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

Original Citation:
Geocoding health data with Geographic Information Systems: a pilot study in northeast Italy for developing a standardized data-acquiring format / Baldovin, Tatjana; Zangrando, Dorelia; Casale, P; Ferrarese, Francesco; Bertoncello, Chiara; Buja, Alessandra; Marcolongo, A; Baldo, Vincenzo. - In: JOURNAL OF PREVENTIVE MEDICINE AND HYGIENE. - ISSN 1121-2233. - ELETTRONICO. - 56:2(2015), pp. E88-E94.

Availability:
This version is available at: 11577/3203307 since: 2016-10-25T16:00:33Z

Publisher:
Pacini Editore

Published version:
DOI:

Terms of use:
Open Access
This article is made available under terms and conditions applicable to Open Access Guidelines, as described at http://www.unipd.it/download/file/fid/55401 (Italian only)

(Article begins on next page)
Geocoding health data with Geographic Information Systems: a pilot study in northeast Italy for developing a standardized data-acquiring format

T. BALDOVIN1, D. ZANGRANDO1, P. CASALE2, F. FERRARESE3, C. BERTONCELLO1, A. BUJA1, A. MARCOLONGO3, V. BALDO1

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

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.
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.

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.GecodeServer 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
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%.

<table>
<thead>
<tr>
<th>Tab. I. Tools employed for manual localization of addresses.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Web Map Service (WMS)</strong></td>
</tr>
<tr>
<td>Google Maps</td>
</tr>
<tr>
<td>Google Street View</td>
</tr>
<tr>
<td>VirgilioMappe</td>
</tr>
<tr>
<td>Tuttocittà</td>
</tr>
<tr>
<td><strong>Online Telephone Directories</strong></td>
</tr>
<tr>
<td>White Pages</td>
</tr>
<tr>
<td>Yellow Pages</td>
</tr>
<tr>
<td>Pronto Comune</td>
</tr>
<tr>
<td><strong>Online City Maps</strong></td>
</tr>
<tr>
<td>Geoplan</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Tab. II. The percentage of addresses geocoded according to match type.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Match type</strong></td>
</tr>
<tr>
<td>U - (Unmatched)</td>
</tr>
<tr>
<td>T - (Tied)</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>T - (Tied)</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>M - (Matched)</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Total</td>
</tr>
</tbody>
</table>
Geocoding health data with GIS

positioned the patient’s location at the geometric center
able to assign a street segment to the address, therefore
patients. That is to say, the geocoding process was not
postcode level match (M), for a total amount of 12,451
had to be carried out for those addresses with a city or
On the other hand, a procedure of manual geocoding
database.
line feature class, being more up-to-date than the address
324 records, equal to the 0.7%, were wrongly matched
correctness of the match was not compromised. Only
phe, article or capital letters in the street’s name, yet the
differences were
double-check it was realized that differences were
as 9,165 patients appeared to have the domicile address
carrying out a SQL query on 48,615 records, as many
set was generated which includes, for each address, the
road segment attributes, and a new point feature data-
case study subjects, with
the permanent address of the case study subjects, with
nicipality geographical layout (Fig. 2 part A), or bound-
to an environmental hazard, such as air pollution or the proximity to dumps, industries,
incinerator, etc.

Successively, addresses matched at a street name level,
considered a good match whilst those with score ≤ 89,
9,700 records, required being checked 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
ture 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 boundary less 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 boundaryless, 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-

---

Tab. III. Parcelling of the addresses for geocoding.

<table>
<thead>
<tr>
<th>Attribute Name</th>
<th>Street Prefix</th>
<th>Street Name</th>
<th>House Number</th>
<th>Unit Number</th>
<th>Postcode</th>
<th>Municipality Code</th>
<th>Municipality Name</th>
<th>Country Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example</td>
<td>Borgo</td>
<td>Contrada</td>
<td>Corso</td>
<td>Largo</td>
<td>Località</td>
<td>Piazza</td>
<td>Vicolo</td>
<td>Zona</td>
</tr>
<tr>
<td></td>
<td>Strada</td>
<td>Viale</td>
<td>Villaggio</td>
<td>Strada</td>
<td>Etc.</td>
<td>Title</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(if available: grade or clerical rank with no abbreviation)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+ space key</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+ Name (in extenso)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+ space key</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+ Surname</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>No article</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>No preposition</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>No apostrophe</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>House number of the Building</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Apartment or sub-unit number in Arabic Numeral or Roman Numeral or Letters of the Alphabet</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Postal Code (ambiguous) Five numbered code</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Provided by the National Institute for Statistics (Istat) (unambiguous) Five numbered code</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Name of the Municipality (in extenso)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Geocoding elements: Yes (as a single string)
Fig. 2. Difference between geocoding founded on a municipality boundary basis and a street level basis.

A

Population Density in Rovigo and Neighboring Municipalities

Legend

Population Density
Subjects attending to ULSS 18

<100
100 - 200
201 - 300
301 - 400
> 401

B

Population Density in Rovigo and Neighboring Municipalities

Legend

Population Density
Subjects attending to ULSS 18

<100
100 - 200
201 - 300
301 - 400
> 401
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 homogeneity 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 distancings 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


