# Web System for Spatial Processing and Analysis: An Application in a Neonatal Screening Program in Southern Brazil

**Augusto Oliveira** (augustocesarfmo@gmail.com)
Universidade Federal Rural de Pernambuco
**Cristiane Kopacek**
Universidade Federal de Ciências Médicas de Porto Alegre
**Simone M. Castro**
Universidade Federal do Rio Grande do Sul
**Guilherme Vilar**
Universidade Federal Rural de Pernambuco
**Moacyr C. Filho**
Universidade Federal Rural de Pernambuco

**Research Article**

**Keywords:** Geographic information systems, Georeferencing, Neonatology

**Posted Date:** November 15th, 2021

**DOI:** https://doi.org/10.21203/rs.3.rs-1048547/v1

## Abstract

# Background

Spatial data analysis refers to the process of finding patterns, detecting anomalies, or testing hypotheses and theories by observing phenomena associated with a specific geographic area or location. The literature in the area presents different studies that seek to understand its phenomena through spatial analysis techniques and methods. However, these studies have several problems, such as the frequent use of only one type of analysis, area or punctual. Furthermore, the studies do not formally describe the process of treatment and organization applied to the data to replicate the spatial analyzes in other research areas. Thus, this work proposes a web system for generating, organizing, and processing data compatible with geographic information systems to construct spatial analysis of area and points.

# Methods

The proposed method was developed with the JavaScript programming language and structured in four sequential steps: data acquisition, processing and organization, data validation, and spatial analysis. Data from three diseases (cystic fibrosis, congenital adrenal hyperplasia and hemoglobinopathies) from a neonatal screening program in southern Brazil were used to validate the proposed method and construct the spatial analyses. The choropleth mapping and kernel density estimation methods were used to build the analyses.

# Results

The results obtained made it possible to georeference the data, validate it to its area of study, associate it with its micro and mesoregions, and cross it with public databases. In addition, the results enabled the construction of scientific maps of area and points to visualize the primary evidence from the spatial distribution of disease cases.

# Conclusions

The developed method showed high replication potential for other study contexts. Also, it proved to be relevant in the context of spatial analysis, enabling speed in processing, data organization and, consequently, in the construction of significant results that can be used in public policies that directly impact people's quality of life and health challenges.

## Background

An application of spatial analysis is seen in the literature in different contexts, from bioenergy, agriculture to land transport management [1–3]. Spatial data analysis is associated with techniques responsible for testing hypotheses and theories or finding patterns and anomalies based on spatial data [4, 5]. Spatial data is data associated with a specific geographic area, or location [6], such as cities, hospitals, and roads.

In the field of spatial analysis, there are three types of investigations: (i) **study of point events or patterns** - which identify points located over space, such as the location of crimes and occurrences of diseases; (ii) **study of continuous surfaces** - which can be regularly or irregularly distributed over space; and (iii) study of areas with counts and aggregated rates - which do not have the exact location of the events, but rather a value per area, such as numbers or values associated with municipalities, neighborhoods or census sectors [4].

The most common way to process and analyze spatial data is using a Geographic Information System (GIS). These are programs or a combination of programs that work together to help users understand their spatial data. GIS includes managing, manipulating, customizing, analyzing, and creating visual presentations [4, 7].

The success of GIS projects critically depends on accessing accurate, timely, and compatible spatial data [8]. The organization and detailing of these data enables the construction of scientific maps that facilitate communication between the visualization of the geographic distribution of events and the challenges in identifying priority areas for intervention or strategic decision-making [9].

In the literature, it is possible to identify different studies that aim to understand real-world phenomena through spatial analyses in health sciences [8, 10, 11]. However, these studies have several problems, such as the frequent use of only one type of analysis, area or punctual. Furthermore, they do not formally describe the process of treatment and organization applied to the data to replicate the spatial analysis in other research areas. To exemplify the problems mentioned, we can reference the works of [8, 10, 11].

In Neves et al. [10], the authors observed the incidence and spatial distribution of chronic myeloid leukaemia, a malignant hematologic disease characterized by a clonal disorder, in Pernambuco, Brazil. They used TerraView 3.3.1 software to compile the geographic data. The study proved to be the basis for monitoring and epidemiological investigation of the disease. However, the study proved limited only to observing the spatial distribution of geographic data by area. If data punctual analysis was included, the authors would determine whether the observed events exhibited any systematic pattern, which could help obtain other types of discussions.

In Mas et al. [11], the authors described the spatial patterns of referrals to the mental health program in western Sydney, Australia. Referral rates were analyzed using spatial autocorrelation and spatial regression indices in ArcGIS 10© and GeoDa 1.8.16.4.3 software. The results and the techniques used helped monitor inequality of care and planning health policy in the city. Although the study uses different spatial statistical techniques, the observation of specific geographic data was not contemplated. Also, observing the punctual distribution of events in the study could complement the analyses, pointing out the construction of possible service centers or corroborating existing ones.

In Murad [8], different applications of GIS for planning healthcare services in Jeddah, Saudi Arabia, were presented. The identified problems were modeled in ArcGIS software using the choropleth mapping, kernel density, and Euclidean distance (straight line) functions. The approaches used proved to be essential for health managers in decision-making and strategic planning in the city. However, the study does not formally present the treatment and organization carried out on the data to achieve the results obtained. The presentation of the data organization process, through flowcharts or algorithms, could facilitate the replication of the study to other research areas.

Given those problems, this work proposes the first method that formally presents generating, organizing, and processing geographic data compatible with GIS. The proposed method presents all this detailing through a graphical interface available on the web available for other studies and similar analyses. Also, the proposed method is the first method that presents two automated mechanisms for constructing point and area scientific maps that generate choropleth maps and Kernel Density Estimation (KDE) maps, also known as thematic and heat maps.

To validate the proposed method, we used data from the diagnosis of three diseases, cystic fibrosis, congenital adrenal hyperplasia, and hemoglobinopathies, from neonatal screening in the state of Rio Grande do Sul, Brazil. Therefore, in addition to contributing with a unique method in the literature, this work also aims to contribute with spatial point and area analyzes for the health area of the state of Rio Grande do Sul, Brazil. The proposed spatial analyzes can support strategic decisions that directly impact the quality of life of people who have the diseases studied in this state.

The discussion of this work is structured as follows: “Methods” section presents the study scenario and the proposed method.“Results” section presents the results obtained in the form of point and area maps using our method.“Discussion” section discusses the benefits and limitations of the proposed method. Finally, the “Conclusions” section concludes this study, presenting the main indications for using the method and the contributions obtained.

## Methods

This section is organized as follows: “Study region” section presents the study concentration region and its main characteristics. “Data collection” section presents the data, diseases, and variables studied in this work. Finally, “Flowchart of the proposed method” section presents the flowchart of the proposed method and its operation in detail.

# Study region

This study was carried out in Rio Grande do Sul (RS), located in the southern region of Brazil. The Rio Grande do Sul is the fourth largest state in Brazil, with over 11 million inhabitants and central Italian and German descent. It is currently distributed in 497 municipalities and seven mesoregions [15], as shown in Figure 4.

In 2001, the Reference Service in Neonatal Screening (SRTN) was established in the state capital, Porto Alegre, at the Presidente Vargas Materno-Infantil Hospital (HMIPV). SRTN is the service responsible for analyzing and diagnosing newborns (NB) with the six diseases screened for in the public heel prick test, which are: cystic fibrosis, congenital hypothyroidism, biotinidase deficiency, hemoglobinopathies, phenylketonuria, and congenital adrenal hyperplasia [16].

The neonatal screening strategy requires that the diagnosis and initiation of treatment occur before the onset of symptoms, thus reducing child morbidity and mortality and the specific sequelae of each disease [17]. The SRTN receives all data from the heel prick test carried out at 1.307 collection points in the state in a volume that corresponds to approximately 75% of NB in Rio Grande do Sul, an average of 105.000 births per year.

## Data collection

This study included data from patients from three diseases, cystic fibrosis, congenital adrenal hyperplasia, and hemoglobinopathies, collected from 2004 to 2020. In total, 405 records were obtained from the three specified diseases. The variables associated with the records and worked on in this study were: patients' address, pathology, and race. Data is in the form of an excel spreadsheet (.xlsx) in the SRTN.

