International Journal of Geo-Information

MDPI

Article

Using an Optimized Chinese Address Matching Method to Develop a Geocoding Service: A Case Study of Shenzhen, China

Qin Tian [1], Fu Ren [1,2,3], Tao Hu [4], Jiangtao Liu [5], Ruichang Li [1] and Qingyun Du [1,2,3,6,*]

1 School of Resource and Environmental Science, Wuhan University, 129 Luoyu Road, Wuhan 430079, China; tianqin@whu.edu.cn (Q.T.); renfu@whu.edu.cn (F.R.); liruichang@whu.edu.cn (R.L.)
2 Key Laboratory of Geographic Information System, Ministry of Education, Wuhan University, 129 Luoyu Road, Wuhan 430079, China
3 Key Laboratory of Digital Mapping and Land Information Application Engineering, National Administration of Surveying, Mapping and Geo-information, Wuhan University, 129 Luoyu Road, Wuhan 430072, China
4 School of Information Engineering, Xuchang University, 88 Bayi Road, Xuchang 461000, China; hutao.cumt@163.com
5 Shenzhen Municipal Planning & Land Real Estate Information Center, 8007 Hongli Road, Shenzhen 518034, China; liujt.ebor@gmail.com
6 Collaborative Innovation Center Of Geospatial Technology, Wuhan University, 129 Luoyu Road, Wuhan 430079, China
* Correspondence: qydu@whu.edu.cn; Tel.: +86-27-8766-4557; Fax: +86-27-6877-8893

Academic Editor: Wolfgang Kainz
Received: 19 January 2016; Accepted: 2 May 2016; Published: 13 May 2016

Abstract: With the coming era of big data and the rapid development and widespread applications of Geographical Information Systems (GISs), geocoding technology is playing an increasingly important role in bridging the gap between non-spatial data resources and spatial data in various fields. However, Chinese geocoding faces great challenges because of the complexity of the address string format in Chinese, which contains no delimiters between Chinese words, and the poor address management resulting from the existence of multiple address authorities spread among different governmental agencies. This paper presents a geocoding service based on an optimized Chinese address matching method, including address modeling, address standardization and address matching. The address model focuses on the spatial semantics of each address element, and the address standardization process is based on an address tree model. A geocoding service application is implemented in practice using a large quantity of data from Shenzhen Municipality. More than 1,460,000 data records were used to test the geocoding service, and good matching rates were achieved with good adaptability and intelligence.

Keywords: Geographical Information System; Chinese address match; address model; address matching; geocoding service

1. Introduction

Increasing amounts of various types of data are being collected in various domains with the coming era of big data and the rapid development and widespread application of Geographical Information Systems (GISs) [1]. It appears to be an urgent necessity to possess the ability to assign coordinates to these data and continue to pursue deeper research into big data from the spatial perspective [2]. One key topic of interest is the spatial-information-containing texts known as addresses and it can be the most resources which citizens used to convey geographical information or

locations [3]. The Global Sourcebook of Address Data Management introduced the address systems of 194 countries [4]. Most urban GIS applications require geocoding technology, which can obtain spatial coordinates from addresses [5]. To date, geocoding has been the most popular and efficient method for translating addresses into spatial coordinates and has been employed in many fields, including public health [6,7], criminology [8], cancer research [9], transportation [10], traffic crashes response [11,12], and others [7,13,14].

Geocoding, which is also referred to as address matching, is the process of mapping a text-based descriptive address to digital geographic coordinates that can be used in spatial analysis and visualization [15–17]. In other words, geocoding is a type of positioning technology that can extract coordinates from a given text or natural-language descriptions of spatial location information, such as an urban address, building numbers, street addresses, postal codes, building names and company names [15]. Geocoding is the most important available means of bridging the gap between non-spatial information and spatial information, and accuracy and certainty are of great importance for geocoding systems and services [5,18–21]. Many researchers are working to improve geocoding matching rates [16,22–26].

Geocoding is a worldwide challenge, and is currently more problematic in China than in most Western countries because the Chinese language is much more complex than the English language regarding segmentation and semantics [27,28]: the English language has delimiters between words, and Western countries have many successful solutions for performing such tasks, such as the Geographic Base File/Dual Independent Map Encoding (GBF/DIME) address geocoding solution, Topologically Integrated Geographic Encoding and Reference (TIGER, or TIGER/Line) Geocoding solution, ESRI address geocoding solution and so on [29]. Almost all commercial GIS software packages (e.g., ArcGIS and MapInfo) and web-based map service vendors (e.g., Google, Bing and MapQuest) provide geocoding modules that function well for non-Chinese text addresses. However, these solutions are designed for Western countries and are unsuitable for application to Chinese addresses. Chinese geocoding remains a persistent problem because of the complexity of Chinese syntax and semantics, as described below [24,27,28]:

(1) Chinese text has no delimiters between words, unlike English [2,28]. It is extremely difficult to parse the string and extract exact address elements, such as administrative region names, road names, street numbers, and building numbers, to match the reference database.

(2) Various descriptive styles exist among Chinese addresses. The regulation of Chinese address structures is poor and received little attention from urban planning and administrative managers before the early 1990s. A large number of addresses that have been collected and stored in basic geographic databases lack a uniform structure and have been defined according to local custom in the absence of explicit authoritative rules to follow. With the increasing availability of such address data, the disordered state of Chinese addresses is becoming more evident.

(3) Urban planning and address management in many areas of China is chaotic. For a long time, city layouts in China have been developed without sufficient urban planning, and building numbers were assigned randomly and without regulation. Consequently, it is difficult to add, delete and update to accommodate for new addresses, and address interpolation and address matching are difficult because, for example, several different roads are named Zhongshan Road. This further increases the difficulty and inaccuracy of the address matching task in China.

(4) The rights for address naming and approval are distributed among several different governmental agencies. The right to name administrative regions, such as provinces, cities and districts, belongs to the State Council, whereas street numbers and building numbers are managed by the Department of Public Safety. Meanwhile, Urban Planning Departments hold the naming rights for roads and streets. For example, in the address "Shenzhen City, Futian District, Hongli West Road Number 8890", the address elements "Shenzhen City" and "Futian District" are determined by the State Council, whereas "Hongli West Road" is defined by the Urban Planning Department and "Number 8890" is assigned by the Department of Public Safety. The lack of united address

planning makes it quite difficult to manage addresses and perform address matching because different agencies have different address formats and descriptive forms. This situation makes it quite difficult to standardize even a single Chinese address.

In this paper, a set of optimized Chinese address matching methods is presented to resolve the situation described above. The proposed methods are intended to: (1) adapt to the complexity of Chinese text; (2) account for the topological relationships among address elements; and (3) improve the efficiency and accuracy of traditional geocoding algorithms. In the remainder of this paper, we first describe the study area and data in section 2, followed by the optimized method in detail, including the structure of the address model and the address standardization and address matching process in Section 3. Then, the conducted experiments and their results are presented in Section 4. A discussion is provided in Section 5, and the conclusions are summarized in Section 6.

