International Journal of Geo-Information

MDPI

Article

Development of an Algorithm to Evaluate the Quality of Geolocated Addresses in Urban Areas

Rafael Sierra Requena [1], José Carlos Martínez-Llario [2], Edgar Lorenzo-Sáez [3,*] and Eloína Coll-Aliaga [3]

[1] Regional Office of Directorate General for Cadastre (Ministry of Finance, Spain), Roger de Lauria nº 26, 46002 Valencia, Spain; rafael.sierra@catastro.hacienda.gob.es or rasiere@upvnet.upv.es
[2] Department of Cartographic Engineering, Geodesy and Photogrammetry, Universitat Politècnica de València, 46003 Valencia, Spain; jomarlla@upv.es
[3] ITACA Research Institute, Universitat Politècnica de València, Camí de Vera s/n, 46022 Valencia, Spain; ecoll@cgf.upv.es
* Correspondence: edlosae@upv.es

Abstract: The spatial and semantic data of geographic addresses are extremely important for citizens, governments, and companies. The addresses can georeference environmental, economic, security, health, and demographic parameters in urban areas. Additionally, address components can be used by users to locate any point of interest (POI) with location-based systems (LBSs). For this reason, errors in address data can affect the geographic location of events, map representations, and spatial analyses. Thus, this paper presents the development of an algorithm for evaluating the quality of semantic and geographic information in any geospatial address dataset. The reference datasets are accessible using open data platforms or spatial data infrastructure (SDI) and volunteered geographic information (VGI), and both have been compared with commercial datasets using geocoding web services. Address quality analysis was developed using several open-source data science code libraries combined with spatial databases and geographic information systems. In addition, the quality of geographic addresses was evaluated by carrying out normalized tests in accordance with International Geospatial Standards (ISO 19157). Finally, this methodology assesses the quality of authorized and VGI address datasets that can be used for geocoding any relevant information in specific urban areas.

Keywords: addresses; spatial data quality; geocoding; open data; volunteered geographic information

Citation: Sierra Requena, R.; Martínez-Llario, J.C.; Lorenzo-Sáez, E.; Coll-Aliaga, E. Development of an Algorithm to Evaluate the Quality of Geolocated Addresses in Urban Areas. *ISPRS Int. J. Geo-Inf.* **2023**, *12*, 407. https://doi.org/10.3390/ijgi12100407

Academic Editor: Wolfgang Kainz and Wei Huang

Received: 26 June 2023
Revised: 14 September 2023
Accepted: 24 September 2023
Published: 4 October 2023

1. Introduction

The geographic component is a strategic attribute in digital data with respect to policy and operational planning [1]. The geographic component of addresses is often linked to multiple data types, such as building, transport, population, marketing, delivery, safety, and health information data [2,3]. There is consensus in the European statistical community on the best methodology for linking statistics to a location using point-based geocoding infrastructure [4]. The address data components on the Internet network are present on most web pages and almost 20% of users' queries are on web browsers [5]. In fact, network users can navigate to other well-known places or shared locations using sensors on mobile devices [6] with location-based system (LBS) applications [7]. For these reasons, addresses are a fundamental component of urban management, smart cities, and urban spatial analytics [8,9].

Addresses are a type of geographic feature that can be collected directly by users or professionals. This dataset can be stored in the spatial databases of urban cartography before use in any geographic information system (GIS). In fact, georeferencing dwellings using postal address references has historically been important for the scientific

community [10] due to the linkage between records of the geographic component and other types of information. For more than 20 years, numerous scientific papers on linking data with geocoded addresses to exploit the geographic component in other disciplines, such as epidemiology [11,12], environment, demography [13], business [14], crimes [15], emergencies [16], and security [17], have been published. In fact, there are numerous academic, industrial, or commercial studies on the positional accuracy [2,18–24] or semantic quality of the addresses obtained using geocoders [25,26].

Authoritative addresses can be downloaded from government web repositories, web services, and open data web platforms (OpenAdresses [27]). In addition, standard addresses are a key element for delivering policies at national and international levels [28] in support of the United Nations (UN) Sustainable Development Goals (SDGs). Currently, data re-usability has an increasingly important place on the agendas of many open data and open government data initiatives [29]. In fact, the European Union is implementing spatial data infrastructure (SDI) [30] to provide harmonized datasets through web service operations (WFS or Atom) and metadata according to the INSPIRE European Directive [31] and international standards from the Open Geospatial Consortium (OGC [32]) and ISO/TC 211 [33]. Nevertheless, many countries lack a government-maintained address database (demonstrated in section *2.2 Authoritative Datasets from INSPIRE Addresses on European SDI*). Furthermore, addresses can be made accessible as volunteer geographic information (VGI) that is georeferenced by the users [34,35] of crowdsourcing map platforms, such as OpenStreetMap (OSM [36]).

On the other hand, there are many companies that provide commercial address datasets using geocoding web services (Google [37], Microsoft [38], Here [39], etc.). The geocoding process has two main ways of querying address data: 1. the direct method, which provides geographic coordinates (latitude and longitude) as a result when there is a match with the text of the requested address of the web service, and 2. the reverse geocoding method, which is the process of extracting a text address (street and number) by providing global position coordinates.

Despite the large volume of address data produced by governments, it is sometimes necessary to improve and update the authoritative geographic information of the addresses using data produced by volunteers [40] or data from commercial geocoding services [19]. However, there are some related studies about the quality of authoritative and crowdsourced geospatial information [41,42–44]. Overall, these studies analyzed residential address datasets compared with commercial addresses obtained from global geocoding web services [19,20,25,26].

Nevertheless, there is no research about automated algorithms that can be used to check the quality and reliability of large authoritative or volunteered address datasets [45] from several countries that are stored on different platforms (public repositories or SDI web services) against commercial addresses using both methods of geocoding. The methodology developed focuses on residential addresses that are used as spatial and semantic references in LBSs to locate buildings, dwellings, businesses, or recreation and leisure venues. The results show statistics about the spatial and semantical quality components of addresses of authoritative, commercial, or crowdsourced datasets. Thus, this paper aims to develop an algorithm to determine the quality of address data from urban areas and several datasets using geocoding web services.

The algorithm extracts well-known random samples of authoritative and VGI addresses as reference data in the main urban areas of Europe; these were automatically requested from commercial geocoding web services. The responses are stored on spatial databases in order to analyze the quality elements (positional and thematic accuracy, completeness, or logical consistency) according to ISO 19157:2013 [46] ("International Standard Geographic information—Data quality," 2013). The spatial and semantic algorithms are developed with the Python language, using libraries for data mining [47,48], machine learning [8], and big data [49] management in cloud computing systems in order to evaluate different address-matching methods [50].

Therefore, the algorithm contributes to checking the quality components of commercial, crowdsourced, and authoritative addresses within the user-defined geographical urban area. The main goal is to automatically evaluate whether the quality of voluntary or commercial geographic information relative to postal addresses in any given area is good enough to improve authoritative address datasets. However, this tool can be also used to design any future geolocated urban data in a GIS using several geocoding address datasets. Finally, the results allow for enhancing strategic address datasets for the benefit of public administration, companies, and citizens.

2. Data

Analyzing the components of an address is an essential preliminary step prior to designing quality control. The developed methodology needs to be able to automatically compare several datasets with different data schema, texts, and positions with respect to whether they are commercial, voluntary, or authorized addresses, and they must necessarily represent the same concepts and similar values.

There are two different techniques for determining the positional component of addresses in geocoding web services: street (linear network analysis) or rooftop (points). Some publications analyze differences in the quality of georeferenced geocoding techniques [19,51], but in this study, the positional origin of addresses is not evaluated because the algorithm does not discriminate between geocoding web service responses. Furthermore, the street method is mainly applied in the USA, and it is important to have a well-defined transport network in order to correctly interpolate the input number. On the other hand, the point method is mainly applied to addresses that are authorized in European countries (the UK, France, Spain, etc.), although, depending on the type of land use, addresses can be georeferenced on a roof, the centroid of a parcel, the entrance to a house, or on a public road in front of a building.

The semantic component of address data basically consists of text (usually the street name), building or door numbers, postal codes, and administrative units. In some cases, addresses provided by government agencies may add other identifiers that may be linked to cadastral parcels, statistical units, demographic censuses, and other types of urban surveys [51].

In addition, the components of semantic addresses for administrative units and settlements must particularly follow a complete common structure and order relative to a hierarchy (settlement or district, municipality, province, state, region, and country). Thus, geocoders can distinguish similar addresses from different areas in the same region or municipality. Some countries use a postal, census, or zip code, which relates to a specific geographical area. In fact, the geographical boundaries of administrative units or urban areas are important in this algorithm in order to extract reference address samples for quality analysis. The datasets used to test the algorithm in European countries are detailed in the following sections.

2.1. Authoritative Dataset from the OpenAddresses Web Platform

The authoritative address datasets must be free and open and obtained from the governments' web portals because they are the most reliable reference source for text address components [52].

The OpenAddresses web platform gathers this information and normalizes it in order to distribute the information in text files with plain format, comma-separated values (CSVs) or geographic formats (GEOJSON and shapefiles) with the following structure: lon, lat, number, street, unit, city, district, region, postcode, id, and hash. Moreover, this open repository has one folder per country with several address files in CSV text format and metadata files in JavaScript Object Notation (JSON) text format.

In this study, 144 datasets from OpenAddresses in the European area were used, with addresses from 21 countries: 14 country-wide, 16 regional, and 114 local datasets. The datasets were inserted into a unique spatial database (PostgreSQL with PostGIS extension)

as point geometries for better quality control algorithm performance. Developed algorithm (developed in section *3.1 Development of a Quality Control Algorithm for Spatial Address Data*) shows the geographical extent of OpenAddresses datasets, which includes the continental area of Europe, with the exception of Greece, Bulgaria, and Hungary.

Datasets need a complete revision of quality components with respect to completeness, logical consistency, and the harmonization of different data schemas across countries. Thus, an initial analysis of addresses was carried out to ensure the initial quality with respect to these measures. Some spatial queries were designed to analyze almost 100 million addresses in the initial 80 European datasets loaded into the database, with almost 1% being duplicated in terms of attributes (lon, lat, street, and house number) and 0.14% being duplicated in terms of attributes (street, house number, and city). Only 19 datasets did not contain some of the initial errors. There are also a few datasets with completeness problems with respect to semantic attributes, with most being caused by data import problems.

The OpenAddresses web portal offers information about products but could be more useful for obtaining metadata product links of authoritative address datasets.

2.2. Authoritative Datasets from INSPIRE Addresses on European SDI

The INSPIRE European Directive ensures the interoperability of European authoritative datasets using common well-known data schemas, metadata, and network service implementation that are accessible without restrictions from a unique geoportal [53].

The INSPIRE Addresses schema [54] is linked to other geographic feature schemas such as buildings, geographical names, administrative units, and transport networks. Address data depend on the other themes in order to be completely useful.

The European geoportal in SDI offers 60 downloadable address datasets, 217 metadata records, and 107 viewable datasets in its catalog. The address download service was implemented in 21 out of 32 countries, but only 18 countries have data in the INSPIRE common schema.

The results from downloadable web services (Table 1) with addresses are as follows: 17 implemented the Web Feature Service and 15 implemented Atom Massive Download links (standard OpenSearch using feeds with the XML language). These data were collected by different public institutions, such as national or regional maps, agencies, cadasters, municipalities, land registries, and statistical offices (Table 1).

Most of these addresses' datasets are identical to those in the OpenAddresses data platform because they have the same government source. However, there are fewer authoritative address datasets in the INSPIRE schema than those provided by OpenAddresses because some countries do not distribute this formatted data in the European SDI.

Table 1. Addresses web services per country and institution in charge of data management.