The Ethics Committee approved this study of the Research and Graduate Studies Group of the Hospital Materno-Infantil Presidente Vargas in Porto Alegre, Rio Grande do Sul, under number 4.397.969.

## Flowchart of the proposed method

The procedure of the method proposed is presented in the flowchart in Figure 5. This flowchart is divided into four sequential steps: **step 1** - data acquisition; **step 2** - data processing and organization; **step 3** - data validation; and finally, **step 4** - spatial analysis. All of these steps are discussed in their respective subsections below.

## Data acquisition

Initially, the algorithm receives a data file in spreadsheet format (.xlsx) with two possible variables, the address and the record identifier. The address is used in the georeferencing process. Georeferencing is the process of converting addresses such as “1600 Parkway Amphitheater, Mountain View, CA” to geographic coordinates such as “latitude 37.423021 and longitude -122.08373”. The record identifier was a single variable used to differentiate one record from the others. In addition, a record identifies the records with their possible relevant attributes, such as race and type of disease, in subsequent processes.

## Data processing and organization

The algorithm iterates over the records, going one by one, calling the Google Maps Geocoding Application Programming Interface (API). The API assigns the address variable (text) as a request parameter. In the computing field, this term request is the intermediary process between the client's communication (web page or app) with the server, that is, between our proposed method and the Google Maps Geocoding API.

If there is a return from the API, that is, success in the georeferencing, the algorithm treats the data set received in JSON (JavaScript Object Notation) format. The treatment was done by organizing and separating API response variables: formatted address, latitude, longitude, street name, house number, neighborhood, city, state, and zip code.

In each algorithm iteration, the processed and organized data are saved in a new spreadsheet (.xlsx). The new spreadsheet generated with the georeferenced data had two sheets, one for failure and one for success. The fault sheet presented addresses where the API did not return georeferencing results. The data that could not be georeferenced undergo manual intervention to correct possible spelling errors and are then submitted again to the automatic georeferencing process.

The choice of the Geocoding Google Maps API for this process was based on the analysis and experimentation of six commercial geocoding services, namely: LocationIQ [18], OpenCage [19], ArcGIS [20], HERE [21], Google [22] and Bing [23].

The Geocoding Google Maps API showed greater precision in georeferencing, correspondence between the local address and its geographic coordinates, and a greater variety of supported response formats. In addition, it has 99% worldwide coverage, reliable and comprehensive data from over 200 countries and territories [24].

The algorithm was developed using the JavaScript web programming language that made it possible to calculate, manipulate and validate data [25]. In addition, we used the React library, built-in JavaScript, which allows creating dynamic user interfaces on web pages [26]. A demo version of this process can be seen at https://spatialworkspace.ddns.net/geocoding.

This method/algorithm was called ASSYRIA, from the acronym of "d**A**ta proce**S**sing **SY**stem fo**R** spat**I**al **A**nalysis". The acronym was generated on the Acronymify website [27].

## Data validation

Bonner et al. [28] showed that geocoding, especially in populated cities, is often very accurate, regardless of the geocoding source used. However, Dominkovics et al. [29] showed that errors in location accuracy for georeferenced addresses are unavoidable and depend on many factors, such as data quality and the inherent accuracy of the commercial georeferencing sources used.

Therefore, to validate the accuracy of our data resulting from the geocoding process, we used a post-manual correction process based on Goldberg et al. [30], who proposed a manual correction of incorrectly geocoded data through an interactive web-based approach. This process was carried out from the dynamic visualization of geocoded data on the Google map. The data visualized outside the area delimited to the study must undergo manual correction. Figure 6 exemplifies a scenario where three geographic coordinates are outside the study area and must go through the manual correction process.

A demo version of this process can be seen at https://spatialworkspace.ddns.net/map. It was possible to import a spreadsheet of georeferenced data and dynamically visualize the data on the Google map. This visualization made it possible to observe data with possible positioning or georeferencing errors. Data identified outside the study area or with possible georeferencing errors underwent manual intervention. Corrections were performed using a demo web version of manual correction. The demo version can be viewed at https://spatialworkspace.ddns.net/address. This process ensures that the data going to the analysis process is accurate and within the study scenario, avoiding possible compilation or interpretation errors in the data spatial analysis process.

# Spatial analysis

This section presents the construction of two types of spatial analysis: (i) **area spatial analysis** and (ii) **point spatial analysis**. In the subsection of “Area spatial analysis”, we describe the process to create two choropleth maps. The first one, the map of the total number of cases of the three diseases. The second one, the maps of prevalence by the municipality, microregion, and mesoregion. In the subsection of “Punctual spatial analysis”, the KDE method for building density maps was presented.

# Area spatial analysis

According to MacEachren and Alan [31], the objective of choropleth mapping is to present numerical data on areas using distinct colors. Lighter colors represent lower numerical values, and darker colors represent higher values of the numerical variation of the phenomenon under study.

Typically, choropleth mapping explores the spatial pattern of attribute distribution between regions visually, for example, demographic attributes such as ”population density and population by sex”, and socioeconomic attributes such as ”per capita income and Human Development Index (HDI)” [8, 14]. In our method, choropleth mapping presents the spatial distribution of the total number of cases of the three diseases and their prevalence by the municipality, microregion, and mesoregion in the state of Rio Grande do Sul, Brazil.

The Associate Regions (AR) algorithm associates the attribute city (text), obtained in “Data processing and organization” section, to their respective regions of the political-administrative division of the state of Rio Grande do Sul. The association was made by automatically searching for the regions corresponding to the city variable in the Localities API of the Brazilian Institute of Geography and Statistics (IBGE) [32]. Figure 7 summarizes the construction process of this analysis.

The association performed generates five more fields associated with the city attribute: city or municipal identifier, microregion name, microregion identifier, mesoregion name, and mesoregion identifier. The data/fields generated are saved in a new spreadsheet (.xlsx) and made available for download. A demo version of this process can be viewed at https://spatialworkspace.ddns.net/region. If any records fail the automatic region association process, they will be available in the failure sheet of the spreadsheet file. Potentially failing data is, in turn, manually searched at https://spatialworkspace.ddns.net/search and then incorporated into the study's success dataset.

Still, for the construction of choropleth maps, we used three indicators: a total number of cases by (i) municipality, (ii) microregion, and (iii) mesoregion. For counting the number of cases, we use a dynamic counting algorithm for this process. A demo version of this process can also be seen at https://spatialworkspace.ddns.net/cases.

To obtain population data, for choropleth maps, we used the IBGE Automatic Recovery System (SIDRA) with 2010 demographic census [33]. The number of the population residing in rural and urban areas using SIDRA. Prevalence, presented in Equation 1, demonstrates the proportion of the frequency of a disease over the population of a region in a given time [34]. The population constants of 10.000, 100.000, and 200.000 were used, respectively, for the municipality, microregion, and mesoregion. The choice for these values was experimented with using the population residing in the state of Rio Grande do Sul, Brazil, and on the characteristics of disease incidence.

$$prevalence = \frac{n.^{\underline{o}}\ of\ cases}{population} * constant \tag{1}$$

Free and open-source GIS QGIS, version 3.10 [35], was used to build the maps. Both maps, number of cases and prevalence, were classified into three categories (low, moderate, and high) using Jenk's Natural Breaks classification [36]. Jenk's Natural Breaks method is a data classification method designed to reduce variance within classes and maximize the variance between classes [37, 38]. We obtained the Shape File (SHP) for mesoregion, microregion, and municipality from IBGE [39].

## Point spatial analysis

Although choropleth mapping is a traditional way of looking at data aggregated by region or area, such as count rates and proportions, this type of mapping is still limited to studying other health challenges [40]. For, the risks and events by area can change in the face of different spatial limits, for example, within the limits of a large metropolis.

