ORIGINAL ARTICLE

Geocoding health data with Geographic Information Systems: a pilot study in northeast Italy for developing a standardized data-acquiring format

T. BALDOVIN[1], D. ZANGRANDO[1], P. CASALE[2], F. FERRARESE[3], C. BERTONCELLO[1], A. BUJA[1], A. MARCOLONGO[2], V. BALDO[1]

[1] Department of Molecular Medicine, Institute of Hygiene, Laboratory of Public Health and Population Studies, University of Padua, Italy; [2] Azienda ULSS 18 Rovigo, Veneto Region, Rovigo, Italy; [3] Department of Historical and Geographic Sciences and the Ancient World, University of Padua, Italy

Key words

Environment and Public Health • Geographic Information Systems • Data matching

Summary

Introduction. *Geographic Information Systems (GIS) have become an innovative and somewhat crucial tool for analyzing relationships between public health data and environment. This study, though focusing on a Local Health Unit of northeastern Italy, could be taken as a benchmark for developing a standardized national data-acquiring format, providing a step-by-step instructions on the manipulation of address elements specific for Italian language and traditions.*

Methods. *Geocoding analysis was carried out on a health database comprising 268,517 records of the Local Health Unit of Rovigo in the Veneto region, covering a period of 10 years, starting from 2001 up to 2010. The Map Service provided by the Environmental Research System Institute (ESRI, Redlands, CA), and ArcMap 10.0 by ESRI® were, respectively, the reference data and the GIS software, employed in the geocoding process.*

Results. *The first attempt of geocoding produced a poor quality result, having about 40% of the addresses matched. A procedure of manual standardization was performed in order to enhance the quality of the results, consequently a set of guiding principle were expounded which should be pursued for geocoding health data. High-level geocoding detail will provide a more precise geographic representation of health related events.*

Conclusions. *The main achievement of this study was to outline some of the difficulties encountered during the geocoding of health data and to put forward a set of guidelines, which could be useful to facilitate the process and enhance the quality of the results. Public health informatics represents an emerging specialty that highlights on the application of information science and technology to public health practice and research. Therefore, this study could draw the attention of the National Health Service to the underestimated problem of geocoding accuracy in health related data for environmental risk assessment.*

Introduction

In the past decade, Geographic Information Systems (GIS) have become an innovative and somewhat crucial tool for analysing relationships between public health data and environment. According to Longley et al. (2005), a GIS is an application-led technology, which can be used, in this instance, for monitoring and understanding observed spatial distribution of attributes such as the geography of environmental health [1]. Thus, it is required to transpose data stored in a health care related database to a spatial related database, assigning to each record a univocal spatial location (X, Y Coordinates). This procedure is known as geocoding. Its role continues to grow and evolve as new forms of geocoding emerge and as geocoded data are applied to an ever-diverse set of spatially based investigations [2]. Geocoding technology has been applied in many fields: social, political, and economic and more recently in public health research and practice. In general, these applications are related to interpolation rather than matching. This is because interpolation requires a lower level of accuracy in data manipulation. Clustering, aggregation, spatial smoothing are typical applications in epidemiology. The literature provides many examples focused on the surveillance of infectious and chronic diseases [3-5], environmental exposures [6, 7], drinking water epidemiology [8, 9] and pharmacoepidemiology [10].

In general, the limitation is the spatial accuracy of the geographic location computed for any particular subject. Accuracy represents an important issue particularly in Italy due to the complexity of the address name and street morphology. Address complexity and street morphology depend on historical heritage of Italy, so georeferencing requires additional manipulation of place names data. In this scenario, the improvement of geocoding accuracy plays a key role in developing a reliable tool for public health research in Italy. The main aim of this paper is to describe the procedures involved in the conversion from database collected data into geocoded dots representing a health event, in order to display the spatial distribution of five major chronic-degenerative diseases within the Rovigo Local Health Unit (ULSS N.18) in the Rovigo

Province (Veneto Region - North Italy). The highly accurate geographic localization of patients, served by the Local Health Unit, will widen the range of opportunities for further spatial analysis and modelling, such as environmental related hazard or monitoring the prevalence of certain diseases through time, gender or age. This paper therefore provides a step-by-step instructions on the standardization of address elements specific for Italian language and traditions.
A description of the methodology involved in the geocoding process will be as crucial as outlining some guidelines for a standardized method, strongly required for the data collection component, involving local addresses. This study, though focusing on the Local Health Unit of Rovigo, could be taken as a benchmark for developing a standardized national data-acquiring format.