2. Study Area and Data

2.1. Study Area

Shenzhen is a major city located in the southeast of China between longitudes of 113°46′ to 114°37′ east and latitudes of 22°27′ and 22°52′ north. Shenzhen covers a total area of 1952 km^2 and it spans a range of is 81.4 km from east to west and 10.8 km from north to south. Shenzhen is connected to Hong Kong in the south, Dongguan in the north, and Huizhou in the east. Shenzhen is one of the special economic zones in China. At present, Shenzhen consists of 10 districts (Luohu, Futian, Nanshan, Yantian, Baoan, Longgang, Guangming, Pingshan, Longhua and Dapeng). Figure 1 shows the location of Shenzhen Municipality and a map of the districts of Shenzhen.

Figure 1. Location of Shenzhen in China and Districts in Shenzhen.

With the rapid development of Shenzhen Municipality, the government of Shenzhen has collected and stored a massive amount of data in recent years. These data can be divided into the two following general categories: data that include spatial coordinates, namely, spatial data, and data that include only text descriptions of spatial information, such as addresses, postal codes, and toponyms. The latter type of data does not have coordinates, and are called non-spatial data. For more effective utilization of these data, the government plans to initiate efforts towards providing open data. However, these plans face the critical problem of how to attach spatial information to these non-spatial data. For this purpose, geocoding is the best solution.

2.2. Data

As mentioned above, large amounts of spatial data and non-spatial data have been collected by the Shenzhen government over the years, including spatial data for roads, point of interest (POI), buildings, street numbers, and building numbers and non-spatial data for hospitals, schools, and companies.

The spatial data that we used in this study to construct a reference database for the development of the proposed geocoding service were obtained from the Urban Planning, Land and Resources Commission of Shenzhen Municipality (SZPL). This database contains administrative data, road data, POI data, street number data, toponym data and house number data (a total of approximately 708,000 data records).

We used non-spatial disease data associated with home addresses and company data associated with company addresses. The disease data contains more than 100,000 records obtained from the Shenzhen Center for Health Information (SCHI), and the company data include approximately 1,360,000 records obtained from the Market and Quality Supervision Commission of Shenzhen Municipality (SZSCJG).

3. Address Match Methods

Figure 2 illustrates the fundamental principles of the geocoding process [26,27], which include address model building, address standardization, address matching and reference database construction. The address model explains how a single address is organized and defines the relationships between address elements. The establishment of an internationally standardized geospatially enabled address model is the purpose of ISO 19160 [30]. Unfortunately, little significant progress in address internationalization has been made in China. Address standardization is the process of parsing an input address string into a list of address elements and then reordering these elements based on their spatial semantics. The process also involves the elimination of errors, the correction of address imperfections and the confirmation of the destination of the input address string. Address matching is the next step of geocoding, and refers to the process of assigning spatial coordinates to a standardized address by searching in a reference database using a particular address matching algorithm.

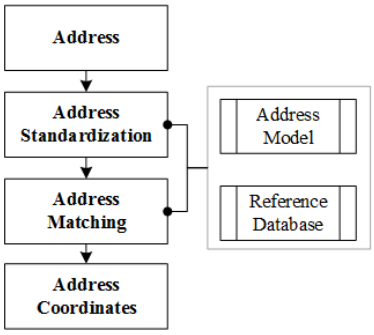


Figure 2. The geocoding process.

Geocoding accuracy is one of the largest concerns for users when seeking the most appropriate geocoding service, because lower geocoding accuracy will result in errors in geocoding applications. Most geocoding solutions employ a general method in which addresses are matched based solely on text comparison and the spatial semantics of the address elements are neglected. By contrast, the geocoding solution proposed in this paper fully considers the relevant spatial semantics, especially the topological relationships between address elements. With the constraints provided by the spatial relationships, the fitting rules used to reconstruct the destination address will be more precise, and spatial operations can be used in the geocoding process.

3.1. Address Model

The address model is an abstract description of how a given address containing certain address elements is organized and expressed. A Chinese address consists of three major components: administrative elements, basic constraint objects and local point locations [31,32]. The organization rules for these address elements are often described as below:

<Standard address> ::=<Administrative name> <Basic constraint object> <Local point location>

The elements are defined as follows:

<Administrative name > ::=<country><province> [district] <county> [village]

<Basic constraint object> ::=<street> | <alley> | <industrial district> | <natural village>

<Local point location> ::=<building numbers> [house numbers] | <landmark> | <point of interest>

As an example, consider the address "No. 8088, Hongli Road, Futian District, Shenzhen", as shown in Figure 3.

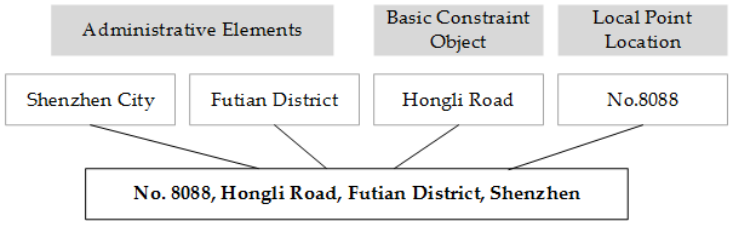


Figure 3. An example of the organization rules for addresses.

The organization rules for addresses provide important guidance regarding how to describe a location in natural language, but they do little to represent the spatial relationships and constraints between address elements, which are important for address parsing and standardization. Certain spatial relationships do exist between address elements: an analysis of Chinese addresses reveals topological relationships (such as containing, adjacency, and adjoining), distance relationships and direction relationships. In most cases, the topological relationships between address elements, especially containing relationships, are very common and obvious; an address model with this hierarchical characteristic is commonly called a hierarchical address model [33]. Figure 4 shows an example of a Shenzhen address with multiple-types of spatial relationships. There are topological (containing) relationships between the address elements "Shenzhen City", "Hongli Road" and "Number 8088"; a direction relationship between the address elements "Number 8088" and "West"; and a distance relationship exists between the address elements "Number 8088" and "200 m".

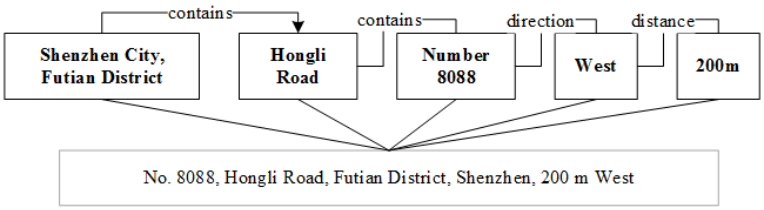


Figure 4. An example of the spatial relationship model.