So, to complement the analysis of this study, we used the non-parametric Kernel Density Estimation (KDE) method to explore the spatial density of our point data. Kernel densities have often been used to analyse point events and explore hotspots in various domains, including criminology [41], spatial epidemiology [8, 42–44], or ecology [45]. Furthermore, it is a method that is easy to use and interpret [4]. Formally this method is given by Equation 2.

$$\hat{\lambda}_{\tau}(u) = \frac{1}{\tau^2}\sum_{i=1}^{n} k\left(\frac{d(u_i, u)}{\tau}\right), d(u_i, u) \le \tau \tag{2}$$

Where $u_1, \ldots, u_n$ are locations of punctual events observed in a given region. The density estimator is computed from the $m$ events $u_i, \ldots, u_{i+m-1}$ contained in a radius of size $\tau$ around $u$ and the distance $d$ between the position and the ith sample. Thus, a circular kernel forms around each generator point (e.g., murder victim locations) with a predefined bandwidth as its radius [43]. Thus, the density of the map surface is produced from the count of all points within the radius of influence, weighting them by the distance of each one to the location of each generated point.

The basic parameters of this method are: a) radius of influence ($\tau \geq 0$) and b) kernel estimation function. The radius of influence defines the area centered on the estimation point $u$, which indicates how many events $u_i$ contribute to estimating the density function $\lambda$. Usually, the bandwidth, or radius size, is decided in an exploratory way. KDE is produced using different bandwidth sizes (radius) for empirical identification of a suitable value for the context under study [43]. Small radius values result in discontinuous surfaces, and values that are too high result in smooth surfaces. Kernel estimation calculation functions are uniform, triangle, Epanechnikov, quartic (biweight), tricube, triweight, Gaussian, quadratic, and co- sine. However, usually, third or fourth-order functions or Gaussian kernels are used for kernel estimation [4].

Free and open-source GIS QGIS, version 3.10 [35], was also used to build this analysis. The kernel density function used visualizes the hotspots or areas with the highest density of cases of the three diseases studied in this work. The parameters defined for this analysis were: bandwidth or radius size in 10 km and the fourth-order function for kernel estimation. In addition, the recording of punctual events used in this analysis resulted from the geocoding process carried out in “Data processing and organization” section. The Shape File of the political-administrative division at the municipal level, the year 2020, was used as a basis for viewing the hotspots generated. The Shape File was obtained from the IBGE municipal mesh collection [39].

## Results

This section is structured in the following subsections: “Data validation” and “Spatial analysis”. The first subsection presents the success and failure result in georeferencing data and association by region. In the second subsection, we present the main results and observations perceived in the spatial distribution of area and point maps.

### Data validation

#### Georeferencing

Using the Geocoding API from Google Maps, the georeferencing of 401 records from the 405 addresses studied in this work was successful. The automatic geocoding processing failed in four records. Two records failed because of the automatic geocoding process. After the correction of the spelling errors, we submitted them again into the manual georeferencing process. The other two records were georeferenced outside the study area. Therefore, the automatic process had 99.01% success in georeferencing and less than 1% failure in this dataset.

# Region association

After the process described in the section "Data processing and organization", the division regions of Rio Grande do Sul were associated with 405 records. Of these data, 398 were successful in the automatic association process, and only seven records failed in the automatic process. Records that failed the association process were corrected and manually associated. The failures occurred due to accentuation errors and special characters in the records obtained from the IBGE Localities API. The automatic association process has 98.27% successful, and less than 2% failed on this dataset.

# Spatial analysis

# Area spatial analysis

Performing the region association process, described in the "Area spatial analysis" section, choropleth maps were generated by mesoregions (Figure 1a), microregions (Figure 1b), and municipalities (Figure 1c) of the total number of cases of the three diseases.

In Figure 1a, it was possible to observe a greater concentration of cases in the Metropolitan mesoregion. In Figure 1b, a gradual decrease in the number of cases was observed, at the micro-region level, in the Northwest and Southeast of the state. In Figure 1c, it was possible to visualize numerous municipalities without the registration of the number of cases in the northern region of the Southwest, Midwest, and Northwest mesoregions, and an intense concentration of the number of cases in the state capital, Porto Alegre.

Also, choropleth maps were generated by mesoregions (Figure 2a), microregions (Figure 2b), and municipalities (Figure 2c) of disease prevalence.

In Figure 2a, it was possible to observe a greater concentration of prevalence in the Metropolitan and Southeast mesoregions. In Figure 2b, it was also possible to visualize a high concentration, at the micro-region level, of prevalence per 100.000 thousand inhabitants, in the Southwest, Southeast, Metropolitan, and Northeast regions of the state. In Figure 2c, we observed that most municipalities in the Southwest and Southeast regions have a low level of prevalence per 10.000 inhabitants, and few municipalities have a high concentration of prevalence.

# Point spatial analysis

In the punctual analysis, we observed a more specific visualization of the phenomenon, thus,complementing the visualization of the spatial distribution shownin Figure 1. In Figure 3, we observed the exact location of the disease cases over the total number of inhabitants per municipality.

In Figure 3b, using the KDE method, in a radius of 10 km, it was possible to visualize the hot areas of the occurrence of cases within the territorial extension of each municipality in the state. The map shows a high concentration of cases, that is, above 5, in at least 14 municipalities of the 497 municipalities of the state, they are: Pelotas, Rio Grande, São Leopoldo, Canoas, Porto Alegre, Alvorada, Viamão, Caxias do Sul, Bento Gonçalves, Lajeado, Santa Cruz do Sul, Passo Fundo, Santa Maria and Uruguaiana.

Furthermore, it was possible to observe in Figure 3b that at least 4 municipalities in the Southwest and Center-East regions of the state have a large concentration of cases under a relatively small population, that is, around 80 thousand inhabitants. The municipalities were: Bagé, Uruguaiana, Santa Cruz do Sul and Lajeado. Also, it was seen that at least 8 municipalities had a large concentration of cases under a relatively large population, that is, approximately 750 thousand inhabitants. The municipalities were: Rio Grande, Pelotas, Santa Maria, Passo Fundo, Caxias do Sul, Porto Alegre, Canoas and São Leopoldo.

## Discussion

The discussions in this study are organized in the following subsections: “Data validation” and “Spatial analysis”. In the first subsection, we present the advantages and limitations of georeferencing and region association methods. The second subsection presents the benefits and limitations of choropleth mapping and KDE methods and indications for using other methods.

# Data validation

# Georeferencing

The construction of the method and algorithm (ASSYRIA) for georeferencing, organization and processing of address data, made it possible to manipulate the different Google Maps Geocoding API returns, such as: street, city and zip code. In other georeferencing services, such as ArcGIS and QGIS, it was impossible to dynamically manipulate the data return, using only the geographic coordinates information, latitude and longitude.

In addition, the construction of the algorithm in a programming language and current technology [12] enabled it to be free and interactive available on the web. The tool was made available at https://spatialworkspace.ddns.net/geocoding. Its availability allowed accelerating the process of organizing data, making it compatible with GIS, building maps, and decision making in other study scenarios.

During the study, we observed that for complete address records (street, number, neighborhood, city, state, zip code), the georeferencing proved accurate, corresponding to the physical address on the global map. As for incomplete address records, for example, only city and state georeferencing pointed to locations in the centrality of locations. Thus, when possible, we indicate the use of complete address records for more excellent reliability of the georeferencing results and, consequently, of the spatial analysis.

The Google Maps Geocoding API has no daily limit on the number of requests that can be made to its servers. However, there is a usage limit in free mode, which was related to the maximum number of queries/requests per second. That is a maximum of 50 requests per second or 3.000 requests per minute. If this limit is exceeded, an “OVER QUERY LIMIT” status code is returned in the response.

The method/algorithm developed for the automatic georeferencing and data organization carried out requests simultaneously and iteratively. In some tests, the usage limit exceeded the maximum number of requests per minute. To get around this limitation, we added a delay of 200 ms to each algorithm iteration. The execution time of the 405 requests/requests from our dataset for the Geocoding API occurred in approximately 6 minutes and 15 seconds, an average of at least 1 minute and 30 seconds per 100 records, based on this adopted approach.

The API tested for a data volume of up to 5.000 records, and we have not observed the potential for big data in the free version of the Google Maps API.

## Region association

The association of municipalities to their regions depended on the process described in the “Data processing and organization” section. Since the ASSYRIA method of data processing allowed obtaining the full address variables (street, number, neighborhood, city) in a standardized way and without spelling errors, for example, the following record: ”rua moneiro lobato, 420, bom conselho”, was obtained formatted and standardized, as: ”Rua Monteiro Lobato, 420, Bom Conselho”. Uppercase, low- ercase, spaces in the text, or spelling errors have all been corrected and formatted.

