sbbd:02

Chapter 2

Geospatial Data: From Theory to Practice

Augusto Cesar Souza Araujo Domingues, Fabrício Aguiar Silva, Antonio Alfredo Ferreira Loureiro

Abstract. *The ever increasing amount of context-aware systems lead us to large volumes of data being generated and stored at every moment. In this scenario, one of the most interesting dimensions currently is geospatial data, that represent the position of an entity (e.g., vehicles) on Earth. Based on that, public and private sectors work to extract useful knowledge, aiming to understand urban mobility behavior, improving services and providing state-of-the-art solutions in areas related to mobility, disease control, and so on. With this in mind, the objective of this chapter is to present the main theoretical concepts together with practical examples related to working with geospatial data including collection, storage, transformation, and visualization.*

2.1. Introduction

Data is one of the major ingredients for the technological, political, and economical advances, as well as an ever-increasing byproduct. By using data, policies and services can be enhanced and personalized to guarantee a better experience [Hess et al. 2015]. For example, when it comes to human mobility specifically, issues such as traffic flow prediction, contagion models, network resource optimization, urban planning, social behavior analysis and even migratory flows can be dealt with data collected at scale from multiple sources [Barbosa et al. 2018]. Another approach, a more traditional one, is done through the construction of mathematical and statistical models to derive the behaviors of the entities being studied with a certain degree of realism. On the other hand, due to the ever-increasing collection of geospatial data through means such as mobile devices and location-based social networks (LBSN), a second approach based on historical mobility data analysis has become notable. These historical mobility data – often called mobility traces – allow the construction of models with a high degree of realism without the need for prior expert knowledge about the entities.

As expected, both private and public sectors take advantage of this kind of data: for the former, we can highlight location-based social networks, ride-sharing services (e.g., Waze Carpool), car-hailing services (e.g., Uber and Lyft) and mobility-based car

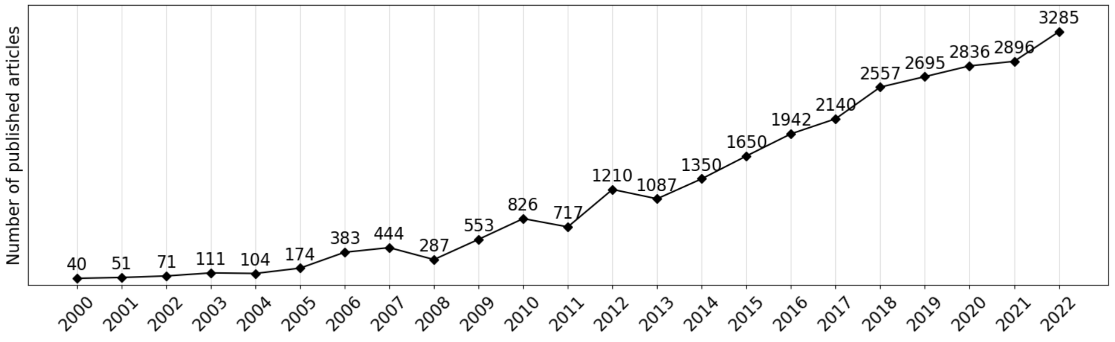


Figure 2.1. Published articles per year with the topic "geospatial data" (Source: Web of Science)

insurances. For the latter, geospatial data collection from public services and systems, such as buses' live locations and the spatial distribution of criminal activities, serve both as a way to increase transparency of resources management, as well as an opportunity for the community, including academics, to generate new knowledge from this data. Cities such as New York[1], Chicago[2], and Rio de Janeiro[3] have large collections of geospatial data openly available to the public. To illustrate this growing interest, Figure 2.1 shows a survey of published articles per year that include the topic *geospatial data*, since the year 2000, where a clear upward trend can be seen.

With this in mind, this work aims to increase the body of knowledge regarding concepts and techniques applied in the analysis of geospatial data – from collecting the data to applying it. We hope that by the end of this chapter, the reader is able to produce relevant results from geospatial data analysis, adopting the most appropriate practices and selecting adequate tools and algorithms.

2.1.1. Why geospatial data?

Next, we present some current and promising applications for which the usage of geospatial data brings benefits, aiming to illustrate to the reader the importance of this kind of data, as well as of its adequate use. For each application, Table 2.1 shows examples of open access geospatial datasets found in the literature that can be used as relevant sources for the development of analysis and applications.

2.1.1.1. Urban Mobility

To understand and model the urban behavior of people, vehicles and other mobile objects is one of the pillars of urban computing [Zheng et al. 2014]. From the knowledge obtained, cities can plan better the future of urban centers, improving the quality of life of its inhabitants. In this context, geospatial data can provide information about mobility dynamics from millions of people, being more precise and cheaper to obtain when

[1] https://opendata.cityofnewyork.us

[2] https://data.cityofchicago.org

[3] https://data.rio

compared to conventional strategies of data collection, such as field surveys and inductive loop counters [Naboulsi et al. 2016].

From the application point of view, we can highlight datasets of large populational scale, such as public mobility data, and logs from mobile network operators. For the former, urban mobility flow, transport demand, as well as points of interest (PoI) can be extracted from public mobility data, such as taxi and buses trips [Castro et al. 2012]. For the latter, network and call logs, referred to as *Call Detail Records*, can be used to plan and to allocate network resources, allowing better services during peak hours and large events [Marques-Neto et al. 2018].

2.1.1.2. Internet of Drones

According to [Motlagh et al. 2016] in a few years millions of drones will be available to work in many economy sectors, performing activities such as package delivery, tracking, surveillance in dangerous or hard to reach locations, agricultural, and even in combat. For this to happen, the mobility and communication between these unmanned aerial vehicles must be enhanced, through the development of communication protocols and orchestration methods. These technologies will make use of (among others) geospatial sensors, such as GPS, Bluetooth, and high-definition cameras, allowing the creation of drones swarms and coordinating their mobility.

2.1.1.3. Mobile and Vehicular Networks

Mobile and Vehicular Ad hoc Networks (MANETs and VANETs, respectively) are networks that enable the communication between mobile entities (in the second case, vehicles) and roadside auxiliary units, aiming to provide services such as traffic and accident alerts, multimedia sharing, and so on. Their main objective is to make mobility safer and more enjoyable for drivers, passengers and pedestrians. To do this, we have to model vehicular mobility so that the applications, systems and network protocols can take advantage of this information to adapt themselves to the vehicles' behavior. Data sources such as taxi, buses and private vehicles mobility are essentially important for the development of these technologies, being used both during behavior analysis, generating mobility models, as well as during the validation of proposed algorithms and protocols which will be used in urban environments.

2.1.1.4. Epidemics and Contagion models

Geospatial data are capable of capturing the human mobility behavior and its characteristics, such as PoI, social interactions, collective mobility patterns and flows. These factors make possible the usage of geospatial data to construct and enhance contagion models, that by themselves allow us to estimate the effects of infectious diseases on a population. By applying the information obtained from geospatial data to contagion models, we can estimate infection rates and model transmission in a given population, aiding in the development of actions and preventive measures. Specially during the pan-

Table 2.1. Open access geospatial datasets found in the literature

Name	Description	Applications	Source
Yellow Taxi Trip Data	Taxi trip records, capturing pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts.	Urban mobility	data.cityofnewyork.us
Transporte Rodoviário - GPS dos Ônibus	Geographical position and status of buses from the Rio de Janeiro city, collected at each minute	Urban mobility, Vehicular Networks	data.rio/
NYPD Complaint Data Historic	Includes all valid felony, misdemeanor, and violation crimes reported to the New York City Police Department from 2006	Safety	data.cityofnewyork.us
Crimes - 2001 to Present	Reported incidents of crime that occurred in the City of Chicago from 2001 to present, extracted from the Chicago Police Department	Safety	data.cityofchicago.org/
Traffic Crashes	Information about each traffic crash on city streets within the City of Chicago	Urban mobility, Safety	data.cityofchicago.org/
Package Delivery Quadcopter Drone	Flight data from drones performing a series of experiments carrying different payloads	Internet of Drones	kilthub.cmu.edu
New South Wales COVID-19 cases by location	COVID-19 cases by notification date and postcode, local health district, and local government area.	Epidemics	data.nsw.gov.au
Ciclovias	Geometries of all cycle paths in the city of Sao Paulo	Urban mobility	dados.prefeitura.sp.gov.br/
Foursquare Dataset	Includes check-in data from users of Foursquare all around the world	Urban mobility, Epidemics, Mobile Networks	paperswithcode.com/dataset/foursquare
Salzburg's 4G Driving Tests	4G measurements via repeated drive tests that covers two years on a typical highway section	Mobile Networks	ieee-dataport.org/
Gowalla	Check-in data from users of Gowalla along with their friendship networks	Mobile Networks Epidemics	snap.stanford.edu/data/loc-gowalla

demics period, geospatial data has been used to monitor population and the formation of agglomerations, allowing scientists to track the evolution of dissemination in almost real-time [Cebrian 2021].

2.1.1.5. Privacy and Safety

At last, geospatial data analysis from multiple sources can help protect the population, both individually and collectively. In vehicular networks, for example, telemetry data (e.g., current speed and engine temperature) from one's vehicle and from the vehicles surrounding it, road and weather conditions can aid the driver in preventing accidents, providing assisted or autonomous driving systems, with autonomous braking, collision and traffic detection, among other possibilities. Additionally, by understanding one's mobility behavior, collection processes can be enhanced to reduce the risks of sharing personal and private data, which can be used to pose threats if in the wrong hands [de Mattos et al. 2019]. Regarding the aspects of public and private safety, geospatial data provide a spatial coverage that together with collective sensing, allows each user in the network to contribute to the general safety. Also, it can contribute to map the behavior of suspicious entities and events, allowing the prediction and resolution of crimes.

