Quality of automatic geocoding tools

Quinteros, Maria Elisa; Blazquez, Carola; Rosas, Felipe; Ayala, Salvador; García, Ximena Marcela Ossa; Delgado-Saborit, Juana Maria; Harrison, Roy M.; Ruiz-Rudolph, Pablo; Yohannessen, Karla

DOI:
10.1590/0102-311X00288920

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (Harvard):
Quinteros, ME, Blazquez, C, Rosas, F, Ayala, S, García, XMO, Delgado-Saborit, JM, Harrison, RM, Ruiz-Rudolph, P & Yohannessen, K 2022, 'Quality of automatic geocoding tools: a study using addresses from hospital record files in Temuco, Chile', *Cadernos de Saude Publica*, vol. 38, no. 1, e00288920. https://doi.org/10.1590/0102-311X00288920

CSP CADERNOS DE SAÚDE PÚBLICA
REPORTS IN PUBLIC HEALTH

ARTIGO
ARTICLE

Quality of automatic geocoding tools: a study using addresses from hospital record files in Temuco, Chile

Calidad de las herramientas automáticas de geocodificación: un estudio usando direcciones de fichas de registro hospitalario en Temuco, Chile

A qualidade das ferramentas de geocodificação automática: um estudo utilizando endereços de prontuários hospitalares em Temuco, Chile

Maria Elisa Quinteros [1]
Carola Blazquez [2]
Felipe Rosas [3]
Salvador Ayala [4]
Ximena Marcela Ossa García [3]
Juana Maria Delgado-Saborit [5,6]
Roy M. Harrison [6]
Pablo Ruiz-Rudolph [4]
Karla Yohannessen [4]

doi: 10.1590/0102-311X00288920

Abstract

Automatic geocoding methods have become popular in recent years, facilitating the study of the association between health outcomes and the place of living. However, rather few studies have evaluated geocoding quality, with most of them being performed in the US and Europe. This article aims to compare the quality of three automatic online geocoding tools against a reference method. A subsample of 300 handwritten addresses from hospital records was geocoded using Bing, Google Earth, and Google Maps. Match rates were higher (> 80%) for Google Maps and Google Earth compared with Bing. However, the accuracy of the addresses was better for Bing with a larger proportion (> 70%) of addresses with positional errors below 20m. Generally, performance did not vary for each method for different socioeconomic status. Overall, the methods showed an acceptable, but heterogeneous performance, which may be a warning against the use of automatic methods without assessing quality in other municipalities, particularly in Chile and Latin America.

Geographic Mapping; Residence Characteristics; Spatial Analysis

Correspondence
P. Ruiz-Rudolph
Universidad de Chile.
Independencia 939, Independencia, Santiago, Chile.
pabloruizr@uchile.cl

[1] *Universidad de Talca, Talca, Chile.*
[2] *Universidad Andrés Bello, Viña del Mar, Chile.*
[3] *Universidad de la Frontera, Temuco, Chile.*
[4] *Universidad de Chile, Santiago, Chile.*
[5] *Universitat Jaume I, Castellón, España.*
[6] *University of Birmingham, Birmingham, U.K.*

Introduction

Knowing the spatial distribution of certain attributes, health determinants or conditions of individuals or populations has helped researchers and policy-makers to monitor and to understand some important relationships between public health and people's environments [1]. In the last decades, Geographic Information Systems (GIS) have been increasingly used in environmental [2,3,4], nutritional [5,6], and social epidemiological studies [7], as well as in public health research and practice [2,8,9,10,11]. Thus, the transformation of a written address into spatial information, i.e., geocoding, is essential and has become an important methodology to locate people and services, among others [4,8,11,12].

Address geocoding describes the process of spatially locating an address by finding the coordinate that best fits its physical location on a map [3,9,10,11,13]. Geocoders are the service providers that receive the query address, process the geocoding task, and output the coordinate results. Recently, several online geocoding applications – including address geocoding – have become widely available with Bing Maps (https://www.bing.com/maps/), Google Maps (https://www.google.com/maps/), and Open Street Maps (https://www.openstreetmap.org/) [11,14]. In general, a set of addresses are queried automatically and the results are retrieved, including metadata indicators of quality along with the coordinates [7,10,11,12,13].