This process implies the compatibility with the municipal meshes of the 26 Brazilian states for the construction of scientific maps by area and possible crossing of data with open databases, such as IBGE and SIDRA. The municipal meshes of Brazilian states are standardized by IBGE and made available in Shape File format [13]. This format can be read by the vast majority of geoprocessing software and geoinformation viewers.

The association algorithm has no usage limit, unlike the georeferencing algorithm. Only one request was made to the IBGE server, recovering all data. Data were iteratively compared and associated with each other, so there is no restriction on the volume of the dataset.

However, the replication of this method was restricted to other countries. The IBGE API only provides information on the political-administrative division of Brazilian states. However, we believe that the method described in the “Area spatial analysis” section, presented in Figure 7, can be a model for the construction of other region association algorithms in other countries in the world. Thus, it also contributes to observing their social or demographic characteristics from their macro and micro areas.

# Spatial analysis

# Area spatial analysis

The combination of georeferencing and region association methods enabled the construction of maps at the regional levels of mesoregion, microregion, and municipality. These maps were organized in (i) choropleth maps of the total number of cases and (ii) maps of the prevalence of the three diseases, cystic fibrosis, congenital adrenal hyperplasia, and hemoglobinopathies.

According to Wei, Tong, and Phillips [14], choropleth mapping is an essential exploratory technique for spatial data analysis that has been widely used to explore the spatial pattern of attribute distributions between regions visually. Furthermore, it is one of the most widely used methods to visually explore the spatial distributions of demographic and socioeconomic data [8].

This analysis enabled the visualization of spatial patterns, that is, places with a higher concentration of cases and especially areas where maybe unattended by the local public health. This work was limited to the observation of indicators of the total number of cases and prevalence. However, the results showed that other demographic and social indicators could also be used from the approach used. For example, indicators of white or black population, education, Human Development Index (HDI), work and income can be used to build other maps that can bring relevant information for strategic planning and decision-making by the government or local competent body.

Another limitation of this study was that it did not individually observe the spatial distribution of each disease. Nevertheless, the purpose of constructing these analyses using georeferencing and region association methods was to highlight the potential for replication of the methods to other study contexts. In addition, point out possibilities for the construction of other choropleth maps using other indicators as well.

# Point spatial analysis

One of the essential characteristics of the georeferencing method is the availability of georeferenced data (latitude, longitude) in a spreadsheet file (.xlsx). In addition, another feature is the dynamic validation of these data in the region of the study area, using the Google map and, consequently, the possibility of compatibility with another point spatial data analysis software.

The KDE method used in this analysis made it possible to observe hotspots areas, that is, places with an intense concentration of cases. This analysis complements the spatial analysis of areas and regions described in the “Area spatial analysis” section. In this way, more information was added to the study, bringing the broader concept of visualization, cases by area, to the cases most specific, punctual location.

Other kernel radius sizes were also tested, such as 15 km and 20 km. However, in some cases, they stayed larger than the territorial extension of some municipalities, making it difficult to visualize the spatial distribution. Therefore, the radius size of 10 km in this study proved ideal to complement the concept of spatial visualization from the broadest to the most specific.

Although this work is limited to the use of the KDE method, organizing and processing data brings insights for using other methods of spatial point analysis and, for example, using inferential methods to predict the construction of a future hospital or care center in a specific area. Also, check that existing call centers are correctly distributed, given the spatial distribution of cases or identifying areas that may still need health care.

## Conclusions

In this study, we developed a method called ASSYRIA. The method allowed the construction of two types of spatial analysis, area and punctual. Data from three neonatal screening diseases in the Rio Grande do Sul, Brazil, were used to validate the algorithms proposed in the ASSYRIA method. According to the studies in the literature analyzed, this is the first work that dynamically processes, organizes, validates, and makes available geographic data through a web interface to construct scientific maps of areas and points.

The results presented in this work highlight the potential for replication of the method to other study contexts. Furthermore, the methods/algorithms used and developed proved to be relevant in spatial analysis. They enabled speed in processing, data organization, and, consequently, in constructing significant results that can be used in public policies that directly impact people's quality of life and health challenges.

The use of other spatial analysis methods, in addition to choropleth mapping and KDE, are also encouraged. Other methods can help identify cases and exposures, characterize spatial trends, correlate different spatial data sets, and test statistical hypotheses by crossing information with large open databases, such as IBGE and SIDRA. Finally, perspectives that address associations with data from the Ministry of Health (MS) or health secretariats are also encouraged.

## Abbreviations

GIS: Geographic Information System; KDE: Kernel Density Estimation; RS: Rio Grande do Sul; SRTN: Reference Service in Neonatal Screening; HMIPV: Presidente Vargas Materno-Infantil Hospital; NB: Newborns; API: Application Programming Interface; JSON: JavaScript Object Notation; HDI: Human Development Index; AR: Associate Regions; IBGE: Brazilian Institute of Geography and Statistics; SIDRA: IBGE Automatic Recovery System; SHP: Shape File; ASSYRIA: Data Processing System for Spatial

Analysis; MS: Ministry of Health.em; KDE: Kernel Density Estimation; RS: Rio Grande do Sul; SRTN: Reference Service in Neonatal Screening; HMIPV: Presidente Vargas Materno-Infantil Hospital; NB: Newborns; API: Application Programming Interface; JSON: JavaScript Object Notation; HDI: Human Development Index; AR: Associate Regions; IBGE: Brazilian Institute of Geography and Statistics; SIDRA: IBGE Automatic Recovery System; SHP: Shape File; ASSYRIA: Data Processing System for Spatial Analysis; MS: Ministry of Health.

# Declarations

### Ethics approval and consent to participate

This study was approved by the Ethics Committee of the Research and Graduate Studies Group of the Hospital Materno-Infantil Presidente Vargas in Porto Alegre, Rio Grande do Sul, under number 4.397.969.

### Consent for publication

Not applicable.

### Availability of data and materials

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

### Competing interests

The authors declare that they have no competing interests.

### Funding

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

### Authors’ contributions

AO developed the method, built the spatial analyses, and was a major contributor in writing the manuscript. AO, SMC, CK and GV contributed to the conception and design of the study. SMC and CK provided the study data and helped to interpret the results. All authors read and approved the final manuscript.

### Acknowledgements

Not applicable.

### Authors' information

[1]Departamento de Estatística e Informática, Universidade Federal Rural de Pernambuco, Recife, Pernambuco, Brazil. [2]Serviço de Referência em Triagem Neonatal, Porto Alegre, Rio Grande do Sul, Brazil. [3]Departamento de Pediatria, Faculdade de Medicina, Universidade Federal de Ciências Médicas de Porto Alegre, Porto Alegre, Rio Grande do Sul, Brazil. [4]Departamento de Análises, Faculdade de Farmácia, Universidade Federal do Rio Grande do Sul, Porto Alegre, Rio Grande do Sul, Brazil.

