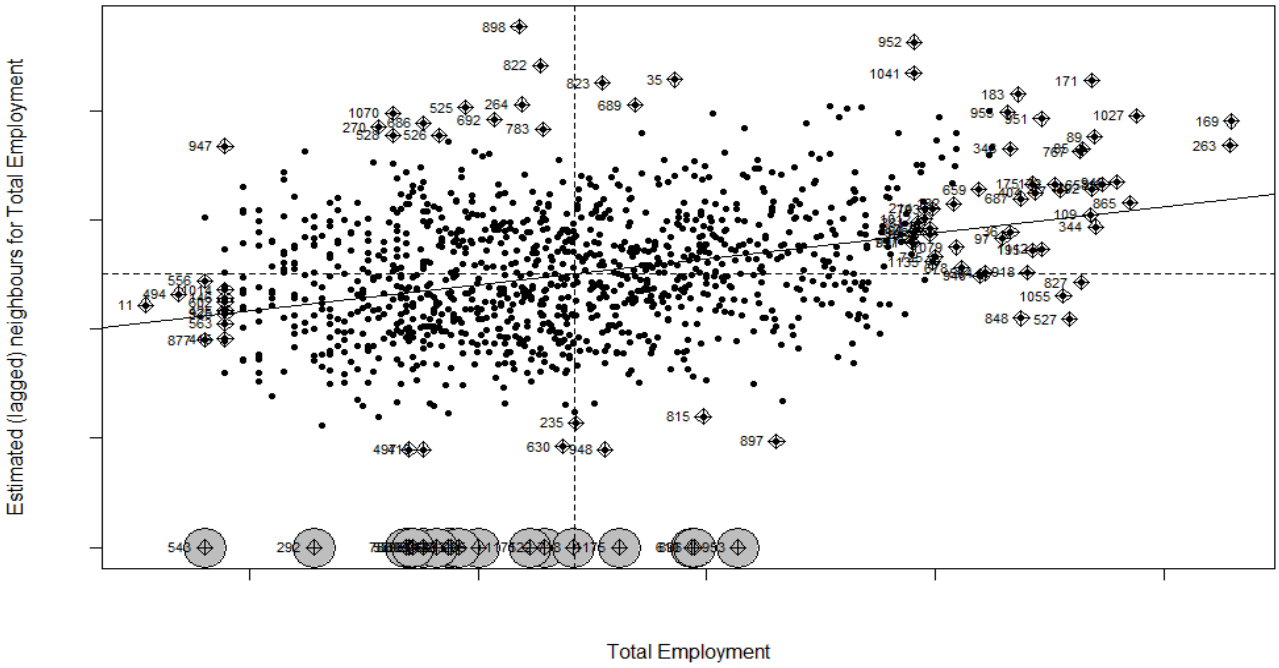

Figure 9 Moran's Plot for the LISA analysis. The scale of this graph is logarithmic.

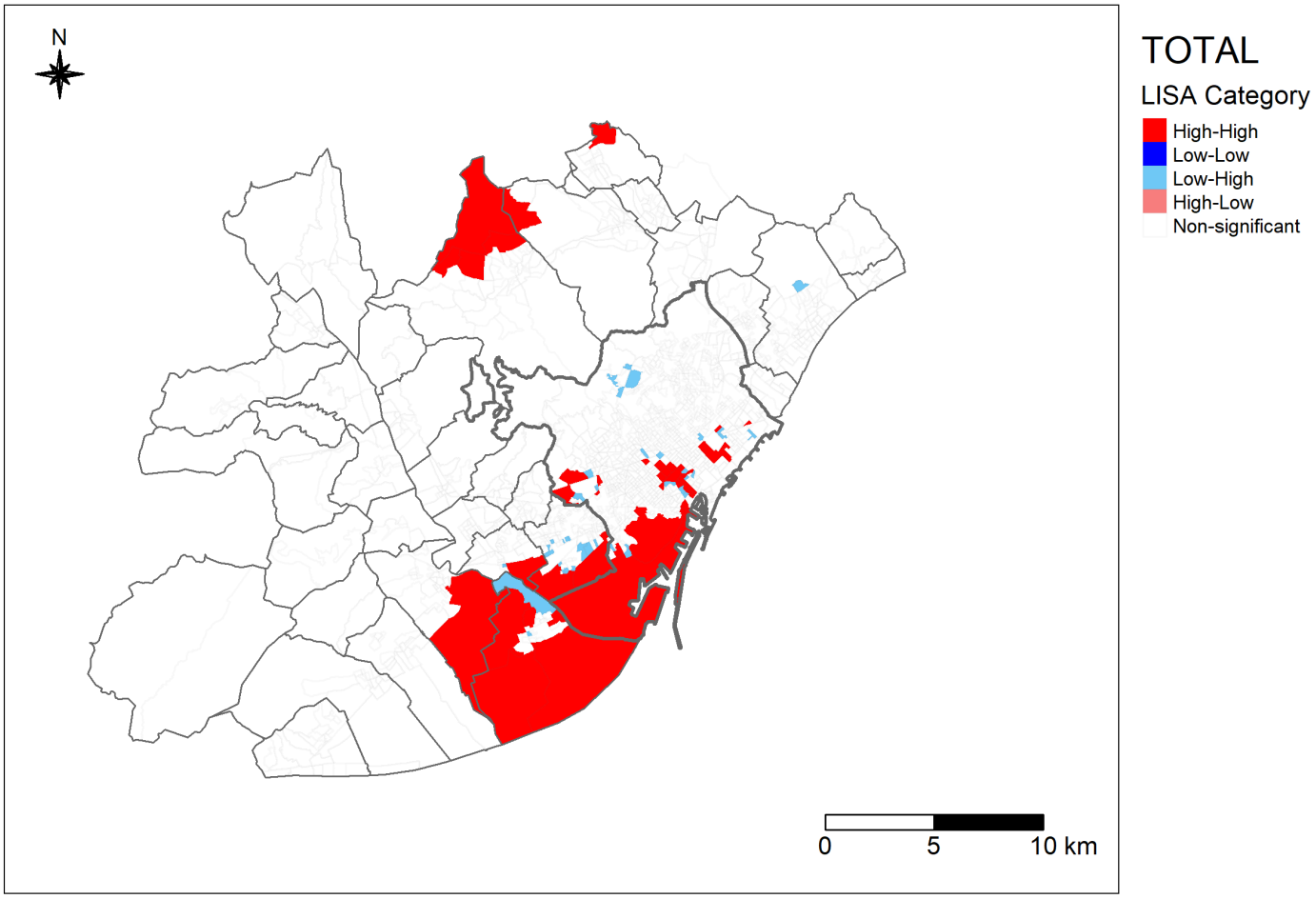

Figure 10 LISA map representing the clusters per census tract for the total employment in the BMA.

Finally, in Figure 11 a set of 6 maps representing the most relevant adapted CNAE classification categories are plotted. Again, the estimation of a LISA is used to map each of the results. They all show different cluster patterns over space. Remind that the HH results in LISA refer to areas with high values surrounded by other areas with high values. The LISA map only shows areas where the Local Moran's I is significant, i.e. the p-value is less than 0.05, this does not mean that in not-significant areas there are no employment centres from each specific category.