Methods

The Study Population

The Local Health Unit of Rovigo (ULSS N.18) has collected data on their catchment area, made up of 41 municipalities, over a period of 10 years, starting from 2001 up to 2010, and stored it in Microsoft *Access*™ database format, for 268,517 records (Fig. 1).
It is crucial to understand that 268,517 is the total amount of records collected in the above-mentioned period, while the population census, provided by the National Institute of Statistics (ISTAT) at 1st of January 2011 was 175.816 persons residing in those 41 Municipalities.
For this study's purpose, the year 2010 has been chosen as sample group, hence the provided data had to be checked and sorted carefully. By means of SQL query language, a sequence of selection criteria has been applied on the original database and the amount of valid records significantly shrank, from 268,517 to 178,183. This decrement is ascribed to the number of subjects whose status was 'deceased' or 'transferred' to a different ULSS up to the 31st of December 2009, as well as to those subjects who have their domicile outside the ULSS of Rovigo. To obtain a consistent sample, it has been decided to remove those subjects having a residence address within the ULSS of Rovigo, yet having their permanent address outside of it. Forthcoming analysis will use this data for mapping and clustering health events associated to environmental hazards, therefore it is assumed that domicile related data has a greater deal of truthfulness compared to residence data [11]. The difference of 2,367 subjects between the census data and the collected data lies in the amount of persons residing in a different municipality, though attending to the Local Health Unit of Rovigo.
For personal data protection policy, a progressive sequential unique identifier, linking to a different database, has replaced all information regarded as sensible data. Further information stored were gender, date of birth and mostly important, an alphanumerical code (Tax Code) for personal and unambiguous identification issued by the National Health Service (NHS). Moreover, a numeric code is included, which identifies the current status of patients within the Local Health Unit (LHU), for instance, if the patient is active, transferred to a different ULSS or dead.

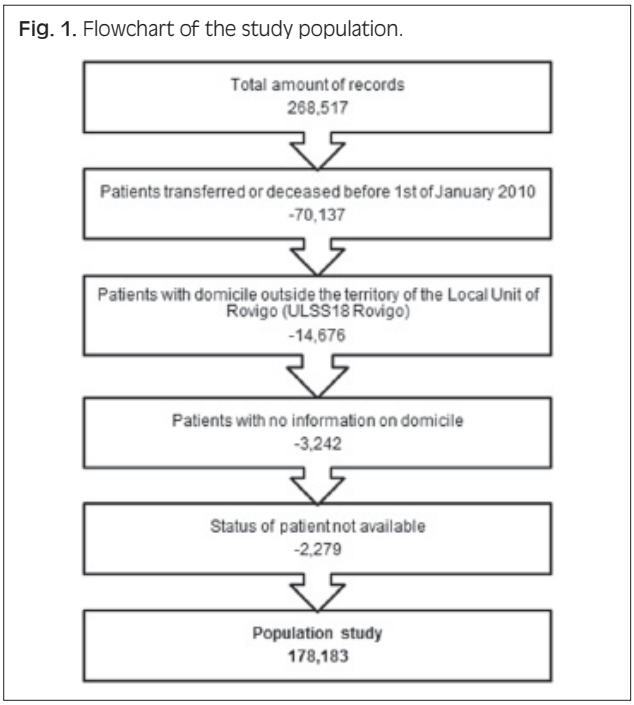


Fig. 1. Flowchart of the study population.