Geocoders – and other online tools – may vary in both match rate, i.e., the rate at which addresses are found in a certain study, and accuracy, i.e., how close to the real location the queried address is placed. These quality estimates are essential for public health research, since differences in match rates across locations and/or spatial displacements of the addresses may bias the study results [4,9,12,13,15]. Geocoders use different databases and algorithms, and, therefore, the quality of geocoding are expected to be different. Many recent studies have attempted to assess the quality of geocoder services in different settings [2,9,10,11,13]. To this date, most studies of geocoding quality have been conducted mainly in North America and Europe, where it was possible to identify differences in quality; and only recently a report of this nature have emerged on Brazil [14], with no other studies known for Chile or other Latin American countries, although the substrate for geocoding in the region could greatly differ. There is a large interest in studying spatial health determinants in the region [16,17], and following quality estimations or recommendations from international studies may introduce biases, as well as under or overestimation of the health effect in the local population, which ultimately might affect the success of implementing local public health policies [17].

For these reasons, a study of the quality of geocoding in municipalities of Latin American need attention. We developed this study as part of a larger research project studying the association between air pollution and pregnancy outcomes in a cohort of women in Temuco, Chile, where residential addresses of pregnant women, obtained from handwritten hospital records, were used to spatially estimate air pollution exposures. This article aims to determine the quality of three automatic online geocoding tools by comparing them with a reference method in a random subsample of the addresses.