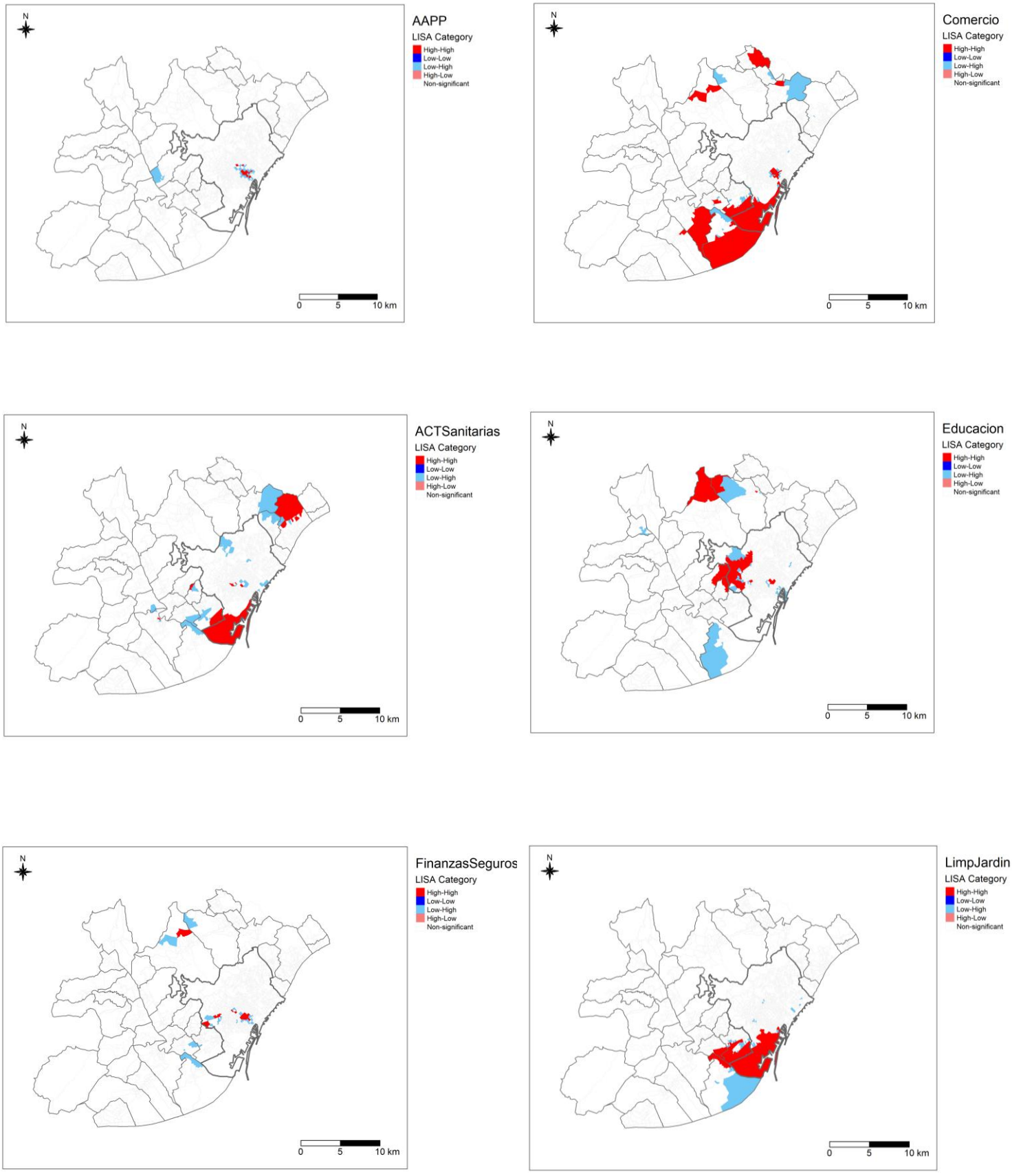

Figure 11 Set of 6 LISA maps representing the clusters per census tract for different adapted CNAE classifications in the BMA.

6. Discussion

This section is divided into two main parts. The first part discusses the methodology and results of the data preparation and the geocoding process. The second part focuses on the results obtained from the spatial data analysis. The discussion follows the research questions order from section 2.

6.1 Geocoding

The primary objective of the first part of the project is to create a high-quality dataset with the maximum number of geolocated points based on the addresses of each employment centre in Catalonia. Significant results have been obtained, which are discussed in detail below. The structure is as follows: First, a discussion of the preprocessing methods used; second, a discussion based on the geocoding methodologies; and finally, a discussion focused on the validation methods.

The main research question (section 2) for this section is: *Is there any method that can be applied to automate in an efficient way the conversion from addresses to geolocated points?*. The following sub-sections will give answer to its sub-questions.

6.1.1 Preprocessing

As stated in the introduction, the importance of good data quality has been emphasized, among many others, by Figueiredo & Pereira (2017) & Frankenfield (2022). This allows for the creation, exploitation, and maintenance of knowledge, as well as the production of valuable research. The ability to draw conclusions from data analysis is crucial for the development of society. Therefore, the creation of the Catalan employment centres' geolocation dataset has been an asset for further analysis and research. Although some results obtained from the geocoding may not be entirely accurate, a significant and reliable geocoded dataset has been obtained which can be utilized for further research and analysis related to this topic.

Multiple data cleaning and standardizing techniques have been applied to the dataset that not only made it easier to work with but also transformed it into a more user-friendly dataset. Initially, the dataset provided contained spelling mistakes, employment centres' individualization errors and lack of geocoding addresses. However, through the process that has been carried out almost all the necessary corrections were made in the mentioned aspects. As the article from Bhaya (2017) emphasizes, data preprocessing is a crucial step to work with data and improve its efficiency.

6.1.2 Geocoding methodologies

This sub-section answers partly all the research questions. The first one: *How can the process be automated in order to accelerate the process?*. It also partly helps to understand the question of *Is it possible to use open software for this process with meaningful results?*. And finally to the question: *Which geocoding services have better performance? Which can geocode more addresses?*. Note that the answer to these questions is not limited to this section, the others are complementary to answer all questions.

Although the results from the used methodologies for geocoding and automating the process can differ in their speed, they all can process the 15091 employment centres in less than an hour. The different methods are explained in the appendices, where all the scripts are detailed to make it easily understandable for the public. Moreover, the use of open software (ICGC, OSM and Google Maps) provided accurate results that highlight the potential of such free tools. Google Maps has performed particularly well in the geocoding process and its results have been similar to ArcGIS Pro, which requires a payment fee. The results for this process are described below. First the open software are explained, and later the paid methodology.

The Geocoder from the ICGC has been difficult to be discussed as it is a local (Catalonia) geocoder and not widely used in the GIS community. For this reason, no comparison or performance analysis research

papers have been found. Although it could seem to be a good geocoder for local addresses, other widely used APIs provided more suitable and accurate results.

Despite OSM is known to be a global and widely used geodatabase (*OpenStreetMap*, 2022), and that it has been used with high-quality results in scientific articles (Lemke et al., 2015), the ones obtained in this research have not been satisfactory. This relatively low performance of OSM, according to the municipality checking, requires further investigation. It is hypothesized that the lack or ambiguity on the address field, either related to the dataset or to the OSM geodatabase, are the cause of this low performance.

On the other hand, the Google Maps Geocoding API has given high-quality results. The format of the address in the variables used highly determines the performance of the geocoding process for Google Maps. The importance of the address format (spelling mistakes, broad and nonspecific addresses…) is also seen in Panasyuk et al. (2019). The address correction process has shown an increase in the number of well-located employment centres' (Table 7), which support the findings from the last mentioned author. Thus, it is crucial to invest time in preparing the correct address data to have a better location specification. This ensures to have greater accuracy in the results obtained. Although the results obtained from Google Maps have been slightly lower than those acquired using the ArcGIS Pro method, they are extremely precise and highlight the potential of the free geocoding tools.

Based on the municipality checking validation, ArcGIS Geocoding Addresses tool has the best geocoding performance among the methods applied. This geocoding method has already been considered a reliable tool for the geocoding processes in other studies (Jordan et al., 2022). Thus, the results obtained from this research reaffirm its excellent performance as a geocoding service application. However, it is worth noting that this method requires of a payment fee in contrast to all the other methods used in this project.

6.1.3 Geocoding validation

This sub-section answers the following questions:

- *Is it possible to use open software for this process with meaningful results?*
- *Which geocoding services have better performance? Which can geocode more addresses?*

The results from the methodologies have been obtained based on the results of the different validators. This section discusses the performance of each validator and their role in all the geocoding process. The validators are discussed one by one following the same order used in the methodology part.

No literature has been found regarding the first validation method, which suggest that it is an innovative approach. Moreover, it has been a key process to validate the accuracy of geocoding points to their respective municipalities, as well as to help determine which have been the best geocoding methods used (ArcGIS and GM3). It also works as a first filter to distinguish between the most useful techniques. Thus, the selected methodologies can be further analysed (by means of the third validation methodology) to end the dataset with the final coordinates.

The results obtained from the second validation method (distance from the centroid) have not been satisfactory for assessing the suitability of the different methods when geocoding. The limited number of nearby centroid locations obtained, which is insufficient for making proper validation decisions has been the main reason of why it has been rejected as a trustworthy validation application. Additionally, numerous well-located values were found when manually checking the values below the centroid-proximity threshold. The process does not provide valuable information for decision-making and therefore, it is not recommended for other researchers or users.

The third validation method revealed how ArcGIS performance is slightly better in quantity and quality compared to the Geocoding API from Google Maps, making it the best method for geocoding locations.

Focusing on this validation process, the manual checking of 95 random employment centres identified some common errors based on how addresses were written. A major part of the kilometric points in employment centres' addresses can be corrected using the name of the employment centre instead of the specific street name. In a related article, Prener, (2021) already make manual and distance comparisons between different geocoders.

The last checks, which involved combining the 3rd and the 1st validation methods, have been useful to determine which methods needed to be applied for each of the common errors. This also show how patterns can be identified by manually reviewing the dataset. Thus, not all the solutions rely on creating different validation variables.

6.1.4 Geocoding Summary

The author recognized that validating the geocoding results must go hand by hand with the data preprocessing to get the most suitable geocoded addresses. Depending on the characteristics, several geocoding and validating processes may be needed. This research section provides guidelines to geocode data and obtain precise results.