2.2. Fundamental Concepts

In this section, we discuss the main concepts related to geospatial data, essential in all steps during analysis. First, we highlight the Earth's geographic characteristics and how they can affect operations with geospatial data. Next, in a broader context, we explore what are reference systems and introduce the definition and properties of spatial projections. Finally, we discuss the different distance measurements used for geospatial data and their characteristics.

2.2.1. Geography and its properties

Geodesic sciences are responsible for studying the shape and surface of the Earth, considering its imperfections and the many existing objects – natural or artificial – over (and under) it. It deals with gathering the information and defining representations and measurements for it. According to Bolstad [Bolstad 2016], to make use of geospatial data and its derived systems in an effective manner, we need to establish a clear understanding of how coordinate systems are defined for the Earth, how these coordinates are measured over its curved surface, and how they can be converted in different projections for its usage. If these factors are not taken into consideration, geospatial data collected will not be precise, and consequently, the operations performed with them can generate wrong results. While this imprecision may appear small and irrelevant for some cases, high risk applications such as the trajectory calculations for airplanes and missiles cannot allow it.

We can define two main factors to be considered in relation to geography: the shape of the Earth and the lack of precision from measurements. Regarding the former, the most common models used to represent the terrestrial surface are the planar projection, the spherical projection, and the elliptical projection. Although they allow for an easier visualization of maps in 2D surfaces, Earth's planar projections cause distortions to its curved geometry. Take, for example, a straight line between any two points in a planar map: it omits the existing curvature of the Earth between them (although for very short distances this curvature is virtually non-existent). On the other hand, while spherical projections eliminate the limitations of planar ones, they lead to imprecise measurements when closer to the poles due to the flattening of the Earth at these regions. Finally, elliptical models are the closest to Earth's geometry. As they become more realistic, models also become more complex, requiring advanced calculations – as we will discuss in Section 2.2.4 – which can affect the efficiency of the proposed solution.

It is worth mentioning that the existing models are simplified representations of Earth's real format, and therefore they present imperfections. It is not feasible to capture all the geographical characteristics of the Earth surface at a given moment, especially considering that it is constantly changing. In practice, we choose the model according to the spatial resolution of the research problem being dealt with.

2.2.2. Coordinate systems

Coordinate systems use coordinates to determine the position of objects in space. This space can be composed of one or more dimensions, and each dimension can contain particular properties such as inferior and superior limits, notations and scales. Thus, for a Cartesian coordinate system with two dimensions, we can define the location of an

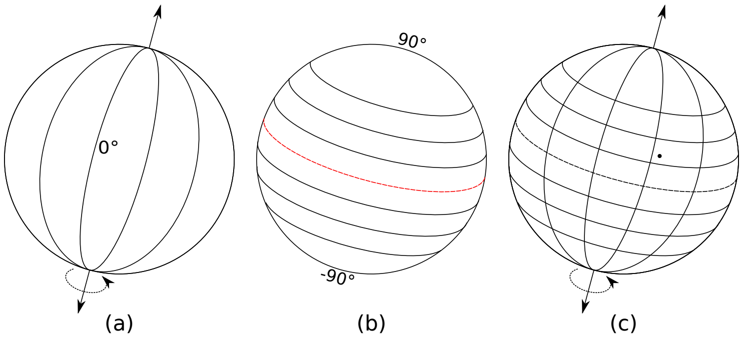


Figure 2.2. Geographic coordinate systems used to locate objects on Earth

object over the Cartesian surface by the pair of coordinates $P = (x, y)$, where each value represents the position in one of the dimensions. Similarly, geographic coordinate systems represent the location of objects on the Earth through geographic coordinates. These can be of two or more dimensions and consider different models for the Earth's shape. Following, we discuss about geographic coordinate systems and their properties.

The most used model of geographic coordinates is based on a spherical coordinate system to locate objects in a surface that resembles the Earth's shape. This system uses two rotational angles to specify positions on the modeled surface. The first rotation angle, called longitude (Figure 2.2(a)), is computed over the imaginary axis where the Earth performs its rotational movement. This axis go through the center of the Earth and has as extremities the North and South Poles. The positional variation over the axis is measured in degrees, with the zero position (0°) located on a imaginary line (a meridian) close to the Royal Observatory Greenwich, in England. The variation is positive when moving East and negative when moving West, reaching the maximum values of 180°and -180°, respectively, exactly at the opposite point of the zero position on the Earth's surface.

The second rotational angle, called latitude (Figure 2.2(b)), is computed over the Equator line, which represents half the distance between North Pole and South Pole. Its zero position (0°) is located exactly on the Equator line, with northbound variations being positive and southbound variations being negative, reaching maximum and minimum values at the Poles of 90 ° and -90 °, respectively. As such, we can define the position of an object on the Earth through a pair of latitude and longitude angles (Figure 2.2(c)). By its turn, each degree can be divided in 60 minutes (and each minute in 60 seconds), allowing geographic coordinates of latitude and longitude to specify the location of an object with precision under 1 meter. By convenience, angles are always specified in the order (*latitude*, *longitude*).

2.2.3. Spatial Projections

Geospatial data provide the precise location of objects on Earth through latitude and longitude angles. However, oftentimes we need to represent these positions on surfaces with different formats, such as a plain map. Plain maps cover bigger surfaces, are easier to visualize on paper, and their creation is simple. On the other hand, it is impossible to apply directly the position of objects in a spherical surface over a plain surface. As such, we apply spatial projections, that use mathematical formulas to transform locations from an original surface to a new one.

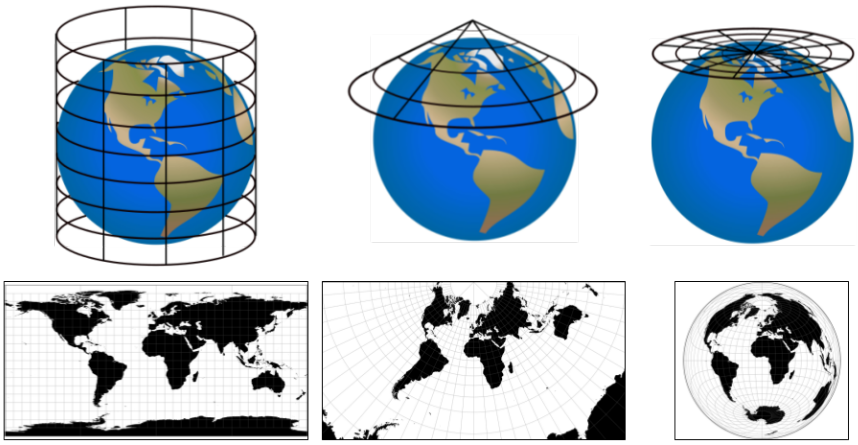

(a) Cilindric (b) Conic (c) Planar

Figure 2.3. Projection types

There are different projections available for the Earth, and although they are all used to represent locations on a planar surface, they vary in type and properties. The projection type refers to the geometric shape used to convert the sphere into a planar surface. This shape can be cylindrical, conic, planar, or a combination of those (Figure 2.3). Regarding properties, they represent the characteristics the projection preserves in relation to the Earth's surface, which can be conformal (preserving angles and shapes), equivalent (or equal-area, preserving area measurements), compromise (a half term between conformal and equivalent), and equidistant (preserving the real distance between points in the map). Table 2.2 presents a comparison of the main existing projections regarding their type and properties.

Different projections distort the Earth's surface differently. To adjust to a planar map, distortions are used to compress or elongate regions of the map. In fact, globe distortions are inevitable in planar projections. It is the case for the *Mercator* projection, that in order to keep the correct shape of continents, distort regions close to the poles, presenting sizes far bigger than their real areas, which can cause inconsistencies between visualizations and numeric results[4]. A variant of *Mercator* projection, the UTM (*Universal Mercator Transverse*) preserves the angles and formats of the regions, at the cost of distorting distances and areas. By the other hand, *Gall-Peters* projection presents surfaces with exact proportion and areas, at the cost of distorting their shapes. Lastly, the equidistant projection preserves the real distance between any two points on the surface, at the cost of distortions in the shape and area of regions.

2.2.4. Measuring distances

Finally, when dealing with geospatial data in the format of geographic coordinates, measuring distances between two or more points requires attention. Due to the Earth's ellipsoidal format, methods such as Euclidean distance will generate errors. In this situation, two other methods can be highlighted: Haversine formula and Vincentys formula. However, all methods have errors, and depending on the application even the Euclidean one can be used.

[4] An interactive graphic of these distortions can be seen at www.thetruesize.com

Table 2.2. Comparison of the main projections found in the literature regarding their types and properties

Projection	Type	Properties
Mercator	Cylindrical	Conformal
UTM	Cylindrical	Conformal
Gall-Peters	Cylindrical	Equivalent
Equidistant	Cylindrical	Equidistant
Equidistant Conic	Conic	Equidistant
Azimuthal Equidistant	Planar	Equidistant

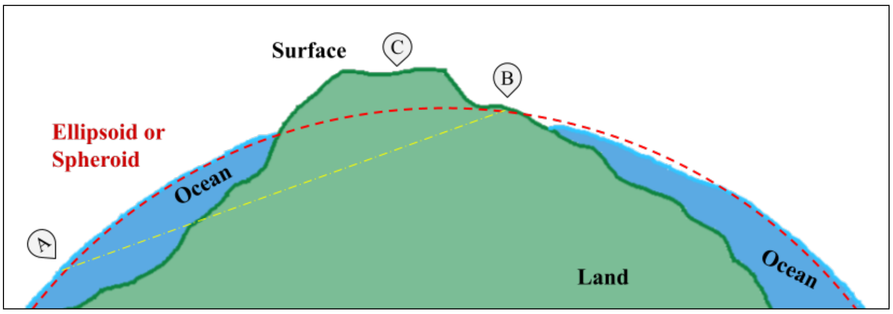


Figure 2.4. Measurements errors caused by irregularities on the Earth's surface

By inferring that the distance between two points is a line, the Euclidean method generates the biggest error of the three. As the distance between two coordinates increase, this error aggravates, impacting negatively its application. On the other hand, if the data analyzed is projected into a plane or if the expected distance between the points is too small, Euclidean distance can be used. By its turn, Haversine formula considers the distance between two points as a curve, better fitting the shape of the Earth, making it one of the most used methods in geospatial data applications. However, by considering the Earth as a sphere and not as an ellipse, this method also generate errors – although smaller ones. Lastly, Vincentys method computes the distance between two points based on an ellipse, being the most accurate one between the three. On the other hand, it is also the most complex, being more compute-intensive. Figure 2.4 shows an example of errors obtained when measuring distances on Earth. While it is clear that the Euclidean approach of measuring with lines produces significant errors, as we can see from the yellow dotted line between points A and B, the ellipsoid and spheroid approaches can also be at fault. Although the Haversine or Vincentys formula produces small errors for the distance between points A and B, Earth's irregularities still allow for larger errors, such as when measuring the distance between points A and C, or when measuring distances closer to the poles (in case of using Haversine formula).