# References

1. Scarlat N, Fahl F, Dallemand JF, Monforti F, Motola V. A spatial analysis of biogas potential from manure in Europe. Renewable and Sustainable Energy Reviews. 2018;94:915–930.
2. Harmanny KS, Malek Zˇ. Adaptations in irrigated agriculture in the Mediterranean region: an overview and spatial analysis of implemented strategies. Regional environmental change. 2019;19(5):1401–1416.
3. Shanmukhappa T, Ho IWH, Tse CK. Spatial analysis of bus transport networks using network theory. Physica A: Statistical Mechanics and its Applications. 2018;502:295–314.
4. Monteiro AMV, Cˆamara G, Carvalho M, Druck S. Análise espacial de dados geográficos. Brasília: Embrapa. 2004;.
5. Goodchild MF. Data Analysis, Spatial. Shekhar S, Xiong H, editors. Boston, MA: Springer US; 2008. Available from: https://doi.org/10.1007/978-0-387-35973-1 236.
6. Câmara G, Casanova MA, Magalhães GC. Anatomia de sistemas de informação geográfica; 1996.
7. Lo CP, Yeung AK, et al. Concepts and techniques of geographic information systems. Pearson Prentice Hall; 2007.
8. Murad A. Planning and location of health care services in Jeddah City, Saudi Arabia: Discussion of the constructive use of geographical information systems. Geospatial health. 2018;13(2).
9. Boquett JA, Zagonel-Oliveira M, Jobim LF, Jobim M, Gonzaga L, Veronez MR, et al. Spatial analyzes of HLA data in Rio Grande do Sul, south Brazil: genetic structure and possible correlation with autoimmune diseases. International journal of health geographics. 2018;17(1):1–12.
10. Neves WBd, Brito AMd, Vasconcelos AP, Melo FCdBC, Melo RAM. Incidence and spatial distribution of Chronic Myeloid Leukemia by regions of economic development in the state of Pernambuco, Brazil. Hematology, transfusion and cell therapy. 2019;41:212–215.
11. Maas C, Salinas-Perez JA, Bagheri N, Rosenberg S, Campos W, Gillespie J, et al. A spatial analysis of referrals to a primary mental health programme in Western Sydney from 2012 to 2015. Geospatial health. 2019;.
12. Stack Overflow — Developer Survey, 2020; 2021. Accessed 13 Aug 2021. https://insights.stackoverflow.com/survey/2020#technology-programming-scripting-and-markup-languages-professional-developers.
13. Malha Municipal: O que é? — Instituto Brasileiro de Geografia e Estatística (IBGE); 2021. Accessed 13 Aug 2021. https://www.ibge.gov.br/geociencias/organizacao-do-territorio/malhas-territoriais/15774-

malhas.html?=&t=o-que-e.

14. Wei R, Tong D, Phillips JM. An integrated classification scheme for mapping estimates and errors of estimation from the American Community Survey. Computers, Environment and Urban Systems. 2017;63:95–103.
15. IBGE. Pesquisa Nacional por Amostra de Domicílios e Contagem da População. Rio de Janeiro: Instituto Brasileiro de Geografa e Estatística (IBGE); 2010.
16. Kopacek C, Castro SM, Chapper M, Amorim LB, Ludtke C, Vargas P. Evolução e funcionamento do Programa Nacional de Triagem Neonatal no Rio Grande do Sul de 2001 a 2015. Boletim Científico de Pediatria-Vol. 2015;4(3):71.
17. Botler J, et al. Avaliação de desempenho do Programa de Triagem Neonatal do Estado do Rio de Janeiro. Escola Nacional de Saúde Pública Sergio Arouca – FIOCRUZ, Departamento de Epidemiologia e Métodos Quantitativos; 2010.
18. LocationIQ: Fast Geocoding and Reverse Geocoding service from Unwired Labs; 2021. Accessed 13 Aug 2021. https://locationiq.com/docs#search-forward-geocoding.
19. OpenCage: Geocoding API Documentation; 2021. Accessed 13 Aug 2021. https://opencagedata.com/api#forward-resp.
20. ArcGIS Developer: Geocoding — Documentation; 2021. Accessed 13 Aug 2021. https://developers.arcgis.com/documentation/mapping-apis-and-services/search/geocoding/.
21. HERE: Geocoding & Search API; 2021. Accessed 13 Aug 2021. https://developer.here.com/documentation/geocoding-search-api/dev guide/topics/quick-start.html.
22. Google Developers: Geocoding Service — Maps JavaScript API; 2021. Accessed 13 Aug 2021. https://developers.google.com/maps/documentation/javascript/geocoding.
23. Bing Maps: Find a Location by Address — Microsoft Docs; 2021. Accessed 13 Aug 2021. https://docs.microsoft.com/en-us/bingmaps/rest-services/locations/find-a-location-by-address.
24. APIs de geolocalização — Google Maps Platform — Google Cloud; 2021. Accessed 22 Sept 2021. https://cloud.google.com/maps-platform/?hl=pt-PT.
25. W3Schools: What is JavaScript?; 2021. Accessed 13 Aug 2021. https://www.w3schools.com/whatis/whatis js.asp.
26. React: A JavaScript library for building user interfaces; 2021. Accessed 13 Aug 2021. https://reactjs.org/.
27. Acronymify! - Automatically generate fun acronyms for your project; 2021. Accessed on 11 Oct 2021. https://acronymify.com/.
28. Bonner MR, Han D, Nie J, Rogerson P, Vena JE, Freudenheim JL. Positional accuracy of geocoded addresses in epidemiologic research. Epidemiology. 2003;14(4):408–412.
29. Dominkovics P, Granell C, Perez-Navarro A, Casals M, Orcau A, Cayla JA. Development of spatial density maps based on geoprocessing web services: application to tuberculosis incidence in Barcelona, Spain. International journal of health geographics. 2011;10(1):1–14.

30. Goldberg DW, Wilson JP, Knoblock CA, Ritz B, Cockburn MG. An effective and efficient approach for manually improving geocoded data. International journal of health geographics. 2008;7(1):1–20.

31. MacEachren AM. Some truth with maps: A primer on symbolization and design. Assn of Amer Geographers; 1994.

32. API de Localidades: Instituto Brasileiro de Geografia e Estatística (IBGE); 2021. Accessed 13 Aug 2021. https://servicodados.ibge.gov.br/api/docs/localidades#api-Municipios-estadosUFMunicipiosGet.

33. Tabela 3175: População residente — Sistema IBGE de Recuperação Automática (SIDRA); 2021. Accessed 13 Aug 2021. https://sidra.ibge.gov.br/Tabela/3175.

34. John M. A dictionary of epidemiology. Oxford university press Oxford, UK; 2001.

35. QGIS: User Guide, version 3.10; 2021. Accessed 13 Aug 2021. https://docs.qgis.org/3.10/en/docs/user manual/.

36. Dent BD, Torguson J, Hodler T. Cartography: thematic map design. Dubuque: Wm. C. C Brown Geographic Designs. 1993;.

37. Jenks GF. The data model concept in statistical mapping. International yearbook of cartography. 1967;7:186–190.

38. McMaster R. In Memoriam: George F. Jenks (1916-1996). Cartography and Geographic Information Systems. 1997;24(1):56–59.

39. Malha Municipal: Downloads — Instituto Brasileiro de Geografia e Estatística (IBGE); 2021. Accessed 13 Aug 2021. https://www.ibge.gov.br/geociencias/organizacao-do-territorio/malhas-territoriais/15774- malhas.html?=&t=downloads.

40. Azevedo L, Pereira MJ, Ribeiro MC, Soares A. Geostatistical COVID-19 infection risk maps for Portugal. International Journal of Health Geographics. 2020;19(1):1–8.

41. Levine N. Crime mapping and the Crimestat program. Geographical analysis. 2006;38(1):41–56.

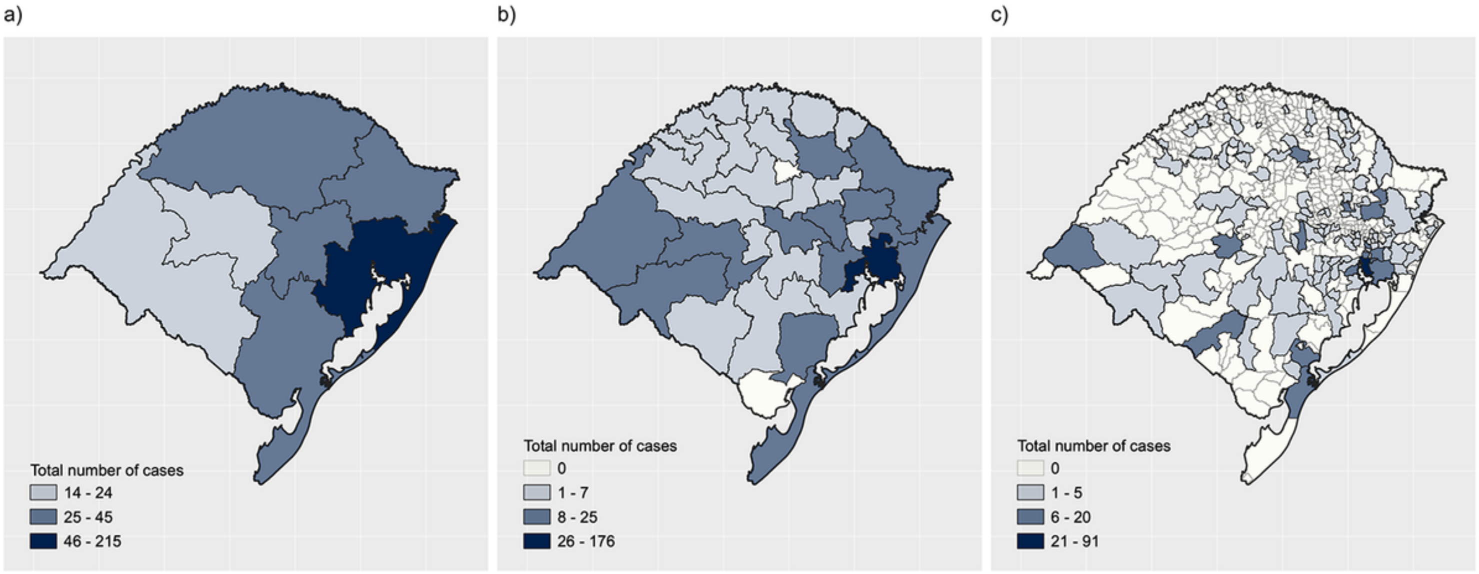
42. Cromley EK, McLafferty SL. GIS and public health. Guilford Press; 2011.