The Geocoding Process

The geocoding process involves converting a string information, such as a street names, town or place name, into geographic features located on the earth's surface, which can be spatially displayed within a GIS. Finding good quality up-to-date reference data becomes a crucial point, hence different commercially available street network databases have been weighed. The Map Service, named World Street Map 2010, provided by the Environmental Research System Institute (ESRI, Redlands, CA), and ArcMap 10.0 by ESRI® were, respectively, the reference data and the GIS software, employed in the geocoding process.
Within the tools available for geocoding, the Geocode Address Tool was implemented as it allows for the geocoding of a table of addresses. However, in order to match the addresses in the input table, this tool needs to link to the reference data using a service provided by the ESRI Address Locator. It was opted for the TA_Address_EU.GeocodeServer locator, specific for the European Zone, where the domicile address is parsed into 4 syntactic components. The fields required by the operation are:

1. Address: Street name with suffix type (e.g. road, avenue, boulevard) and house number.
2. City: the name of the municipality.
3. Postcode: a 5 number code (related to one or more Municipalities).
4. Country: the code of the country of domicile, in this instance ITA.

The result of a geocoding process is an output table returning the addresses with a score of the probability of

Tab. I. Tools employed for manual localization of addresses.

Web Map Service (WMS)	Google Maps Google Street View	http://maps.google.it/ http://maps.google.it/	GoogleMaps - ©2012 Google GoogleMaps - ©2012 Google
	VirgilioMappe	http://mappe.virgilio.it/	Matrix® S.p.A.
	Tuttocittà	http://www.tuttocitta.it/	Navteq© Ximage®
Online Telephone Directories	White Pages	http://www.paginebianche.it/	Seat Pagine Gialle© S.p.A.
	Yellow Pages	http://www.paginegialle.it/	Seat Pagine Gialle© S.p.A
	Pronto Comune	http://www.prontocomune.it/	Società Editrice Europea® Srl
Online City Maps	Geoplan	http://www.geoplan.it/	Geoplan© S.r.l.

having matched the correct location. In fact, geocoding is a probabilistic system, where each field participating in the linkage comparison is subject to error and is measured by the probability that the field agrees versus the probability of chance agreement of its values [11]. Consequently, two fields are generated in the output table showing the match type and the score: the former indicates whether there was a match (M), an unmatched result (U), or tied results (T), which requires to be manually checked by the operator. On the other hand, the score is expressed as the percentage of having identified the best possible candidate of the address within the reference data.
For the manual localization of those addresses not matched automatically, a wide range of open source resources have been employed (Tab. I).

Results

The first attempt of geocoding produced a poor quality result having about 60% of the addresses tied while only 40% matched. After having manually verified those records it was noticed that although the street name was present in both, the reference data and the input table, it was spelled the other way round. A procedure of manual standardization was performed in the input table, so that all the streets name were spelled correspondingly to those in the reference data.
After carrying out the previously mentioned adaptations, the result of the geocoding had significantly improved, reaching almost 98% of matched (M) records and 2% of tied (T) records, yet this outcome does not reflect the real precision of the result. In fact, as previously explained, the address is parsed into 4 components, and a match (M) result is achieved every time just 2 of these are met by the query. As a result, 3 different levels of matching precision have been identified, depending on the number of the address components available during the geocoding process. Therefore, when merely the City and Country fields are matched the M type is specified as EU_City.ITA; likewise EU_PostCity.ITA identifies those records where only the Postcode and Country proved to correspond, whilst EU_StreetName.ITA refers to those record matched at a street name level.
As shown in Table II, the percentage of EU_StreetName.ITA addresses matched is 90.9% while the addresses geocoded at a city and postcode level are respectively 0.18% and 6.8%.

Tab. II. The percentage of addresses geocoded according to match type.

Match type	Match level	Match score (%)	No. of addresses	% of addresses
U - (Unmatched)	none	none	none	
T - (Tied)	EU_City.ITA	100	24	0.01%
T - (Tied)	EU_Street_Name.ITA	≤ 69 70-99 100	423 45 3,137 **3,605**	2.02%
M - (Matched)	EU_City.ITA	≥ 90	333	0.18%
M - (Matched)	EU_PostCity.ITA	≤ 99 100	137 11,981 **12,118**	6.8%
M - (Matched)	EU_Street_Name.ITA	≤ 69 70-79 80-89 90-99 100	3,859 2,885 3,038 1,100 151,221 **162,103**	90.98%
			Total 178,183	100%

Tab. III. Parcelling of the addresses for geocoding.