Based on the above discussion, a geocoding order has been established that depends on the type of addresses being geocoded: Addresses with kilometric points are geocoded first using the Google Maps last approach (name of the employment centre, city, and country). Second, the addresses without number use the GM3 geocoding process (street name and number, city, and country). Finally, the ArcGIS Pro method is used for the remaining, and more common addresses. This order can be beneficial for researchers and geocoders looking to perform similar operations. However, locations should be checked after the processes to validate the performance of each method.

6.2 Data Analysis

Barcelona, the capital of Catalonia, is the second largest city in Spain (see section 3 for more information). Metropolitan areas are key places for the country's economy development as evidenced by Longhi et al., (2014). This creates an opportunity to analyse quantitatively the most significant employment sectors in the BMA and their distribution. This matter forms the basis of this second part of the discussion. Nevertheless, it is challenging to discuss the results with existing literature as the literature obtained from the employment in the BMA is not completely aligned with this study.

The main research question (section 2) for this section is: *Which is Barcelona's metropolitan area economic spatial distribution?* The following sub-sections will give answer to its sub-questions.

6.2.1 Spatial Join

The Catalan Institute of Statistics (Idescat) published socioeconomic data from Catalan individuals in Census Tracts format to preserve their privacy (Lagonigro et al., 2017) which usually range from 500 to 2000 individuals over 18 years. Moreover, these areal measures have been widely used in research (Aune et al., 2020; MacIndoe & Oakley, 2022) as it is also the provided format for many of the socioeconomic data. For this reason, this research has only used this areal units. However, as mentioned in Li & Monzur (2018), there are differences in area between each census tract. The authors have created several grid cells nets (with different sizes) for all the study area based on the census tracts to assess their size effects. Moreover, using *fishnets* can help decrease the issues caused by what is known as Modifiable Areal Unit Problem (MAUP). This dilemma still needs to be further studied in this specific case; other articles have also transformed the data from census tracts to grid cells, an example is the paper from (Depsky et al., 2022). Not only significant differences appear in the area size but also between the 3rd and 4th quartile which may explain the clustering of data in some areas of the BMA. This highlights that data distribution can not be understood without mapping, and that a small percentage of tracts may include a high number of employees.

6.2.2 Basic Exploratory Analysis

This sub-section answers the research question *How is the BMA intraurban employment distribution? Which spatial employment pattern does it follow?*

Spatial analysis enhances the understanding of data by generating various maps and data visualizations. A map representing the absolute number of employees give a first insight of the employment distribution among all the tracts and provides a first overview of the most important employment areas in the BMA. The results section provides a list the municipalities that are part of these areas, which can be useful for the researchers, employment centres, or the public administration workers. Nevertheless, readers can not extrapolate the results of this map to all the employment sectors, as some specific sectors may only be present in tracts with a low number of employees. For more sector-specific information the researcher should create similar maps using the variables of interest (i.e., specific employment sectors) as described in the APPENDIX XII.

The density map offers an alternative view of how data is distributed in space. Density and employee's data has been used in other studies, such as the ones on the Tokyo Metropolitan Area and Los Angeles (Giuliano et al., 2007; Li & Monzur, 2018). Representation of the density is different between these articles. First, Giuliano et al., (2007) followed a similar process as the one carried out here obtaining data (employment and population) from census tracts and calculating and plotting the density based on these areal structures. However, as earlier mentioned, Li & Monzur (2018) went a step further and created grids over the census tracts data, reducing the influence of the census tracts areas. Density representation is then finer compared to the census tracts representation and can exemplify better delimitation results for the employment centre representation. It can either help to determine a monocentric or a polycentric structure for the BMA region.

6.2.3 Autocorrelation Analysis

This sub-section answers the following research questions:

- *How is the BMA intraurban employment distribution? Which spatial employment pattern does it follow?*
- *Are there any clusters per sectors in certain specific regions of the BMA?*

Moran plots provide a useful initial overview of data and its neighbours' trend. Their axes indicate which is the variable value for each record (x axis) and the estimated variable values of its neighbours (lagged) based on a previously computed weight matrix (y axis). Thus, you get a first understanding of how data may be plotted in space. At first, LISA results differentiate between HH, LL, LH and HL clusters (without considering their p-value/significance level). Later, the resulting LISA map can reveal important employment areas in the BMA based on their significance level. Given this is an innovative study, there is no literature for comparison in this specific area. However, based on the clusters' location, the areas with highest employment (represented as HH) are near the airport (el Prat de Llobregat), the harbour (Zona Franca), Barcelona centre and the 22@Barcelona district, which is a Barcelona area that promotes technology and innovation. The two first areas can be considered important industrial and logistic hubs, and the latter two as more centred to Knowledge-intensive business services (KIBS). Clustered industrial and logistic sectors are located in larger census tracts, far from the city centre (less population density), and close to transportation cores. On the other hand, KIBS cluster sectors are mainly located in small census tracts (higher population density) from centric areas where non-commuting transportation is less important. Moreover, there are other important clusters in the edges of the BMA where chemical and pharmaceutical products, wholesale and retail trade, financial services, and computer and programming services have the highest number of employment.

Another interesting category is the LH which represent tracts with low employment but high values in their surroundings. These locations may lead to high employment places due to their proximity with important economic areas. As mentioned earlier, LISA map only displays significant results based on

the neighbours of each census tract. Therefore, other locations with high employment values may not be represented on the map. To gain a more detailed overview of how employment is distributed across the region, it is necessary to use the absolute employee's map together with the LISA map.

Furthermore, LISA analysis can be performed in a sectoral scale using the adapted CNAE classification which allows the determination of spatially autocorrelated employment clusters for each class. This method offers flexibility and can be customized depending on the study needs. To demonstrate the capabilities of this developed tools, the six sectoral maps have been created. Bibliographic research provided insights on different approaches to sectoral classifications. Some studies focus on all the sectors without any specification (Giuliano et al., 2007), others have created classes for all the employment sectors (Li & Monzur, 2018), and others are specialized in specific employment sectors (Kaygalak & Reid, 2016; Rivza et al., 2018). This variability shows the importance of selecting the appropriate division limits depending on the research focus. The datasets together with the R scripts and programs create a powerful tool that allow researchers choose between any of these study groupings for the BMA.

6.2.4 Map creation

This sub-section give answer to the question of *How can maps be displayed in a way that are easy to understand and provide significant and accurate results about the analytical study?*

There is the need to think how the public, including experts and non-experts, can understand maps and draw conclusions out of them. The final maps results of this study demonstrate a simple yet effective way to understand how the employment data is distributed in the space. The well-designed colour classifications (Red's ramp and the LISA specific categorization) with their respective intensities helped differentiate between groups and clusters. Moreover, fine lines have been used to improve the display of census tracts and to clearly distinguish between municipalities. This research demonstrate how easy-to-understand maps can be created using the R software for research purposes.

Moreover, using similar mapping codes facilitates the reproduction of different data, making map replicates easy to be created. The most efficient map creation has been performed with the final LISA and the *lapply* function which created a total of 58 LISA maps without interruption.

6.2.5 Data Analysis Summary

After obtaining the main dataset with census tracts data and employment, the potential of research is huge. For this reason, the spatial data analytical part provides an overview of how data is distributed across space, as well as different ways to interpret it based on the required level of detail. The resulting maps offer significant and interesting layouts which facilitate the comprehension of employment distribution within the BMA. These findings are part of an innovative study which has the potential to be continuously expanded, as outlined in the Future Research section (section 8).

Overall, regarding to the main question *Is it possible to geolocate all the employment centre's addresses from Catalonia and understand the spatial employment distribution in Barcelona Metropolitan Area?*, almost all employment centre addresses from Catalonia have been able to be geocoded with good accuracy. Although a first understanding of the spatial employment distribution in the BMA has been done through a LISA analysis, there is the need to do further research using the dataset provided to obtain more accurate and interesting results.

7. Conclusion

The Barcelona Metropolitan Area is characterized by a complex employment structure that has been the focus of various studies, i.e., the ones from Coll-Martínez et al. (2017), Maddah et al. (2021), Garcia-López & Muñiz (2010) and Marmolejo et al. (2010). However, none could examine its structure in high detail. This research has provided a promising avenue for future research in the employment structure, either for BMA, and for Catalonia, as is the first time a complete dataset about employment centre locations is available. The original dataset contains 15,091 addresses.

Address geocoding has been a topic of interest for many researchers. There are numerous methods which can provide satisfactory results. However, the large number of addresses available (15,091) gave difficulties in assessing the performance of the geocoding methods applied. A total of four geocoders and two validation methods have been selected to assess the most accurate method for geocoding the addresses. The results showed how the combination of two high-quality performing geocoders has been the best approach, achieving almost the 99% accuracy in geocoding the addresses. Thus, by trying different geocoders and analysing their performance, users can determine the most suitable methods for geocoding addresses, always considering the different patterns the addresses may follow.