Authorative Addresses from UE Published by Public Organisms				**INSPIRE Web Services**			
Geographical Area	**Territorial Scope**	**Organism**	**Country**	**WFS**	**ATOM**	**Restrictions**	**Check**
Spain	National	Cadastre	Spain	1	1	no	yes
Navarra	Regional	Regional Map Agency	Spain	1	0	no	yes
Guipuzcoa	Province	Cadastre - Statistical	Spain	1	1	no	yes
Catalunya	Regional	Regional Map Agency	Spain	0	1	no	yes
Portugal	National	Statistical	Portugal	0	1	no	yes
Azores Island	Local	Municipalities	Portugal	1	0	no	yes
Bruxelles	Regional	Regional Map Agency	Belgium	0	1	no	yes
Wallonie	Regional	Regional Map Agency	Belgium	0	1	no	yes
Denmark	National	National Map Agency	Denmark	0	1	no	yes

Iceland	National	Land Registry	Iceland	1	0	no	no
Norway	National	National Map Agency	Norway	1	1	no	yes
Finland	National	National Map Agency	Finland	1	0	no	yes
Estonia	National	National Map Agency	Estonia	1	0	no	yes
Letonia	National	Land Registry	Letonia	1	0	no	yes
Lituania	National	Land Registry	Lituania	0	1	no	yes
Poland	National	National Map Agency	Poland	1	0	no	yes
Berlin	Regional	Regional Statistics	Germany	0	1	no	yes
Mecklemburgo-Pomerania	Regional	Cadaste	Germany	1	0	no	yes
Thueringen	Regional	Regional Map Agency	Germany	1	1	no	no
Hamburg	Regional	Regional Map Agency	Germany	1	0	no	yes
Sachsen	Regional	Regional Map Agency	Germany	0	1	no	yes
Nordrhein-Westfalen	Regional	Regional Map Agency	Germany	1	0	no	yes
Czech Republic	National	Cadaste - NMA	Chequia	1	0	no	no
Slovakia	National	National Map Agency	Slovakia	1	0	no	yes
Austria	National	National Map Agency	Austria	0	1	no	yes
Slovenia	National	National Map Agency	Slovenia	1	1	no	yes
Romania	National	Cadaste	Romania	0	1	no	no

2.3. Crowdsourcing Dataset from the OpenStreetMap Project

This address reference dataset was collected by volunteers of the OpenStreetMap (OSM) crowdsourcing project. Basically, this VGI comes from a collaborative map of streets, points of interest, and other geographic information principally produced by users.

These crowdsourced addresses can be collected by users on site and at the street level using a mobile device with a global navigation satellite system (GNSS) sensor or by carrying out digitalization using a map platform such as JOSM. OSM address data, such as house numbers, are collected by users and can be geometrically associated ("node") with any geographic feature and establish relationships with another element [55]. The semantic component of the address can then be linked to the geographic position in front of an entrance or at the top of a building, and some studies analyze the quality relation of crowdsourced information. This VGI on OSM frequently has poor information with respect to metadata in datasets and geocoding responses, although it could make a relevant contribution to evaluating address quality.

In this study, 35 million OSM addresses were downloaded using queries in the osmosis application. Principally, building nodes contain house numbers, but there are some point-of-interest tags with addresses (1.8% amenity; 1.4% shops; 0.1% tourism; 0.02% public transport). In addition, a preliminary quality analysis of OSM addresses identified serious problems with respect to completeness (17%—street without a name or code; 2%—house number without text) and logical consistency (3%—duplicated addresses). Nevertheless, the algorithm using the reverse method will use the spatial component of OSM data to locate nearby addresses via commercial geocoding web services.

On the other hand, OSM's geocoding web service, Nominatim, can geocode addresses and place names using direct and inverse methods. Nominatim has been used to develop algorithms with direct methods in order to obtain 10 semantic address components in the GeoJSON geographic format (house number, road, hamlet, town, village, city, state district, state, postcode, and country). The returned addresses are parsed and compared with authoritative addresses. This can be useful for public administrations in order to improve upkeep. In fact, some parts of the OSM database were massively imported from authoritative sources such as the TIGER database of the U.S. Census Bureau [56].

2.4. Commercial Datasets from Geocoding Web Services

These datasets were obtained in response to requests made to geocoding web services. Hence, they have no availability with respect to examining the complete features of address data in a specific urban area to discover their density, geographic coverage, or data quality. However, these geocoding services are quite reliable as they are consulted by two of the world's leading web service, LBS, and software providers, i.e., Google and Microsoft (Bing), and by long-standing address data providers such as Here (Nokia, formerly Navteq).

Commercial geocoding companies were also chosen following the criteria of intensive use or demand from users. Moreover, we queried some of the literature that reviewed quality analysis using some geocoders [57,58] in order to choose free geocoders that have the most responses and matches and the highest accuracy. However, it might be interesting to compare their data against other LBS address data with respect to accurate commercial geocoding systems because they usually use multiple sources of reference datasets [59].

Basically, all chosen geocoders have REST architecture and are limited in use because they must be paid for and return data in XML and JSON formats. In addition, in all of them, the number of possible responses per request can be limited. Currently, the algorithm only works when using a single option as a response from the geocoder in JSON format, which is similar to plain texts with JavaScript codes chosen for their simplicity in parsing the data contained in lists within the Python library. These data are not completely validated by a data schema, such as the XML INSPIRE schema, but this is not necessary as the algorithm is developed to automate structure and check data responses:

- Google: The Geocoding API is requested by the algorithm. The response is a JSON list with seven semantic components (street number, route, locality, two administrative areas, country, and postal code), locations (latitude and longitude), and geocoding type (rooftop).
- Bing (Microsoft): The REST location service by address provides six semantic components (address line, two administrative districts, country region, locality, and postal code), locations (latitude and longitude), and usage type (route).
- Here: The Geocoding & Search API returns nine semantic components (street, house number, city, postal code, district, subdistrict, county, state, and country), positions (latitude and longitude), and house number type (PA or building, interpolated along street).

3. Methods

3.1. Development of a Quality Control Algorithm for Spatial Address Data

The algorithm developed uses the extract, transform, and load (ETL) process, where the reference dataset was applied as the input and authoritative addresses for the direct geocoding method, and the VGI dataset was applied as the input for reverse geocoding (Figure 1).

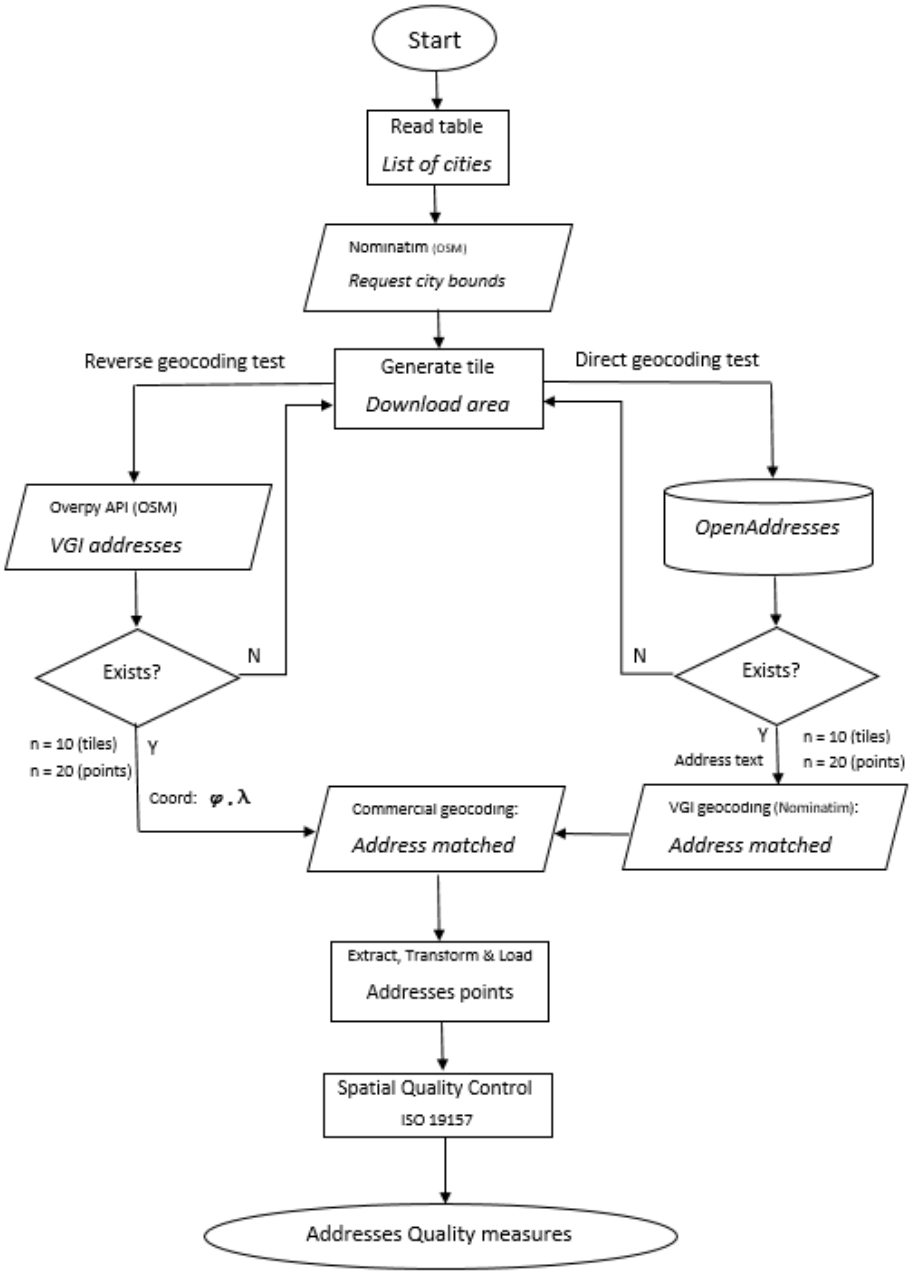

Figure 1. Algorithm workflow with respect to the quality control of spatial addresess data from extraction to quality measures result.

The algorithm is configurable in that the user chooses the number of sample portions and addresses to extract from each area. To obtain these parameters, the population of the dataset must first be calculated so that the distribution is optimal for producing statistics. Therefore, the address data reference for the quality assessment will be extracted using the input parameter settings and random samples. The algorithm uses OSM as a data reference for the inverse geocoding method and OpenAddresses or INSPIRE for the direct method (Figure 1).

The first part of the algorithm (Figure 1) can import a file with a list of urban places or geographical names to verify the quality of address data within an urban area. In particular, upon the pilot application, a list with some cities in the European Union was drawn up in order to verify the specific quality of national, regional, and local addresses that are available on OpenAddresses.

Figure 1 shows a diagram of the algorithm that implements both geocoding methods before measuring quality parameters for address datasets. First, the quality of the commercial and voluntary addresses is checked by contrasting them against official addresses using a geocoding method. Once they are verified to be reliable, the existence or not of discrepancies between the unofficial and the official ones are checked with the other geocoding method. Therefore, the fact that the algorithm has 2 branches means that the functionality checks the quality and identifies and corrects errors (Figure 1).

The developed algorithm can automatically request samples of the reference addresses from geocoding web services, which allows for identifying the accurate dimensions of urban areas in order to extract authoritative addresses from geographical sample zones. The algorithm provisionally implements the OSM geocoding service Nominatim for these tasks because it returns the coordinates of a bounding box for the requested city. Sometimes, there are insufficient automatically extracted address samples with respect to any tile. In this case, the algorithm generates another tile in order to request addresses from web services or spatial databases, as shown in Figure 1.

This sometimes provides overly large boundaries that correspond to large evolving metropolitan areas, which include nearby rural areas or dormitory towns. To solve this problem, several urban settlement datasets were examined to achieve a more suitable geometric area. The Global Human Settlement layer made by the European project Copernicus was considered as a suitable dataset for these tasks. It provides delimitations of urban areas in images using remote sensing techniques. Nevertheless, the spatial resolution and the lack of a direct download web service have made searching for other options necessary. The Corine Land Cover dataset [60] and European Urban Atlas [61] products were also considered, but their use could complicate the methodology because they require prior hard processing tasks to extract the geographical extent. Finally, the algorithm automatically obtains randomly sampled mosaic portions or tiles with a dimension of 70 Ha (1 km × 0.7 km) within the determined urban areas' geographical extent for each city returned by the Nominatim web service. Afterward, several addresses from reference datasets inside this area were selected as samples for geocoding web service requests. Figure 2 shows the address sample areas.

Figure 2. The samples (purple dots) inside an area of extraction (tile, red) and the city boundaries (blue) are provided automatically by the Nominatim web service.

The open public administration address datasets were downloaded from the OpenAddresses repository on the "Github" platform to evaluate their spatial and semantic quality [62] on direct geocoding tests using this algorithm. Although the European SDI is still not completed by all country members, the algorithm can incorporate its reference datasets, extracting spatial databases from previous downloads using the Atom web

service or directly querying the web feature download service inside the bounding box (tile) and count (points numbers) parameters.

The crowdsourced addresses were downloaded using the "osmosis" application, which extracts selected data from the OSM spatial database to reverse the geocoding test method in this algorithm. Therefore, a script loads addresses from OpenAddresses and OSM into the PostGIS spatial database for evaluation processes using the algorithm. The geographical distribution from the loaded reference datasets is shown in Figure 3.