2.3. Data Collection

Geospatial data often represent a simplified vision of the relation between one or more physical entities, such as a person or a vehicle, in relation to one or more locations, such as roads, cities, geographical coordinates, PoI, and so on. By capturing a check-in of a LBSN user in a restaurant, for example, we extract the information needed to represent this event as a geospatial datum, with the timestamp of when the check-in occurred, and the name

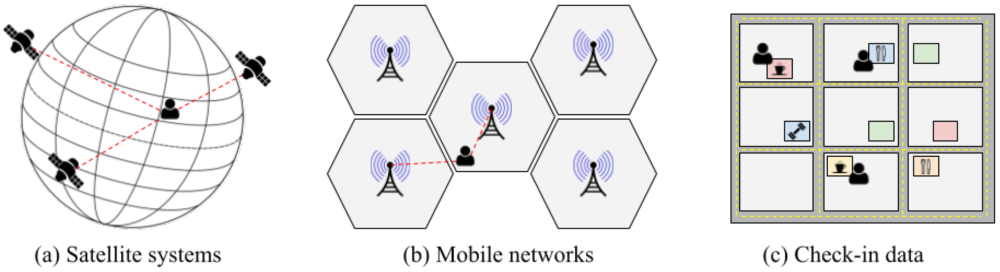

(a) Satellite systems (b) Mobile networks (c) Check-in data

Figure 2.5. How different geospatial data sources collect their data

and geographical coordinates of the location (when available). The definition of which information to collect must occur according to the analysis to be done, considering also the limitations of the technology applied for the collection. Also, issues such as sensed users privacy and the ethic of the proposed analysis must be taken into consideration.

Nowadays, the majority of geospatial data is collected using automatic measurements, due to its scalability, as well as for being less intrusive than manual collection methods. Such measurements are made using location sensors, and the collected data can be streamed in real-time (as some applications demand, such as to live track vehicle and passenger's position in *Uber*) or stored for later access. These sensors can be under possession of the sensed entity or not, e.g., smartphones and drones, respectively.

In this section, we discuss about the process of geospatial data collection, providing the reader with the concepts and tools needed to do so. Existing data sources, their properties, and examples are shown in Section. Then, we introduce different aspects related to the quality of the collected data in Section.

2.3.1. Data Sources

Geospatial data collection is a complex activity with high costs involved. Therefore, a clear scope of the data usage is needed in order to obtain the expected results without incurring additional costs with collection and post-processing steps. One of the main steps is selecting the data source, which must take into consideration the trade-off between the quality of the data obtained and the application costs of the sensing technology. Consider, for example, a scenario where we want to map points of interest within a city region. In this case, location data collected from Call Detail Records (CDR) do no have similar accuracy such as GNSS data; however, check-in data from LBSN may produce similar results with inferior collection costs. These suppositions are also valid to previously collected data from third-party agents that may have acquisition costs.

2.3.1.1. Global Navigation Satellite Systems

Global Navigation Satellite Systems (GNSS), such as GPS[5], GLONASS, and BeiDou, are satellite-based technologies that provide precise information about objects location on the

[5] Although this term is commonly used in the literature, GPS is a specific implementation of a GNSS. Most modern sensing devices (e.g., smartphones) are capable of connecting to multiple GNSS networks.

Earth's surface (Figure 2.5(a)). To do this, we need receptors that capture the signals from the satellites, calculating their position in degrees of latitude and longitude. To obtain its position, a receptor captures the signal from three satellites and performs a process called triangulation. In the presence of a fourth satellite, a receptor can also obtain the current date and time with high precision. GNSS are robust, capable of operating uninterruptedly, independently of weather conditions, and virtually in any exterior environment on the Earth's surface.

On the one hand, GNSS sensors are the most advanced geospatial data collection devices, obtaining accurate positioning with high frequency. Adding to that, nowadays these sensors are found in the majority of mobile devices such as smartphones and smartwatches, allowing a high-scale collection with low-cost. On the other hand, issues such as users' privacy, high energy consumption, and loss of signal must be addressed. First, the accuracy and frequency of positioning sensing of a user allows obtaining personal information such as home and work location [Kang et al. 2004], as well as work times [Gu et al. 2016]. Next, the elevated energy consumption must be considered, given that mobile devices have limited sources of energy and that most of the times the sensing devices have other functions sharing the same source; thus, the frequency of update must be just enough for the purpose. Together with that, physical barriers such as buildings and mountains can interrupt the signal reception, generating spatial and temporal gaps in the sensing [Silva et al. 2015] as well as positioning errors, issues that must be addressed using post-processing techniques such as filling and calibrating data [Celes et al. 2017].

2.3.1.2. Call Detail Records and Wireless Networks

Wireless networks – such as mobile telecommunications networks and Wireless access points (AP) – used by mobile devices, computers, and even vehicles, can be employed as low-cost sources of geospatial data. Mobile networks obtain location information for a user by the closest base stations, even if the user is not actively interacting (for example, making a call) with the network at the moment. The reported position corresponds to the transmission range of the contacted tower (Figure 2.5(b)). By its turn, the location of devices connected to access points is given by the unique identifier of the AP and the timestamp of when the access started. Like in mobile networks, the positioning precision corresponds to the transmission range of the AP. For both, multiple towers or multiple access points can be used to obtain a preciser positioning through triangulation and signal reception angles [Naboulsi et al. 2016].

There are certain advantages in using wireless networks to collect geospatial data. First, the energy consumption is low, given that the collection depends only on the connection of the devices to the network, a fundamental activity for their usage. Moreover, by being less precise, this collection is also less intrusive in comparison to the one using GNSS sensors, which reduces the rejection of the sensed individuals to provide their location. Finally, we can highlight the larger number of devices capable of connecting to wireless networks in comparison to devices with GNSS sensors. On the other hand, the biggest drawback of this source is the reduced precision, which can influence the quality of the geospatial data collected. Additionally, in comparison to GNSS, collection can suffer from the lack of coverage in remote areas, where network signal cannot be found.

2.3.1.3. Check-in Data

Check-in records represent the presence of a user in a location of interest (also called point of interest) during a determined time interval. These can be public spaces, such as parks, restaurants, and shops, as well as private and individual spaces, such as home and workplaces (Figure 2.5(c)). Check-in records are collected through location-based social networks such as *Facebook*, *Twitter*, and *Foursquare*, that capture information about the user, the visited location and the time of the visit, and may or may not contain the geographical coordinates of the PoI. The LBSN can use location sensors, such as GNSS and AP's identifiers - to automatically detect (or suggest) the user location, instead of requiring manual input. Like CDRs, check-in records present coarse temporal and spatial granularity, because the visits performed by a user during their activities are only registered if they voluntarily share it through their LBSN, which does not always happen due to issues such as privacy (i.e., the user does not want others to know where he is) and safety (i.e., the user fears that reporting his location may put himself in danger).

2.3.2. Collection Quality

Besides the technologies used and the sources of geospatial data, other issues must be taken into consideration during the definition of the data collection scope. Next, we introduce the main questions that arise, discussing how they can affect, not only the data collection, but its processing and the resulting analysis as well.

2.3.2.1. Location accuracy and precision

Accuracy, in terms of geospatial data collection, refers to the proximity of the collected location measurement to the real position of an entity. Therefore, the bigger the accuracy, closer the measurement is to the real value. It can vary from a few millimeters (e.g., high-capacity GNSS sensors) to kilometers (e.g., mobile networks in remote areas), thus comprehending accuracy is fundamental to select the most suitable source to the analysis subject. Precision, on its turn, represents the variance of the collection, and the bigger its value, more centered are the samples around a single point. To obtain preciser measurements, more powerful location sensors can be applied, nonetheless resulting in higher energy consumption and a higher collection cost overall.

Although they are related, a higher accuracy does not necessarily translate into a higher precision, and vice-versa. In the presence of data with low accuracy or precision, two approaches can be applied. The first is using more powerful sensors – when possible. The second is applying techniques to calibrate the collected measurements. This can occur during the collection, such as in the usage of additional signals in GNSS and wireless networks, as well as during data processing [Newson and Krumm 2009, Hoteit et al. 2016].

2.3.2.2. Privacy

Geospatial data can be used to report the location of various entities, in special those capable of moving such as humans and vehicles. Therefore, by providing their location, an individual allows the access to a real and valuable information. By the one hand, numerous benefits are provided from mobility data analysis and its applications. For example, services such as *Uber* and *Pokemon Go* are only possible by sharing location data. By the other hand, this information can be used by malicious agents to construct privacy attacks for the users being sensed, which can in turn demotivate them to share their data. The compromise between the utility of geospatial data and the risk to their privacy must be considered when collecting their data.