Geocoding the locations of the companies in Catalonia has provided valuable insights into its spatial employment distribution. Moreover, data can be joined to other datasets to obtain more relevant results that help understand the location of specific employment centres or sectors. Spatial distribution of employment data, for example, can provide good insights for future urbanisation, household location and distribution by income, patterns of commuting, transport networks and other social or economic actions that can be used by companies or administrative bodies.

With the dataset created, the employment data format is now available in single point features, which allows it to be used in any type of feature or raster and offer a more comprehensive and systematic picture of employment distribution. Socioeconomic data is usually provided in census tracts which although this format has been widely used, it suffers from Modifiable Area Unit Problems.

LISA results have shown how employment data present spatial autocorrelation, and therefore helped to understand the spatial distribution of employment in the BMA. This study has generated some map outputs and tools which can be used for studying this spatial distribution and obtain significant results from it. However, the study did not define any polycentric or monocentric structures in the BMA. Future research using the new dataset will allow to study if there are more and significant employment centres that would challenge the monocentric urban model mainly used to analyse urban economics. To define other possible sub-centres of employment, there is the need to obtain more accurate data and define an employment and areal threshold, as it is done in Li & Monzur (2018). High employment clusters show the most significant results in Barcelona Municipality, the main industrial and logistic hub (Zona Franca), and some peripherical areas. Although Barcelona municipality concentrates high economic activity and many employment centres, not all high employment locations are found there.

This research has succeeded in providing the first companies' geolocation dataset. Moreover, it used it to lay the foundations to study accurately the employment distribution in the BMA. From this point, further research needs to be done.

8. Future Research

This section describes and discusses possible future applications and improvements aimed at obtaining more accurate employment and socioeconomic results, all of which are based on the datasets created.

Firstly, the MAUP has been discussed in section 6.2.1 but no detailed solution to this problem has been provided. It has been seen that census tract data have different sizes which can affect the interpretation of data in a map. As already described, the article from Li & Monzur (2018) could inspire the usage of a grid cell system to minimize the areal size differences. However, this process estimates that the socioeconomic variables are constant across each of the census tracts, which implies giving the same value to smaller grids (not getting finer socioeconomic data). In addition, if a grid cell is found between two census tracts of different sizes (one large and one small) and the mean is calculated between both, it is considered that the data from the big census tract is close to the grid cell that is being used: It is probable that the data used is far away from the grid cell. In conclusion, obtaining more accurate data and results requires relying on public bodies to provide increasingly precise information in the future.

Secondly, monocentric and polycentric structures have not been defined yet. Articles like the ones from Fernández-Maldonado et al. (2014), Giuliano et al. (2007) and Li & Monzur (2018) suggest possible applications of how these urban structures have been determined from census tracts data. Thus, a first identification of possible sub-centres of employment in the BMA can be defined.

Finally, Madariaga et al. (2014) show how regression models can be applied to better understand of how, for instance, the distance to the main CBD or subcentres affects the income. Spatial Lag and Spatial Error models can be used for these analyses. Moreover, other analyses based on spatially modelling the relation between employment quantity and other socioeconomic variables can be performed. All these analyses have to consider the spatial autocorrelation (obtained from the Moran's I) to not make errors.

While some example applications have been provided above, it is up to the researchers to explore further methodologies to get a more comprehensive understanding of the employment distribution and its socioeconomic context in the BMA.

9. References

Abbafati, C., Machado, D. B., Cislaghi, B., Salman, O. M., Karanikolos, M., McKee, M., Abbas, K. M., Brady, O. J., Larson, H. J., Trias-Llimós, S., Cummins, S., Langan, S. M., Sartorius, B., Hafiz, A., Jenabi, E., Mohammad Gholi Mezerji, N., Borzouei, S., Azarian, G., Khazaei, S., … Zhu, C. (2020). Global age-sex-specific fertility, mortality, healthy life expectancy (HALE), and population estimates in 204 countries and territories, 1950–2019: a comprehensive demographic analysis for the Global Burden of Disease Study 2019. *The Lancet*, *396*(10258), 1160–1203. https://doi.org/10.1016/S0140-6736(20)30977-6

Abozeid, A. S. M., & AboElatta, T. A. (2021). Polycentric vs monocentric urban structure contribution to national development. *Journal of Engineering and Applied Science*, *68*(1), 1–18. https://doi.org/10.1186/s44147-021-00011-1

Alidadi, M., & Dadashpoor, H. (2018). *Beyond monocentricity : examining the spatial distribution of employment in Tehran metropolitan region , Iran*. *5934*. https://doi.org/10.1080/12265934.2017.1329024

Allaire, J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., Wickham, H., Cheng, J., Chang, W., & Iannone, R. (2023). *rmarkdown: Dynamic Documents for R*. *R package*. https://github.com/rstudio/rmarkdown

Àrea Metropolitana de Barcelona. (n.d.). Retrieved March 2, 2023, from https://www.amb.cat/s/home.html

Armstrong, M. P., Rushton, G., & Zimmerman, D. L. (1999). Geographically masking health data to preserve confidentiality. *Statistics in Medicine*, *18*(5), 497–525. https://doi.org/10.1002/(sici)1097-0258(19990315)18:5<497::aid-sim45>3.0.co;2-%23

Aune, K. T., Gesch, D., & Smith, G. S. (2020). A spatial analysis of climate gentrification in Orleans Parish, Louisiana post-Hurricane Katrina. *Environmental Research*, *185*. https://doi.org/10.1016/j.envres.2020.109384

Bajak, A. (2017). *How to geocode a csv of addresses in R – storybench*. https://www.storybench.org/geocode-csv-addresses-r/

Bhaya, W. S. (2017). *Review of Data Preprocessing Techniques in Data Mining*.

Bivand, R. S., Pebesma, E. J., & Gómez-Rubio, V. (2008). *Spatial Analysis with R*.

Brewer, C. A. (2006). Basic mapping principles for visualizing cancer data using geographic information systems (GIS). *American Journal of Preventive Medicine*, *30*(2 SUPPL.), 25–36. https://doi.org/10.1016/j.amepre.2005.09.007

Coll-Martínez, E., Moreno-Monroy, A. I., & Arauzo-Carod, J. M. (2019). Agglomeration of creative industries: An intra-metropolitan analysis for Barcelona. *Papers in Regional Science*, *98*(1), 409–431. https://doi.org/10.1111/pirs.12330

Depsky, N. J., Cushing, L., & Morello-Frosch, R. (2022). High-resolution gridded estimates of population sociodemographics from the 2020 census in California. *PLoS ONE*, *17*(7 July), 1–21. https://doi.org/10.1371/journal.pone.0270746

Diana, S., Christina, A., Agung, A., & Gunawan, S. (2021). Fiscal decentralization analysis that affect economic performance using geographically weighted regression (GWR). *Procedia Computer Science*, *179*(2020), 399–406. https://doi.org/10.1016/j.procs.2021.01.022

Duncan, D. T., Castro, M. C., Blossom, J. C., Bennett, G. G., & Steven, L. (2011). *Evaluation of the positional difference between two common geocoding methods*. *5*(2), 265–273.

Fan, C., Chen, M., Wang, X., Wang, J., & Huang, B. (2021). A Review on Data Preprocessing

Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data. *Frontiers in Energy Research*, *9*(March), 1–17. https://doi.org/10.3389/fenrg.2021.652801

Fernández-Maldonado, A. M., Romein, A., Verkoren, O., & Parente Paula Pessoa, R. (2014). Polycentric Structures in Latin American Metropolitan Areas: Identifying Employment Sub-centres. *Regional Studies*, *48*(12), 1954–1971. https://doi.org/10.1080/00343404.2013.786827

Figueiredo, M. S. N., & Pereira, A. M. (2017). Managing Knowledge – The Importance of Databases in the Scientific Production. *Procedia Manufacturing*, *12*(December 2016), 166–173. https://doi.org/10.1016/j.promfg.2017.08.021

Frankenfield, J. (2022). *Data Analytics: What It Is, How It's Used, and 4 Basic Techniques*. https://www.investopedia.com/terms/d/data-analytics.asp

Garcia-López, M. À., & Muñiz, I. (2010). Employment decentralisation: Polycentricity or scatteration? the case of Barcelona. *Urban Studies*, *47*(14), 3035–3056. https://doi.org/10.1177/0042098009360229

García, M. Á., & Muñiz, I. (2005). Employment Decentralisation: Polycentric compaction or sprawl? The case of the Barcelona metropolitan region. *Departament d'Economia Aplicada UAB*.