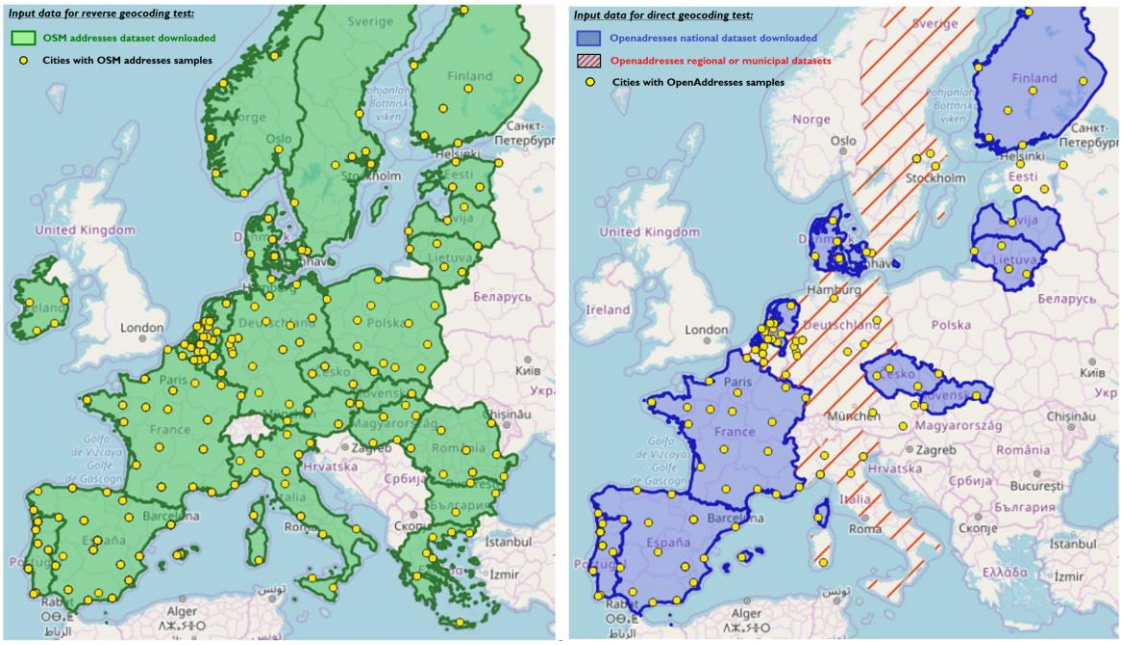

Figure 3. (**Left**): Geographical extent of loaded OSM addresses and extracted samples from European cities (yellow dots) that are compared against commercial geocoders. Green area represents countries analised (**Right**): Reference data source download from OpenAddresses showing samples extracted from cities (yellow dots), with blue representing the complete dataset relative to the countries and partial data (regional or local). Diagonal red lines: Countries with partial address data.

Thus, the developed algorithm is used to perform a quality analysis check of both reference datasets by comparing them with the responses obtained from crowdsourced and commercial geocoding web services. In this case, the use of different datasets as inputs for the algorithm has some advantages, as the results can show different quality aspects of both address data sources compared to the responses from the same geocoding web services. Then, the positional or semantical component is requested from three geocoding web services provided by multinational corporations in order to obtain commercial address datasets. These geocoding web service requests were configured using the text components of addresses in UTF-8 encoding. The language of the addresses can be the official national language or English depending on the geocoding web service and the reference dataset. All chosen geocoders have REST architecture and limitations of use since they must be paid for, with the exception of Nominatim, which is free. All web services return data in XML and JSON formats, and Nominatim also adds the GeoJSON geographic format.

The designed algorithm obtains a geocoding response in JSON format that is similar to plain text with JavaScript codes, and this was chosen due to its simplicity in parsing the data contained in lists within the Python library. These data are not completely validated by a data schema such as XML, but this is not necessary as the algorithm needs to perform these checks for quality control.

There are some code libraries in Python language for geocoding, such as Geocoder, but we chose to directly implement the process in Python, which allows control of the

answers, making it very easy to integrate this process with other processes. Once the response is transformed into the designed schema, it is inserted directly in the PostGIS spatial database as point geometries using the Python library psycopg2. Once the data are loaded, the final step is to carry out the quality control of the information and extract statistical results.

3.2. The Address Quality Control, Method, and Measures

3.2.1. Approaches to Spatial Quality Control according to ISO 19157

The quality of data can be assessed using a homogeneous method for any product, but quality can also be evaluated using a spatial or semantical criterion and information that describes the usage of products from a user's point of view [63].

The quality evaluation of datasets was performed according to ISO 19157 [46], which cites the following: "a data quality evaluation can be applied to dataset series, a dataset or a subset of data within a dataset, sharing common characteristics so that its quality can be evaluated". Thus, a set of authoritative address databases was used as a baseline, but it was also necessary to have good metadata (ISO 19115) on the products [64] that describe the main characteristics of the addresses including dataset data size, capture method, and resources, dates, interoperability, linkage, etc. The downside of this approach is that we have little knowledge—or none—of data origins, heterogeneous locations, theme characteristic system management, and data updating [63].

Since the local administration oversees maintaining the names and numbers of the roads, the city is the minimum unit of quality analysis used by the algorithm. Therefore, the quality analysis was performed relative to the city; then, the results obtained relative to the country were aggregated in order to obtain spatial data quality product parameters. ,mThe different implementations assess the reliability of the OSM address positions and text components [42] introduced by volunteers and check commercial address positions that introduce OpenAddresses semantic components.

The spatial data quality elements [65] given by ISO 19157 evaluate positional accuracy, thematic accuracy, completeness, logical consistency, and usage relative to different datasets. The algorithm determines the spatial quality values of the responses of geocoding web services compared with authoritative (direct method) or (inverse method) crowdsourced address datasets.

There are some sub-elements of quality that were not checked in this analysis because they were not representative of these datasets (i.e., topological consistency or classification correctness). The temporal element cannot be assessed yet, but it might not be suitable in random samples from different datasets. Additionally, some measures of quality are difficult to assess using the automated methodology because the analysis would require interpretation by a technical specialist who could evaluate errors in a GIS or the implementation of a machine learning algorithm [66,67].

3.2.2. Quality Measures for Completeness, Logical Consistency, and Usage

The repeatability of address data values relative to text attributes causes logical consistency problems and concrete commission errors, and these are controlled using the quality procedure. Development using Python checks address duplicates in the geocoding web services' address responses (spatial and text). In addition, there are measures designed to assess the completeness of separate spatial and semantic addresses, such as the street, house number, postal code, administrative units, etc. The lack of response from geocoding web services to a requested point generates an omission error. To prevent this, the algorithm does not allow more than one candidate answer to be received from each specific requested address point. In urban areas, there are rarely several position candidates relative to one address, and evaluating the programming code to determine whether an address point is a candidate can sometimes be difficult.

On the other hand, the algorithm checks if semantic address components are complete in geocoding responses when evaluating logically consistent quality parameters.

Finally, the usage quality element can be checked with the user experience acquired in this study using all datasets. In all cases, commercial and crowdsourced geocoding web services exhibit good performance, and the algorithm collects the time responses of geocoding web services.

3.2.3. Semantic Quality Measures for Thematic Accuracy

Thematic accuracy was mainly assessed using measurements related to matching the same semantic attributes in different datasets. Furthermore, the theme attributes (street, number, and administrative units) in geographic information addresses must be unique.

This approach required implementing a similarity measure control that takes two input values and returns a match rate. The result is a code value that represents the similarity of an address's text component [68]. The algorithm was developed using different Levenshtein implementations in Python packages, but there are many algorithms [69] that can be implemented and explored in future studies.

There is a table in the spatial database that stores the results of the Levenshtein similarity text algorithm for each address's text component. Then, these Levenshtein results in the database are aggregated and analyzed using user-configurable thresholds to evaluate the full-address matching procedure. Finally, the algorithm returns a zero value for similarity acceptance, and the mismatching error returns a value of 1. Table 2 illustrates this point with the following address example from OpenAddresses: "Route de saint-priest 2 Lyon, France". The method to classify the number of errors in every semantic address component can be observed in Table 3.

Table 2. Example of the method's application in assessing the number of text matching errors in address components between different requested geocoding web services. This example shows response for coordinates request (lat = 45.68112049; lon = 4,9491959), and distance on meters.

Pr.City_Test	tile	Num_Point	Geocode_Comp	Full_Address_api	Distance
Lyon, France	1	1	Google Places	2 Route de Saint-Priest, 69960 Corbas, France	4095.01
Lyon, France	1	1	Bing Maps	2 Route de Saint-Priest, 69960 Mions, France	16.857
Lyon, France	1	1	Here Technologies	2 Route de Lyon, 69800 Saint-Priest, France	2761.84
Lyon, France	1	1	Nominatim (OSM)	2, Route de Saint-Priest, Mions, Lyom, Métropole de Lyon, 11epart. Rhône, Auvergne-Rhône-Alpes, 69780, France	0.067

Table 3. Method to classify the number of errors in every semantic address component. The numeric values on addresses semantical attributes correspond to the match between geocoder responses: 1 = Nominatim-Here; 2 = Nominatim-Bing; 3 = Nominatim-Google; 4 = Nominatim-Here; 5 = Here-Bing; 6 = Here-Google; 7 = Bing-Google.

			Match Street							Match Housenumber							Match Postcode							Match City							Match Address						
City_Test	Tile	n_Point	1	2	3	4	5	6	total	1	2	3	4	4	6	Total	1	2	3	4	5	6	Total	1	2	3	4	5	6	Total	1	2	3	4	5	6	Total
Lyon, France	1	1	1	0	0	1	1	0	3	0	0	0	0	0	0	0	1	0	1	1	1	1	5	1	0	1	1	1	1	5	1	1	1	1	1	1	6

3.2.4. Quality Measures for Positional Accuracy

The evaluation of positional accuracy of geocoded locations can be performed using absolute global geodetic coordinates, latitude, and longitude in contrast to the geometric position of the address dataset's features.

Positional error is a two-dimensional vector with a north–south component and west–east component [70], and the algorithm uses the norm of this vector or the equivalent Euclidean distance, similar to geodetic distances within this spatial proximity range between the compared addresses [71]. Therefore, residual errors can be calculated using the geographic distance in meters and by carrying out SQL queries on PostGIS spatial databases.

Other authors [72] also applied measures that introduce a bearing parameter in degrees to improve the adjustment between geocoding web service coordinates and the reference address's coordinates. This orientation parameter with angular values is not currently implemented in the algorithm, although it is prepared for applications that choose the best candidate among several geocoding responses.

A set of lists in database tables with distances between positions (spatial errors) is introduced in a matrix object (a "dataframe" object in the "pandas" Python code library) to evaluate the positional accuracy. Finally, the algorithm extracts measures of positional accuracy between authoritative, crowdsourced, and commercial geocoders using direct and reverse methods and attempts to find a correlation pattern among all geocoding services.

This algorithm evaluates positional measures before checking the other quality measures in order to apply a previous filter that does not contain large outliers in the proposed addresses' spatial positions (Figure 4).

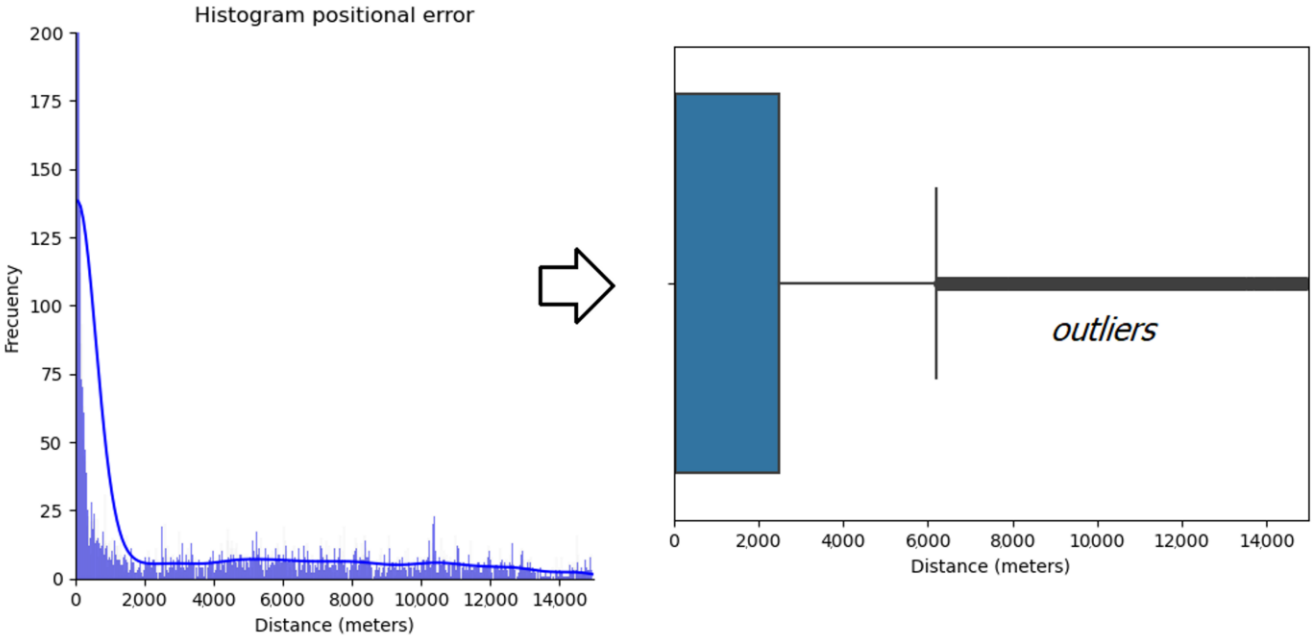

Figure 4. Graphics showing raw positional error data. Most errors are minimal; thus, they accumulate near the distance of zero, but there is a large queue of outliers depending on previous filtering.