Attribute Name	Street Prefix	Street Name	House Number	Unit Number	Postcode	Municipality Code	Municipality Name	Country Code
Example	Borgo Contrada Corso Galleria Largo Località Piazza/Piazzale Via/Viale Vicolo Villaggio Strada Zona Etc..	Title (if available: grade or clerical rank with no abbreviation) + space key + Name (in extenso) + space key + Surname No article No preposition No apostrophe	House number of the Building	Apartment or sub-unit number in Arabic Numeral or Roman Numeral or Letters of the Alphabet	Postal Code (ambiguous) Five numbered code	Provided by the National Institute for Statistics (Istat) (unambiguous) Five numbered code	Name of the Municipality (in extenso)	ITA
Data Type	String	String	Short Integer	String	Short Integer	Short Integer	String	String
Geocoding elements	✓ (as a single string)		✓	×	✓	×	✓	✓

Successively, addresses matched at a street name level, were weighed against the score achieved during geocoding. Locations that yield a score of ≥ 90 were considered a good match whilst those with score ≤ 89, approximately 9,700 records, required be checking individually and adjusting by hand. Though assuming the correctness of those 151,221 results having a match at a street name level and a score ≥ 90, it was opted to verify if the addresses did actually coincide with the true location on the map. For this purpose, the Municipality of Rovigo was chosen as the sample unit, since it is the largest municipality with the highest number of people attending to the LHU. The 32% of the above mentioned addresses, that is roughly 48,000 patients, fall within the administrative boundaries of Rovigo and their geometry has been checked using the Intersect tool of ArcInfo®.

The point feature class, representing the patients' domicile, was intersected with the line feature class of all the road segment attributes, and a new point feature dataset was generated which includes, for each address, the name of the street segment that was overlapped. After carrying out a SQL query on 48,615 records, as many as 9,165 patients appeared to have the domicile address matched to the wrong street segment. However, after a double-check it was realized that differences were caused mainly by the presence/absence of the apostrophe, article or capital letters in the street's name, yet the correctness of the match was not compromised. Only 324 records, equal to the 0.7%, were wrongly matched since the error was caused by the street names, in the line feature class, being more up-to-date than the address associated to the patient's domicile stored in the LHU database.

On the other hand, a procedure of manual geocoding had to be carried out for those addresses with a city or postcode level match (M), for a total amount of 12,451 patients. That is to say, the geocoding process was not able to assign a street segment to the address, therefore positioned the patient's location at the geometric center of the Municipality. This entailed to seek for the correct coordinates by means of several web mapping service applications, online telephone directories and, in the most difficult cases, even by contacting the Municipality office.

Furthermore, all results with a tied (T) match type were assessed individually to remove any uncertainty; for this, only 447 proved wrongly geocoded and the right coordinates have been assigned manually.