Generalitat de Catalunya. (2019). *Trets de l'economia catalana*. 122. http://economia.gencat.cat

Geocode Addresses (Geocoding)—ArcGIS Pro | Documentation. (n.d.). Retrieved October 13, 2022, from https://pro.arcgis.com/en/pro-app/latest/tool-reference/geocoding/geocode-addresses.htm

Getis, A. (2007). Reflections on spatial autocorrelation. *Regional Science and Urban Economics*, *37*(4), 491–496. https://doi.org/10.1016/j.regsciurbeco.2007.04.005

Giuliano, G., Redfearn, C., Agarwal, A., Li, C., & Zhuang, D. (2007). Employment concentrations in Los Angeles, 1980-2000. *Environment and Planning A*, *39*(12), 2935–2957. https://doi.org/10.1068/a393

Google Cloud Platform. (2022). https://cloud.google.com/

Hacienda, M. de E. y. (2007). 28 de abril, por la que se aprueba la Clasificación Nacional de Actividades Económicas 2009 (CNAE-2009). *Boletín Oficial Del Estado*, *102*(28 de abril de 2007,), 18574-18593 (Anexo).

Hellner, J. (2021). *Digitalization in the rail industry : Localizing damaged cargo wagons using spatial operations and big data*. *560*.

Huai, Y., Lo, H. K., & Ng, K. F. (2021). Monocentric versus polycentric urban structure: Case study in Hong Kong. *Transportation Research Part A: Policy and Practice*, *151*(April), 99–118. https://doi.org/10.1016/j.tra.2021.05.004

ICGC. (2020). *Document tècnic de client de geocodificació massiva*.

Indriyani, R. A. A., & Widaningrum, D. L. (2021). A spatial equity assessment of the public facilities in the greater Jakarta area using Moran's i spatial autocorrelation. *IOP Conference Series: Earth and Environmental Science*, *794*(1). https://doi.org/10.1088/1755-1315/794/1/012090

Información de Empresas Españolas | eInforma. (n.d.). Retrieved March 2, 2023, from https://www.einforma.com/

Jordan, L., Elessawy, M., & Munro-ludders, G. (2022). Geographical dataset of firearms manufacturing in the United States : *Data in Brief*, *45*, 108626. https://doi.org/10.1016/j.dib.2022.108626

Kahle, D., Wickham, H., Jackson, S., & Korpela, M. (2019). *Package ' ggmap .'*

Kaygalak, I., & Reid, N. (2016). The geographical evolution of manufacturing and industrial policies in Turkey. *Applied Geography*, *70*(May), 37–48. https://doi.org/10.1016/j.apgeog.2016.01.001

Kounadi, O., & Leitner, M. (2015). Defining a threshold value for maximum spatial information loss of masked geo-data. *ISPRS International Journal of Geo-Information*, *4*(2), 572–590. https://doi.org/10.3390/ijgi4020572

Küçük Matci, D., & Avdan, U. (2018). Address standardization using the natural language process for improving geocoding results. *Computers, Environment and Urban Systems*, *70*(January), 1–8. https://doi.org/10.1016/j.compenvurbsys.2018.01.009

Lagonigro, R., Martori, J. C., & Apparicio, P. (2018). Environmental noise inequity in the city of Barcelona. *Transportation Research Part D: Transport and Environment*, *63*(June), 309–319. https://doi.org/10.1016/j.trd.2018.06.007

Lagonigro, R., Martori, J. C., & Apparicio, P. (2020). Understanding Airbnb spatial distribution in a southern European city: The case of Barcelona. *Applied Geography*, *115*(August 2019), 102136. https://doi.org/10.1016/j.apgeog.2019.102136

Lagonigro, R., Oller, R., & Martori, J. C. (2017). A quadtree approach based on European geographic grids: Reconciling data privacy and accuracy. *Sort*, *41*(1), 139–158.

Lemke, D., Mattauch, V., Heidinger, O., & Hense, H. W. (2015). *Wer trifft ins Schwarze ? Ein qualitativer Vergleich der kostenfreien Geokodierungsdienste von Google und OpenStreetMap Who Hits the Mark ? A Comparative Study of the Free Geocoding Services of*. 160–165.

Li, Y., & Monzur, T. (2018). The spatial structure of employment in the metropolitan region of Tokyo: A scale-view. *Urban Geography*, *39*(2), 236–262. https://doi.org/10.1080/02723638.2017.1308182

Liboreiro, P. R., Fernández, R., & García, C. (2021). The drivers of deindustrialization in advanced economies : A hierarchical structural decomposition analysis. *Structural Change and Economic Dynamics*, *58*, 138–152. https://doi.org/10.1016/j.strueco.2021.04.009

Longhi, C., Musolesi, A., & Baumont, C. (2014). Modeling structural change in the European metropolitan areas during the process of economic integration. *Economic Modelling*, *37*(October 2017), 395–407. https://doi.org/10.1016/j.econmod.2013.10.028

Lovelace, R., Nowosad, J., & Muenchow, J. (2019). Geocomputation with R. *Chapman & Hall*.

MacIndoe, H., & Oakley, D. (2022). Encouraging a Spatial Perspective in Third Sector Studies: Exploratory Spatial Data Analysis and Spatial Regression Analysis. *Voluntas*. https://doi.org/10.1007/s11266-022-00459-6

Madariaga, R., Martori, J. C., & Oller, R. (2014). Income, distance and amenities. An empirical analysis. *Empirical Economics*, *47*(3), 1129–1146. https://doi.org/10.1007/s00181-013-0772-8

Madariaga, R., Martori, J. C., & Oller, R. (2019). Wage income inequality in Catalonian second-rank cities. *Annals of Regional Science*, *62*(2), 285–304. https://doi.org/10.1007/s00168-019-00896-0

Maddah, L., Arauzo-Carod, J. M., & López, F. A. (2021). Detection of geographical clustering: cultural and creative industries in Barcelona. *European Planning Studies*, *0*(0), 1–22. https://doi.org/10.1080/09654313.2021.2020218

Marmolejo, C., Aguirre, C., & Roca, J. (2010). Revisiting employment density as a means to detect metropolitan sub-centres: An analysis for Barcelona and Madrid. *Architecture, City and Environment*, *23*, 33–63. https://doi.org/10.581/ace.8.23.2596

Martori, J. C., Madariaga, R., & Oller, R. (2016). Real estate bubble and urban population density: six Spanish metropolitan areas 2001–2011. *Annals of Regional Science*, *56*(2), 369–392. https://doi.org/10.1007/s00168-016-0743-z

McKinney, W. (2013). *Python for Data Analysis*.

Nominatim. (2022). https://nominatim.org/

Oller, R., Martori, J. C., & Madariaga, R. (2017). *Monocentricity and Directional Heterogeneity : A Conditional Parametric Approach*. 343–361. https://doi.org/10.1111/gean.12119

OpenStreetMap. (2022). https://www.openstreetmap.org

Panasyuk, A., Yu, E. S., & Mehrotra, K. G. (2019). Improving Geocoding for City-level Locations. *2019 IEEE 13th International Conference on Semantic Computing (ICSC)*, 416–421. https://doi.org/10.1109/ICSC.2019.00081

Prener, C. G. (2021). *Creating open source composite geocoders : Pitfalls and opportunities*. 1868–1887. https://doi.org/10.1111/tgis.12741

Rivza, B., Kruzmetra, M., & Sunina, L. (2018). Changes in composition and spatial distribution of knowledgebased economy in rural areas of Latvia. *Agronomy Research*, *16*(3), 862–871. https://doi.org/10.15159/AR.18.147

Saputra, H. Y., & Radam, I. F. (2022). Accessibility model of BRT stop locations using Geographically Weighted regression (GWR): A case study in Banjarmasin , Indonesia. *International Journal of Transportation Science and Technology*, *xxxx*. https://doi.org/10.1016/j.ijtst.2022.07.002

Schootman, M., Sterling, D. A., Struthers, J., Yan, Y. A. N., Laboube, T. E. D., Emo, B., & Higgs, G. (2006). *Positional Accuracy and Geographic Bias of Four Methods of Geocoding in Epidemiologic Research*. *13*. https://doi.org/10.1016/j.annepidem.2006.10.015

Shi, H., Su, R., Xiao, J., & Goulias, K. G. (2022). Spatiotemporal analysis of activity-travel fragmentation based on spatial clustering and sequence analysis. *Journal of Transport Geography*, *102*(April), 103382. https://doi.org/10.1016/j.jtrangeo.2022.103382

Spatial joins by feature type—ArcMap | Documentation. (n.d.). Retrieved March 2, 2023, from https://desktop.arcgis.com/en/arcmap/latest/manage-data/tables/spatial-joins-by-feature-type.htm

Team, R. C. (2022). R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria*. https://www.r-project.org/

Tran, A. B. (2018). *Geolocating data :: Journalism with R*. https://learn.r-journalism.com/en/mapping/geolocating/geolocating/

United Nations Development Group. (2017). *UNDG Guidance Note on Big Data for Achievement of the 2030 Agenda : Data Privacy, Ethics, and Protection*. 16. https://unsdg.un.org/sites/default/files/UNDG_BigData_final_web.pdf

Zhan, F. B., Brender, J. D., Lima, I. D. E., Suarez, L., & Langlois, P. H. (2006). *Match Rate and Positional Accuracy of Two Geocoding Methods for Epidemiologic Research*. https://doi.org/10.1016/j.annepidem.2006.08.001

10. APPENDICES

APPENDIX I

This part has been performed by the DAM group from UVic-UCC. It is composed of one function and a while loop. Section 4.1.1.