The non-normal distribution of positional errors in spatial data has implications beyond spatial data accuracy standards since most error propagation techniques for spatial data are also based on an assumption of normality [21].

The designed algorithm automatically removes outliers using the interquartile range (IQR) and transforms positional data using least squares to reduce values relative to the skewness and kurtosis of spatial data distributions. The objective is to fit a Gaussian model in order to estimate the RMSE (3σ interval) for every dataset or sample. However, it is only an estimation because transformed data cannot be obtained using a normality test or Gaussian model (Figure 5).

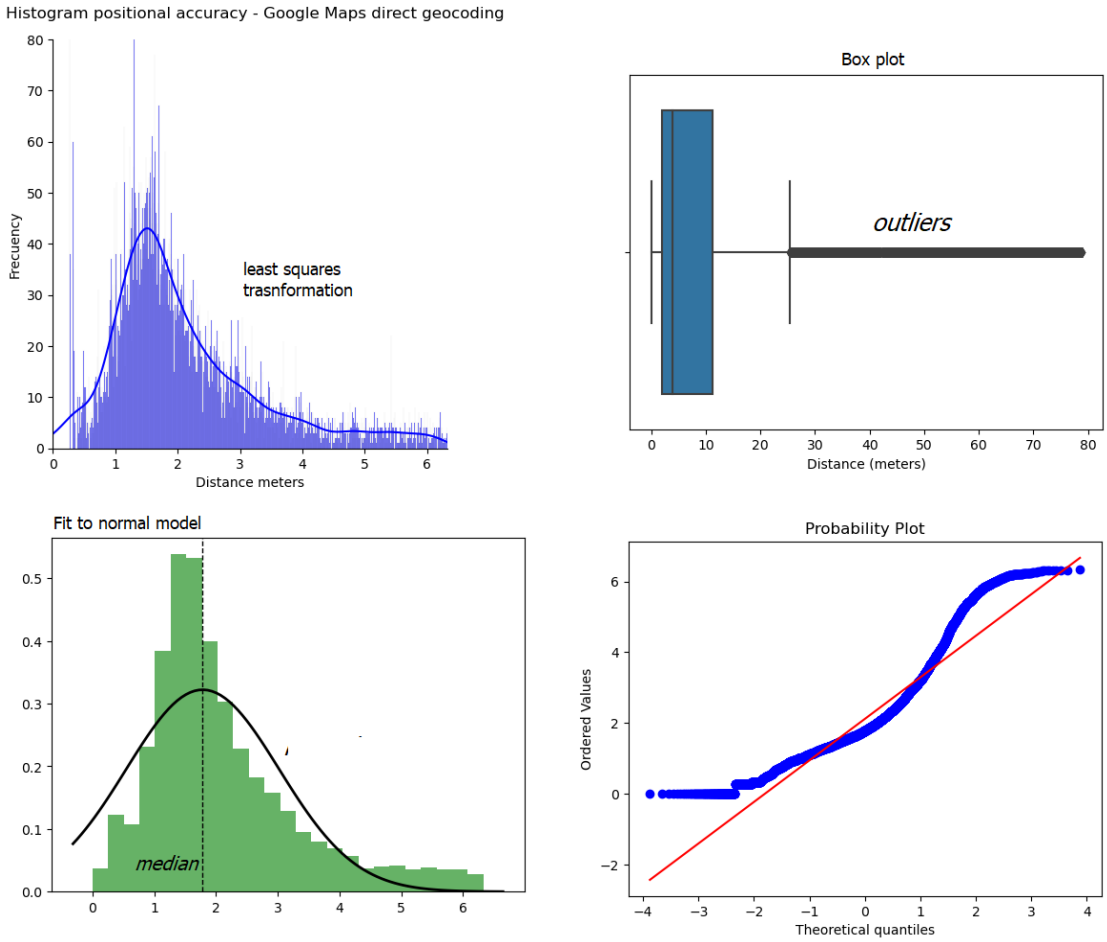

Figure 5. Graphics showing the positional error data sample with respect to the direct geocoding responses from Google Maps. The raw data shown in Figure 4 are transformed using least squares to approximate them in a Gaussian model, but the figure demonstrates the inability to accomplish a normal distribution model.

The positional error analysis in ISO 19157 and positional accuracy standards (NMAS, EMAS, and NSSDA) conform to a parametric statistical distribution function that uses normal or Gaussian models for the majority of situations.

However, some working documents in ISO 19157 [73] do not require fitting the positional error to an underlying parametric statistical distribution function in non-parametric models. In this case, the error distribution can be given by the observed data. Therefore, the statistics are based on the percentiles and proportions that are reported in the Results section. Moreover, the algorithm can extract descriptive statistics (median, proportions, and percentiles), removing outliers using spatial thresholds in order to compare results between the datasets.

4. Results

The application of the algorithm to the European region shows that the number of responses from geocoding web services was over 83% for the direct method and 74% for the reverse method. Spatial accuracy is one of the most important spatial quality elements, and the results confirm our initial expectations: Geocoders have similar spatial magnitudes within a range of 10 m. However, it is necessary to filter the responses of outliers using an estimated threshold: 45% in the direct method and 5% in the reverse method.

The case study was designed to test the algorithm, where both geocoding methods extract random address samples inside random geographic areas (tiles) in some urban zones of European cities. Then, if the algorithm does not find any addresses inside their

bounds, it continues to search for addresses inside the tiles for up to 10 tiles. Thus, the number of tiles is a good indicator of the available geographic reference address datasets in both OpenAddresses and OSM. Above all, the number of address samples in every city can demonstrate the availability of address data within this area (Figure 6).

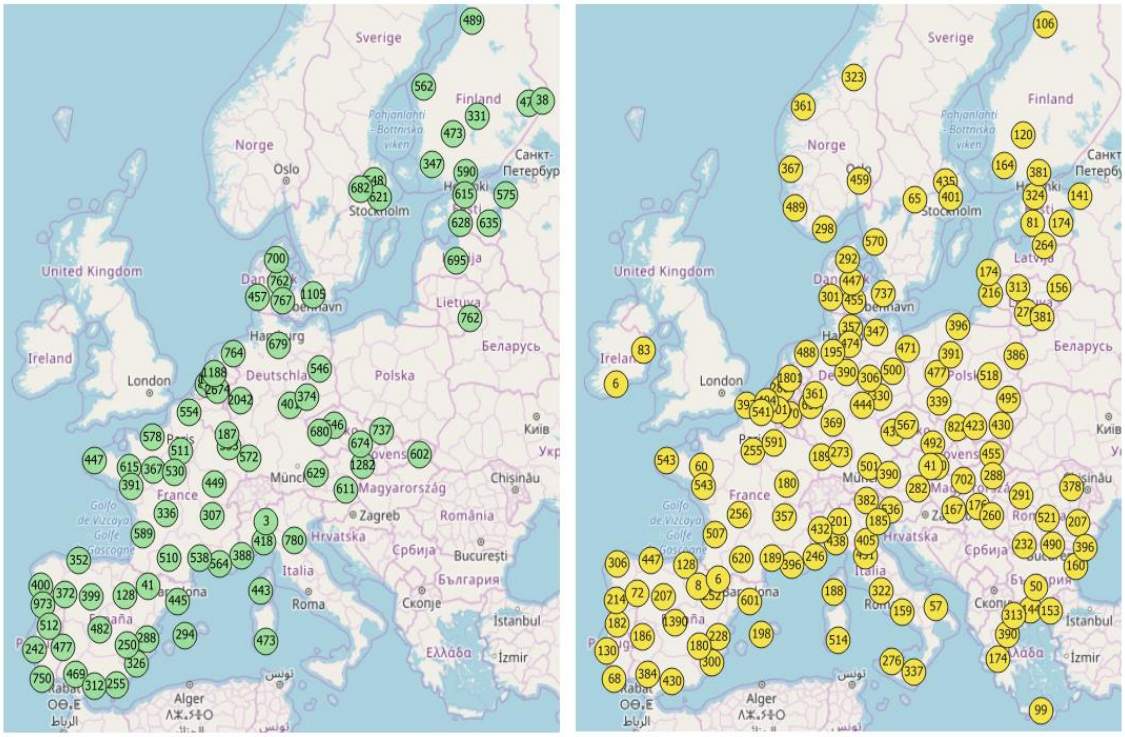

Figure 6. (**Left**): Spatial distribution of address points (green)extracted using the direct method. (**Right**): Spatial distribution of address points (yellow) extracted using the reverse method.

The first obtained result is the influence of semantic quality on the text components in address datasets. If some components are incomplete (street name, number, city, region, postcode, etc.), the match rate will be low with respect to any geocoding method. If the web service finds a matching address but its associated spatial coordinates are far from the real ones, then an outlier is introduced in the spatial data quality sample (Figure 7).

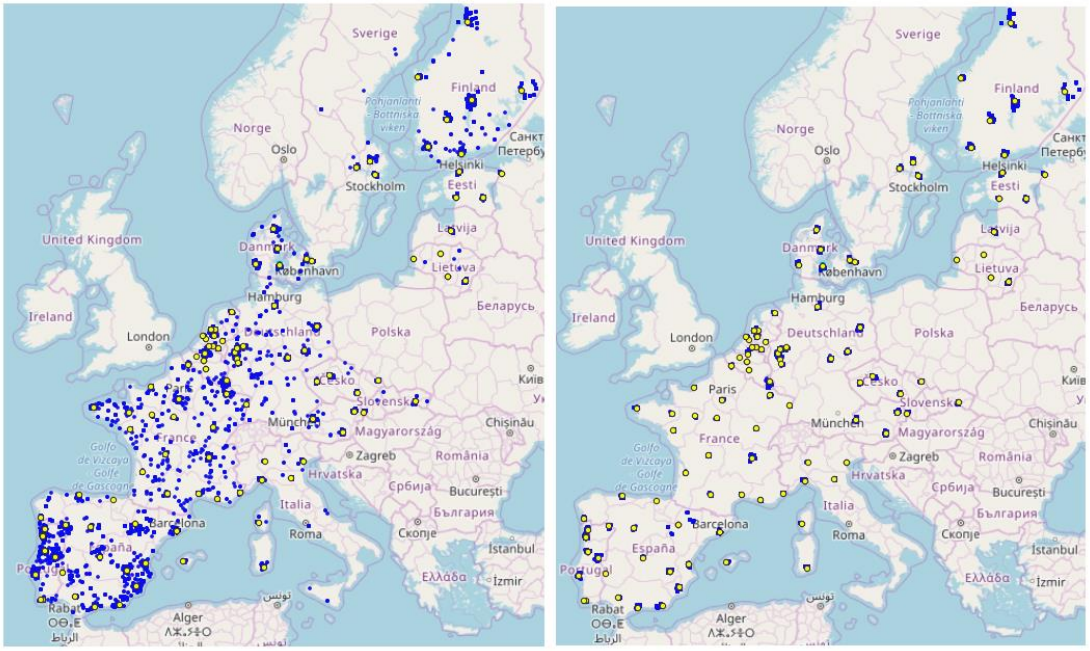

Figure 7. (**Left**): Extracted spatial samples from direct geocoding web services. (**Right**): Positions filtered by distances of less than 15 km relative to the referenced OpenAddresses dataset.

The case study shows that the high number of positional accuracy outliers in the direct method is due to the high coincidence response rate of commercial geocoders (82% Table 4) with respect to incomplete text addresses in OpenAddresses, which often lack correct information for settlements, residential districts, or municipalities in the requests to geocoders. However, this algorithm does not analyze the confidence or quality of address data references. Sometimes, this information is a unique source for locating any feature in the world. In addition, a possible cause of the large positional error results could be that a random assessment area may contain addresses in rural areas adjacent to large cities, which are not assessed in this study (lower spatial resolution). Once filtered by geocoding error distances and the number of matches using geocoders, this problem could be identified and solved. Sometimes, placing an address on a roof or street can also change the degree of positional accuracy of the product. Moreover, sometimes, the cause of positional issues is building typology and the type of urban planning, as open areas with large plots and single-family construction units can generate a higher number of positional errors.