According to the organization rules mentioned above, containing relationships exist among the three classes of address elements. *<Administrative name>* elements contain *<Basic constraint object>* elements; for example, province elements contain city elements, and district elements or county elements are contained in city elements. *<Basic constraint object>* elements contain *<Local point location>*

elements; for example, street numbers or building numbers are located on certain roads or blocks. In certain cases, direction relationships and distance relationships may simultaneously exist among <*Local point location*> elements; for example, in the local point location description "No. 8088 West 200 m", "West" indicates direction and "200 m" indicates distance. The combination of the local point descriptor "No. 8088", the direction "West" and the distance "200 m" specifies an accurate location; none of these elements can be omitted. Figure 5 illustrates the detailed relationships about the sample address.

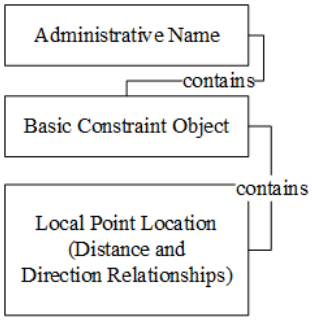


Figure 5. Relationships between address elements.

3.2. *Address Standardization*

Address standardization is the most important process involved in geocoding. In this process, an input address string parsed and then exported in a standard format. This process usually involves two steps [34]: address parsing and address normalization. The purpose of address parsing is to segment the input address string into meaningful address elements with exact spatial semantics; in address normalization, any informal or abbreviated address elements are converted into a standard format, and address elements that are written incorrectly are re-recorded in the correct format.

In this study, we employed a standardization process based on an address tree model, which was first proposed in our previously published paper [35]. The process is defined based on the following three rigorous definitions:

(1) An address is a collection of address elements and is allowed to point to multiple different spatial entities.
(2) Each address element possesses certain address semantics. The semantic meaning of an address element is the spatial object that is its real destination; this definition reflects the phenomenon that the same position may have multiple different addresses.
(3) The semantic level of an address element is defined by its class. According to this definition, administrative elements have the highest semantic level, and detailed local elements have the lowest semantic level.

As Figure 6 shows, the purpose of address standardization is to find a connected path in the semantic collection that has the correct spatial constraint relationships, and each connected path that is found can be regarded as a subtree of an address tree model. Address standardization can be suitably performed using a tree model based on the semantic characteristics of that model. The detailed standardization process is described as follows:

(1) Parse the input address string and organize it as a collection of address elements X and a semantic collection S.
(2) Create the root node, extract address element X_1, and then traverse all of the semantic elements in S_1 associated with X_1 to create address semantic nodes and connect them to the root node.
(3) Continue to extract address element X_i and traverse all of its child nodes. For example, consider the comparison of S_{i1} with the current leaf node L_i. First, perform the semantic level comparison,

and if the semantic level of S_{i1} is lower than that of L_i, then evaluate the consistency of the spatial constraint relationship between these two nodes. If this relationship is consistent, then connect S_{i1} to the current leaf node L_i.

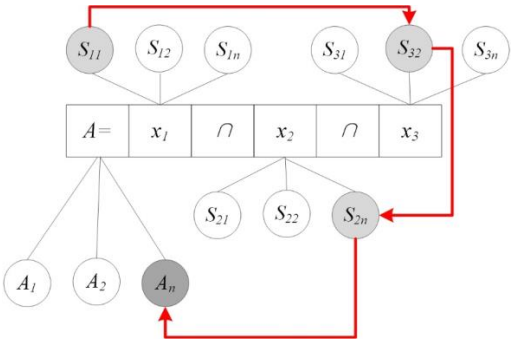


Figure 6. Relationship among addresses, address elements and address semantics.

If the consistency is not correct, the operation must be repeated, tracing back along the current subtree until another node $L_i^{'}$ is found that has a higher semantic level than S_{i1} and the consistency of these two nodes can be confirmed. Thus, Step 3 is an iterative process of comparing *Si1* with each subsequent node $L_i^{'}$; if they are consistent, S_{i1} is inserted into the current position in this subtree, and otherwise, S_{i1} must be compared with the next leaf node of the address tree.

For the same address element, if AddrLevel(S_i) $\neq$ *AddrLevel*(S_j) ($i \neq j$) and S_j has been designated as a new leaf node of the address model, then skip this leaf node.

Through this process, an input address string with confused, disorganized and even incorrect descriptions can be correctly reorganized to allow the spatial semantic structure of the address to be extracted for subsequent processing.

3.3. Address Matching Algorithm

Address matching is the process of translating an input address string into spatial coordinates [7], which can then be downloaded as a formatted file or be displayed on maps. Address matching technology is of great importance for integrating non-spatial data with spatial data and provides the ability to perform address-based mapping and spatial analysis to discover new spatial patterns and spatial correlations that are difficult to see from statistical data, for example, disease mapping and crime mapping. Address matching strives to return the most accurate matching results for any input address string. First, an attempt is made to accurately match the input address at the house number level; if no match result is found, matching will be performed at the next higher level represented in the address, such as the community, street or district level, until a result is found. Finally, the input address is assigned spatial coordinate information for mapping and spatial analysis.

Clearly, there are two possibilities when given a particular set of input address data: the existence or absence of corresponding address data in the reference database. In the first case, address matching is quite simple. In this case, the corresponding address and its geographical coordinates are found in the reference database. However, sometimes the desired address data do not exist in the reference database, perhaps because the input is a new address. In the second case, interpolation is required to find the most likely correct address. In addition, the numbers of the neighboring houses and their geographical coordinates must be determined, and the following formulas must be used to calculate the approximate geographical coordinates of the desired address [22,36]:

$$X = X_1 + \frac{X_2 - X_1}{N_2 - N_1} \times (N - N_1) \quad (1)$$

$$Y = Y_1 + \frac{Y_2 - Y_1}{N_2 - N_1} \times (N - N_1) \quad (2)$$

The variables in the formulas above are defined as follows: X_1 is the longitudinal coordinate of the first house to the right of the given house number, X_2 is the longitudinal coordinate of the first house to the left of the given house number, Y_1 is the latitudinal coordinate of the first house to the right of the given house number, Y_2 is the latitudinal coordinate of the first house to the right of the given house number, N_1 is the number of the first house to the right, N_2 is the number of the first house to the left, and N is the given house number.

If there is no house number to the right of the given house number, then N_2 is assigned the closest possible value to the given number and $X = X_2$. If there is no house number to the left of the given house number, then N_1 is the assigned the closest possible value to the given number and $X = X_1$ and $Y = Y_1$. The flow chart of the interpolation algorithm is shown in Figure 7.