To be able to share their location information, we must guarantee users' privacy with the usage of techniques that reduce the amount of details or that make it harder to access the shared data. We can highlight two techniques: data anonymization and obfuscation. For the former, we replace the users' identifiers by randomly-generated pseudonyms, which can be done during data processing and even directly in the sensing devices [Krumm 2009]. However, even with this anonymization, attacks may re-identify users by detecting mobility patterns [Maouche et al. 2017]. The role of obfuscation is to prevent this by creating small distortions in the data in an attempt to break existing patterns, without affecting its quality [Duckham and Kulik 2005a, Duckham and Kulik 2005b].

2.3.2.3. Sampling rate

The interval between two consecutive location reports is also an important aspect to be discussed. We call this interval sampling rate, in a way that data with higher sampling rates present a smaller time interval between two samples. Therefore, during a given interval, collections with higher sampling rates will produce larger amounts of data. As expected, more data implies in more details and bigger utility, allowing, for example, the analysis of detailed mobility trajectories. On the other hand, collecting data with high sampling rates leads to higher usage of resources such as power and storage, which are limited in portable and mobile devices. When implementing a geospatial data collection process, we need to specify the collection sampling rate to guarantee the coverage of the activities or behaviors which are subject of our analysis.

While some data sources allow us to set the sampling rate (e.g., GNSS), others depend on the interaction of the users with the system (e.g., LBSNs), and thus their rates cannot be controlled directly. Even when possible, adjusting the sampling rate can be costly, as discussed above. From that, spatial and temporal gaps will occur, intervals in which there is no information regarding the whereabouts of the user. These gaps can be filled using techniques such as interpolation [Hoteit et al. 2016] and algorithms based on historical data [Silva et al. 2015, Chen et al. 2017]. The enrichment of geospatial data transforms sparse into dense in an artificial manner, with no need to change the methods and tools of collection.

2.3.2.4. Scale

Finally, we discuss the collection scale. To represent the simulated environment and produce meaningful results, the set of entities contained in the data must be significant. The scale can refer to the number of distinct users being sensed, to the number of time intervals (hours, days or weeks), and to the dimensions of the area being monitored. The scale of these dimensions must comprise a scenario in which the resulting analysis does not present bias due to: user limitations (the sample of individuals does not represent the population); time limitations (the period does not contains all the expected situations); and space limitations (the dimensions of the covered area do not represent the real environment).

When the collected geospatial data are not enough to produce results, some approaches are used to increase the data volume. Data fusion [Rettore et al. 2020] is a technique that combines two or more datasets producing as output a single set containing all the data. However, for very distinct datasets, its application can be compute-intensive. Another approach is the generation of synthetic data, which uses statistical [Kosta et al. 2012] and machine learning methods to generate data that is similar to the real behavior. Although it demands a precise modeling of real data, this technique has as benefit the capacity of generating synthetic data on demand and in large-scale.

2.4. Data Storage

This section will discuss concepts, techniques and existing tools for storing geospatial data, which is fundamental for dealing with massive amounts of data. It is important to discuss this topic, given that in its majority, traditional relational databases are not fit for efficiently storing geospatial data. This is due to the different forms of representing geospatial data that may not be compatible with tabular storage. Moreover, we must consider data manipulation, i.e., inserting and querying data, and the need to perform geospatial filters.

2.4.1. Spatial components structure

The real world is too complex to be fully represented by a data structure, and thus we must select the relevant characteristics (e.g., roads, buildings) for each scenario. To digitally represent geospatial data, there are two primary structures: vector and raster. Vector structure is based on using points, lines, and polygons to define the location and limits of an object. By its turn, raster structure uses a regular cell grid to define objects. Each structure has its own advantages and drawbacks in data modeling. Moreover, we can combine the two approaches in a single project aiming to get the best of both. Table 2.3 presents a comparison between the two data structures.

The characteristics of each structure impact directly on the details they are capable of capturing. Figure 2.6 shows the transformations of a real world representation into vector and raster structures. The vector one represents the existing entities considering their dimensions and shapes. To do so, we must select which geometric shapes will be used in the representation. On the other hand, the raster structure reduces every feature contained in a single cell into a basic identification according to a codification criteria. In the example shown in Figure 2.6, we used the dominant area criterion, in which the label corresponds to the feature that occupies the majority of the cell. Another possible criterion is using the feature located at the center of the cell.

Taking into consideration the characteristics mentioned above, it is clear that there is not a superior structure. Raster structure has the advantage of being simpler to store,

Table 2.3. Comparison between vector and raster structures

Characteristic	Vector	Raster
Data structure	Generally complex	Generally simple
Storage requirements	Small, for most data	Big, for most uncompressed data
Coordinate systems conversion	Simple	Can be slow due to the volume, may require resampling
Positional precision	Limited	Depends on the resolution adopted
Accessibility	Often complex	Easy to modify by using specific programs
Visualization and output	Similar to maps, with continuous curves; poor for images	Good for images, but can produce jagged effects
Spatial relations between objects	Topological relationships between available objects	Spatial relationships must be inferred
Modeling and analysis	Map algebra is limited	Easier superposition and modeling

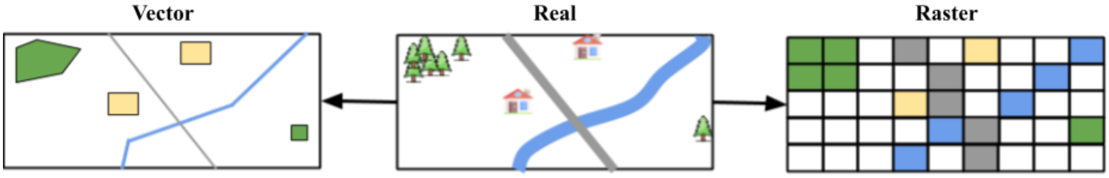


Figure 2.6. Vector and raster structures (Adapted from [Lisboa Filho and Iochpe 2001])

specially when dealing with digital images, such as aerial pictures and satellite images. On its turn, vector structure tends to be more accurate, providing better visualizations and efficient calculations of topological operations. At last, vector structure stores only the essential elements, reducing the amount of storage needed, while raster codifies the entire grid, which can be unnecessary.

2.4.2. Data Compression

As it can be noted, geospatial datasets are used to represent large amounts of information, demanding considerable storage capacity. Just as in traditional datasets, compression algorithms can be applied in geospatial datasets, resulting in a more efficient storage. Compression algorithms can be classified into lossy and lossless: while the former obtain high compression levels at the cost of reducing the data quality, the latter preserves its quality but obtains lower compression levels. Although for certain applications lossy compression is acceptable, when dealing with geospatial data it can severely impact the analysis' results. Thus, using lossy compression algorithms when processing or analyzing geospatial data is not recommended.

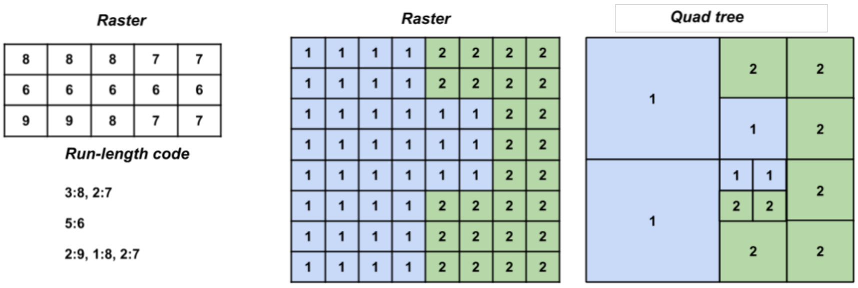


Figure 2.7. Run-length and Quad Tree compression of raster data

Given their representation characteristics, most compression algorithms for geospatial data are focused on raster structured data. A common method for compressing raster data is Run-length code. This compression technique is based on codifying a sequence of cells to optimize space when there are large sequences of cells with the same value. This coding is represented by two numbers, with the first indicating the amount of cells with the same identification and the second being the identification itself. Another well-known coding is a bi-dimensional version of Run-length code called Quad Tree [Finkel and Bentley 1974]. In this method, areas with same values are represented with a single identifier. To do this, the grid is divided recursively into increasing square blocks until the division is not possible anymore, resulting in squares where all identifiers inside them are equal. An application example of both methods can be seen on Figure 2.7.

2.4.3. Databases

One of the most important components for geospatial data analysis tools, such as Geographic Information Systems (GIS), are Spatial Database Management Systems (Spatial DBMS). Besides having the conventional functionalities found in traditional database management systems, Spatial DBMS accept different geospatial reference systems, providing functions for querying and manipulating this type of data. Additionally, they are capable of indexing geospatial data both using coordinates as well as using polygons, improving efficiency. However, without indexing, both querying locations and filtering regions are inefficiently, mainly when dealing with large volumes of data.

Some examples of commonly used DBMS for storing geospatial data are *MySQL*, *PostgreeSQL* with *PostGis* extension, and *Oracle Spatial*. They use the same spatial data standard, called Simple Feature Specification - Structured Query Language (SFS-SQL), which describes a common storage and access model for geometries (points, lines, and polygons). SFS-SQL is a standard defined by Open Geospatial Consortium (OGC) [ogc 2023] that, besides describing the geometries used by GIS, present the definition of operations between geometries.

2.4.4. Indexing

Finally, indexes are data structures used to increase the performance of queries in database systems, allowing data to be retrieved in more efficient ways than linear searches. Indexing (i.e., the creation of indexes) in geospatial databases is useful not only to efficiently query data, but also to perform many spatial operations. We can cite the identification of k-nearest neighbors, geocoding (obtaining coordinates from an address), and reverse geocoding (obtaining an address from coordinates).

The most applied structure for indexing geospatial data is called R-Tree [Guttman 1984], that indexes geometries by using a balanced tree structure. This strategy has as advantage the capacity of querying data with logarithmic time complexity ($O(log_{|M|}|N|)$), justifying its usage in diverse databases and geospatial analysis tools, such as *PostGis*, *Oracle* and *GeoPandas*. On the other hand, we must take into consideration the processing costs for generating the tree, which may limit its usage.