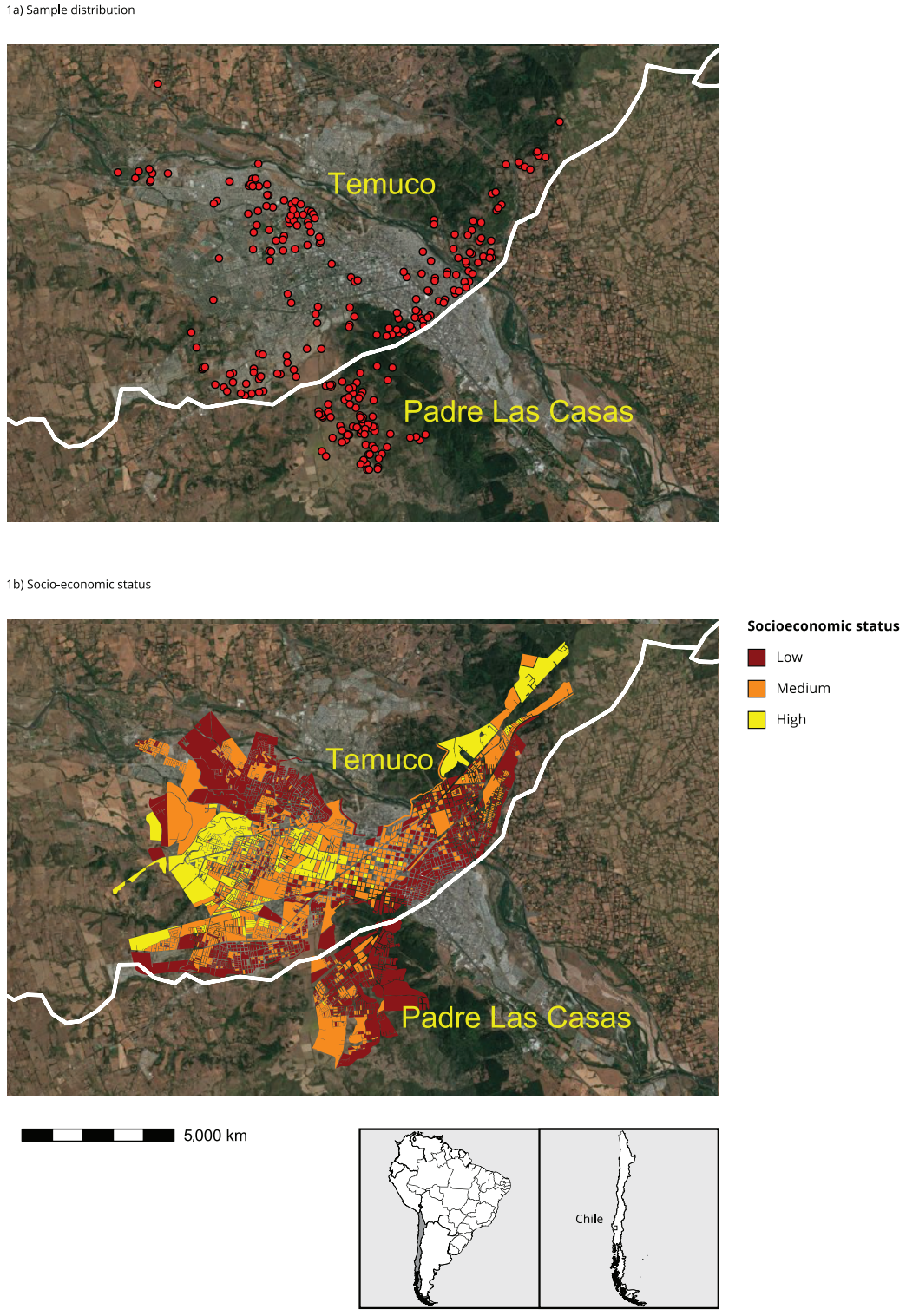
Methods

Study site

Temuco and Padre Las Casas are neighboring cities, separated by the Cautín river (Figure 1), that belong to a conurbation, known as Temuco; located at 38°44' S and 72°35' W in the Araucanía Region in the Southern Chile. Temuco was founded at the end of the 19th century and is the most populated city in the region with a surface area of 464km^2 and a population of 290,000 inhabitants [18]. Padre Las Casas was founded in 1995 and has a surface area of 400km^2 and 80,000 inhabitants [18]. Most of the population (93%) in Temuco live in urban areas, whereas in Padre Las Casas, a larger proportion (40%) reside in rural zones also presenting a larger share of indigenous people [19]. The main economic activities of the region are agriculture and services. This region is the poorest in Chile with 17% of the population living below the poverty line [20].

Figure 1

Study site and location of reference points. Municipalities of Temuco and Padre Las Casas, Chile, 2009-2015.



Note: the white line represents the limit between the municipalities of Temuco and Padre Las Casas.

Study design and data collection

Addresses for geocoders testing were drawn from a retrospective pregnancy cohort study including 15,500 childbirths at the Dr. Hernan Henriquez Aravena Hospital in Temuco, the reference health center for the municipality from 2009 to 2015. Maternal sociodemographic characteristics, obstetrics, and newborn variables were collected from hospital records. The main study was approved by the Araucanía Sur Local Ethics Committee with nationwide accreditation (Servicio de Salud Araucanía Sur, *Resolución Exenta n. 1,179,* March 6, 2014). The study attempted to link air pollution from a spatiotemporal model with maternal information. To achieve this, handwritten addresses – from hospital records – were automatically converted into spatial points.

To evaluate the quality of three different automatic geocoders, the geocoding results from a subsample of addresses were compared with a reference method. A total of 300 handwritten addresses were randomly selected from the cohort database but stratified by municipality to ensure that each municipality was adequately represented in the subsample (200 selected in Temuco and 100 in Padre Las Casas). The number of addresses evaluated was limited to 300 for it was a feasible amount to process using the reference method, allowed adequate comparison, and because similar numbers have been used in previous studies [2,7,21]. Addresses were limited to urban areas within the municipalities using the same inclusion criteria as the larger study. To assure strict confidentiality criteria, a geocoding team was established inside the hospital and an identification number was assigned to each address. Thus, a reduced database was generated for geocoding, including only the identification number and address, with no other personal data available.

Reference method

The reference method consisted of manually geocoding all addresses in the subsample, conducted by a trained technician, who did not belong to the hospital or the research team and was blinded to the address source. The process was conducted in two steps. In the first step, addresses were located using Google Street View (https://streetview.gosur.com/), assuring that the actual street address and number were observed on the screen. Then, the point was located in the middle of the sidewalk in front of the household. If the address was not found with Google Street View, the technician would personally explore the zone until identifying the address, and subsequently using Global Positional System (GPS) receiver (Garmin 60CSx, Garmin Ltd.; http://garmin.com/) to obtain the address coordinates. Due to the high accuracy of the GPS (4.2 and 5.3m for the Easting and Northing coordinates, respectively) [22], no differential correction was employed in this study. Both systems located the points using the WGS84 coordinate reference system. All referencing was achieved in September 2018. As we used two different techniques to build the reference method, we explored the differences between positional errors of GPS and Street View by locating the now known locations found in GPS in Street View. Figure S1 and Table S1 (Supplementary material: http://cadernos.ensp.fiocruz.br/static//arquivo/suppl-e00288920_6701.pdf) show a small positional error between both techniques, with 90% of the points within an error of 20m in both municipalities.