The **function** is found below:

```
nominatim_osm <- function(address = NULL)
{
  if(suppressWarnings(is.null(address)))
    return(data.frame())
  tryCatch(
    d <- jsonlite::fromJSON(
      gsub('\\@addr\\@', gsub('\\s+', '\\%20', address),
'http://nominatim.openstreetmap.org/search/@addr@?format=json&addressdetails=0&lim
it=1&countrycodes=es')
    ), error = function(c) return(data.frame())
  )
  if(length(d) == 0) return(data.frame(lon=NA, lat=NA))
  return(data.frame(lon = as.numeric(d$lon), lat = as.numeric(d$lat))) }
```

And below the **while loop**:

```
inici<-1
while (inici<length(CensLaboralCat_ray$id)) {
  final <- min( inici + 99, length(CensLaboralCat_ray$id))
  cat(inici, " <-> ", final)
  addresses <- CensLaboralCat_ray[inici:final,]$address2
  d <- suppressWarnings(lapply(addresses, function(address) {
    api_output <- nominatim_osm(address)
    return(data.frame(address = address, api_output))
    }) %>%
    bind_rows() %>% data.frame())
  CensLaboralCat_ray[inici:final,]$lon<-d$lon
  CensLaboralCat_ray[inici:final,]$lat<-d$lat
  cat(" Fet", "\n" )
  inici <- final + 1
  Sys.sleep(1.25) }
```

APPENDIX II

The code below was done by the DAM group from UVic-UCC. Section 4.1.1.

```
Private Sub Comprovar_Click()
    On Error GoTo 0
    Dim url, address As String
    Dim json As Object
    If (ActiveCell.Column = 6) Then
        address = ActiveCell + ", " + ActiveCell.Offset(0, columnOffset:=1)
        url = "https://nominatim.openstreetmap.org/search/?q=" & address & "&forma
t=json&addressdetails=0&limit=1&countrycodes=es"
        Set json = ParseJSON(Application.WorksheetFunction.WebService(url))
        If (IsEmpty(json("obj(0).lon"))) Then
            MsgBox ("Coordenades no trobades")
        Else
            ActiveCell.Offset(0, columnOffset:=-2) = json("obj(0).lat")
            ActiveCell.Offset(0, columnOffset:=-1) = json("obj(0).lon")
        End If
    Else
        MsgBox ("Cal situar-se a la columna F")
    End If
    ActiveCell.Select
End Sub
```

APPENDIX III

This process has been done for the **3 different Google Maps approaches**. With the corresponding changes on the variable field for each of it. Section 4.1.1.

```
knitr::opts_chunk$set(eval = FALSE)
register_google(key = "YOURKEY", write = TRUE)
for(i in 1:nrow(df_GoogleMaps))
{
  # Print("Working...")
  result <- geocode(df_GoogleMaps$ADRECA.CENTRE.TREBALL[i], output = "latlona", so
urce = "google")
  df_GoogleMaps$lon[i] <- as.numeric(result[1])
  df_GoogleMaps$lat[i] <- as.numeric(result[2])
}
```

APPENDIX IV

Section 4.1.1.

```
for(i in 1:nrow(df_GoogleMaps_v1.2))
{
 df_GoogleMaps_v1.2$ADRECA.COMPLETA[i] <- paste0(df_GoogleMaps_v1.2$ADRECA.CENTRE.
TREBALL[i], ", ", df_GoogleMaps_v1.2$POBLACIO.NOM[i], ", SPAIN")
}
```

APPENDIX V

This process has been done for all the methodologies (**ICGC, OSM, ArcGIS and Google Maps**). Section 4.1.2.

```
for(k in 1:nrow(CensProva)){
  #print(CensProva[k,]$POBLACIO.NOM.C)
  try(
    if(st_covered_by(CensProva[k,], municipis2[municipis2$MUNICIPI==CensProva[k,]$
POBLACIO.NOM.C, ])[[1]]>0)
      CensProva[k,]$Valid <- as.integer(1), silent = T
  )

}
```

APPENDIX VI

Section 4.1.2.

```
centroids2 <- st_centroid(municipis2)
CensProva10 <- st_as_sf(CensLaboralCatGeocodificat_V1.7, coords=c(18,17))
CensProva10['Valid_Centroid_GM3'] <- NA
CensProva10$Valid_Centroid_GM3 <- as.integer(CensProva10$Valid_Centroid_GM3)

CensProva10 <- st_set_crs(CensProva10, 4326)
CensProva10 <- st_transform(CensProva10, 4326)
centroids2 <- st_set_crs(centroids2, 4326)
centroids2 <- st_transform(centroids2, 4326)


for(t in 1:nrow(CensProva10)){
  result2 <- st_distance(CensProva10[t, ], centroids2[centroids2$MUNICIPI==CensPro
va10[t, ]$POBLACIO.NOM.C, ])
  CensProva10[t, ]$Valid_Centroid_GM3 <- as.numeric(result2[1])
}
```

APPENDIX VII

This part is divided in two. An **R Script** to know which are the *names written differently* in the specific columns between two datasets. And then some **Python** coding examples of how the errors related to names differences is fixed. Section 4.1.3.

The first approach is following:

```
CensProva[ !(CensProva$POBLACIO.NOM.x %in% municipis2$MUNICIPI),]$POBLACIO
.NOM.x
```

The second approach is the following:

```
df['POBLACIO.NOM.2'] = np.where(df['POBLACIO.NOM'].str.contains(", el"),"el ","")
df['POBLACIO.NOM'] = df['POBLACIO.NOM'].str.replace(', el','')
df['POBLACIO.NOM.2'] = np.where(df['POBLACIO.NOM'].str.contains(", la"),"la ",df['
POBLACIO.NOM.2'])
df['POBLACIO.NOM'] = df['POBLACIO.NOM'].str.replace(', la','')
df['POBLACIO.NOM.2'] = np.where(df['POBLACIO.NOM'].str.contains(", l'"),"l'",df['P
OBLACIO.NOM.2'])
df['POBLACIO.NOM'] = df['POBLACIO.NOM'].str.replace(", l'",'')
df['POBLACIO.NOM.2'] = np.where(df['POBLACIO.NOM'].str.contains(", els"),"els ",df
['POBLACIO.NOM.2'])
df['POBLACIO.NOM'] = df['POBLACIO.NOM'].str.replace(', els','')
df['POBLACIO.NOM.2'] = np.where(df['POBLACIO.NOM'].str.contains(", les"),"les ",df
['POBLACIO.NOM.2'])
df['POBLACIO.NOM'] = df['POBLACIO.NOM'].str.replace(', les','')
df['POBLACIO.NOM.C'] = df['POBLACIO.NOM.2'] + df['POBLACIO.NOM']

df['POBLACIO.NOM.C'] = df['POBLACIO.NOM.C'].str.replace('Berà','Barà')
df['POBLACIO.NOM.C'] = df['POBLACIO.NOM.C'].str.replace('el Prats de Reis','els Pr
ats de Rei')
df['POBLACIO.NOM.C'] = df['POBLACIO.NOM.C'].str.replace('el Guiametss','els Guiame
ts')
df['POBLACIO.NOM.C'] = df['POBLACIO.NOM.C'].str.replace('el Pallaresoss','els Pall
aresos')
df['POBLACIO.NOM.C'] = df['POBLACIO.NOM.C'].str.replace('el Hostalets de Pierolas'
,'els Hostalets de Pierola')
df['POBLACIO.NOM.C'] = df['POBLACIO.NOM.C'].str.replace('el Hostalets de Pierolas'
,'els Hostalets de Pierola')
df['POBLACIO.NOM.C'] = df['POBLACIO.NOM.C'].str.replace('Vimbodí i Poblet','Vimbod
í')
df['POBLACIO.NOM.C'] = df['POBLACIO.NOM.C'].str.replace('Saus, Camallera i Llampai
es','Saus')
df['POBLACIO.NOM.C'] = df['POBLACIO.NOM.C'].str.replace('Brunyola i Sant Martí Sap
resa','Brunyola')
```

The CENS.LABORAL refers to the number of employees.