For this reason, the quality algorithm begins with an analysis of completeness or logical consistency for every measure of quality in both geocoding methods. These measures must be indicated by the number of errors according to ISO 19157. Table 3 lists an average of errors, but it may not be a good method for explaining the lack of reliability in some elements for readers.

Table 4. This table lists some quality components and their measures by the average of the number of errors.

	Method 1: OpenAddresses	**Method 2: OpenStreetMap**
Subelements - Measures	**Direct Geocoding Assess**	**Reverse Geocoding Assess**
num_cities	100	166
num_tiles	999	1,480
avg_tiles	99.90%	89.16%
num_points	19,957	17,463
num_responses	66,324	52,036

avg_responses	82.91%	74.40%
Logical Consistency		
postcode_null	17.67%	2.26%
street_null	2.89%	1.50%
housenumber_null	27.13%	30.15%
cityapi_null	0.68%	0.05%
district_null	0.68%	0.05%
province_null	10.69%	3.28%
state_null	45.76%	34.74%
country_null	0.00%	0.00%
numbers_in_cities	0.47%	0.45%
Completeness		
address_duplicate	2.06%	5.40%
coordinate_duplicate	3.43%	10.07%
fulladd_duplicate	2.05%	5.35%
gmaps_omision	0.26%	5.88%
bing_omision	0.00%	4.23%
here_omision	1.25%	4.25%
osm_omision	62.67%	
Thematic Accuracy		
match_fulladdress	39.77%	46.51%
match_street	12.78%	11.03%
match_housenumber	5.39%	7.14%
match_postcode	13.35%	13.72%
match_city	4.25%	0.25%
match_addressapi	12.57%	10.22%

In the original sample for the direct geocoding method, 19,957 address samples were randomly selected to analyze downloaded data from OpenAddresses in order to match their address text. We obtained 66,324 responses, which accounted for 82% of responses on geocoders. For the reverse geocoding method, 17,463 samples were extracted from the OSM "overpass" API, with 74% of responses from commercial geocoders. Using the reverse method, we found OSM addresses in 89% of tiles, and authoritative addresses were extracted from OpenAddresses, making up 71%. The mean response time of the geocoders was 0.54 s using the direct method and 0.65 s using the reverse geocoding method with a 4G broadband connection.

Overall, a lack of consistency in geocoding address data was detected in the attributes related to administrative units. In particular, a different representation of their scope was detected at the population, neighborhood, district, and municipality levels. Reaching such a level is essential because there are streets with the same name within the same municipal administrative area, and it is necessary to know the actual name of a settlement in order to be able to geocode it correctly.

Sometimes, errors are a result of uncertainties in abbreviated street names and the lack of residential district or settlement names in the addresses' data reference (OpenAddresses). Incomplete text components with respect to addresses can produce significant errors during geocoding because geocoding web services try to provide some response to the candidate, while the Nominatim matching algorithm (OSM geocoder) does not force responses to any candidate when requests are incomplete.

This methodology was carried out to extract the average semantic match for each address request made to the geocoding web service using both methods. Aggregating

these results, the algorithm allows for obtaining quality measures with thematic accuracy Table 5 for address datasets from different countries, cities, or areas.

The statistical distribution of geocoding positional accuracies (error distance) in the raw data was positively skewed to the right (around cero), as shown in descriptive statistics and in studies published by other authors [18,19]. Despite the evidence of non-normal behavior [21] with respect to geocoding positional errors, most existing results follow a log-normal distribution [24,26].

The algorithm assumes a non-normal distribution of spatial errors but carries out statistical evaluations with an adjustment using the least squares method to fit distance values and extract the statistical parameters of positional errors. In the future, these spatial accuracy values can be used to filter the spatial matching of the address responses of geocoders, datasets, or subdatasets (regions, cities, etc.). In the case study, we reduced previously filtered commercial geocoder data responses by 30%, applying a transformation with least squares, and we estimated outliers from 6 m, as shown in Figure 6. The results for commercial geocoder positional accuracy values had a median of 1–2 m without ambiguity and an RMSE of 4–5 m, with 99.8% positional feasibility (2.5 sigma) upon matches. However, with respect to the crowdsourced geocoder, the median value was similar, but a better RMSE of 2.44 m was observed, as shown in Figure 8.

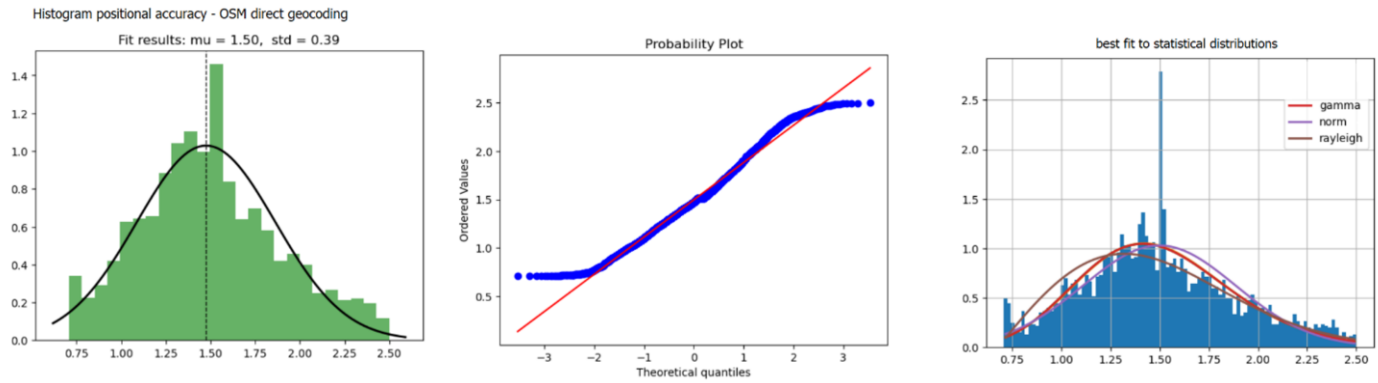

Figure 8. Left and center: Least squares transformation of the positional errors in meters from the crowdsourced (OSM) addresses and the fit to normal distributions. Right: Test showing a better fit of the positional errors relative to some statistical distributions. Some similarities are shown, but non-normal behavior is observed.

OSM positional errors exhibited a better fit relative to the normal distribution, and there were fewer spatial position coincidences. Moreover, this test demonstrated that the OSM spatial data source is unique and does not share address data with geocoders. However, sometimes, OSM can be completed using authoritative address data from some governmental organizations.

However, some studies with respect to geocoding quality characterized this non-normal distribution using statistics such as the mean, median, standard deviation, or percentiles. Therefore, the following case study, which tests the algorithm using some geocoders, datasets, and different spatial domains, shows positional errors, comparing raw data, threshold filtering, and interquartile values.

The address geocoding response data from the experiment using the algorithm are listed in Table 4. The results show raw positional accuracy values such as distance and the length or distance between the real position and that provided by the geocoding web service.

Table 5. Comparison of positional errors considering different maximum errors in order to drop outliers. There are two datasets for both methods: raw data and data filtered by geographic distance in meters.

	OpenAddresses (Direct Method)		OSM (Reverse Method)	
	Raw	**dist_geo < 50 m**	**dist_geo_api**	**dist_geo < 50 m**
count	66324	36,972	52,036	49,033
mean	47,538.65	7.06	25.16	13.34
std	499,310.83	10.37	280.06	10.63
min	0.00	0.00	0.00	0.00
25%	1.52	0.52	5.35	5.01
50%	21.76	2.14	11.81	10.96
75%	6126.63	9.16	20.86	18.66
90%	18,089.53	22.19	36.53	28.62
95%	56,533.29	31.64	54.60	35.62
99%	554,412.14	45.34	115.56	45.80
max	11,059,140.09	49.98	18,187.29	49.98

Figure 5 and Table 5 show how the raw positional data had outliers (atypical values) that introduced distortions in our descriptive statistics and probabilistic results. These atypical values can be observed in the box plots, scatter graphics, and histograms shown in Figures 4 and 5.

The spatial accuracy obtained using all datasets falls within the expected range of values for this geographical data type, which is obtained from urban areas. Following this study and previous studies [2,23,74], we estimate that the acceptable positional error of geocoding responses must not be around 50 m in urban areas because, in most cases, the distance values between geocoded global positions are around 5–20 m.

An interesting result of this study can be observed in Table 3, which demonstrates that the median can be used as a more robust centrality measure in this type of positional analysis. In fact, there is a positional coincidence rate of 33% in 19,957 geocoded text addresses using the direct method. Analyzing the results in depth (Table 5), Google Places has incorrect locations relative to 1837 addresses (9.2%), while the other geocoding services have better results. The positional errors found in geocoders comprised: Bing Maps—123 incorrect addresses (0.6%); Here—776 incorrect addresses (3.8%); and Nominatim (3.2%). VGI yields fewer requests, with no uncertainties in 23% of cases (4632 addresses), but commercial geocoders have errors that amount to more than 50 m compared with OpenAddresses. On the other hand, the reverse geocoding method has a positional matching rate of 90%, without uncertainties relative to the 17,463 addresses requested from the coordinates. In this case, nearly 2% of the geocoding errors originate from Google Places and Here, while the rest obtain correct positions. Bing Maps yields 0.3% of the errors, while the other geocoders provide correct global positions (Table 6).

Table 6. Comparison of positional errors in meters between all geocoders requested using the direct and reverse methods.

	Raw	**Count**	**Mean**	**Median**	**std**	**25%**	**75%**	**90%**	**95%**	**99%**
Direct Geocoding (OpenAddresses)	Bing Maps - Microsoft	19,957	35,716	18.58	380,921	1.00	6834	23,095	95,105	484,421
	Google Places	19,787	10,126	50.04	822,536	5.79	6028	15,211	32,506	7,033,801
	Here Technologies	19,131	20,770	43.62	90,225	1.25	8624	22,499	72,012	431,851

	Nominatim (OSM)	7449	52,471	2.84	37,051	0.06	15	7704	16,814	108,241
	dist_geo < 50 m									
	Bing Maps - Microsoft	11,265	6.23	1.28	9.93	0.50	7.70	20.77	29.49	44.53
	Google Places	9893	10.30	5.79	11.62	1.99	14.24	28.66	37.60	47.17
	Here Technologies	9693	6.35	1.28	10.19	0.51	7.69	21.17	30.95	45.02
	Nominatim (OSM)	6121	4.48	1.10	7.77	0.05	5.26	12.95	21.11	39.85
Reverse Geocoding (OSM)	**raw**	**count**	**mean**	**median**	**std**	**25%**	**75%**	**90%**	**95%**	**99%**
	Bing Maps - Microsoft	17,457	37.83	12.08	481.05	5.25	20.24	33.83	48.58	98.49
	Google Places	17,135	16.10	7.04	32.72	3.03	17.10	36.02	55.70	138.67
	Here Technologies	17,444	21.39	14.93	33.13	9.38	23.73	39.50	57.87	114.02
	dist_geo < 50 m									
	Bing Maps - Microsoft	16,636	13.33	11.40	10.50	4.96	18.64	28.13	34.96	45.54
	Google Places	16,110	10.40	6.42	10.62	2.86	14.24	26.66	34.48	45.71
	Here Technologies	16,287	16.26	14.19	9.93	8.94	21.21	30.57	37.10	46.15

However, Bing has similar accuracies and responses compared to Here using the direct method, but when using the reverse method, the results have more spatial differences (Table 6). The reason for this similar magnitude of errors is because Here is a provider of geographic data and LBS data to Microsoft. OSM has better accuracy than Bing Maps because OSM has fewer filtered address responses and because the Nominatim geocoder does not always try to find responses to the requested addresses.

In addition, the direct method algorithm (identifying text addresses from OpenAddresses) obtains a higher average of outliers in terms of absolute positional accuracy compared with the reverse method (matching location with OpenStreetMap) (Table 7). These results show geocoder problems when matching the semantic components of addresses. Furthermore, this observation explains the relationship between the high number of geocoding responses (95%) and the substantial number of matching errors (providing addresses far from the requested urban area).

Table 7. Some examples of positional errors by city using the direct method and reverse method algorithm.