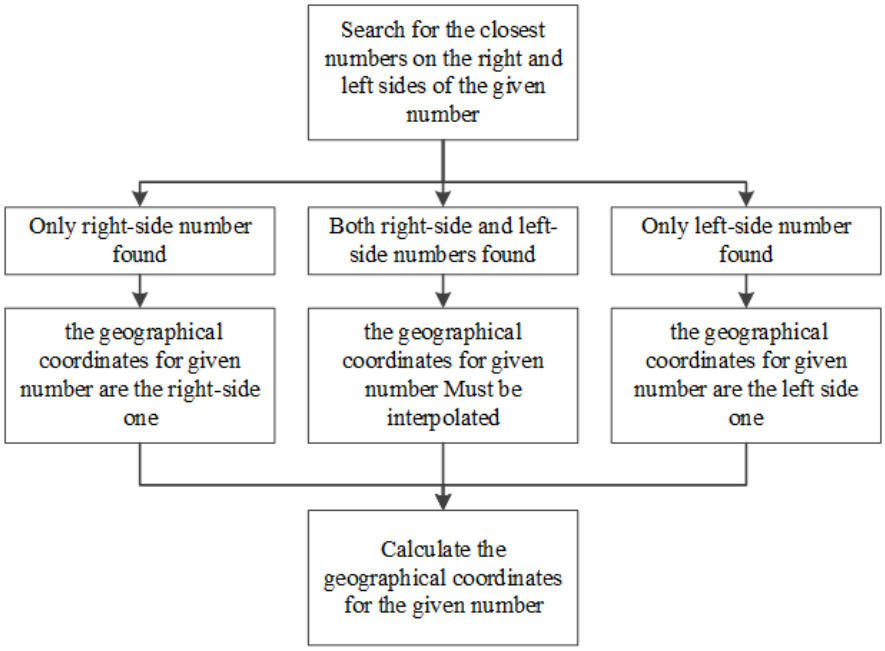


Figure 7. The flow chart of the interpolation algorithm for address matching.

4. Geocoding Service Application

To verify the addressing match methods proposed in this paper, a geocoding service system was developed for Shenzhen Municipality, China. This geocoding system uses the methods described in the previous section, including the address model, the address standardization procedure and the address matching algorithm, for Shenzhen Municipality.

The geocoding service system contains datasets collected from the SZPL. Totally, there are approximately 708,000 records. Specifically, the dataset contain 714 records of administrative names, 9475 records of district names, 2000 records of landmarks, 340,000 records of POIs, 25,000 records of roads, 250,000 records of building numbers and 81,000 records of house numbers.

4.1. Developing the Geocoding Service

The geocoding system was built for the city of Shenzhen, China using the address matching methods presented above. The system consists of five components: a reference database, an address-matching engine, a geocoding service based on the address matching-engine, an online geocoding display system, and a geocoding database management system, which are visually summarized in Figure 8.

The reference database is based on Oracle 11g Release 2 and contains an administrative region dataset, a road dataset, a POI dataset, a building numbers dataset, index data, rules data and some other auxiliary datasets.

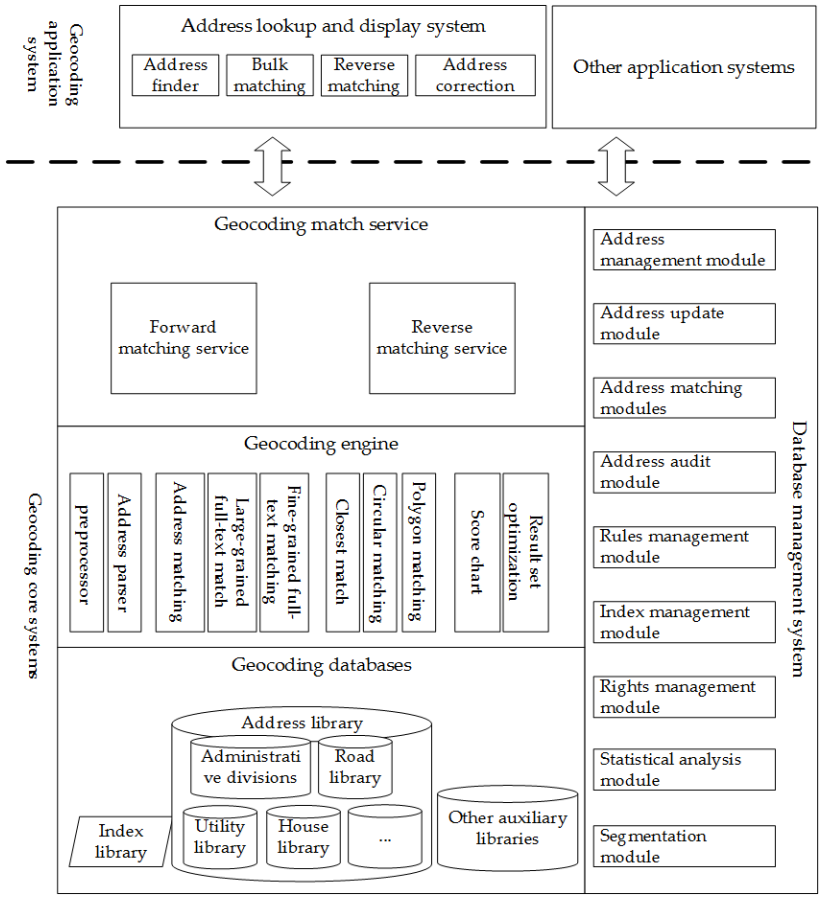
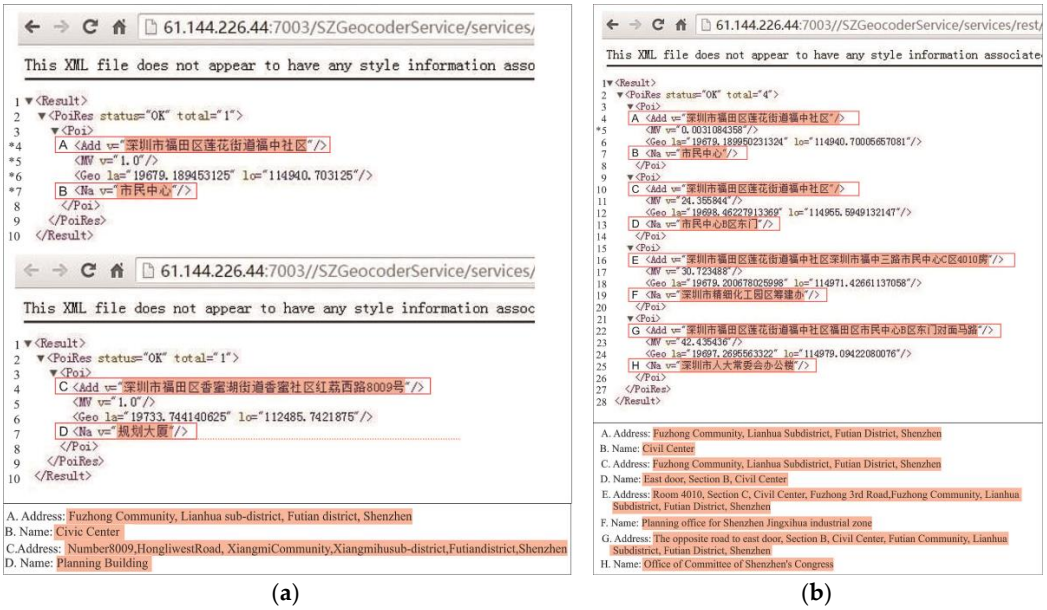


Figure 8. Architecture of the geocoding service system.