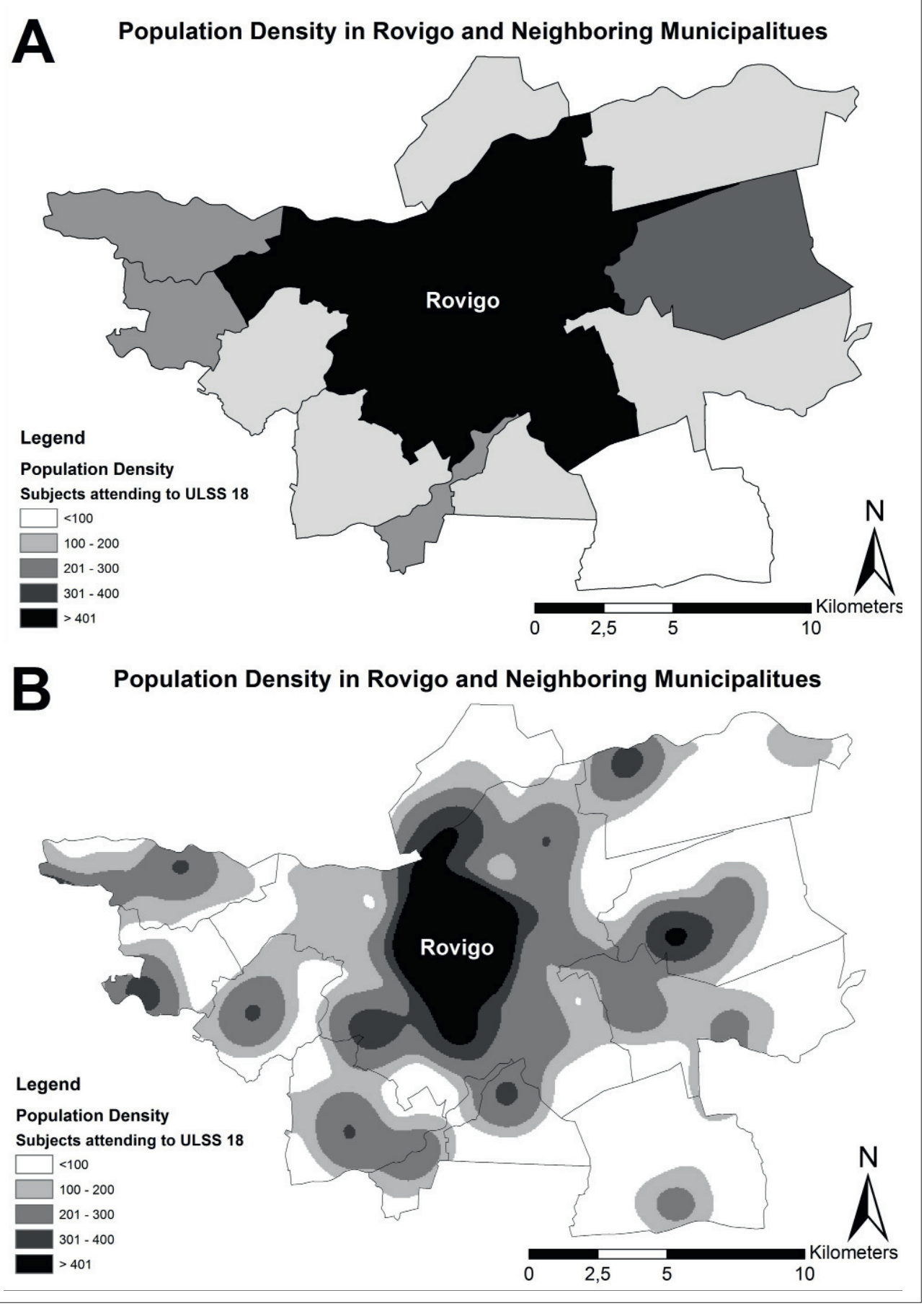
The guiding principles, which should be pursued for preparing data for geocoding, are summarized and exposed in Table III. These, however, do referred strictly to the ESRI World Street Map web service, used as the main reference data, to which addresses should conform.

In a health related prospect, the result of having a high detail geocoded population could increase the perception and comprehension of the distribution of any given health event, thus allowing for an administrative boundless view. As showed in Figure 2, the population density can be represented with neat lines, according to the Municipality geographical layout (Fig. 2 part A), or boundless, that is to say representing any event, in this case the permanent address of the case study subjects, with no constrains associated to human made frontiers (Fig. 2 part B). This instance could be applied when mapping the spatial distribution of some chronic-degenerative diseases, such as asthma, and analyzing if a given pattern could be linked to an environmental hazard, such as air pollution or the proximity to dumps, industries, incinerator, etc.

Discussion

Ecological studies are not based on individual but on aggregated disease and exposition data [12]. The prospect of a wide range of spatially related Health analysis, such as disease clustering and risk exposure to environmental hazards, on a large number of patients, was the leading endeavor of this project. Out of the 286,517 records pro-

Fig. 2. Difference between geocoding founded on a municipality boundary basis and a street level basis.