Besides R-Tree, we can use grid-based systems, which are easier to implement, such as *Geohash* [Morton 1966] and *H3* [Uber 2015]. Grid-based systems can analyze massive amounts of geospatial data through the division of larger areas into uniquely identifiable cells. *H3* provides an hierarchical spatial index based on hexagons, grouping points in hexagons of different sizes according to the precision needed in the analysis. *H3* index levels determine the area of hexagons and its selection is essential for a better precision during indexing. On the one hand, hexagons too big will group distant points in a same cell. On the other hand, hexagons too small will result in a very large number of indexes, decreasing performance. *Geohash*, on its turn, uses rectangles instead of hexagons. However, the indexing approach is the same as H3, with the disadvantage that each rectangle has eight neighbors while each hexagon has six – more neighbors means more locations to search.

2.5. Data Transformation and Knowledge Extraction

This section presents the core of the process of geospatial data analysis, which comprehends the steps of data transformation and knowledge extraction.

2.5.1. Data Transformation

The transformation of geospatial data involves five steps: formatting, sampling, cleaning, filtering, and aggregating. Although they are all essential in preparing data to analyze, the need to apply each one is defined by the initial conditions of the input data and the characteristics of the analysis. Moreover, multiple transformation iterations can occur with the aim to refine and validate the obtained results.

2.5.1.1. Formatting

When working with geospatial data, it is essential to observe formatting, since it can be represented in different forms. When dealing with geographical coordinates (latitude and longitude), for example, there are three basic formats in which they can be found: degrees, seconds, and minutes; degrees, seconds, and decimal seconds; and degrees and decimal degrees. The choice of which format to use will depend of the application, given that a larger number of decimal places allows representing locations with higher precision. Tools such as Geospatial Databases, GIS software, and sources of official data often use the last format, that represents coordinates as degrees and decimal degrees (e.g., 38.8897°). Another form of representation is the set of coordinates, creating polygons or lines. Platforms such as OpenStreetMaps represent polygons as a sequence of coordinates; The *Shapely* library, on the other hand, adopts the Well-Known Text (WKT) format to represent sequences of coordinates. Other approaches are Shapefiles, sheets, and JSON. Besides certifying that all data are with the same format, it is important to verify if all records are under the same projection and datum, since different projections will lead to wrongful analysis.

Other data sources, such as check-in and wireless networks, may require data formatting as well. For the former, different LBSNs may refer to a same location with different names (by abbreviating words, for example), requiring the identification of all the possible formats and the conversion to a unique identifier. Additionally, during the collection time span, locations can change names or addresses, which also must be considered during formatting. Similarly, for the latter, access points and towers can change their identifiers over time, as well as each device can reproduce this identifier in a particular way, according to different technologies and software specifications.

2.5.1.2. Cleaning

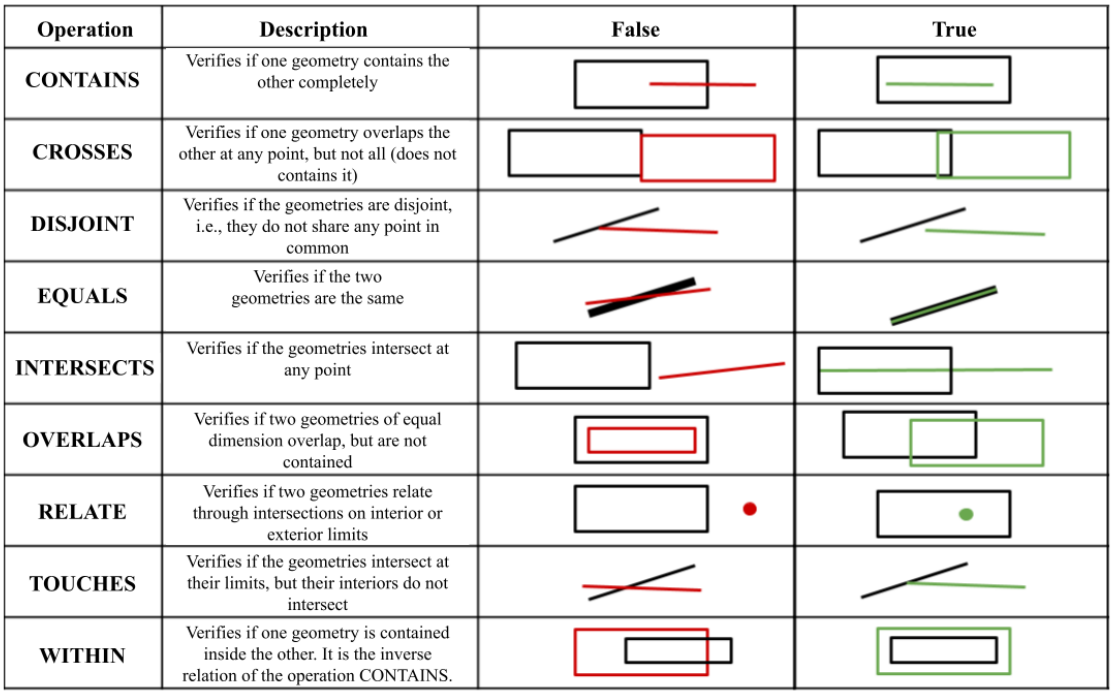
Besides undesired locations, the usage of data collected from urban areas can bring some challenges, such as the imprecision of geographical coordinates. This occurs due to the large number of obstacles, such as buildings, that obstruct the line of sight of satellites, a phenomenon called urban canyons [Johnson and Watson 1984], reducing their capacity to attribute a location with the required accuracy. For this reason, it is important to analyze the impact of the affected locations in the collected data, and if needed, remove them. The main consequence of using erroneous data (or data with low location accuracy) is reducing the significance of geospatial analysis, specially when dealing with distances between points and users' density.

Even if they do not refer to geospatial data, other data dimensions must be considered during cleaning. These, besides not being affected by the issues presented above, can also contain irregularities such as null values or outliers. Therefore, cleaning them is crucial to guarantee trustful results.

2.5.1.3. Filtering

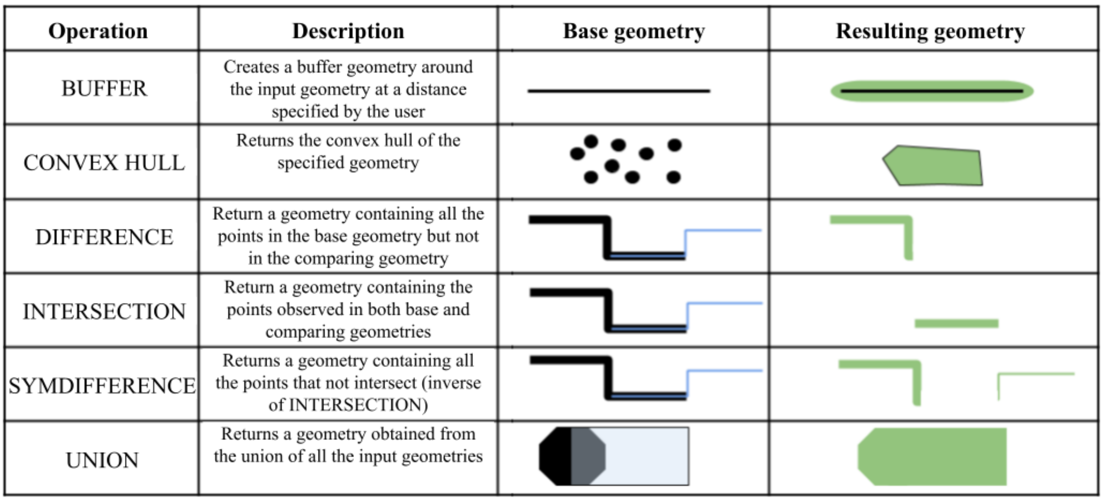
While the cleaning step aims to remove wrongfully sensed information, the filtering step aims to select, given a cleaned dataset, a subset that meets the specified rules for analysis. While here we specifically refer to rules applied to geospatial data, it is valid to point out that filtering can include other data dimensions as well. Thus, according to the application, we can remove data located at unwanted regions, e.g., outside the city limits where we want to focus our analysis.

It is important to highlight that records from geospatial data are frequently represented as a single point, that is, a single latitude and longitude pair. To filter records within a location of interest – which can be a street, a place, or even a city – we can consider the geometric representation of this area (i.e., a polygon or a set of polygons encompassing the whole area). To do this, various geographic operations can be applied to evaluate the relationship between geometries. Such operations can be found in GIS software and databases with support for geospatial data, as well as in libraries such as *Shapely*



Operation	Description	False	True
CONTAINS	Verifies if one geometry contains the other completely		
CROSSES	Verifies if one geometry overlaps the other at any point, but not all (does not contains it)		
DISJOINT	Verifies if the geometries are disjoint, i.e., they do not share any point in common		
EQUALS	Verifies if the two geometries are the same		
INTERSECTS	Verifies if the geometries intersect at any point		
OVERLAPS	Verifies if two geometries of equal dimension overlap, but are not contained		
RELATE	Verifies if two geometries relate through intersections on interior or exterior limits		
TOUCHES	Verifies if the geometries intersect at their limits, but their interiors do not intersect		
WITHIN	Verifies if one geometry is contained inside the other. It is the inverse relation of the operation CONTAINS.		

Figure 2.8. Examples of operations that verify a relation between two geometries



Operation	Description	Base geometry	Resulting geometry
BUFFER	Creates a buffer geometry around the input geometry at a distance specified by the user		
CONVEX HULL	Returns the convex hull of the specified geometry		
DIFFERENCE	Return a geometry containing all the points in the base geometry but not in the comparing geometry		
INTERSECTION	Return a geometry containing the points observed in both base and comparing geometries		
SYMDIFFERENCE	Returns a geometry containing all the points that not intersect (inverse of INTERSECTION)		
UNION	Returns a geometry obtained from the union of all the input geometries		

Figure 2.9. Examples of spatial operations on geometries

and *GeoPandas*. Figure 2.8 details the main operations, with each returning a Boolean value (true or false) by comparing two geometries. Additionally, there are operations that do not analyze the relationship between two geometries, but perform spatial operations, returning values or new geometries as output, as it can be seen in Figure 2.9.