The NUM.EMPRESES refers to the number of Employment Centres.

Section 4.2.

```
joined_table = st_join(SeccionsCensals, new_DF_geom_BMA, join = st_contains_proper
ly) |> #make the spatial join
  group_by(CUSEC) |> #group code by Census Tract
  summarize(NUM.EMPRESES=length(CUSEC), CENS.LABORAL = sum(CENS.LABORAL))
joined_table = st_drop_geometry(joined_table)

SeccionsCensals_CENS = merge(x = SeccionsCensals, y = joined_table, by = "CUSEC",
all = TRUE)



#give 0 to all NA values
SeccionsCensals_CENS["CENS.LABORAL"][is.na(SeccionsCensals_CENS["CENS.LABORAL"])]
= 0
SeccionsCensals_CENS["NUM.EMPRESES"][is.na(SeccionsCensals_CENS["NUM.EMPRESES"])]
= 0
#give 0 value to the NUM.EMPRESES field where the CENS.LABORAL is 0, as they mean
they don't have any employment centre
for (a in 1:nrow(SeccionsCensals_CENS)){
  try(
    if(SeccionsCensals_CENS[a,]$CENS.LABORAL == "0")
      SeccionsCensals_CENS[a,]$NUM.EMPRESES = "0"
  )
}

#Get the map with only the census tracts that have employment centre
SeccionsCensals_CENS_ambCENS = SeccionsCensals_CENS[SeccionsCensals_CENS$CENS.LABO
RAL>0, ]
```

APPENDIX IX

In here the adapted BMA classification groups based on the CNAE is found. Section 4.2.

Nomenclature	Section	Division	Group	Description
AAPP	O		841	Administration of the State and the economic and social policies of the community
ACTBibliotecasArchivos	R	91		Libraries, archives, museums and other cultural activities
ACTCreacion	R	90		Arts creation and performing arts activities
ACTDeportRecreativas	R	93		Sports activities and amusement and recreation activities
ACTExtraterritoriales	U	99		Activities of extraterritorial organizations and bodies
ACTHogares	T	97:98		Activities of households as employers of domestic personnel and Undifferentiated goods- and service-producing activities of private households for own use
ACTJuegosApuestas	R	92		Gambling and betting activities
ACTResidencias	Q	87		Residential care activities
ACTSanitarias	Q	86		Human health activities
ACTSerSocialNoAloj	Q	88		Social work activities without accommodation
AgenciasViajes	N	79		Travel agency, tour operator and other reservation service and related activities
AgriculturaGanaderia	A	01		Crop and animal production, hunting and related service activities
AlimBebidasTabaco	C	10:12		Manufacture of food products, beverages and tobacco products
Almacenamiento	H	52		Warehousing, storage and support activities for transportation
Alojamiento	I	55		Accommodation
Alquiler	N	77		Rental and leasing activities
ArquitecturaIngenieria	M	71		Architectural and engineering activities; technical testing and analysis
Asociaciones	S	94		Activities of membership organizations
AuxOfi	N	82		Office administrative, office support and other business support activities
CaptRecogAguas	E	36:37		Water collection, treatment and supply, and sewerage
CauchoPlastNoMetal	C	22:23		Manufacture of rubber and plastic products, and of other non-metallic mineral products
CineTelesonido	J	59		Motion picture, video and television programme production, sound recording and music publishing activities
Comercio	G	46:47		Wholesale and retail trade, except motor vehicles and motorcycles
ComidasBebidas	I	56		Food and beverage service activities
Construccion	F	41:43		Construction of residential and non-residential buildings, civil engineering, and specialized construction activities
Consultoria	M	70		Activities of head offices and management consultancy
Correos	H	53		Postal and courier activities
Edicion	J	58		Publishing activities
Educacion	P	85		Education
EmisionRadioTele	J	60		Programming, broadcasting, news agency and other content distribution activities
Empleo	N	78		Employment activities
FinanzasSeguros	K	64:66		Financial services activities, insurance, reinsurance and pension funds except compulsory social security, and activities auxiliary to financial services and insurance activities
ID	M	72		Scientific research and development

Nomenclature	Section	Division	Group	Description
Immobiliarias	L	68		Real estate activities
IndustriaExtractiva	B	08		Other mining and quarrying (see others from classification to better understand)
InfoElectro	C	26:27		Manufacture of computer, electronic and optical products, and manufacture of electrical equipment
InformaticaInformacion	J	62:63		Computer programming, consultancy and related activities, and computing infrastructure, data processing, hosting and other information service activities
JuridicasContabilidad	M	69		Legal and accounting activities
LimpJardin	N	81		Services to buildings and landscape activities
MaderaPapelArtesgraf	C	16:18		Manufacture of wood and of products of wood and cork, except furniture; manufacture of articles of straw and plaiting materials. Manufacture of paper and paper products, and Printing and reproduction of recorded media
MaquinaVehiculosTrans	C	28:30		Manufacture of machinery and equipment n.e.c., Manufacture of motor vehicles, trailers and semi-trailers, and Manufacture of other transport equipment
Metal	C	24:25		Manufacture of basic metals and Manufacture of fabricated metal products, except machinery and equipment
OtrasProfVeterinaria	M	74:75		Other professional, scientific and technical activities, and veterinary activities
OtrasReparacion	C	31:33		Manufacture of furniture, other manufacturing, and Repair, maintenance and installation of machinery and equipment
OtrosServiciosPers	S	96		Personal service activities
PescaAcuicultura	A	03		Fishing and aquaculture
PrestacionServicios	O		842	Provision of services to the community as a whole
PublicidadMKT	M	73		Activities of advertising, market research and public relations
QuimFarma	C	20:21		Manufacture of chemicals and chemical products, and Manufacture of basic pharmaceutical products and pharmaceutical preparations
Reparaciones	S	95		Repair and maintenance of computers, personal and household goods, and motor vehicles and motorcycles
ResiduosDescontaminacion	E	38:39		Waste collection, recovery and disposal activities, and Remediation activities and other waste management service activities
SeguridadInvestig	N	80		Investigation and security activities
SS	O		843	Compulsory social security activities
SuministroEnergetico	D	35		Electricity, gas, steam and air conditioning supply
Telecos	J	61		Telecommunication
TextilConfeccionCuero	C	13:15		Manufacture of textiles, wearing apparel, and leather and related products of other materials
Transporte	H	49:51		Land transport and transport via pipelines, Water transport, and Air transport
Vehiculos	G	45		Sale and repair of motor vehicles and motorcycles

The description of the above table has been translated based on the Eurostat database standards (NACE). This allows comparison between different countries and easier understanding.

APPENDIX X

Functions for making the adapted classifications of the CNAE. Section 4.2.

```
#Classification for the ranges that contain 2 digits
Class_CNAE_v2 = function(Min, Max, Nom, Seccions) {
  join_CNAE = st_join(SeccionsCensals_CENS_ambCENS, new_DF_geom_BMA[between(new_DF
_geom_BMA$CNAE.ID_2v_Num, Min, Max), ], join = st_intersects) |>
    group_by(CUSEC) |>
    summarize(NUM.EMPRESES.Nom=length(CUSEC), CENS.LABORAL.Nom = sum(CENS.LABORAL)
)
  join_CNAE = st_drop_geometry(join_CNAE)
  names(join_CNAE)[2:3]<-c(paste0("NUM.EMPRESES.", Nom), paste0("CENS.LABORAL.", N
om))
  SeccionsCensals_prova = merge(x = Seccions, y = join_CNAE, by = "CUSEC", all = T
RUE)
  return(SeccionsCensals_prova)
}


#Classification for the ranges that contain 3 digits
Class_CNAE_v3 = function(Min, Max, Nom, Seccions) {
  join_CNAE = st_join(SeccionsCensals_CENS_ambCENS, new_DF_geom_BMA[between(new_DF
_geom_BMA$CNAE.ID_3v_Num, Min, Max), ], join = st_intersects) |>
    group_by(CUSEC) |>
    summarize(NUM.EMPRESES.Nom=length(CUSEC), CENS.LABORAL.Nom = sum(CENS.LABORAL)
)
  join_CNAE = st_drop_geometry(join_CNAE)
  names(join_CNAE)[2:3]<-c(paste0("NUM.EMPRESES.", Nom), paste0("CENS.LABORAL.", N
om))
  SeccionsCensals_prova = merge(x = Seccions, y = join_CNAE, by = "CUSEC", all = T
RUE)
  return(SeccionsCensals_prova)
}
```

APPENDIX XI

Section 4.2.

```
library(sf)
SeccionsCensals_Classificada = "CENSUS_TRACT_PATH_IN_PC”
SeccionsCensals_Classificada = st_read(SeccionsCensals_Classificada)
selected_variables = SeccionsCensals_Classificada[4:22] #get the columns of intere
st
selected_variables = st_drop_geometry(selected_variables) #drop geometry
library(skimr)
skim(selected_variables)
```

APPENDIX XII

The code below show how to create maps with tmap, representing the quantiles by deciles. To create maps by quartiles change the n from 10 to 4.