Automated geocoding services

The three geocoding methods used were Bing, Google Earth, and Google Maps. Both Bing and Google Earth were implemented using a code in R software (http://www.r-project.org), while Google Maps geocoding was implemented using GIS software (https://www.qgis.org/). The solution output included metadata and quality indicators besides the coordinates. The results may include more than one solution, and some solutions may be erroneous (i.e., in other cities or even in other countries). To ensure the selection of a result that was likely the actual address in question and filter the inadequate ones, some quality criteria were established for each geocoder using the returned indicators.

(a) Bing. Addresses were automatically supplied to Bing Map using a code in R software. A typical query was "street name + street numbering, city, Chile". Six criteria were established based on the metadata: (i) confidence must be "high"; (ii) entity type must be "address" or "roadblock"; (iii) accuracy must be "rooftop" or "interpolation"; (iv) match code must be "good"; (v) the city must match the one

in the record ("Temuco" or "Padre Las Casas"); and (vi) a street number must be found. An address was considered found and selected when all six criteria were met. Usually, only one result matched the six criteria.

(b) Google Earth. The process was similar to Bing except that the platform used was Google Earth and the criteria were adjusted as follows. Five criteria were used: (i) one component must be a "route"; (ii) another component must be "street number"; (iii) the found city must match the one in the record; (iv) type of point must be "rooftop" or "range-interpolated"; and (v) result must match the city. As with Bing, usually, only one result satisfied the five criteria.

(c) Google Maps. Addresses were loaded to Google Maps in batch mode using the MMQGIS plugin of QGIS. The plugin employs an attribute table in CSV format with the addresses (street number, street, city, and country) to obtain a geocoded point layer. Three criteria were used to evaluate the performance of the geocoding method: (i) accuracy must be "rooftop" or "interpolated range"; (ii) address type must be "street" and "house number"; and (iii) district must be "Temuco" or "Padre Las Casas". Google Maps provided only one result per query.

All automatic geocoding presented were initially performed in September 2018, to be comparable to the reference method. New searches were repeated at a later date, yielding similar results to those obtained in 2018.

Data analysis

Firstly, the match rate of the reference method was calculated by dividing the number of geocoded addresses by the total number of submitted addresses [23,24,25]. Then, match rates of the automatic methods were estimated using only the addresses previously found by the reference methods. The positional error was calculated as the Euclidean distance, in meters, between the results of the automated tools and the reference method to compare the accuracy of the results. To do this, all locations were first projected to a UTM zone 18H south coordinate system. The positional error was characterized and compared by using descriptive statistics (mean, median, standard deviation, and percentiles) and plotting the cumulative frequency distribution of positional errors. The outcome was also analyzed by socioeconomic status. ADIMARK [26] is a common instrument used in Chile to evaluate socioeconomic status, dividing the population into five groups: ABC1, C2, C3, D, and E, according to income and purchasing power, with the first being higher-income group, whereas the last being the one with lower income. The variable was calculated for each block in the cities based on data measured at the household level using the 2002 Census of the Chilean National Institute of Statistics, which included the education level of the head of the household and possession of assets. To facilitate the analyses, the addresses were grouped in blocks of high (ABC1), medium (C2+C3), and low socioeconomic status (D+E) and matched to each address by block.

Results

Performance of the reference method

Figure 1a shows the spatial distribution of the addresses that could be located by the reference method. Notably, the distribution of these addresses was spread across both cities, although less represented in sectors with higher socioeconomic status (Figure 1b). This occurred because the hospital performs approximately 80% of the cities childbirths, mostly of lower and medium socioeconomic status mothers. From the 300 addresses, 90% were successfully found by the reference method (Table 1). Geocoding using the reference method required approximately 24 hours of the technician's time, compared to automatic methods that were executed in few minutes. Most addresses were found in the initial step using Google Street View (63%) with rates slightly higher in Temuco compared to Padre Las Casas. Regarding those addresses not found, the technician reported having the address number not matching actual street numbering as the main reason. Furthermore, four addresses located in Padre Las Casas in the clinical record were found in Temuco. This emphasizes the initial difficulties faced when working with transcribed handwritten addresses.

Table 1

Success rate of the reference method. Municipalities of Temuco and Padre Las Casas, Chile, 2009-2015.