43. Shiode N, Shiode S, Rod-Thatcher E, Rana S, Vinten-Johansen P. The mortality rates and the space-time patterns of John Snow's cholera epidemic map. International journal of health geographics. 2015;14(1):1–15.

44. Chaikaew N, Tripathi NK, Souris M. Exploring spatial patterns and hotspots of diarrhea in Chiang Mai, Thailand. International Journal of Health Geographics. 2009;8(1):1–10.

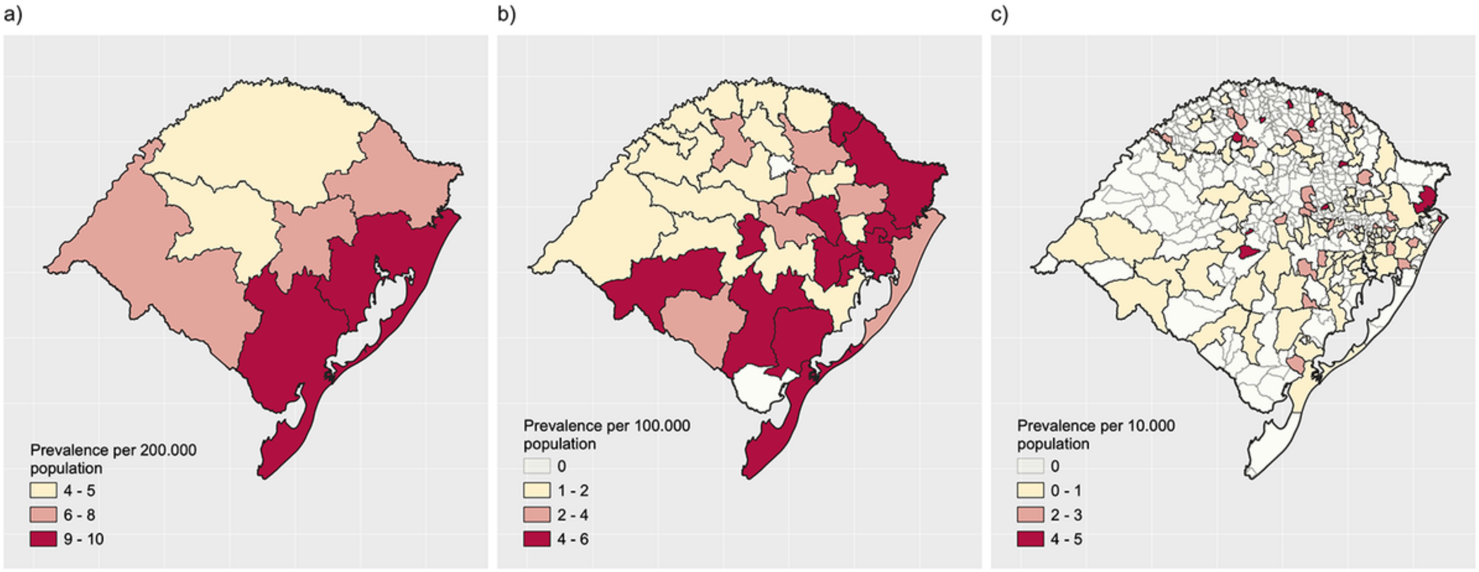
45. Worton BJ. Kernel methods for estimating the utilization distribution in home-range studies. Ecology. 1989;70(1):164–168.

# Figures



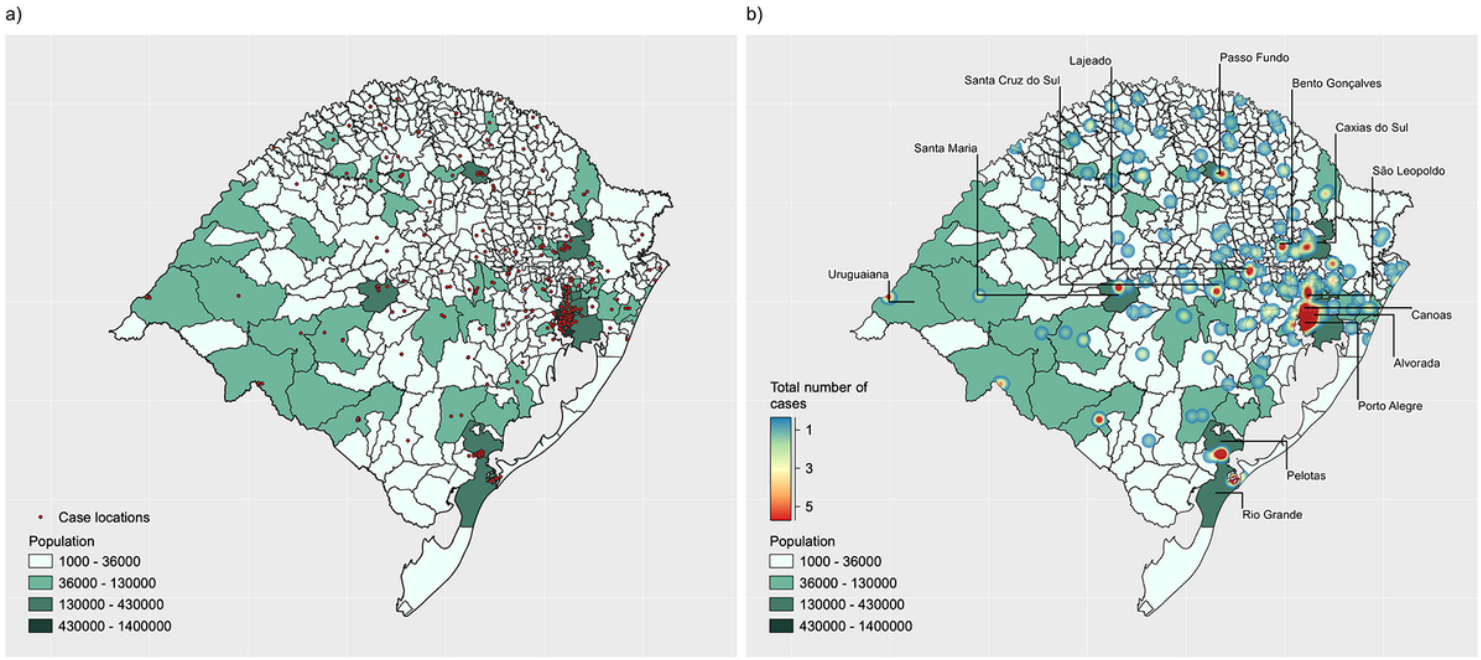
## Figure 1

Choropleth maps of the total number of cases. a mesoregions; b microregions; c municipalities


a)
b)
c)
Prevalence per 200.000 population
4 - 5
6 - 8
9 - 10
Prevalence per 100.000 population
0
1 - 2
2 - 4
4 - 6
Prevalence per 10.000 population
0
0 - 1
2 - 3
4 - 5