2.5.1.4. Sampling

When analyzing large amounts of data, we may not possess the computational capacity or even the interest in the whole dataset. Frequently, a small sample is enough to understand and to generate knowledge about all the data. Once data is formatted, cleaned and filtered, we can consider using samples of data. Sometimes, data sampling is the first step in the data transformation process, making it easier to clean and filter the dataset, since we are dealing with a reduced volume. For this, we must guarantee that the sample is representative of the population, otherwise it will introduce bias, invalidating the results. In a valid sample, records are selected randomly to guarantee that no biased record gets picked on purpose. In the resulting sample, the distribution of the data inside it must originate from the same distribution of the population.

Regarding geospatial data, we must also observe if the sample represents the existing variations regarding space and time. For the former, entities in different regions may behave differently, and thus it is important to capture all this variability. For the latter, the original dataset can cover a time interval with comprising many weekdays, holidays, and even multiple seasons. Considering the change in the users' routine caused by these periods, we must consider splitting the sampling process or creating a stratified sampling, generating one or more samples to meet the needs of our analysis.

2.5.1.5. Aggregation

Finally, with the data ready to be used, we must analyze if its current granularity is enough for the desired analysis. If not, then we must aggregate our data, according to defined aggregating rules and aggregating regions. For example, if our analysis is focused on populational metrics by neighborhood, then aggregating records allows us to proceed while also reducing the processing needed by decreasing the amount of data. Aggregation can also happen due to privacy considerations: in certain scenarios, analyzing the raw data can reveal personal information. By aggregating, individual traces can disappear or become indistinguishable from one another. Finally, aggregation can also occur to enable fusing with other data sources. Weather forecast, safety, and traffic indicators are examples of data sources frequently used in geospatial data analysis, each one with different granularity. In order to combine these sources in an efficient way, aggregation can be used with few or no information loss.

2.5.2. Knowledge Extraction

Next, we showcase applications of knowledge extraction from geospatial data. For each one, we give examples of its functioning and highlight its importance in data analysis.

2.5.2.1. Radius of gyration

When working with geospatial data, we frequently assume that the analyzed users have a reference location, which can be their home, work place or any other point of interest. Based on that, we can calculate a users' radius of gyration, that can be defined as the maximum distance between its reference location and the other locations visited by them.

The radius of gyration provides information about the mobility of users, allowing their classification and evaluation according to their radius' value, for example.

To compute a users' radius of gyration, we must first define the criteria to identify their reference location through geospatial data. Some approaches found in the literature are using a random point [Kosta et al. 2012], the average of all points, the most visited location, and the first point reported in a day [Ekman et al. 2008]. We must consider the characteristics of the dataset, such as its granularity and sensors used, to select a criteria that identifies reasonable reference locations. Additionally, we must select a distance function (haversine or euclidean, for example) to perform the measurements.

Knowing the radius of gyration of users has contributed to research about urban mobility. Previous studies have used this information to argue that users tend to explore ever increasing regions over time [González et al. 2008], as well as shown that some users tend to explore more than others [Pappalardo et al. 2015]. Additionally, radius of gyration analysis has been used to investigate locations visited by social networks users [Jurdak et al. 2015]. Finally, it has aided in understanding escape routes during natural disasters [Wang and Taylor 2014] and understanding how communication in social networks helps disseminating actions in global scale [Morales et al. 2017].

2.5.2.2. Spatial Clustering

Other common activity when dealing with geospatial data is the need of identifying groups that present similar behavior patterns. Such patterns can be represented by users that frequent a same region [Sakai et al. 2014], providing information to identify regions with higher demands for services, locations where diseases can be spread easier, among others. To identify these regions or groupings, we use a technique called spatial clustering.

There are several clustering algorithms found in the literature, using different approaches to obtain groupings, such as algorithms based on distance (K-means), based on density (DBScan), and based on distribution (GMM). Although they are algorithms frequently used in common datasets, using them for clustering geospatial data can lead to errors. Algorithms based on distance frequently work with euclidean distance, that fails to compute the correct distance between points on the Earth's surface [Ingole and Nichat 2013]. While algorithms based on distribution can also work, the most popular algorithms for spatial clustering are the ones based on density, since they can be applied for most types of data. However, parameters must be well-calibrated so the groupings produced are representative.

There are many different applications of spatial clustering: adjusting network resources to accommodate demand of a region, feeding recommender systems for indicating similar locations, and personalizing marketing campaigns according to the visited location [Tran et al. 2013] are some examples.

2.5.2.3. Social relationships

Inferring social relationships between individuals in a dataset can help us better understand daily activities, such as mobility patterns that change due to others [Cho et al. 2011], or even the projection of disease propagation in a pandemic [Firestone et al. 2011]. For this to happen, correctly identifying social links between two users is essential. Social links derived from an analysis can indicate if two individuals are friends, acquaintances, neighbors, workmates or housemates, for example. However, inferring theses links is not a simple task: we must define rules – using the data – to affirm that two individuals were in contact with each other. To define a contact, one may use the spatial proximity between users' reported positions (in case of GNSS data) or the presence at a same location (in case of LBSN data). In both scenarios, the temporal proximity must also be considered, to guarantee that users were in the same area at the same time. Moreover, commonly used rules to identify social links are high encounter frequency, the intersection between users' social links sets, and the detection of communities.

Regarding the relation of social links and geospatial data, there are works in the literature to identify the probability of users having similar tastes by the occurrence of similar routes [Hung et al. 2009]. Other works also analyze the occurrence of communities from social links obtained from geospatial data [de Melo et al. 2015], as well as using these links to disseminate data in opportunistic networks [Domingues et al. 2022].

2.6. Visualization

Visualizing data is essential in all steps during analysis. We start visualizing data right after collecting it, with the objective to validate and verify, identifying issues that can be corrected through new collections or processing steps. Next, different visualizations are created to depict the distribution of the data, allowing the detection of outliers and the discovery of the population characteristics. During analysis, visualizations aid in decision making and presenting partial results. Finally, the end results are also presented through visualizations, that ease the understanding and illustrate the proposed ideas.

In this way, it is fundamental to create easy to read graphics with no considerable time or effort. These two factors guarantee that the usage of graphics become a tool for the process, and not another problem to be solved. For this to happen, we must know the different forms of visualization of geospatial data and the results they provide. Different from numerical and categorical data, in which graphics such as barplots and lineplots are enough to transmit the information, the visualization of geospatial data generally involves the need of a map over which the location records are drawn. Additionally, we need to consider aggregation strategies due to the large volume of data. Drawing massive amounts of data can be inefficient in terms of computing resources, and also produce polluted results. Finally, we are frequently more interested in visualizing existing patterns than the behavior of unique individuals.

To aid in understanding geospatial data graphics, we draw maps from the represented regions to approximate the figure to the real environment. To draw a map, first we define its type and projection. Map types define the characteristics it represents, such as terrain, territorial borders, streets and roads. Projection, on its turn, refers to representing data in two or three dimensions. Both factors must be selected in a way to increase the understanding of the generated graphic, while also non essential characteristics must be unconsidered to avoid pollution. For example, using a three-dimension projection for two dimensional data will only add uncertainty to a figure.

Zoom levels refer to magnify the region covered by the figure, approximating (zooming in) or distancing (zooming out). A graphic must cover the whole area where the geospatial data are located, unless the objective is to highlight a specific region. However, in the occurrence of outliers, we may generate visualizations that are distant from the main region of interest, and thus, zooming in may be considered.

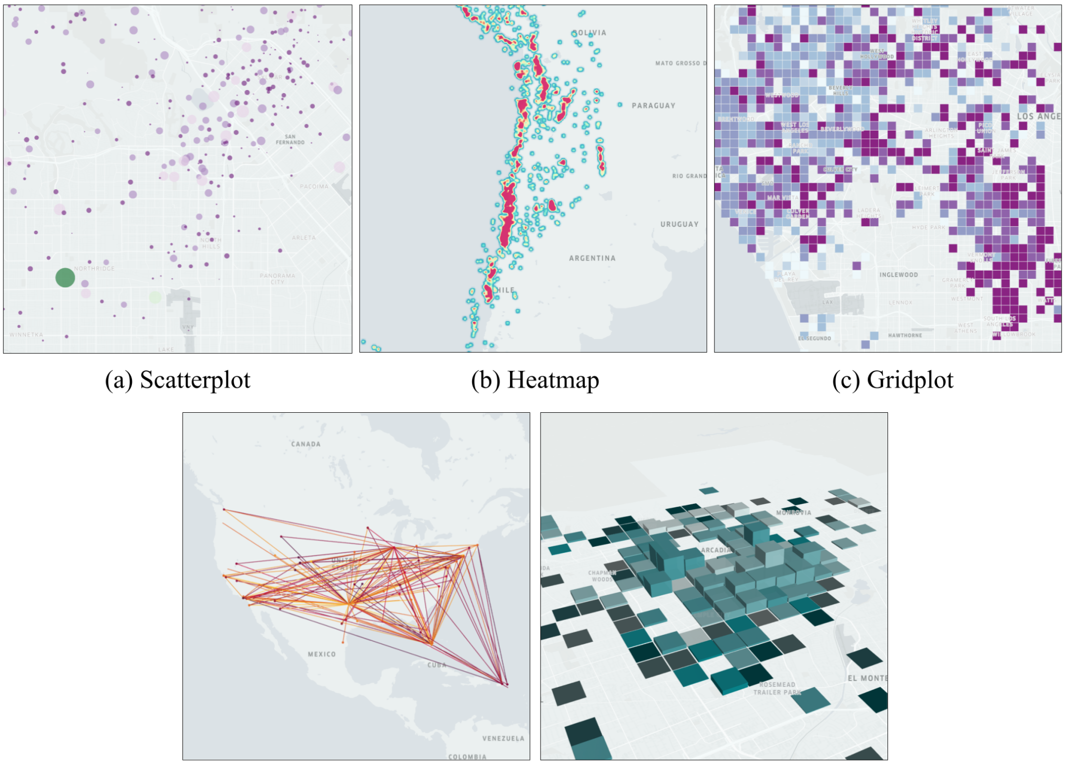
Finally, beyond aesthetic reasons, color scales in a graphic can be used to indicate intensity levels (such as altitude) and to separate categories found in data (such as device type or PoI category). Whenever possible, we must opt for color scales with high contrast (considering that the figure may also be visualized in grayscale) and that refer to the categories being represented (e.g., using a red and blue scale to indicate temperature). In the presence of an elevated number of categories, colorscales can produce confusing visualizations, and in this case, they can be replaced by textual (e.g., written values) and visual (e.g., different symbols for each category) subtitles.