	Direct							Reverse					
City	**Count**	**Outliers**	**Mean**	**Median**	**std**	**95%**	**99%**	**Count**	**Outliers**	**Mean**	**Median**	**std**	**95%**	**99%**
Amsterdam	595	23.91%	2.97	0.39	7.52	19.29	44.08	359	10.25%	16.28	14.38	11.61	36.82	45.61
Bologna	669	14.78%	5.41	1.94	8.81	27.68	43.92	366	9.63%	14.51	11.03	11.50	37.36	48.00
Bratislava	563	22.34%	4.41	2.82	5.30	10.64	31.02	389	1.77%	14.33	13.67	8.50	30.36	40.15
Copenhague	441	36.64%	2.39	0.95	4.68	11.93	20.19	483	1.63%	12.97	11.76	9.77	29.62	33.87
Dortmund	330	50.75%	4.60	1.62	6.88	20.92	35.17	240	4.76%	12.88	10.23	10.07	34.35	41.28
Düsseldorf	194	68.35%	7.94	2.70	11.49	36.55	47.08	231	6.10%	12.08	8.65	10.44	33.61	44.44
Eindhoven	422	40.14%	1.99	0.53	4.40	8.46	21.07	516	4.44%	17.48	16.40	11.08	36.68	47.78
Helsinki	425	43.63%	16.24	11.89	14.18	46.43	48.17	349	8.40%	17.27	16.62	9.34	35.14	46.63
Lisbon	180	69.95%	10.51	6.30	9.76	22.54	42.05	128	1.54%	11.53	11.02	7.10	23.16	34.51
Paris	401	38.96%	7.74	2.81	11.02	36.68	46.41	240	0.00%	10.73	9.22	8.59	27.31	38.48
Prague	509	28.81%	2.28	0.66	4.90	10.74	24.94	554	2.29%	13.34	11.29	10.75	38.14	44.93
Sevilla	344	42.57%	11.63	9.30	10.43	36.12	46.16	380	1.04%	11.08	8.76	9.09	27.98	42.51
Stockholm	607	7.61%	3.56	0.70	6.39	15.18	30.92	376	6.23%	16.47	13.38	12.34	41.56	48.01
Tallin	544	21.16%	9.68	6.20	10.83	32.86	45.74	291	10.19%	20.37	21.84	12.79	40.58	46.90
Vienna	607	16.04%	14.51	11.16	10.43	37.63	46.01	163	1.21%	11.08	10.29	7.06	24.19	33.38

Moreover, the developed algorithm allows for controlling the relative errors in the positional accuracy between geocoders. The results allow for an analysis of geographical relationships or correlations in the spatial positions of address data responses between one geocoder and another geocoder. Table 8 shows that Bing Maps often uses Here address data to reference their address in the geocoding web service.

Table 8. This table shows filter samples relative to the spatial threshold of relative positional errors between commercial geocoding web services.

	Absolute Measures			Relative Measures		
Country	**osm_gmaps**	**osm_bing**	**osm_here**	**gmaps_bing**	**gmaps_here**	**bing_here**
Netherlands	34.88	32.56	32.56	18.99	18.99	0.00
Belgium	9.38	28.76	28.76	42.20	42.20	0.00
Italy	3.29	5.45	9.67	3.20	8.79	5.63
Norway	0.11	8.39	8.39	17.09	17.09	0.00
Deutschland	1.04	1.96	6.25	4.70	9.62	13.20
Poland	16.06	4.83	4.83	19.29	19.29	0.00
Spain	3.71	9.81	17.16	9.05	18.57	14.48

Although no data on temporal quality were measured directly, we considered how often dataset updating can affect the matching rate of semantic address data. These results show a lack of harmonization with respect to the addresses' data attributes, which comprise the type of values and order and the form of representation in different geocoders; moreover, the language or codification of text characters can also generate errors during geocode matching in some cases.

5. Discussion

The quality of an address dataset affects the hit rate in geocoding web services [17,75] and the geospatial subproducts made using geocoded data. The developed algorithm

allows for analyzing quality following the standard rules denoted in ISO 19157, evaluating specific aspects with respect to geographic information products such as addresses. Therefore, the developed algorithm allows for assessing the quality of any address dataset before its use in georeferencing any spatial information.

Concretely, this paper examines the quality of authoritative and crowdsourced address datasets in urban areas across Europe, including datasets with different languages and schemas. There are similar studies using crowdsourced data from OSM, but they have other themes, such as land use [76], road networks [77], and similar specially made experiments [78] with respect to buildings. These were published years before this algorithm was used to extract results. Furthermore, the algorithm is implemented to automatically compare geographic information with commercial datasets from geocoding web services, which is carried out in other studies [19,25,49]. However, the implemented algorithm allows for automatically choosing random sampling areas inside urban zones, as requested by a city relative to crowdsourced datasets (OSM). However, in order to improve analyses, the number of samples for extraction can be chosen by users, and the geographical data reference can be chosen to define specific urban sample areas or city bounds.

The responses of geocoding web services from principal companies using geospatial data (Google, Microsoft, and Here) were stored in a spatial database in order to check spatial quality. In general, there is a considerable number of web service responses (98% matching) that agree with other authors [19], but these have a high number of mistakes. This is due to the incorrect behavior of the geocoder as it tries to provide possible answers for unknown addresses. Sometimes, the problem is caused by a position error in the reference dataset or an incomplete text component in the reference dataset. However, the results show that common errors are mainly present in the names of administrative units, small residential districts, and settlements, and sometimes, these errors are a result of the multilingualism present in their texts. Most addresses from geocoders need to normalize the parsed components, which agrees with [79]. Other related studies [54,80–82] use applications that implement the requested geocoders (ArcGIS, MMQGIS, and Batch; Easergeocoder), but they do not examine the raw data from providers compared to the developed algorithm.

In fact, the broad geographic scope used in this study is not used in other relevant and similar studies [13,18,24,45,83–86], as these studies focus on specific urban areas inside one country or region. Furthermore, our approach analyzes the implemented algorithm by forcing the automatic extraction of about 40,000 sample addresses for geocoding. In contrast, similar research studies developed their work using fewer sampling data and the same reference dataset source [19,57,72].

Furthermore, the implementation of the developed algorithm assists in evaluating the availability and completeness of the VGI dataset within the European Territory. If this is consistent with other studies in the European Union with respect to crowdsourced data quality compared with authoritative datasets [76,43-44, 87], then the OpenAddresses dataset can be concretely [62] used. An analysis of semantic similarity in address text components could improve the testing of new word-matching techniques. The algorithm implements a variant of the Levenhstein distance, similar to other authors [25], but it improves the analysis because it discriminates the order of words. Some studies introduced other specific algorithms that were derived from the Levenshtein algorithm [88]. Other studies [89] implemented other similar text algorithms and deep learning algorithms for words (Word2Vec), which demonstrated that the improvements did not benefit developments or increase implementation costs.

On the other hand, the positional accuracy results obtained using the OSM geocoder, i.e., Nominatim, are similar to the results of [89]. However, the results have poor semantic quality, as demonstrated by [44], and there are a high number of omissions, as [90] and [18] reported. Other studies [91] reveal the impacts of demographic biases, voluntary response, and community contribution. Nevertheless, the Nominatim geocoding OSM service has better positional accuracy with respect to geocoding responses upon this

algorithm's implementation compared with a similar test [80]. Furthermore, most related papers that analyze geocoding responses in the USA have good performance due to the normalization of the data reference TIGER/Line from the U.S. Census Bureau [13,24,51, 57,81,92,93]. In fact, other papers that examine geocoding address quality in European countries have worse positional accuracy results [80,86,87,94] than those obtained using our algorithm. However, experiments on positional accuracy in specific datasets must consider estimating the sampling size for a given population [64] and check the results relative to the product's specifications.

Moreover, the positional accuracy obtained in geocoding responses does not conform to the normal distribution as required to extract common statistical values. The designed algorithm transforms accuracy values to fit the Gaussian function and extracts the probability of spatial errors for an empirical evaluation of address matching. However, the results with respect to transformed positional error probabilities are estimated for algorithm users because they did not pass the normality test (Kolmogorov–Smirnov, D'Agostino's K-squared, and Anderson–Darling) and thus rejected the null hypothesis. Therefore, the positional accuracy measures of the quality extracted with the algorithm comprise descriptive statistics, such as the median and percentiles (figures), which are filtered by the spatial threshold, similar to other related studies. Thus, the algorithm can be used to compare statistics with a non-parametric test based on ranks such as Wilcoxon or Friedman [19]. Other studies apply a spatial Monte Carlo simulation [95,96] to find spatial patterns for points in area-based tests. In the future, the algorithm could be implemented to extract non-parametric statistics, or improved technologies could be used in machine learning processes. However, the relative cost of the algorithm's computation must be considered before quality control datasets.

The results of our algorithm's implementation could show the EU's effort toward authoritative data standardization and distribution using the common INSPIRE SDI platform. Currently, there are many geocoding applications from the European Public Administration, and only QGIS 3.28 LTR software has geocoding complements from different countries, such as France, Finland, Norway, Germany, and the Netherlands, and cities, such as Barcelona. In the future, this methodology can be implemented as a QGIS complement to benefit community users.

Finally, the statistical results show sufficient positional accuracy (about 10–40 m) when there is a match between address datasets. This positional accuracy value is similar to other related studies [18,69,72,81,83] that confirm accuracy values. The reverse method produced better positional results with respect to its raw address responses without a positional threshold, but the results were slightly worse within 50 m. The accuracy of addresses obtained with geocoding is essential because errors can be propagated to geospatial subproducts [97].

6. Conclusions

An algorithm that evaluates the quality of geolocated addresses in urban areas was developed and tested, and we obtained good results. The developed address quality algorithm was tested using VGI and authoritative datasets from repositories in some European countries as inputs in order to compare the results with crowdsourced (OSM—Nominatim) and commercial (Google, Microsoft, and Here) geocoding web services. In fact, the developed algorithm allows for choosing either geocoding method, whether direct or reverse, to check the spatial quality of a reference address dataset.

In addition, the spatial quality of address datasets and the geocoders' usability were examined. User-configurable input parameters were included so that the address quality analysis could be performed according to data size, the number of areas to be checked (tiles), the number of points per sample, a priori errors, and control points. The implementation with open technologies, such as the Python language and PostGIS spatial database, allows for including third-party developments or packages, easy sharing or updating

procedures by the developer's community, optimal processing times, and integration in main GIS applications.

The quality output measures are useful for checking the reliability of the semantic and spatial components in authoritative address datasets for any selected urban area. In addition, the algorithm can provide estimated probability parameters relative to semantic and positional accuracy, and it can be used for future address-matching processes.

The normalization of the quality analysis based on ISO 19157 was ensured in the algorithm's development. The normalization allows for the reproduction of different analyses to establish similar comparisons between several datasets and territories. In this respect, the methodology proposed by INSPIRE via SDI can greatly improve the quality of postal addresses. INSPIRE's geographic data rules propose a complete address data schema linked to other spatial data themes using a common system to structure, distribute, share, and maintain information. The algorithm also includes the possibility of obtaining authoritative data from standard web services (WFSs) implemented following the OGC and INSPIRE European directive, but this methodology is not appropriate for testing the algorithm's performance in this research study. However, SDI and OGC technologies are starting to use modern technologies and simpler formats in order to collect geographical information from web services using cloud computing systems.

Finally, the quality results of the implementation confirm that crowdsourced addresses could be integrated to improve and update authoritative datasets. On the other hand, the results confirm that the address datasets of geospatial companies have sufficient accuracy for the quality control of authoritative data, but filtering some responses is necessary. Therefore, this algorithm, which uses geocoding web services in order to check authoritative addresses, could optimize spatial data quality control in public administration. In fact, the automatic evaluation of address correspondence using this algorithm could increase and link thematic data (health, energy, demographics, housing, etc.) to public spatial databases. Thus, the developed algorithm can test the reliability of an address dataset in an urban area, city, or region using commercial or collaborative data, and the results can be used to detect semantic or positional errors and complete or update any address dataset.

Author Contributions: Conceptualization, R.S.R., J.C.M.-L. and E.C.-A.; methodology, R.S.R., J.C.M.-L. and E.C.-A.; software, R.S.R.; validation, E.L.-S. and J.C.M.-L.; formal analysis, R.S.R.; investigation, R.S.R., J.C.M.-L. and E.L.-S.; resources, E.C.-A.; data curation, R.S.R.; writing—original draft preparation, R.S.R.; writing—review and editing, R.S.R., E.L.-S., J.C.M.-L. and E.C.-A.; visualization, R.S.R.; supervision, J.C.M.-L. and E.C.-A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here:

- OpenAddresses: https://batch.openaddresses.io/data
- OpenStreetMap: https://wiki.openstreetmap.org/wiki/Downloading_data
- Google Places: https://developers.google.com/maps/documentation/places/web-service/overview?
- Bing: https://learn.microsoft.com/en-us/bingmaps/rest-services/locations/
- Here: https://developer.here.com/documentation/geocoding-search-api/

In addition, the locations of the study cities and the aggregate results of the analysis in each of them can be consulted here: https://figshare.com/ndownloader/files/42507916

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Longley, P.A.; Goodchild, M.F.; Maguire, D.J.; Rhind, D.W. *Geographic Information Science and Systems*; John Wiley & Sons: Hoboken, NJ, USA, 2015.