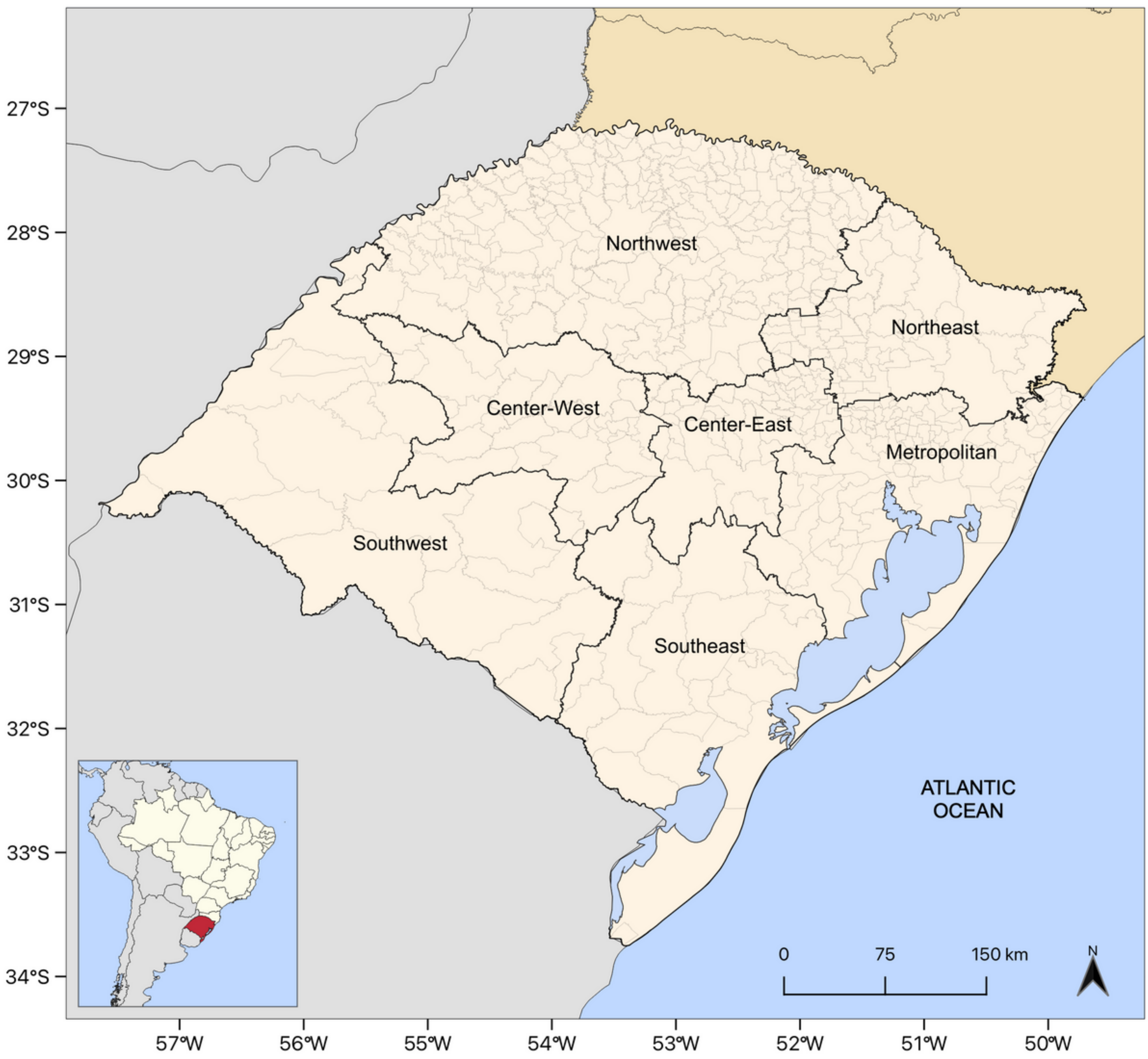
## Figure 2

Choropleth maps of disease prevalence. a mesoregions; b microregions; c municipalities



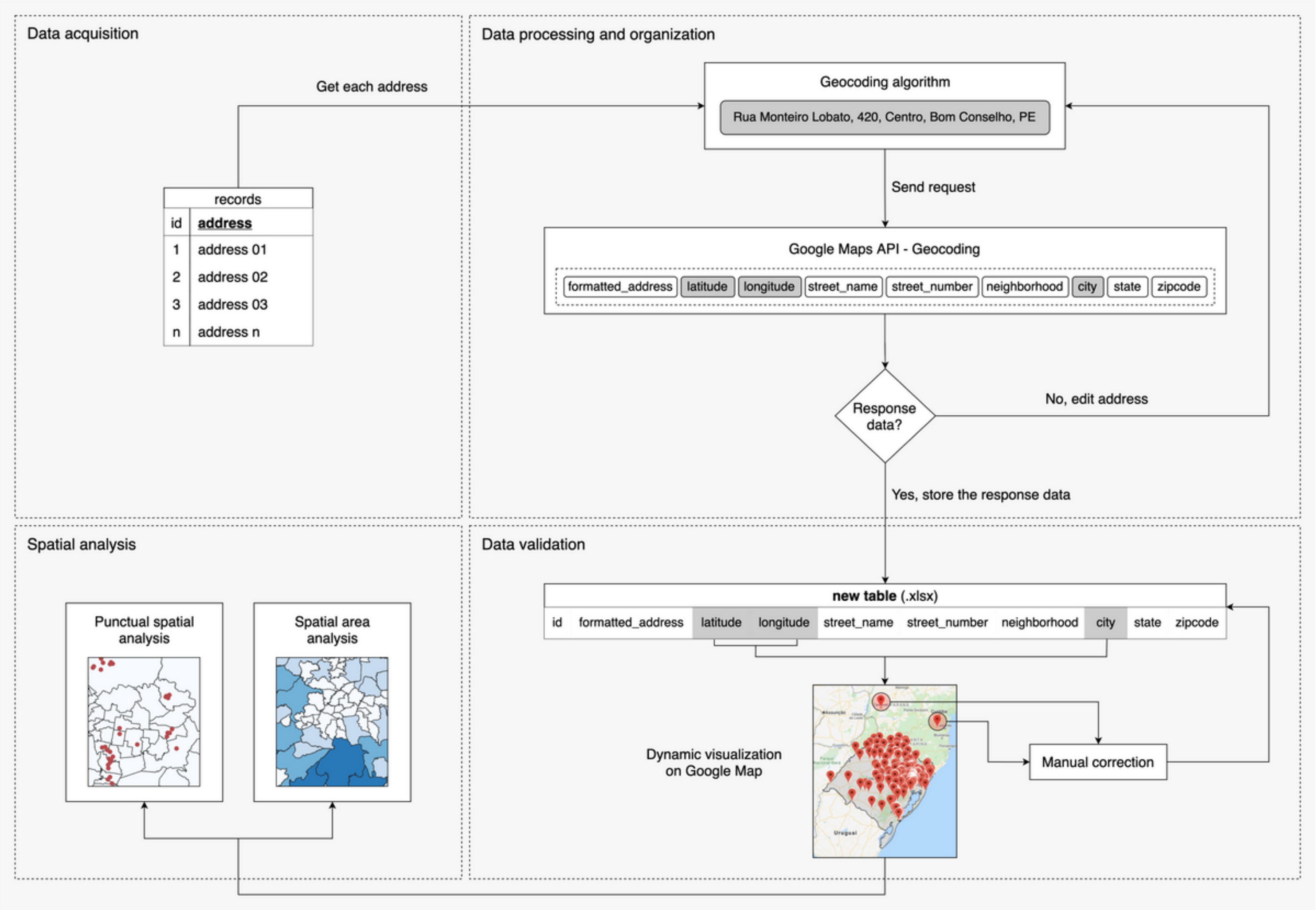
## Figure 3

Point maps of cases over the total population. a specific location of cases; b kernel density estimation of cases