Method	Padre Las Casas		Temuco		Overall	
	n	%	n	%	n	%
Found	80	83	190	93	270	90
Street View	40	50	130	68	170	63
GPS	40	50	60	32	100	37
Not found	16	17	14	7	30	10
Total	96	100	204	100	300	100

Performance of the automated methods

• Match rate

Table 2 shows the match rates of the three automated methods compared to the reference one. We observed large significant differences in the match rates between methods and between cities (Table 1). We also observed better overall performance for Google Maps with rates above 90% for both cities, followed by Google Earth with rates above 80% for both cities, with statistical differences between methods (Table 3). Finally, Bing had rates above 80% only for Temuco and no matches in Padre Las Casas. Considering socioeconomic status, we found large and significant differences in the match rates between the methods for addresses in the low and medium socioeconomic status (Tables 4 and 5), whereas we found no differences in match rates when comparing socioeconomic status for each method.

• Positional errors

Table 6 shows the distribution of positional errors for the three methods. We found significant differences among methods (Table 7). Overall, Bing showed a lower positional error with a higher proportion (88%) of the observations with positional errors in smaller ranges, i.e., < 20m, and lower proportion in the larger ranges (1%), i.e., ≥ 100m, compared to the other methods, in the order of 70% for smaller positional errors (< 20m) and 6%-10% for larger positional errors (≥ 100m). We observed significant differences between Bing and Google Earth and Bing and Google Maps (Table 7), but not for Google Earth and Google Maps. This was more evident when inspecting the cumulative distribution of positional errors plot (Figure 2), in which it was clear that Bing had a better performance followed by Google Earth and Google Maps.

Moreover, Table 8 shows some very large errors (> 1,000m) observed for some cases (p98) in Google Earth and Google Maps. When analyzing each city separately, the trends in Temuco were similar to overall results, whereas Bing presented no matches in Padre Las Casas and the performance of the other two methods was slightly worse than in Temuco.

Finally, we found significant differences when considering socioeconomic status (Tables 9 and 10). Bing showed lower positional error in low socioeconomic status, with a higher proportion (92%) of the observations with positional errors in smaller ranges, i.e., < 20m, compared to the other socioeconomic status. We observed no significant differences between methods by socioeconomic status (Table 10).

Table 2

Comparison of match rates of the automated methods against the reference method. Municipalities of Temuco and Padre Las Casas, Chile, 2009-2015.

Method	Padre Las Casas		Temuco		Overall	
	n	%	n	%	n	%
Bing	0	0	155	83	155	57
Google Earth	69	82	160	86	229	85
Google Maps	73	91	171	90	244	90
Total	80	100	190	100	270	100

Table 3

Pearson's chi-squared test of the match rate of the automated methods against the reference method. Municipalities of Temuco and Padre Las Casas, Chile, 2009-2015.

Variables	Statistic	Parameter	p-value
Methods			
3 methods	96.5	2	**< 0.01**
Bing/Google Earth	48.0	1	**< 0.01**
Bing/Google Maps	74.3	1	**< 0.01**
Google Earth/Google Maps	3.3	1	0.07
Within Temuco			
Bing/Google Earth	0.3	1	0.57
Bing/Google Maps	3.1	1	0.08
Google Earth/Google Maps	1.1	1	0.30
Within Padre Las Casas			
Google Earth/Google Maps	2.2	1	0.14
Compare cities			
All	82.7	1	**< 0.01**
Bing	161.0	1	**< 0.01**
Google Earth	0.4	1	0.52
Google Maps	0.0	1	0.93

Note: the bold p-values mean statistical significance.

Table 4

Match rate for each method and socioeconomic status. Municipalities of Temuco and Padre Las Casas, Chile, 2009-2015.

Method	Low		Medium		High		Overall	
	n	%	n	%	n	%	n	%
Bing	113	61	40	49	2	100	155	57
Google Earth	158	85	69	84	2	100	229	85
Google Maps	165	89	77	94	2	100	244	90

Table 5

Pearson's chi-squared test of the match rate by socioeconomic status. Municipalities of Temuco and Padre Las Casas, Chile, 2009-2015.

Label	Statistic	Parameter	p-value
Method comparison			
All socioeconomic status	96.5	2	**< 0.01**
Low socioeconomic status only	50.1	2	**< 0.01**
Medium socioeconomic status only	50.1	2	**< 0.01**
Comparison by socioeconomic			
Bing only	4.8	2	0.09
Google Earth only	0.4	2	0.82
Google Maps only	2.0	2	0.37

Note: the bold p-values mean statistical significance.