2. Bonner, M.R.; Han, D.; Nie, J.; Rogerson, P.; Vena, J.E.; Freudenheim, J.L. Positional Accuracy of Geocoded Addresses in Epidemiologic Research. *Epidemiology* **2003**, *14*, 408–412. https://doi.org/10.1097/01.EDE.0000073121.63254.c5.
3. Chainey, S.; Ratcliffe, J. *GIS and Crime Mapping*; John Wiley & Sons: Hoboken, NJ, USA, 2013.
4. Haldorson, M.; Moström, J. Developing a Statistical Geospatial Framework for the European Statistical System. In *Service-Oriented Mapping: Changing Paradigm in Map Production and Geoinformation Management*; Döllner, J., Jobst, M., Schmitz, P., Eds.; Springer International Publishing: Cham, Switzerland, 2019; pp. 185–206. https://doi.org/10.1007/978-3-319-72434-8_9.
5. Delboni, T.M.; Laender, A.H.F.; Borges, K.A.V.; Davis, C.A. Semantic Expansion of Geographic Web Queries Based on Natural Language Positioning Expressions. *Trans. GIS* **2007**, *11*, 377–397.
6. Roick, O.; Heuser, S. Location Based Social Networks—Definition, Current State of the Art and Research Agenda. *Trans. GIS* **2013**, *17*, 763–784. https://doi.org/10.1111/tgis.12032.
7. Jiang, S.; Alves, A.; Rodrigues, F.; Ferreira, J.; Pereira, F.C. Mining Point-of-Interest Data from Social Networks for Urban Land Use Classification and Disaggregation. *Comput. Environ. Urban Syst.* **2015**, *53*, 36–46. https://doi.org/10.1016/j.compenvurbsys.2014.12.001.
8. Lin, Y.; Kang, M.; Wu, Y.; Du, Q.; Liu, T. A Deep Learning Architecture for Semantic Address Matching. *Int. J. Geogr. Inf. Sci.* **2020**, *34*, 559–576. https://doi.org/10.1080/13658816.2019.1681431.
9. Xiao, Y.; Zhang, P.; Wang, T.; Li, T.; Song, Z. A Study of the Framework of Smart City Management System Construction. In Proceedings of the 2021 2nd International Conference on Artificial Intelligence and Computer Engineering, ICAICE 2021, Hangzhou, China, 5–7 November 2021; Institute of Electrical and Electronics Engineers Inc.: 2021; pp. 566–569. https://doi.org/10.1109/ICAICE54393.2021.00112.
10. McLeod, K.S. Our Sense of Snow9e Myth of John Snow in Medical Geography. *Soc. Sci. Med.* **2000**, *50*, 923–935. https://doi.org/10.1016/S0277-9536(99)00345-7.
11. Krieger, N. Place, Space, and Health: GIS and Epidemiology. *Epidemiology* **2003**, *14*, 384–385. https://doi.org/10.1097/01.ede.0000071473.69307.8a.
12. Donovan, G.H.; Michael, Y.L.; Butry, D.T.; Sullivan, A.D.; Chase, J.M. Urban Trees and the Risk of Poor Birth Outcomes. *Health Place* **2011**, *17*, 390–393. https://doi.org/10.1016/j.healthplace.2010.11.004.
13. Edwards, S.E.; Strauss, B.; Miranda, M.L. Geocoding Large Population-Level Administrative Datasets at Highly Resolved Spatial Scales. *Trans. GIS* **2014**, *18*, 586–603. https://doi.org/10.1111/tgis.12052.
14. Kebe, A.M.; Faye, R.M.; Lishou, C. Multi Agent-Based Addresses Geocoding for More Efficient Home Delivery Service in Developing Countries. In *e-Infrastructure and e-Services for Developing Countries*; Mendy, G., Ouya, S., Dioum, I., Thiaré, O., Eds.; Springer International Publishing: Cham, Switzerland, 2019; pp. 294–304.
15. Haberman, C.P.; Hatten, D.; Carter, J.G.; Piza, E.L. The Sensitivity of Repeat and near Repeat Analysis to Geocoding Algorithms. *J. Crim. Justice* **2021**, *73*, 101721. https://doi.org/10.1016/j.jcrimjus.2020.101721.
16. Li, D.; Cova, T.J.; Dennison, P.E.; Wan, N.; Nguyen, Q.C.; Siebeneck, L.K. Why Do We Need a National Address Point Database to Improve Wildfire Public Safety in the U.S.? *Int. J. Disaster Risk Reduct.* **2019**, *39*, 101237. https://doi.org/10.1016/j.ijdrr.2019.101237.
17. Ratcliffe, J.H. Geocoding Crime and a First Estimate of a Minimum Acceptable Hit Rate. *Int. J. Geogr. Inf. Sci.* **2004**, *18*, 61–72. https://doi.org/10.1080/13658810310001596076.
18. Chow, T.E.; Dede-Bamfo, N.; Dahal, K.R. Geographic Disparity of Positional Errors and Matching Rate of Residential Addresses among Geocoding Solutions. *Ann. GIS* **2016**, *22*, 29–42. https://doi.org/10.1080/19475683.2015.1085437.
19. Roongpiboonsopit, D.; Karimi, H.A. Comparative Evaluation and Analysis of Online Geocoding Services. *Int. J. Geogr. Inf. Sci.* **2010**, *24*, 1081–1100. https://doi.org/10.1080/13658810903289478.
20. Goldberg, D.W. Advances in Geocoding Research and Practice. *Trans. GIS* **2011**, *15*, 727–733. https://doi.org/10.1111/j.1467-9671.2011.01298.x.
21. Zandbergen, P.A. Positional Accuracy of Spatial Data: Non-Normal Distributions and a Critique of the National Standard for Spatial Data Accuracy. *Trans. GIS* **2008**, *12*, 103–130.
22. Whitsel, E.A.; Quibrera, P.M.; Smith, R.L.; Catellier, D.J.; Liao, D.; Henley, A.C.; Heiss, G. Accuracy of Commercial Geocoding: Assessment and Implications. *Epidemiol. Perspect. Innov.* **2006**, *3*, 8. https://doi.org/10.1186/1742-5573-3-8.
23. Ward, M.H.; Nuckols, J.R.; Giglierano, J.; Bonner, M.R.; Wolter, C.; Airola, M.; Mix, W.; Colt, J.S.; Hartge, P. Positional Accuracy of Two Methods of Geocoding. *Epidemiology* **2005**, *16*, 542–547. https://doi.org/10.1097/01.ede.0000165364.54925.f3.
24. Cayo, M.R.; Talbot, T.O. Positional Error in Automated Geocoding of Residential Addresses. *Int. J. Health Geogr.* **2003**, *2*, 10. http://www.ij-healthgeographics.com/content/2/1/10.
25. Kiliç, B.; Gülgen, F. Accuracy and similarity aspects in online geocoding services: A comparative evaluation for Google and Bing maps. *Int. J. Eng. Geosci.* **2020**, *5*, 109–119. https://doi.org/10.26833/ijeg.629381.
26. Karimi, H.A.; Sharker, M.H.; Roongpiboonsopit, D. Geocoding Recommender: An Algorithm to Recommend Optimal Online Geocoding Services for Applications. *Trans. GIS* **2011**, *15*, 869–886. https://doi.org/10.1111/j.1467-9671.2011.01293.x.
27. OpenAddresses. Available online: https://openaddresses.io/ (accessed on 26 September 2023).

28. McKenzie, D.; Jonas, M.; Coetzee, S.; Body, C.; Smith, M.; Blake, M.; Abhayaratna, J.; Judd, M.; Roos, M. The Role of Geospatial Information Standards for Sustainable Development. *Sustain. Dev. Goals Connect. Dilemma* **2019**, 223–241. https://doi.org/10.1201/9780429290626-14.
29. Benitez-Paez, F.; Comber, A.; Trilles, S.; Huerta, J. Creating a Conceptual Framework to Improve the Re-Usability of Open Geographic Data in Cities. *Trans. GIS* **2018**, *22*, 806–822. https://doi.org/10.1111/tgis.12449.
30. Vancauwenberghe, G.; Valečkaitė, K.; van Loenen, B.; Welle Donker, F. Assessing the Openness of Spatial Data Infrastructures (SDI): Towards a Map of Open SDI. *Int. J. Spat. Data Infrastruct. Res.* **2018**, *13*, 88–100. https://doi.org/10.2902/1725-0463.2018.13.art9.
31. The European Parliament and the Council of the European Union. INSPIRE European Directive. Available online: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32007L0002&qid=1687602142795 (accessed on 26 September 2023).
32. Open Geospatial Consortium. Available online: https://www.ogc.org/ (accessed on 26 September 2023).
33. *ISO/TC 211*; Geographic Information/Geomatics. Swedish Institute for Standards: Stockholm, Sweden, 1994. Available online: https://committee.iso.org/sites/tc211/home/standards-in-action/addressing.html (accessed on 26 September 2023).
34. Goodchild, M.F. Citizens as Sensors: The World of Volunteered Geography. *GeoJournal* **2007**, *69*, 211–221. https://doi.org/10.1007/s10708-007-9111-y.
35. Gómez-Barrón, J.P.; Manso-Callejo, M.Á.; Alcarria, R. Volunteered Geographic Information Systems: Technological Design Patterns. *Trans. GIS* **2019**, *23*, 976–1007. https://doi.org/10.1111/tgis.12544.
36. OpenStreetMap. Available online: https://www.openstreetmap.org/about (accessed on 26 September 2023).
37. Google. Geocoding API Google Maps. Available online: https://developers.google.com/maps/documentation/geocoding/overview?hl=en (accessed on 26 September 2023).
38. Microsoft. Bing API Locations. Available online: https://learn.microsoft.com/en-us/bingmaps/rest-services/locations/ (accessed on 26 September 2023).
39. Here. Here Geocoding Services. Available online: https://www.here.com/platform/geocoding (accessed on 26 September 2023).
40. Sangiambut, S.; Sieber, R. The V in VGI: Citizens or Civic Data Sources. *Urban Plan.* **2016**, *1*, 141–154. https://doi.org/10.17645/up.v1i2.644.
41. Heipke, C. Crowdsourcing Geospatial Data. *ISPRS J. Photogramm. Remote Sens.* **2010**, *65*, 550–557. https://doi.org/10.1016/j.isprsjprs.2010.06.005.
42. Fonte, C.C.; Antoniou, V.; Bastin, L.; Estima, J.; Arsanjani, J.J.; Bayas, J.-C.L.; See, L.; Vatseva, R. Assessing VGI Data Quality. In *Mapping and the Citizen Sensor*; Ubiquity Press: London, UK, 2017. https://doi.org/10.5334/bbf.g.
43. Jokar Arsanjani, J.; Mooney, P.; Zipf, A.; Schauss, A. Quality Assessment of the Contributed Land Use Information from OpenStreetMap Versus Authoritative Datasets. In *OpenStreetMap in GIScience: Experiences, Research, and Applications*; Jokar Arsanjani, J., Zipf, A., Mooney, P., Helbich, M., Eds.; Springer International Publishing: Cham, Switzerland, 2015; pp. 37–58. https://doi.org/10.1007/978-3-319-14280-7_3.
44. Antoniou, V.; Skopeliti, A. Measures and Indicators of Vgi Quality: An Overview. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2015**, *2*, 345–351. https://doi.org/10.5194/isprsannals-II-3-W5-345-2015.
45. Senaratne, H.; Mobasheri, A.; Ali, A.L.; Capineri, C.; Haklay, M. (Muki). A Review of Volunteered Geographic Information Quality Assessment Methods. *Int. J. Geogr. Inf. Sci.* **2017**, *31*, 139–167. https://doi.org/10.1080/13658816.2016.1189556.
46. *ISO 19157:2013*; Geographic Information—Data Quality. British Standards Institution: London, UK, 2013. Available online: https://www.iso.org/standard/32575.html (accessed on 26 September 2023).
47. Zhang, C.; He, B.; Guo, R.; Ma, D. A Graph-Based Approach for Representing Addresses in Geocoding. *Comput. Environ. Urban Syst.* **2023**, *100*, 101937. https://doi.org/https://doi.org/10.1016/j.compenvurbsys.2022.101937.
48. Andrienko, G.; Andrienko, N.; Weibel, R. Geographic Data Science. *IEEE Comput. Graph. Appl.* **2017**, *37*, 15–17. https://doi.org/10.1109/MCG.2017.3621219.
49. Baru, C.; Institute of Electrical and Electronics Engineers; IEEE Computer Society. In Proceedings of the 2019 IEEE International Conference on Big Data, Los Angeles, CA, USA, 9–12 December 2019.
50. Cruz, P.; Vanneschi, L.; Painho, M.; Rita, P. Automatic Identification of Addresses: A Systematic Literature Review. *ISPRS Int. J. Geo-Inf.* **2022**, *11*, 11. https://doi.org/10.3390/ijgi11010011.
51. Zandbergen, P.A. A Comparison of Address Point, Parcel and Street Geocoding Techniques. *Comput. Environ. Urban Syst.* **2008**, *32*, 214–232. https://doi.org/10.1016/j.compenvurbsys.2007.11.006.
52. Coetzee, S.; Odijk, M.; van Loenen, B.; Storm, J.; Stoter, J. Stakeholder Analysis of the Governance Framework of a National SDI Dataset—Whose Needs Are Met in the Buildings and Address Register of the Netherlands? *Int. J. Digit. Earth* **2020**, *13*, 355–373. https://doi.org/10.1080/17538947.2018.1520930.
53. European Commission. European INSPIRE Geoportal. Available online: https://inspire-geoportal.ec.europa.eu/ (accessed on 26 September 2023).
54. European Comission. INSPIRE Addresses Specification. Available online: https://inspire.ec.europa.eu/file/1728/download?token=K1Jh4B5h (accessed on 26 September 2023).