(**a**) (**b**)

Figure 9. The online geocoding service: (**a**) address matching and (**b**)reverse address matching. Note: In (**a**), * 4 shows the address matched, * 5 shows the matching degree, * 6 gives the coordinates, and * 7 gives the place name. In (**b**), * 5 gives the distance between the matched point and the point provided by the user.

The address matching engine is the most important part of the system and runs on the database described above. It consists of a preprocessor, an address parser and an address-matching module. The

preprocessor removes meaningless symbols, and the address parser attempts to parse each address input by the user in accordance with the address model and the characteristics of Chinese addresses. The address matching engine is coded in Java and uses Lucene for fuzzy and intelligent matching.

Figure 10. The online geocoding display system (accurate matching).

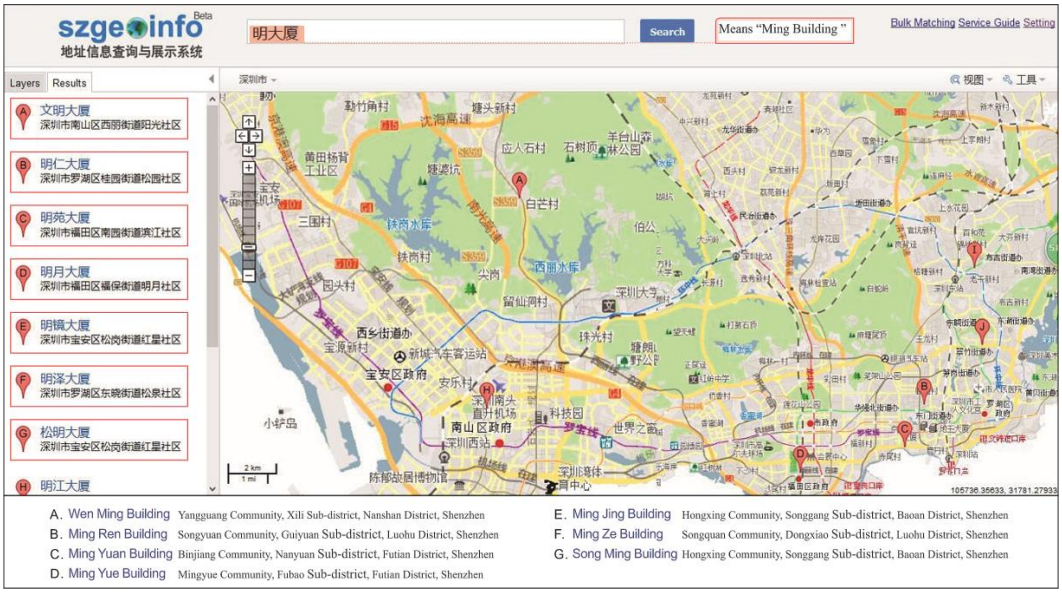


Figure 11. The online geocoding display system (fuzzy matching).

The geocoding service is based on the address-matching engine and was built using standard web services. Figure 9 shows an example of the geocoding service output. The online geocoding display system is based on the geocoding service and uses the OpenLayers library for map visualization; Figures 10 and 11 show the interface of the online geocoding display system. Finally, a management system that includes integrated mapping and data-reading functionalities, which were implemented by embedding the ArcObjects product offered by ESRI, is proven to be able to manage the data. Figure 12 shows the main user interface of the database management system.

According to the address model described in the previous section, we identified the most suitable address model for Shenzhen Municipality, which is shown in Figure 13. When accounting for the

address model, the higher stable location indicated in a descriptive address was chosen. At a finer level of detail, building numbers are usually the most important address information, followed by landmarks and POIs. Once the address model was created, it was used in the address standardization and address matching processes.

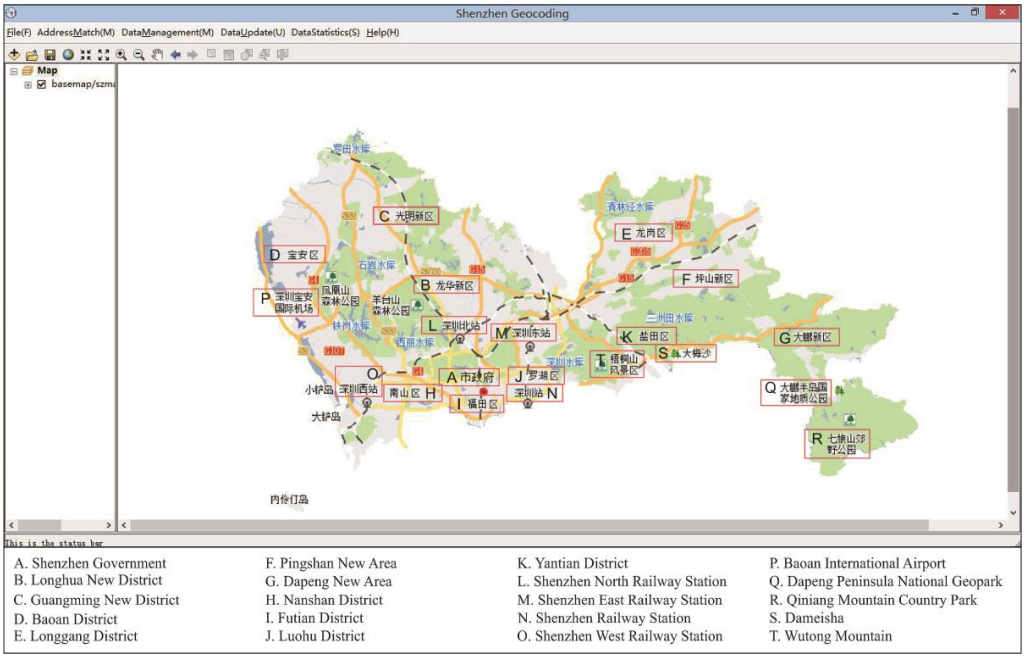


Figure 12. The geocoding management system.