2.6.1. Visualization Types

The visualization types presented next are the basic structure to construct graphics for analyzing geospatial data. From these approaches, it is possible to develop new visualizations, adapting and adding characteristics to meet the requirements of the desired result. By creating the first sketches, the reader will notice that modifications are needed to make the proposed graphic as clear as possible, which is essential for its understanding. Adjusting the map format, zoom levels, color scale, and visual and textual subtitles are some examples of modifications. Figure 2.10 shows an example for each of the visualizations presented next.

Scatterplot. Scatterplots constitute the simplest form of visualizing geospatial data. In it, locations are drawn as points over the map according to their coordinates. Points can have different sizes, colors and formats to depict different characteristics found in data. Scatterplots are easy to comprehend, to implement and modify. On the other hand, visualizing large amounts of data in scatterplots will lead to point superposition, causing loss of details. Additionally, in this scenario their construction will be compute-intensive. With this in mind, they should be used to visualize the dispersion of the collected data, but we should avoid them when we want to explore the details in the population.

Heatmap. Heatmaps are used to represent the density of a variable by means of intensity curves and color scales. When combined with geospatial data, both the variable's density and its spatial dispersion (i.e., how density changes according to space) can be shown in a single visual. In this combination, high-density regions can be formed, that can represent a topology or PoI, for example. Heatmaps are recommended when data does not present a uniform location distribution, leading to the existence of regions with higher densities. Because they are base in density regions, heatmaps can produce poor results for datasets with sparse locations, that is, when points are too far away from each other, with no aggregation.



(a) Scatterplot (b) Heatmap (c) Gridplot

(d) Flowplot (e) Barplot

Figure 2.10. Examples of geospatial data visualizations

Gridplot. Gridplots are created by dividing the analyzed region into a grid with defined dimensions, where the minimum location unit are the cells that compose the grid. For each cell, the data located in the coordinates within it are aggregated through an aggregation function (e.g., summation, average, minimum or maximum value) and the result of this function represents the cell. Besides the aggregation function, another parameter that must be defined is the cell size, that is, the area covered by it. Smaller cells are capable of capturing more details, while larger cells can lose relevant information. On the other hand, it is easier to have an empty cell (with no data inside it) as its size reduces. Eventually, a compromise between the level of details and the minimum number of records inside each cell may be needed.

Flow plot. Flow plots are used to represent the movement flow of entities (that can be people, vehicles, drones, or others) between two or more regions. This displacement is represented by arcs that connect the origin and destination regions, together with an intensity indicator, which can be done using the arc's dimensions (e.g., a bolder arc indicates a bigger flow), or through a color scale. Arcs may not be symmetric, that is, the flow's intensity from point A to point B can be different of the one from point B to point A. In this way, multiple graphics can be drawn to cover all cases and avoid superposition that leads to confusion. Other forms of representing flows between regions can be found in the literature, such as transition graphs and transition matrices. However, they do not communicate the spatial distribution as well as flow plots.

Barplot. Barplots for geospatial data project bars over a spatial region, where each bar represent the impact of a variable over that specific location. Like gridplots, they are useful when geospatial data have information sensitive to location. By its turn, these graphics allow readers to observe regions of interest easily, due to the projection of the bars in a third dimension. However, we must consider the limits of representing data in third dimension through non-interactive means, such as in printing: the projection may cover some regions, causing loss of information. Therefore, barplots for geospatial data should be preferred for digital and interactive means.

2.6.2. Visualization Tools

Next, we discuss some libraries and tools commonly used to create geospatial data visualizations. While most plotting tools (e.g., gnuplot and matplotlib) can be used to draw geospatial data, our aim here is to focus on those that provide specific resources for this kind of data, thus reducing the amount of work needed and speeding up the process.

Bokeh. Bokeh is a library to build and visualize graphics. Built with Python, it is capable of generating interactive visualizations, allowing users to change parameters and scales and add new data to an existing graphic in real time. This makes Bokeh an interesting alternative for publishing results in websites. Additionally, it is capable of rendering large amounts of data. Lastly, users can find an extensive collection of visualizations available for use, including the ones discussed above, as well as more complex ones.

Kepler. Kepler is a geospatial data analysis tool developed by *Uber*. It can be used through a Web interface that allows loading data, performing aggregations, filtering, and projecting over a detailed map of the Earth's surface using many different visualizations, such as scatterplots, heatmaps, flowplots and barplots. Besides that, the user can select the map type and projection, draw additional geometries to complement the visualization, observe data over time (when a time variable is available) and export the results to different formats. However, Kepler can only be used to visualize geospatial data with locations based on latitude and longitude coordinates.

OSMNx. OSMNx is a Python library for building geospatial data visualizations focused on road maps. Using data from OpenStreetMaps, it is capable of generating personalized visualizations of road mesh of a determined region, over which the user can draw additional geospatial data. Besides that, it is capable of constructing the road network by using graphs, allowing the mapping of geospatial locations to their closest roads, that in turn allows computing distances, road-based shortest paths, as well as metrics and complex networks algorithms.

QGis. QGis is a multiplatform software to visualize and edit geospatial data for analysis. It is a robust system, capable of loading large amounts of data and that supports different input types. Due to its advanced capacities, it can produce high-quality visualization. On the other hand, it demands a stepper learning curve. Finally, by being a standalone tool, its interaction with other geospatial data analysis (e.g., Python scripts for processing) can be complex, making it less preferable for building quick visualizations.

2.7. Conclusion

This chapter presented an in-depth study about geospatial data and its applications in knowledge extraction, generating new products and services and allowing the capture of new sources of revenue, as well as advancing the state-of-the-art in areas related to mobility, Internet of Things (IoT), urban computing, among others. For this, we presented theoretical concepts and the main techniques and tools applied in the steps of collection, storage, transformation, knowledge extraction, and visualization of geospatial data.

We highlighted the importance of mobility to the development of new techniques and technologies applied to issues such as traffic flow prediction and control, contagion models, network resource optimization, among others. In this scenario, the interest for studying and applying geospatial data has become bigger, due to its capacity to represent the entities' mobility behavior, specially in the case of humans and vehicles. However, although there is an increasing demand for research involving this type of data, still there is not a consensus regarding the adequate methodologies for analyzing it, due to the lack of references that introduce in a clear way the concepts and techniques to be applied.

Aiming to fill this gap, Section 2.2 introduced the main concepts related to geospatial data, such as geographic characteristics, reference systems, and spatial projections. Section 2.3 discussed the process of data collection, highlighting the most commonly used existing sources (GNSS devices, networks and LBSNs). Additionally, characteristics from the collected data such as accuracy and precision, granularity, and entities' privacy were discussed. Next, Section 2.4 presented the particularities of storing geospatial data, discussing spatial database management systems, compression methods for geospatial data, and spatial indexing structures, required for an efficient storage. Section 2.5 presented the steps of data transformation and knowledge extraction. The former is composed by the sub tasks of formatting, cleaning, filtering, sampling, and data aggregation. For the latter, we introduced some examples of applications, such as radius of gyration and spatial clustering. Lastly, data visualization techniques and the tools used to create them were discussed in Section 2.6.

References

(2023). The Home of Location Technology Innovation and Collaboration | OGC. [Online; accessed 7. Aug. 2023].

Barbosa, H., Barthelemy, M., Ghoshal, G., James, C. R., Lenormand, M., Louail, T., Menezes, R., Ramasco, J. J., Simini, F., e Tomasini, M. (2018). Human mobility: Models and applications. *Physics Reports*, 734:1–74.

Bolstad, P. (2016). *GIS Fundamentals: A First Text on Geographic Information Systems*. Eider Press, 5 edition.

Castro, P. S., Zhang, D., e Li, S. (2012). Urban traffic modelling and prediction using large scale taxi gps traces. In *International Conference on Pervasive Computing*, pages 57–72. Springer.

Cebrian, M. (2021). The past, present and future of digital contact tracing. *Nature Electronics*, 4(1):2–4.

Celes, C., Silva, F. A., Boukerche, A., d. C. Andrade, R. M., e Loureiro, A. A. F. (2017). Improving vanet simulation with calibrated vehicular mobility traces. *IEEE Transactions on Mobile Computing*, 16(12):3376–3389.

Chen, G., Viana, A. C., e Sarraute, C. (2017). Towards an adaptive completion of sparse call detail records for mobility analysis. In *2017 IEEE international conference on pervasive computing and communications workshops (PerCom workshops)*, pages 302–305. IEEE.

Cho, E., Myers, S. A., e Leskovec, J. (2011). Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1082–1090.

de Mattos, E. P., Domingues, A. C., e Loureiro, A. A. (2019). Give me two points and i'll tell you who you are. In *2019 IEEE Intelligent Vehicles Symposium (IV)*, pages 1081–1087. IEEE.

de Melo, P. O. V., Viana, A. C., Fiore, M., Jaffrès-Runser, K., Le Mouël, F., Loureiro, A. A., Addepalli, L., e Guangshuo, C. (2015). Recast: Telling apart social and random relationships in dynamic networks. *Performance Evaluation*, 87:19–36.