vided by the Local Health Unit of Rovigo, 178,183 had the prerequisites for being assessed in this study, having lost 2% due to void entries in the database and 31.6% of the data being irrelevant for the study area.
The standardization procedures of the address database, produced a result of 162,103 matched and 3,605 tied addresses, both at a street name level, equal to the 93% of all records. For the remaining 7%, that is 12,475 addresses, matched or tied at a city or postcode level, the geocoding had to be performed manually by the operator, using a wide range of open source data and, if necessary, with the aid of the Municipalities involved.
The geocoding process revealed a dearth of homogeny between the address labeling used by the Local Health Unit, during the registration procedures of patients, and the label attributes of the street segments in the reference data. In particular, there was a conflicting approach in using abbreviations for streets named after Saints, clergy characters and military ranks, as well as the reverse writing of historical figures names, for instance, surname-name order in the input data and name-surname order in the reference data. Furthermore, streets named in memory of historical dates were written in Arabic numeral and in Roman numeral, respectively.
Overall, a lack of consistency in the approach of storing personal data has emerged. In particular, the street prefixes were stored with a variety of abbreviations leading to ambiguity.
As far as the street name is concerned, the name and the title, when included, should not be shortened, as it will result in misspelling errors or in homonymy. Furthermore, neither article, nor preposition, nor apostrophe should be included.
With respect to the municipality details, a few points should be considered: firstly, the name should be written in full length to avoid false mismatch and, secondly, it should always be coupled to the postcode. Unlike in the United States, where the U. S. Postal Service (USPS) uses a zoning improvement plan (ZIP) code as a postal addressing standard [13], in Italy the post code does not serve as an unambiguous identifier, hence more than one municipality can have the same post code.
Health data should be collected originally with compliance to a set of well-defined parameters, if possible using a menu-driven interface with drop-down lists to facilitate users by selecting among a list of pre-compiled values. Misspelling errors of streets, for instance, could be reduced considerably, as well as gross inconsistencies between the municipalities' name and postcodes. The National Health Service (SSN) should consider acquiring a common program and standardize parameters to collect health data.
The main achievement was to outline some of the difficulties encountered during the geocoding of Health data and to put forward a set of guidelines that could be useful to facilitate the process and enhance the quality of the results.
On the other hand, some limitations of this study should be considered. First, given the massive amount of records that had to be geocoded, it was opted to ignore the house number, as it would introduce an additional time consuming and labor-intensive effort to locate manually the wrongly matched addresses. Georeferencing with street centerline data can affect location accuracy, since it introduces many assumptions including the equal parsing of addresses along a road network and the uniform distancing of houses from the road network [14].
Second, there was no possibility to account for the positional accuracy of the results obtained by the use of the ESRI StreetMap as reference data. Accordingly to Zhan et al. (2006), the validity of epidemiologic research depends on the match rate of geocoding (the percentage of addresses geocoded), as well as the positional accuracy of locations of geocoded addresses. Thus, in this study it was not possible to calculate the positional accuracy, defined as the difference between the geographic location of a geocoded address and the ''true'' ground location of that address determined by using a field survey method, i.e., surveying using a global positioning system (GPS) device [15].
According to a recent study, the current state-of-practice lacks of standard resources for geocoding, geocoding accuracy assessment, and for evaluating the impacts of geocoding error on public health decisions [16]. Even though in the last decade several studies have been carried out on the accuracy of geocoded data [14, 17, 18]. No studies have addressed the completeness and accuracy of the reference street network database in Italy.
As a matter of fact, no research has been found in literature which evaluates the topic of geocoding methods in this country, although implementing address coded data in epidemiology research is becoming rather frequent [19, 20].
The research project will now proceed by evaluating risk exposure to environmental hazards, for instance air pollution, and the spatial distribution of some chronic-degenerative diseases, such as asthma, linking health data to the georeferenced patients in the Local Health Unit of Rovigo. Forthcoming results will be soon expounded.

Conclusions

The main achievement of this study was to outline some of the difficulties encountered during the geocoding of Health data and to put forward a set of guidelines that could be useful to facilitate the process and enhance the quality of the results. Health data should be collected originally with compliance to a set of well-defined parameters, if possible using a menu-driven interface with drop-down lists to facilitate users by selecting among a list of pre-compiled values and avoid misspelling bias. Public health informatics represents an emerging specialty that highlights on the application of information science and technology to public health practice and research. Therefore, this study could draw the attention of the National Health Service of Italy to the underestimated problem of geocoding accuracy in health related data for environmental risk assessment.