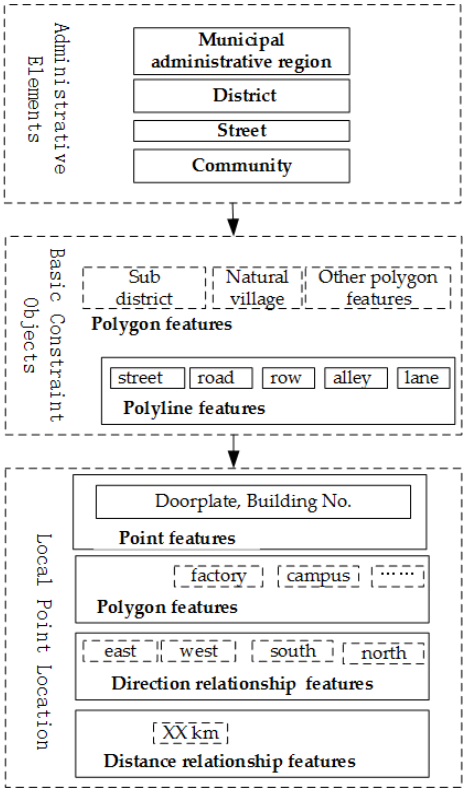


Figure 13. Address Model for Shenzhen.

4.2. Experiment and Results

The address matching degree refers to the fit of a matching address with its destination address. The candidate address of an original address can be obtained and chosen based on the candidate address that best matches the matching address. The matching address degree can be expressed as follows:

$$D = \frac{M_s}{O_s} = \frac{s\left(t_1 + t_2 + t_3 + \cdots + t_n\right)}{t_1 + x_1 + t_2 + t_3 + x_2 + \cdots + t_n + x_i} \approx \frac{s_1 + s_2 + s_3 + \cdots + s_n}{t_1 + x_1 + t_2 + t_3 + x_2 + \cdots + t_n + x_i} = \frac{s_n}{t_1 + x_1 + t_2 + t_3 + x_2 + \cdots + t_n + x_i} \quad (3)$$

where D is the matching degree, M_s is the matching address of the original address, O_s is the destination address of the original address, M is the spatial semantic set of the original address segmentation results and could be described as $s\left(t_1 + t_2 + t_3 + \ldots + t_n\right)$ or $s_1 + s_2 + s_3 + \ldots + s_n$. The most detailed address is s_n. The original address can be divided into some address elements, t means that the address element can be identified and x means that the address cannot be identified. We can calculate the matching degree from Equation (1).

As a better calculation method, we adopted the modified vector space model (M-VSM) to calculate the matching degree, which is expressed as follows:

$$D = \cos\theta = (M \cdot O) / (||M||||O||)$$

where M is the weight vector of the matching address element and O is the weight vector of original address segmentation.

In the results, a matching degree of 100% means that the address description is quite standard and fits for the address model regulation. However, several factors can reduce the matching degree, such as an incomplete reference database and unformulated addresses without relative locations or spatial direction relations.

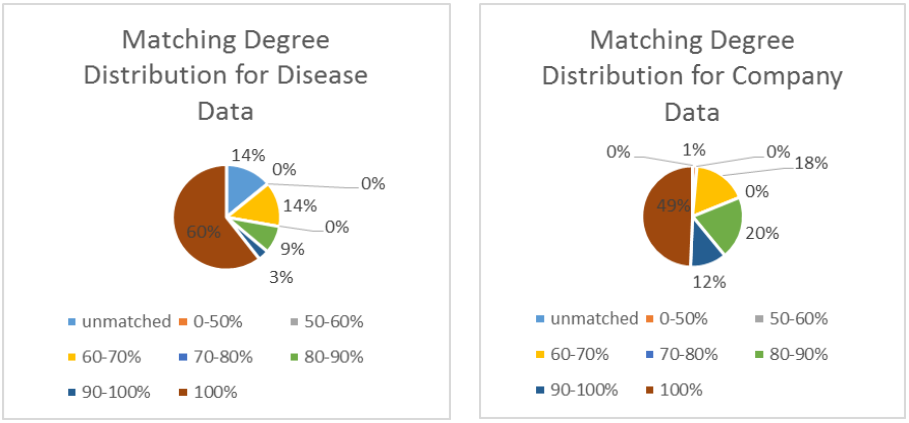
More than 100,000 records of disease data from the SCHI and approximately 1,360,000 records of company data from the SZSCJG were used to test the developed geocoding service system. In this study, we consider the building data as the reference dataset and disease data and company data as the target data. If the address of target data can be searched in the reference dataset, we assume that the target data are completely matched with a matching degree of 100%. If some elements of the target data can be searched in the reference dataset, we consider the target data as partial matching, then we calculate the matching degree for each match process and choose the best one for the target data with a matching degree between 0%–100%. However if no elements of the target data can be found in the reference dataset, the result is unmatched, with a matching degree of 0%. Thus, the higher matching degrees correspond with more accurate matching processes.

The experimental results are shown in Table 1. Figure 14 shows the matching degree distribution. Overall, the results reveal the following: (1) For the disease data, 61.1% of the records were matched with full accuracy, 3.3% of the records were matched with a matching degree of 90%–100%, 8.5% of the records were matched with a match degree of 80%–90%, and only 13.9% of the addresses had a matching degree of less than 60%. (2) For the company data, 98.6% of the records were matched with full accuracy, 81.1% of the records were matched with a matching degree of 90%–100%, 17.4% of the records were matched with a matching degree of 70%–80%, and only 1.4% of the addresses had a matching degree of less than 60%.

For a nonstandard address, such as "Hongli Road, Nanshan District, Shenzhen" (Hongli Road is not located in Nanshan District), the address standardization module will standardize and correct the address (in this case, to "Hongli Road, Futian District, Shenzhen").

Table 1. Experimental results.

Matching Degree		Unmatched	0%–50%	50%–60%	60%–70%	70%–80%	80%–90%	90%–100%	100%
Match rate	Company data	13.8%	0.1%	0%	13.88%	0%	8.5%	3.3%	60.4%
	Disease data	0.1%	1.3%	0%	17.4%	0%	20.2%	11.6%	49.3%



(**a**) (**b**)

Figure 14. Matching degree distributions: (**a**) matching degree result for the disease data; and (**b**) matching degree results for the company data.

5. Discussions

In accordance with the proposed address matching methods, a geocoding service system was developed for Shenzhen, China. To measure the effectiveness of this method, the matching degree, the matching rate, and measures of adaptability and intelligence were used to assess the quality of the address match method. The matching degree indicates the quality of a geocoded record. A record is considered to be accurately matched, inaccurately matched or unmatched. This system provides a concise score for every matched address; a higher score indicates more accurate matching, Table 2 summarizes the interpretation of the matching degree.

Table 2. Matching degree and its interpretation.

Matching Degree	Matching Level Information
100%	Matched with full accuracy
90%–100%	Matched at the building level (building numbers); Building numbers interpolation is possible
80%–90%	Matched at the block level
70%–80%	Matched at the community level
60%–70%	Matched at the road level
50%–60%	Matched at the sub-district level
50%–0%	Matched at the district level
unmatched	Unmatched at any level

The matching rate is the proportion of addresses matched successfully at a given matching degree. Figure 15 shows the matching rates determined from the experimental results.