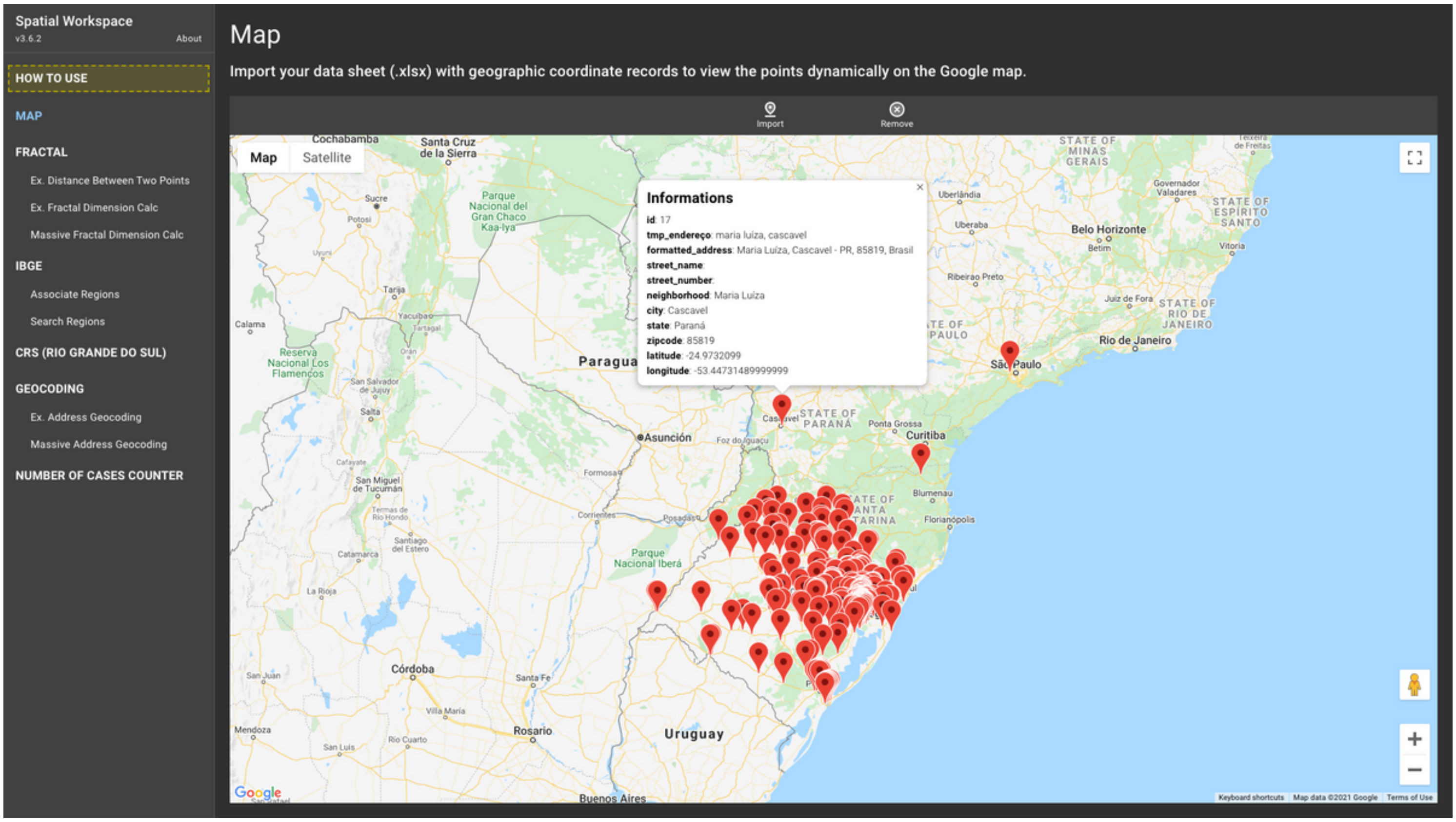
## Figure 4

Location map of the state of Rio Grande do Sul, Brazil, with its mesoregions



## Figure 5

Flowchart of the proposed method's operation



Map

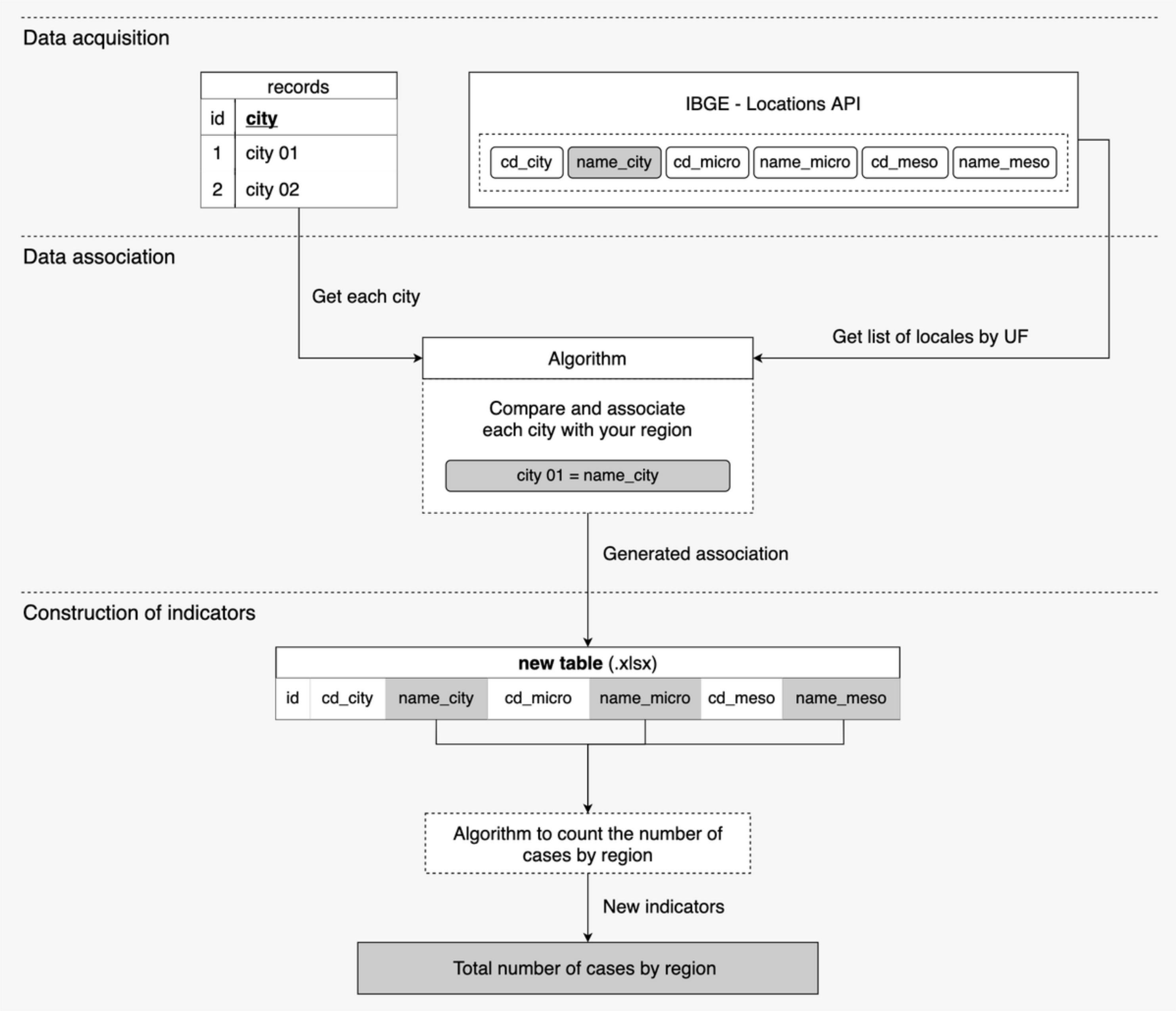
Import your data sheet (.xlsx) with geographic coordinate records to view the points dynamically on the Google map.

**Informations**

**id**: 17
**tmp_endereço**: maria luiza, cascavel
**formatted_address**: Maria Luíza, Cascavel - PR, 85819, Brasil
**street_name**:
**street_number**:
**neighborhood**: Maria Luiza
**city**: Cascavel
**state**: Paraná
**zipcode**: 85819
**latitude**: -24.9732099
**longitude**: -53.44731489999999

## Figure 6

Import and validation of geographic data



## Figure 7

Region association method flowchart