References

[1] Longley PA, Goodchild MF, Maguire DJ, et al. *Geographic Information Systems and Science*. Second Edition. England: John Wiley & Sons Ltd. 2005.

[2] Goldberg DW, Jacquez GM. *Advances in geocoding for the health sciences.* Spat Spatiotemporal Epidemiol 2012;3: 1-5.

[3] Musa GJ, Chiang PH, Sylk T, et al. *Use of GIS Mapping as a Public Health Tool-From Cholera to Cancer.* Health Serv Insights 2013;6:111-6.

[4] Miranda ML, Casper M, Tootoo J, et al. *Putting chronic disease on the map: building GIS capacity in state and local health departments.* Prev Chronic Dis 2013;10:E100.

[5] Carroll LN, Au AP, Detwiler LT, et al. *Visualization and analytics tools for infectious disease epidemiology: a systematic review.* J Biomed Inform 2014;51:287-98.

[6] Xie Y, Chen TB, Lei M, et al. *Spatial distribution of soil heavy metal pollution estimated by different interpolation methods: accuracy and uncertainty analysis.* Chemosphere 2011;82:468-76.

[7] Wang W, Ying Y, Wu Q, et al. *A GIS-based spatial correlation analysis for ambient air pollution and AECOPD hospitalizations in Jinan, China.* Respir Med 2015;109:372-8.

[8] Dangendorf F, Herbst S, Reintjes R, et al. *Spatial patterns of diarrhoeal illnesses with regard to water supply structures - a GIS analysis.* Int J Hyg Environ Health 2002;205:183-91.

[9] Nas B, Berktay A. *Groundwater quality mapping in urban groundwater using GIS.* Environ Monit Assess 2010;160:215-27.

[10] Dijkstra A, Hak E, Janssen F. *A systematic review of the application of spatial analysis in pharmacoepidemiologic research.* Ann Epidemiol 2013;23:504-14.

[11] Zandbergen PA. *A comparison of address point, parcel and street geocoding techniques.* Comput Environ Urban Syst 2008;32:214-32.

[12] Kistemann T, Dangendorf F, Schweikart J. *New perspectives on the use of Geographical Information Systems (GIS) in environmental health sciences.* Int J Hyg Environ Health 2002;205: 169-81.

[13] USPS, 2000. Postal Addressing Standards. Unites States Postal, Service. Publication 28. Updated April 2010; Available from: http://pe.usps.gov/cpim/ftp/pubs/Pub28/pub28.pdf

[14] Dearwent SM, Jacobs RR, Halbert JB. *Locational uncertainty in georeferencing public health datasets.* J Expo Anal Environ Epidemiol 2001;11:329-34.

[15] Zhan FB, Brender JD, De Lima I, et al. *Match rate and positional accuracy of two geocoding methods for epidemiologic research.* Ann Epidemiol 2006;16:842-9.

[16] Jacquez GM. *A research agenda: does geocoding positional error matter in health GIS studies?* Spat Spatiotemporal Epidemiol 2012;3:7-16.

[17] Bonner MR, Han D, Nie J, et al. *Positional accuracy of geocoded addresses in epidemiologic research.* Epidemiology 2003;14:408-12.

[18] Zandbergen PA, Hart TC, Lenzer KE, et al. *Error propagation models to examine the effects of geocoding quality on spatial analysis of individual-level datasets.* Spat Spatiotemporal Epidemiol 2012;3:69-82.

[19] Girardi P, Marcon A, Rava M, et al. *Spatial analysis of binary health indicators with local smoothing techniques The Viadana study.* Sci Total Environ 2012;414:380-6.

[20] Nuvolone D, Della Maggiore R, Maio S, et al. *Geographical information system and environmental epidemiology: a cross-sectional spatial analysis of the effects of traffic-related air pollution on population respiratory health.* Environ Health 2011;10:12.

■ Received on April 15, 2015. Accepted on May 27, 2015.

■ Correspondence: Tatjana Baldovin, Department of Molecular Medicine, Institute of Hygiene, Laboratory of Public Health and Population Studies, University of Padua, Via Loredan 18, 35131 Padova, Italy - Tel. +39 049 8275390/96 - Fax +39 049 8275392 - E-mail: tatjana.baldovin@unipd.it