From the experimental results, we can see that for the company data 60.4% of the addresses were matched accurately, 86% of the addresses were matched successfully at the matching degree greater than 60%, and the corresponding matching rates for the disease data were 49.3% and 98.6%, respectively, at the same matching degrees. After analyzing the matching logs, we identified two main reasons for these findings: First, the company data were recorded in a more regular manner, and the address descriptions in the disease were are highly irregular because they were recorded by different

people with different understandings of address formatting. Second, many of the addresses in the disease data are not located in Shenzhen. However, the results are accurate enough to do some research regarding public health [37,38].

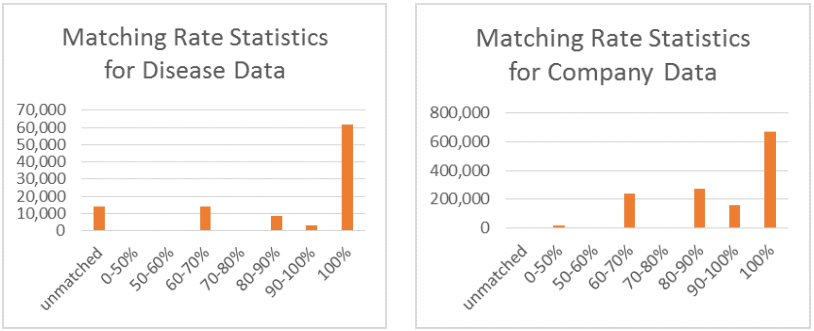


Figure 15. Matching rate statistics: (**a**) for the disease data; and (**b**) for the company data.

The developed geocoding service exhibits good performance in terms of adaptability. For example, consider the address "Hongli Road, Nanshan District, Shenzhen" (Hongli Road is not located in Nanshan District). The system can correct this address to "Hongli Road, Futian District, Shenzhen" while parsing the address using the address tree model.

Furthermore, by virtue of the address tree model (described in Section 2.2) and Apache Lucene, the solution can provide some support for fuzzy and intelligent matching. For instance, when a user inputs the address"KFC, Futian District, Shenzhen", the system will return all of the KFCs in Futian District.

Upon analyzing the results with low matching degree, we observed that most of the results with low matching degrees are house numbers, with some addresses that have changed or have a different identifier (alias). House number descriptions are highly chaotic; different people often use different formats for describing house numbers. Table 3 shows several examples of complex house numbers that make it difficult for a computer to understand the address.

Table 3. Examples of complex house numbers.

Complex House Number
深圳市南山区桃源街道龙光社区南山区龙井路光前村西区14号A\14号B En: No.14A\No.14B, West Area, GuangQian Village, LongJing Road, NanShan District, LongGuang Community, TaoYuan Street, NanShan District, Shenzhen City
深圳市南山区南头街道田厦社区南新路2108号南头影剧院综合大楼4F402 En: 4F402, Comprehensive building of Nantou Theater, No.2108, NanXin Road, TianSha Community, NanTou Street, NanShan District, Shenzhen City
福田区华发北路桑达新村22-5-202 En: 22-5-202,SangDa New Village, North Hua Fa Road, FuTian District
光明新区光明街道笔架山临时住宅37栋1号整 En: No.1, 37 Building, Temporary residence in Beacon Hill, GuangMing Street, GuangMing New District

Another deficiency of the system arises in the case of aliases or historical addresses. With the rapid development of a city, many place names and addresses may change. In addition, some administrative boundaries may change; however, the historical records still use the old place names and addresses. This issue continues to pose a significant challenge.

6. Conclusions

This paper introduced the concept of geocoding Chinese addresses and described the challenges encountered in Chinese geocoding. Then, this paper proposed an optimized address matching method for Chinese geocoding with three components: address modeling, address standardization and address

matching. Based on the proposed method, we created a geocoding service system for Chinese addresses using more than 708,000 address records from the SZPL. Finally, we conducted an experiment by using more than 100,000 records of disease data and approximately 1,360,000 records of company data. In these experiments, the average matching rate was greater than 90% for a matching degree higher than 60%. The geocoding service system has been widely used in many fields, indicating that optimized address matching method can be highly effective and accurate. The fact that our proposed system can parse not only standard addresses but also some nonstandard addresses illustrates its good adaptability. Furthermore, the proposed optimized method provides some degree of support for fuzzy and intelligent matching capabilities. Future improvements will focus on the correct handling of aliases and historical addresses as well as on better parsing of the sometimes chaotic descriptions of building numbers and house numbers.

Acknowledgments: This study was supported by the National Natural Science Foundation of China (Project No. 41271455 and 41371427). The authors would like to thank Mengjun Kang, Haijing Zhang, Yuexin Lu and Tianqi Qiu from Wuhan University for their valuable suggestions.

Author Contributions: Qin Tian, Fu Ren, Tao Hu, Jiangtao Liu, Ruichang Li and Qingyun Du worked collectively. Specifically, Qingyun Du and Fu Ren developed the original idea for the study and conducted the organization of the content. Qin Tian and Tao Hu conducted the geocoding experiments and performed the analysis of the results. Jiangtao Liu and Ruichang Li provided literature guidance and proposals for the study. All of the co-authors drafted and revised the article collectively, and all authors read and approved the final manuscript.

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

References

1. Coetzee, S.; Bishop, J. Address databases for national SDI: Comparing the novel data grid approach to data harvesting and federated databases. *Int. J. Geogr. Inf. Sci.* **2009**, *23*, 1179–1209. [CrossRef]
2. Jing, Z.; Qi, L. Research on the application of geocoding. *Geogr. Geo-Inf. Sci.* **2003**, *19*, 22–25.
3. Eichelberger, P. The importance of addresses: The locus of GIS. In Proceedings of the URISA Annual Conference, Atlanta, GA, USA, 25–29 July 1993; pp. 212–222.
4. Rhind, G.R. *Global Sourcebook of Address Data Management: A Guide to Address Formats And Data in 194 Countries*; Gower: Aldershot, UK, 1999; Volume 615.
5. Davis, C.A.; Fonseca, F.T. Assessing the certainty of locations produced by an address geocoding system. *Geoinformatica* **2007**, *11*, 103–129. [CrossRef]
6. Shah, T.I.; Bell, S.; Wilson, K. Geocoding for public health research: Empirical comparison of two geocoding services applied to canadian cities. *Can. Geogr.* **2014**, *58*, 400–417. [CrossRef]
7. Baldovin, T.; Zangrando, D.; Casale, P.; Ferrarese, F.; Bertoncello, C.; Buja, A.; Marcolongo, A.; Baldo, V. Geocoding health data with geographic information systems: A pilot study in northeast italy for developing a standardized data-acquiring format. *J. Prev. Med. Hyg.* **2015**, *56*, 88–94.
8. Ratcliffe, J.H. Geocoding crime and a first estimate of a minimum acceptable hit rate. *Int. J. Geogr. Inf. Sci.* **2004**, *18*, 61–72. [CrossRef]
9. Rushton, G.; Armstrong, M.; Gittler, J. Geocoding in cancer research: A review. *Am. J. Prev. Med.* **2006**, *30*, 16–24. [CrossRef] [PubMed]
10. Goodchild, M. GIS and transportation: Status and challenges. *GeoInformatica* **2000**, *4*, 127–139. [CrossRef]
11. Qin, X.; Parker, S.; Liu, Y.; Graettinger, A.J.; Forde, S. Intelligent geocoding system to locate traffic crashes. *Accid. Ana. Prev.* **2013**, *50*, 1034–1041. [CrossRef] [PubMed]
12. Mammadrahimli, A. Assessment of crash location improvements in map-based geocoding systems and subsequent benefits to geospatial crash analysis. In Proceedings of 94th Transportation Research Board Annual Meeting, Washington, DC, USA, 11–15 January 2015.
13. Krieger, N.; Chen, J.T.; Waterman, P.D.; Soobader, M.-J.; Subramanian, S.V.; Carson, R. Geocoding and monitoring of us socioeconomic inequalities in mortality and cancer incidence: Does the choice of area-based measure and geographic level matter?: The Public Health Disparities Geocoding Project. *Am. J. Epidemiol.* **2002**, *156*, 471–482. [CrossRef] [PubMed]