55. Pruvost, H.; Mooney, P. Exploring Data Model Relations in OpenStreetMap. *Future Internet* **2017**, *9*, 70. https://doi.org/10.3390/fi9040070.
56. Zielstra, D.; Hochmair, H.H.; Neis, P. Assessing the Effect of Data Imports on the Completeness of Openstreetmap—A United States Case Study. *Trans. GIS* **2013**, *17*, 315–334. https://doi.org/10.1111/tgis.12037.
57. Prener, C.G.; Fox, B. Creating Open Source Composite Geocoders: Pitfalls and Opportunities. *Trans. GIS* **2021**, *25*, 1868–1887. https://doi.org/10.1111/tgis.12741.
58. Singh, S.K. Evaluating Two Freely Available Geocoding Tools for Geographical Inconsistencies and Geocoding Errors. *Open Geospat. Data Softw. Stand.* **2017**, *2*, 11. https://doi.org/10.1186/s40965-017-0026-3.
59. Goldberg, D.W.; Cockburn, M.G. Improving Geocode Accuracy with Candidate Selection Criteria. *Trans. GIS* **2010**, *14* (Suppl. 1), 149–176. https://doi.org/10.1111/j.1467-9671.2010.01211.x.
60. European Union's Space Programme. CORINE Land Cover from Copernicus EU Programme. Available online: https://land.copernicus.eu/pan-european/corine-land-cover (accessed on 26 September 2023).
61. Montero, E.; Van Wolvelaer, J.; Garzón, A. The European Urban Atlas. In *Land Use and Land Cover Mapping in Europe: Practices & Trends*; Manakos, I., Braun, M., Eds.; Springer: Dordrecht, The Netherlands, 2014; pp. 115–124. https://doi.org/10.1007/978-94-007-7969-3_8.
62. Stark, H.-J. Quality Assessment of VGI Based on Open Web Map Services and ISO/TC 211 19100-Family Standards. Available online: https://gispoint.de/fileadmin/user_upload/paper_gis_open/GI_Forum_2011/537509015.pdf (accessed on 26 September 2023).
63. Talhofer, V.; Hošková-Mayerová, Š.; Hofmann, A. *Quality of Spatial Data in Command and Control System*; Springer: Cham, Switzerland, 2019. https://doi.org/https://doi.org/10.1007/978-3-319-94562-0.
64. Ureña-Cámara, M.A.; Nogueras-Iso, J.; Lacasta, J.; Ariza-López, F.J. A Method for Checking the Quality of Geographic Metadata Based on ISO 19157. *Int. J. Geogr. Inf. Sci.* **2019**, *33*, 1–27. https://doi.org/10.1080/13658816.2018.1515437.
65. Van Oort, P. *Spatial Data Quality: From Description to Application*; Wageningen University and Research: Wageningen, The Netherlands, 2006.
66. Lee, K.; Claridades, A.R.C.; Lee, J. Improving a Street-Based Geocoding Algorithm Using Machine Learning Techniques. *Appl. Sci.* **2020**, *10*, 5628. https://doi.org/10.3390/app10165628.
67. Lin, Y.; Kang, M.; He, B. Spatial Pattern Analysis of Address Quality: A Study on the Impact of Rapid Urban Expansion in China. *Environ. Plan. B Urban Anal. City Sci.* **2021**, *48*, 724–740. https://doi.org/10.1177/2399808319895272.
68. Koumarelas, I.; Kroschk, A.; Mosley, C.; Naumann, F. Experience: Enhancing Address Matching with Geocoding and Similarity Measure Selection. *J. Data Inf. Qual.* **2018**, *10*, 1–16. https://doi.org/10.1145/3232852.
69. Kilic, B.; Gülgen, F. Investigating the Quality of Reverse Geocoding Services Using Text Similarity Techniques and Logistic Regression Analysis. *Cartogr. Geogr. Inf. Sci.* **2020**, *47*, 336–349. https://doi.org/10.1080/15230406.2020.1746198.
70. Zimmerman, D.L.; Li, J. The Effects of Local Street Network Characteristics on the Positional Accuracy of Automated Geocoding for Geographic Health Studies. *Int. J. Health Geogr.* **2010**, *9*, 10. https://doi.org/10.1186/1476-072X-9-10.
71. Martínez-Llario, J.C.; Baselga, S.; Coll, E. Accurate Algorithms for Spatial Operations on the Spheroid in a Spatial Database Management System. *Appl. Sci.* **2021**, *11*, 5129. https://doi.org/10.3390/app11115129.
72. Whitsel, E.A.; Rose, K.M.; Wood, J.L.; Henley, A.C.; Liao, D.; Heiss, G. Accuracy and Repeatability of Commercial Geocoding. *Am. J. Epidemiol.* **2004**, *160*, 1023–1029. https://doi.org/10.1093/aje/kwh310.
73. Pan American Institute of Geography and History. *Guide for the Positional Accuracy Assessment of Geospatial Data*; Occasional Papers: London, UK, 2021.
74. Zimmerman, D.L.; Fang, X.; Mazumdar, S.; Rushton, G. Modeling the Probability Distribution of Positional Errors Incurred by Residential Address Geocoding. *Int. J. Health Geogr.* **2007**, *6*, 1. https://doi.org/10.1186/1476-072X-6-1.
75. Briz-Redón, Á.; Martinez-Ruiz, F.; Montes, F. Reestimating a Minimum Acceptable Geocoding Hit Rate for Conducting a Spatial Analysis. *Int. J. Geogr. Inf. Sci.* **2020**, *34*, 1283–1305. https://doi.org/10.1080/13658816.2019.1703994.
76. Dorn, H.; Törnros, T.; Zipf, A. Quality Evaluation of VGI Using Authoritative Data-a Comparison with Land Use Data in Southern Germany. *ISPRS Int. J. Geo-Inf.* **2015**, *4*, 1657–1671. https://doi.org/10.3390/ijgi4031657.
77. Minaei, M. Evolution, Density and Completeness of OpenStreetMap Road Networks in Developing Countries: The Case of Iran. *Appl. Geogr.* **2020**, *119*, 102246. https://doi.org/https://doi.org/10.1016/j.apgeog.2020.102246.
78. Biljecki, F.; Chow, Y.S.; Lee, K. Quality of Crowdsourced Geospatial Building Information: A Global Assessment of OpenStreetMap Attributes. *Build. Environ.* **2023**, *237*, 110295. https://doi.org/10.1016/j.buildenv.2023.110295.
79. Xu, S.; Flexner, S.; Carvalho, V. Geocoding Billions of Addresses: Toward a Spatial Record Linkage System with Big Data. Available online: https://api.semanticscholar.org/CorpusID:15956962 (accessed on 26 September 2023).
80. Cetl, V.; Kliment, T.; Jogun, T. A Comparison of Address Geocoding Techniques–Case Study of the City of Zagreb, Croatia. *Surv. Rev.* **2018**, *50*, 97–106. https://doi.org/10.1080/00396265.2016.1252517.
81. Rashidian, S.; Dong, X.; Avadhani, A.; Poddar, P.; Wang, F. Effective Scalable and Integrative Geocoding for Massive Address Datasets. In Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Redondo Beach, CA, USA, 7–10 November 2017; Association for Computing Machinery: New York, NY, USA, 2017. https://doi.org/10.1145/3139958.3139986.

82. Kinnee, E.J.; Tripathy, S.; Schinasi, L.; Shmool, J.L.C.; Sheffield, P.E.; Holguin, F.; Clougherty, J.E. Geocoding Error, Spatial Uncertainty, and Implications for Exposure Assessment and Environmental Epidemiology. *Int. J. Environ. Res. Public Health* **2020**, *17*, 5845. https://doi.org/10.3390/ijerph17165845.
83. Duncan, D.T.; Castro, M.C.; Blossom, J.C.; Bennett, G.G.; Gortmaker, S.L. Evaluation of the Positional Difference between Two Common Geocoding Methods. Available online: http://dx.doi.org/10.4081/gh.2011.179 (accessed on 26 September 2023).
84. Davis, C.A.; de Alencar, R.O. Evaluation of the Quality of an Online Geocoding Resource in the Context of a Large Brazilian City. *Trans. GIS* **2011**, *15*, 851–868. https://doi.org/10.1111/j.1467-9671.2011.01288.x.
85. Hart, T.C.; Zandbergen, P.A. Reference Data and Geocoding Quality: Examining Completeness and Positional Accuracy of Street Geocoded Crime Incidents. *Policing* **2013**, *36*, 263–294. https://doi.org/10.1108/13639511311329705.
86. Kounadi, O.; Lampoltshammer, T.J.; Leitner, M.; Heistracher, T. Accuracy and Privacy Aspects in Free Online Reverse Geocoding Services. *Cartogr. Geogr. Inf. Sci.* **2013**, *40*, 140–153. https://doi.org/10.1080/15230406.2013.777138.
87. Faure, E.; Danjou, A.M.N.; Clavel-Chapelon, F.; Boutron-Ruault, M.C.; Dossus, L.; Fervers, B. Accuracy of Two Geocoding Methods for Geographic Information System-Based Exposure Assessment in Epidemiological Studies. *Environ. Health* **2017**, *16*, 15. https://doi.org/10.1186/s12940-017-0217-5.
88. Cheng, R.; Liao, J.; Chen, J. Quickly Locating POIs in Large Datasets from Descriptions Based on Improved Address Matching and Compact Qualitative Representations. *Trans. GIS* **2022**, *26*, 129–154. https://doi.org/10.1111/tgis.12838.
89. Haklay, M. How Good Is Volunteered Geographical Information? A Comparative Study of OpenStreetMap and Ordnance Survey Datasets. *Environ. Plan. B Urban Anal. City Sci.* **2010**, *37*, 682–703. https://doi.org/10.1068/b35097.
90. Goodchild, M.F.; Li, L. Assuring the Quality of Volunteered Geographic Information. *Spat. Stat.* **2012**, *1*, 110–120. https://doi.org/10.1016/j.spasta.2012.03.002.
91. Basiri, A.; Haklay, M.; Foody, G.; Mooney, P. Crowdsourced Geospatial Data Quality: Challenges and Future Directions. *Int. J. Geogr. Inf. Sci.* **2019**, *33*, 1588–1593. https://doi.org/10.1080/13658816.2019.1593422.
92. Goldberg, D.W. Improving Geocoding Match Rates with Spatially-Varying Block Metrics. *Trans. GIS* **2011**, *15*, 829–850. https://doi.org/10.1111/j.1467-9671.2011.01295.x.
93. Mazeika, D.; Summerton, D. The Impact of Geocoding Method on the Positional Accuracy of Residential Burglaries Reported to Police. *Policing* **2017**, *40*, 459–470. https://doi.org/10.1108/PIJPSM-03-2016-0048.
94. Küçük Matci, D.; Avdan, U. Address Standardization Using the Natural Language Process for Improving Geocoding Results. *Comput. Environ. Urban Syst.* **2018**, *70*, 1–8. https://doi.org/10.1016/j.compenvurbsys.2018.01.009.
95. Andresen, M.A. Testing for Similarity in Area-Based Spatial Patterns: A Nonparametric Monte Carlo Approach. *Appl. Geogr.* **2009**, *29*, 333–345. https://doi.org/10.1016/j.apgeog.2008.12.004.
96. Andresen, M.; Malleson, N.; Steenbeek, W.; Townsley, M.; Vandeviver, C. Minimum Geocoding Match Rates: An International Study of the Impact of Data and Areal Unit Sizes. *Int. J. Geogr. Inf. Sci.* **2020**, *34*, 1306–1322. https://doi.org/10.1080/13658816.2020.1725015.
97. Zandbergen, P.A.; Hart, T.C.; Lenzer, K.E.; Camponovo, M.E. Error Propagation Models to Examine the Effects of Geocoding Quality on Spatial Analysis of Individual-Level Datasets. *Spat. Spatio-Temporal Epidemiol.* **2012**, *3*, 69–82. https://doi.org/10.1016/j.sste.2012.02.007.