Section 4.2.1.

```
#create bounding box for better fitting everything
bbox_new = st_bbox(SeccionsCensals_Classificada)

xrange = bbox_new$xmax - bbox_new$xmin # range of x values
yrange = bbox_new$ymax - bbox_new$ymin # range of y values

bbox_new[1] = bbox_new[1] - (0.1 * xrange) # xmin - left
bbox_new[3] = bbox_new[3] + (0.25 * xrange) # xmax - right
# bbox_new[2] = bbox_new[2] - (0.25 * yrange) # ymin - bottom
bbox_new[4] = bbox_new[4] + (0.1 * yrange) # ymax - top

bbox_new <- bbox_new |>  # take the bounding box ...
  st_as_sfc() # ... and make it a sf polygon

#Create some background layers
  #BMA borders
MunBMA = st_read("SHAPEFILE_PATH")
class(MunBMA)

MunBMA_agg = MunBMA |> #create the aggregated variable to have all municipalities
in BMA (not by Census Tracts)
  group_by(NMUN) |> #agrupar codi per seccio censal
  summarize(mean(Salari))

  #Create Barcelona Border
Barcelona = MunBMA_agg[MunBMA_agg$NMUN == "Barcelona",]


#create the total absolute number map
EmployeesDeciles = tm_shape(SeccionsCensals_Classificada, bbox = bbox_new) +
  tm_fill(col = "CENS.LABORAL.TOTAL",
          palette = "Reds",
          style = "quantile",
          n = 10,
          title="") +
  tm_layout(title = "Employees by Deciles",
            title.size = 3,
            legend.outside = TRUE) +
  tm_borders(alpha = 0.05) +
  tm_compass(type = "8star", position = c("right", "top"), size = 1.8) +
  tm_scale_bar(breaks = c(0, 5, 10), text.size = 0.9) +
  tm_shape(MunBMA_agg)+
  tm_borders(lwd = 1) +
  tm_shape(Barcelona) +
  tm_borders(lwd = 3)

EmployeesDeciles
```

```
DensityEmployeesDeciles = tm_shape(SeccionsCensals_Classificada, bbox = bb
ox_new) +
  tm_fill(col = "Den_CENS",
          palette = "Reds",
          style = "quantile",
          n = 10,
          title="") +
  tm_layout(title = paste0("Density Employees","\n", "(Sq.Km) by Deciles")
,
            title.size = 3,
            legend.outside = TRUE) +
  tm_borders(alpha = 0.05) +
  tm_compass(type = "8star", position = c("right", "top"), size = 1.8) +
  tm_scale_bar(breaks = c(0, 5, 10), text.size = 0.9) +
  tm_shape(MunBMA_agg)+
  tm_borders(lwd = 1) +
  tm_shape(Barcelona) +
  tm_borders(lwd = 3)


DensityEmployeesDeciles
```

APPENDIX XIII

Final LISA function.

Section 4.2.1.

```
#THE FOLLOWING PACKAGES ARE NECESSARY FOR USING THIS FUCNTION:
library(sf)
library(spdep)
library(dplyr)
library(tmap)



LISA_CNAE_tmap = function(CensusTractMap, employees_variable, CNAE_name, Significa
nce_Level){
  CensusTracts = CensusTractMap
  CensusTracts |>
    spdep::poly2nb(c('CUSEC'), queen = FALSE) |>
    spdep::nb2listw(zero.policy = TRUE) -> weights_CensBMA
  message("Global Moran's I test result for -", CNAE_name,  "- in BMA below:")
  print(spdep::moran.test(CensusTracts[[employees_variable]], weights_CensBMA, zer
o.policy = TRUE))
  plot.title1 = paste0("Ocupacio for ", CNAE_name)
  plot.title2 = paste0("Estimated neighbor plot for ", CNAE_name)
  png(filename = paste0("Destination_PATH_", CNAE_name, ".png"),
      w = 10,
      h = 6,
      units = "in",
      res = 300)
  spdep::moran.plot(log(CensusTracts[[employees_variable]]), weights_CensBMA,zero.
policy=TRUE,xlab = plot.title1, ylab = plot.title2, pch = 20)
  dev.off()
  message("Moran plot for -", CNAE_name, "- CREATED and SAVED")
  lisaRslt = spdep::localmoran(CensusTracts[[employees_variable]], weights_CensBMA
, zero.policy=TRUE, na.action = na.omit)
  lisaRslt2 = data.frame(lisaRslt)
  lisaRslt2$lag<-spdep::lag.listw(weights_CensBMA, var = CensusTracts[[employees_v
ariable]], na.action = na.omit) #get all the moran.plot values.   Compare the valu
e of one variable with the ones from its surroudning
  names(lisaRslt2)[5] <- 'pvalue'
  x_mi = cbind(CensusTracts, lisaRslt2)
  x_mi <- x_mi %>%
    mutate(raw_std = as.numeric(scale(CensusTracts[[employees_variable]])), # scal
e means standardize to mean 0, SD 1
           lag_std = as.numeric(scale(lag)),
           lisa_category = factor(case_when(  # All of this is assigning labels ba
sed on values
             raw_std >= 0 & lag_std >= 0 & pvalue < Significance_Level ~ 'High-Hig
h',
             raw_std <= 0 & lag_std <= 0 & pvalue < Significance_Level ~ 'Low-Low'
,
             raw_std <= 0 & lag_std >= 0 & pvalue < Significance_Level ~ 'Low-High
',
             raw_std >= 0 & lag_std <= 0 & pvalue < Significance_Level ~ 'High-Low
',
             pvalue >= Significance_Level ~ 'Non-significant'),
             levels = c('High-High','Low-Low','Low-High','High-Low','Non-significa
nt')))
  x_mi <- subset(x_mi, !is.na(lisa_category))     #To delete the NA values so they
```

```
do not appear on the map
  # Plot LISA categories on a map using tmap
  #Define a Bounding Box
  bbox_LISA <- st_bbox(x_mi)# current bounding box
  xrange <- bbox_LISA$xmax - bbox_LISA$xmin # range of x values
  yrange <- bbox_LISA$ymax - bbox_LISA$ymin # range of y values
  bbox_LISA[1] <- bbox_LISA[1] - (0.15 * xrange) # xmin - left
  bbox_LISA[3] <- bbox_LISA[3] + (0.25 * xrange) # xmax - right
  bbox_LISA[2] <- bbox_LISA[2] - (0.15 * yrange) # ymin - bottom
  bbox_LISA[4] <- bbox_LISA[4] + (0.15 * yrange) # ymax - top
  bbox_LISA <- bbox_LISA %>%  # take the bounding box ...
    st_as_sfc() # ... and make it a sf polygon

  #Create the map based on the bbox area
  tmap_LISA = tm_shape(x_mi, bbox = bbox_LISA) +
    tm_fill(col = "lisa_category",
            breaks = c("Low-Low", "Low-High", "Not Significant", "High-Low", "High
-High"),
            palette = c("red", "blue", "#6FC8F5", "#F77D7D", "white"),
            title = "LISA Category") +
    tm_layout(title=CNAE_name) +
    tm_borders(alpha = 0.05) +
    tm_legend(outside = TRUE) +
    tm_compass(type = "8star", position = c("left", "top"), size = 1.8) +
    tm_scale_bar(breaks = c(0, 5, 10), text.size = 0.9) +
    tm_shape(MunBMA_agg)+
    tm_borders(lwd = 1) +
    tm_shape(Barcelona) +
    tm_borders(lwd = 2)
  tmap_save(tm = tmap_LISA,
            filename = paste0("Destination_PATH_", CNAE_name, ".png"),
            height = 10,
            width = 10,
            units = "in")
  message("LISA map for -", CNAE_name, "- CREATED and SAVED")
  message("              ")
  message("Go on with the research!")
  message("GOOD LUCK :)")
  message("...................................................................
.................")
  message("                    ")
  return(tmap_LISA)
}
```

The lapply iterator for automating the LISA function is below:

```
lapply(classificacionsCNAE, function(actual_CNAE){
  nomCLASS<-sub(".*\\..*\\.(.*)", "\\1", actual_CNAE) #Get the name of the adapted
CNAE classification from the variable. Extract the word after the second point
  LISA_CNAE_tmap(SeccionsCensals_Classificada, actual_CNAE, nomCLASS,0.05)
})
```