Domingues, A. C., de Souza Santana, H., Silva, F. A., de Melo, P. O. V., e Loureiro, A. A. (2022). Socialroute: A low-cost opportunistic routing strategy based on social contacts. *Ad Hoc Networks*, 135:102949.

Duckham, M. e Kulik, L. (2005a). A formal model of obfuscation and negotiation for location privacy. In *International conference on pervasive computing*, pages 152–170. Springer.

Duckham, M. e Kulik, L. (2005b). Simulation of obfuscation and negotiation for location privacy. In *International conference on spatial information theory*, pages 31–48. Springer.

Ekman, F., Keränen, A., Karvo, J., e Ott, J. (2008). Working day movement model. In *Proceedings of the 1st ACM SIGMOBILE workshop on Mobility models*, pages 33–40.

Finkel, R. A. e Bentley, J. L. (1974). Quad trees a data structure for retrieval on composite keys. *Acta informatica*, 4(1):1–9.

Firestone, S. M., Ward, M. P., Christley, R. M., e Dhand, N. K. (2011). The importance of location in contact networks: Describing early epidemic spread using spatial social network analysis. *Preventive Veterinary Medicine*, 102(3):185 – 195. Special Issue: GEOVET 2010.

González, M. C., Hidalgo, C. A., e Barabási, A.-L. (2008). Understanding individual human mobility patterns. *Nature*, 453(7196):779–782.

Gu, Y., Yao, Y., Liu, W., e Song, J. (2016). We know where you are: Home location identification in location-based social networks. In *2016 25th International Conference on Computer Communication and Networks (ICCCN)*, pages 1–9. IEEE.

Guttman, A. (1984). R-trees: A dynamic index structure for spatial searching. In *Proceedings of the 1984 ACM SIGMOD international conference on Management of data*, pages 47–57.

Hess, A., Hummel, K. A., Gansterer, W. N., e Haring, G. (2015). Data-driven human mobility modeling: a survey and engineering guidance for mobile networking. *ACM Computing Surveys (CSUR)*, 48(3):1–39.

Hoteit, S., Chen, G., Viana, A., e Fiore, M. (2016). Filling the gaps: On the completion of sparse call detail records for mobility analysis. In *Proceedings of the Eleventh ACM Workshop on Challenged Networks*, pages 45–50. ACM.

Hung, C.-C., Chang, C.-W., e Peng, W.-C. (2009). Mining trajectory profiles for discovering user communities. In *Proceedings of the 2009 International Workshop on Location Based Social Networks*, pages 1–8.

Ingole, P. e Nichat, M. M. K. (2013). Landmark based shortest path detection by using dijkestra algorithm and haversine formula. *International Journal of Engineering Research and Applications (IJERA)*, 3(3):162–165.

Johnson, G. T. e Watson, I. D. (1984). The determination of view-factors in urban canyons. *Journal of Climate and Applied Meteorology*, 23(2):329–335.

Jurdak, R., Zhao, K., Liu, J., AbouJaoude, M., Cameron, M., e Newth, D. (2015). Understanding human mobility from twitter. *PloS one*, 10(7):e0131469–e0131469.

Kang, J. H., Welbourne, W., Stewart, B., e Borriello, G. (2004). Extracting places from traces of locations. In *Proceedings of the 2nd ACM international workshop on Wireless mobile applications and services on WLAN hotspots*, pages 110–118. ACM.

Kosta, S., Mei, A., e Stefa, J. (2012). Large-scale synthetic social mobile networks with swim. *IEEE Transactions on Mobile Computing*, 13(1):116–129.

Krumm, J. (2009). A survey of computational location privacy. *Personal and Ubiquitous Computing*, 13(6):391–399.

Lisboa Filho, J. e Iochpe, C. (2001). Modelagem de bancos de dados geográficos. In *Apostila do XX Congresso Brasileiro de Cartografia, Porto Alegre*.

Maouche, M., Mokhtar, S. B., e Bouchenak, S. (2017). Ap-attack: a novel user reidentification attack on mobility datasets. In *Proceedings of the 14th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, pages 48–57. ACM.

Marques-Neto, H. T., Xavier, F. H., Xavier, W. Z., Malab, C. H. S., Ziviani, A., Silveira, L. M., e Almeida, J. M. (2018). Understanding human mobility and workload dynamics due to different large-scale events using mobile phone data. *Journal of Network and Systems Management*, 26(4):1079–1100.

Morales, A. J., Vavilala, V., Benito, R. M., e Bar-Yam, Y. (2017). Global patterns of synchronization in human communications. *Journal of the Royal Society Interface*, 14(128):20161048.

Morton, G. M. (1966). A computer oriented geodetic data base and a new technique in file sequencing.

Motlagh, N. H., Taleb, T., e Arouk, O. (2016). Low-altitude unmanned aerial vehicles-based internet of things services: Comprehensive survey and future perspectives. *IEEE Internet of Things Journal*, 3(6):899–922.

Naboulsi, D., Fiore, M., Ribot, S., e Stanica, R. (2016). Large-scale mobile traffic analysis: a survey. *IEEE Communications Surveys & Tutorials*, 18(1):124–161.

Newson, P. e Krumm, J. (2009). Hidden markov map matching through noise and sparseness. In *Proceedings of the 17th ACM SIGSPATIAL international conference on advances in geographic information systems*, pages 336–343.

Pappalardo, L., Simini, F., Rinzivillo, S., Pedreschi, D., Giannotti, F., e Barabási, A.-L. (2015). Returners and explorers dichotomy in human mobility. *Nature communications*, 6(1):8166.

Rettore, P. H., Santos, B. P., Lopes, R. R. F., Maia, G., Villas, L. A., e Loureiro, A. A. (2020). Road data enrichment framework based on heterogeneous data fusion for its. *IEEE Transactions on Intelligent Transportation Systems*, 21(4):1751–1766.

Sakai, T., Tamura, K., e Kitakami, H. (2014). Extracting attractive local-area topics in georeferenced documents using a new density-based spatial clustering algorithm. *IAENG International Journal of Computer Science*, 41(3):185–192.

Silva, F. A., Celes, C., Boukerche, A., Ruiz, L. B., e Loureiro, A. A. (2015). Filling the gaps of vehicular mobility traces. In *Proceedings of the 18th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, pages 47–54.

Tran, K. A., Barbeau, S. J., e Labrador, M. A. (2013). Automatic identification of points of interest in global navigation satellite system data: A spatial temporal approach. In *Proceedings of the 4th ACM SIGSPATIAL international workshop on geostreaming*, pages 33–42.

Uber (2015). *H3: A hexagonal hierarchical geospatial indexing system.*

Wang, Q. e Taylor, J. E. (2014). Quantifying human mobility perturbation and resilience in hurricane sandy. *PLoS one*, 9(11).

Zheng, Y., Capra, L., Wolfson, O., e Yang, H. (2014). Urban computing: concepts, methodologies, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(3):1–55.

Sobre os autores

Augusto C.S.A. Domingues é bacharel (2016) em Ciência da Computação pela Universidade Federal de Viçosa e mestre (2018) em Ciência da Computação pela Universidade Federal de Minas Gerais. Atualmente, é doutorando em Ciência da Computação pela UFMG, onde realiza pesquisas na grande área de Redes de Computadores, com ênfase em Computação Ubíqua, Computação Urbana, Redes Móveis e Mobilidade Humana. Durante o mestrado, trabalhou com pesquisas relacionadas a caracterização da escolha de rotas em *traces* de mobilidade veicular, propondo um algoritmo para gerar trajetórias de grão fino, enriquecendo dados existentes na literatura e permitindo novas oportunidades de pesquisa na área. Também trabalhou com a caracterização de *traces* de mobilidade em geral, propondo uma ferramenta que permite a extração de um conjunto de métricas sociais, espaciais e temporais. Como resultado, publicou trabalhos em diversas conferências nacionais e internacionais e em periódicos. No doutorado, além dos tópicos já descritos acima, também realiza pesquisas relacionadas a mecanismos de proteção de privacidade de localização de usuários em redes.

Fabrício A. Silva é doutor (2015), mestre (2006) e bacharel (2004) em Ciência da Computação pela UFMG. Durante a graduação e mestrado, trabalhou em projetos de pesquisa na área de redes de sensores sem fio. Entre 2006 e 2010, trabalhou em uma Startup desenvolvendo projetos inovadores na área de processamento de linguagem natural e aprendizagem de máquina. Desde 2010 é professor na UFV-Campus Florestal. Em 2015, obteve o grau de doutor, tendo desenvolvido sua tese na área de redes veiculares. Durante o doutorado, atuou como pesquisador visitante na Universidade de Ottawa, Canadá, sob a orientação do professor Azzedine Boukerche. Sua tese foi escolhida como a segunda melhor do Brasil no Concurso de Teses e Dissertações da Sociedade Brasileira de Computação. Atualmente, é Bolsista de Produtividade em Pesquisa 2 e tem trabalhado com pesquisa nas áreas de sistemas distribuídos e computação ubíqua, principalmente no estudo e caracterização de grandes volumes de dados desses sistemas.

Antonio A. F. Loureiro possui graduação em Ciência da Computação pela Universidade Federal de Minas Gerais (1983), mestrado em Ciência da Computação pela Universidade Federal de Minas Gerais (1987) e doutorado em Ciência da Computação pela University of British Columbia, Canadá (1995). Atualmente é Professor Titular do Departamento de Ciência da Computação da Universidade Federal de Minas Gerais. Tem experiência na área de Ciência da Computação, com ênfase em sistemas distribuídos, atuando principalmente nos seguintes temas: algoritmos distribuídos, computação móvel/ubíqua, comunicação sem fio, gerenciamento de redes, redes de computadores, redes de sensores sem fio. Atualmente, é Bolsista de Produtividade em Pesquisa 1A.