14. Shi, X. Evaluating the uncertainty caused by post office box addresses in environmental health studies: A restricted monte carlo approach. *Int. J. Geogr. Inf. Sci.* **2007**, *21*, 325–340. [CrossRef]
15. Goldberg, D.W.; Wilson, J.P.; Knoblock, C.A. From text to geographic coordinates: The current state of geocoding. *URISA J.* **2007**, *19*, 33–46.
16. 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. [CrossRef]
17. Goldberg, D.W. Advances in geocoding research and practice. *Trans. GIS* **2011**, *15*, 727–733. [CrossRef]
18. 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. [CrossRef] [PubMed]
19. Goldberg, D.W.; Ballard, M.; Boyd, J.H.; Mullan, N.; Garfield, C.; Rosman, D.; Ferrante, A.M.; Semmens, J.B. An evaluation framework for comparing geocoding systems. *Int. J. Health Geogr.* **2013**, *12*. [CrossRef] [PubMed]
20. Roongpiboonsopit, D.; Karimi, H.A. Comparative evaluation and analysis of online geocoding services. *Int. J. Geogr. Inf. Sci.* **2010**, *24*, 1081–1100. [CrossRef]
21. 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*. [CrossRef] [PubMed]
22. Goldberg, D.W. Improving geocoding match rates with spatially-varying block metrics. *Trans. GIS* **2011**, *15*, 829–850. [CrossRef]
23. Lovasi, G.S.; Weiss, J.C.; Hoskins, R.; Whitsel, E.A.; Rice, K.; Erickson, C.F.; Psaty, B.M. Comparing a single-stage geocoding method to a multi-stage geocoding method: How much and where do they disagree? *Int. J. Health Geogr.* **2007**, *6*. [CrossRef] [PubMed]
24. Ran, W.; Xuehu, Z.; Linfang, D.; Haoming, M.; Qi, L. A knowledge-based agent prototype for chinese address geocoding. In Proceedings of the Geoinformatics 2008 and Joint Conference on GIS and Built Environment: Advanced Spatial Data Models and Analyses, Guangzhou, China, 28–29 June 2008.
25. Weihong, L.; Ao, Z.; Kan, D. An efficient bayesian framework based place name segmentation algorithm for geocoding system. In Proceedings of the Fifth International Conference on Intelligent Systems Design and Engineering Applications (ISDEA), Zhangjiajie, China, 15–16 June 2014; pp. 141–144.
26. Zandbergen, P.A. A comparison of address point, parcel and street geocoding techniques. *Comput. Environ. Urban Syst.* **2008**, *32*, 214–232. [CrossRef]
27. Zhang, L.; Yi, J. A brief analysis of geocoding. In *Software Engineering and Knowledge Engineering: Theory and Practice: Volume 2*; Wu, Y., Ed.; Springer: Berlin, Germany, 2012; pp. 987–993.
28. Zhang, X.; Ma, H.; Li, Q. An address geocoding solution for Chinese cities. In Proceedings of the Geoinformatics 2006: Geospatial Information Science, Wuhan, China, 28–29 October 2006.
29. Ratcliffe, J. On the accuracy of tiger-type geocoded address data in relation to cadastral and census areal units. *Int. J. Geogr. Inf. Sci.* **2010**, *15*, 473–485. [CrossRef]
30. Gaitanis, H.; Winter, S. Is a richer address data model relevant for lbs? In *Principle and Application Progress in Location-Based Services*; Liu, C., Ed.; Springer: Berlin, Germany, 2014; pp. 121–137.
31. Chinese Academy of Surveying and Mapping. *GB/T 23705-2009: The Rules of Coding for Address in the Common Platform for Geospatial Information Service of Digital City*; Standards Press of China: Beijing, China, 2009.
32. Huanju, Y.; Qingwen, Q.; Yunling, L. Study on city address geocoding model based on street. *J. Geo-Inf. Sci.* **2013**, *15*, 175–179.
33. Haibo, Y. Techniques on Geocoding in Digital Cities and Their Applications. Master Thesis, China, University of Petroleum, Beijing, China, 2009.
34. Zhaoting, M.Z.; Zhigang, L.; Wei, S.; Jie, Y. An automatic geocoding algorithm based on address segmentation. *Bull. Surv. Map.* **2011**, *2*, 59–63.
35. Mengjun, K.; Qingyun, D.; Mingjun, W. A new method of chinese address extraction based on address tree model. *Acta Geod. Cartogr. Sin.* **2015**, *44*, 99–107.
36. Bakshi, R.; Knoblock, C.; Thakkar, S. Exploiting online sources to accurately geocode addresses. *Int. J. Geogr. Inf. Sci.* **2004**. [CrossRef]

37. Hu, T.; Du, Q.; Ren, F.; Liang, S.; Lin, D.; Li, J.; Chen, Y. Spatial analysis of the home addresses of hospital patients with hepatitis b infection or hepatoma in shenzhen, China from 2010 to 2012. *Int. J. Environ. Res. Public Health* **2014**, *11*, 3143–3155. [CrossRef] [PubMed]
38. Wang, Z.; Du, Q.; Liang, S.; Nie, K.; Lin, D.N.; Chen, Y.; Li, J.J. Analysis of the spatial variation of hospitalization admissions for hypertension disease in shenzhen, China. *Int. J. Environ. Res. Public Health* **2014**, *11*, 713–733. [CrossRef] [PubMed]