Table 6

Distribution of the positional error for the different methods. Municipalities of Temuco and Padre Las Casas, Chile, 2009-2015.

Method/Distance	Padre Las Casas		Temuco		Overall	
	n	%	n	%	n	%
Bing						
< 5m	NA	NA	51	33	51	33
≥ 5m and < 10m	NA	NA	55	35	55	35
≥ 10m and < 20m	NA	NA	31	20	31	20
≥ 20m and < 50m	NA	NA	15	10	15	10
≥ 50m and < 100m	NA	NA	2	1	2	1
≥ 100m	NA	NA	1	1	1	1
Google Earth						
< 5m	6	9	15	9	21	9
≥ 5m and < 10m	25	36	55	34	80	35
≥ 10m and < 20m	17	25	61	38	78	34
≥ 20m and < 50m	12	17	19	12	31	14
≥ 50m and < 100m	4	6	2	1	6	3
≥ 100m	5	7	8	5	13	6
Google Maps						
< 5m	6	8	13	8	19	8
≥ 5m and < 10m	24	32	55	33	79	32
≥ 10m and < 20m	19	25	58	35	77	32
≥ 20m and < 50m	14	18	21	12	35	14
≥ 50m and < 100m	8	11	8	5	16	7
≥ 100m	5	7	13	8	18	7

NA: not available.

Table 7

Pearson's chi-squared test of the distance of the automated methods against the reference method. Municipalities of Temuco and Padre Las Casas, Chile, 2009-2015.

Variables	Statistic	Parameter	p-value
Method comparison			
Both cities	74.0	10	**< 0.01**
Temuco	62.9	10	**< 0.01**
Padre Las Casas	1.3	5	0.94
Compare methods, both cities			
Bing/Google Earth	42.6	5	**< 0.01**
Bing/Google Map	55.5	5	**< 0.01**
Google Earth/Google Map	5.2	5	0.39
Compare methods, Temuco			
Bing/Google Earth	35.3	5	**< 0.01**
Bing/Google Map	45.2	5	**< 0.01**
Google Earth/Google Map	4.9	5	0.43
Compare methods, Padre Las Casas			
Google Map/Google Earth	1.3	5	0.94
Compare cities			
All methods	21.2	5	**< 0.01**
Google Earth	8.0	5	0.16
Google Map	5.6	5	0.35

Note: the bold p-values mean statistical significance.

Discussion

Our results reveal that the quality of geocoding methods greatly varies regarding match rate and accuracy. Concerning match rates, Google Earth and Google Maps showed a good performance compared to Bing, which completely failed in one of the studied areas (Padre Las Casas). However, Bing presented a much lower positional error once the address was found, and with even better performance in lower socioeconomic status. For some years, researchers have been studying the quality of geocoding methods [15,23,27,28,29], but recently, automated methods have been evaluated. Considering five of the most recent methods [2,7,9,21,24], the observed match rate in our study was on the higher end (above 90%) compared to what other authors found, at least for some of the methods. Similarly, all methods in our study had positional errors mostly in the smaller range (i.e., less than 20m), particularly Bing, with only some excursions above 100m. Google Maps and Google Earth, on the other hand, presented relatively larger positional errors more frequently, according to other international studies. Only a fraction (1%-5%) of the addresses presented large errors, similar to previous reports. Surprisingly, the performance did not vary for a given method with socioeconomic status.

The results from this study are far from being generalizable to other cities in Chile or Latin America. On other hand, it warns against the massive use of automated methods without knowing the quality of the outputs, which may result in large differences among cities, or neighborhoods, potentially leading to biases [15]. Locally, it seems advisable to automatically geocode the addresses using all three methods following a tiered protocol based on the quality criteria previously established. For Temuco, the protocol suggests geocoding the addresses automatically with Bing first, proceeding with Google Earth first and Google Maps later. Whereas in Padre Las Casas, we proposed to start with Google Earth, followed by Google Maps. We estimate that using this procedure would ensure an overall 93% match rate and about 80% of positional error below 20m and less than 5% above 100m, thus minimizing biases.

Figure 2

Cumulative frequency distribution of positional errors, overall and for the municipalities of Temuco and Padre Las Casas, Chile, 2009-2015.

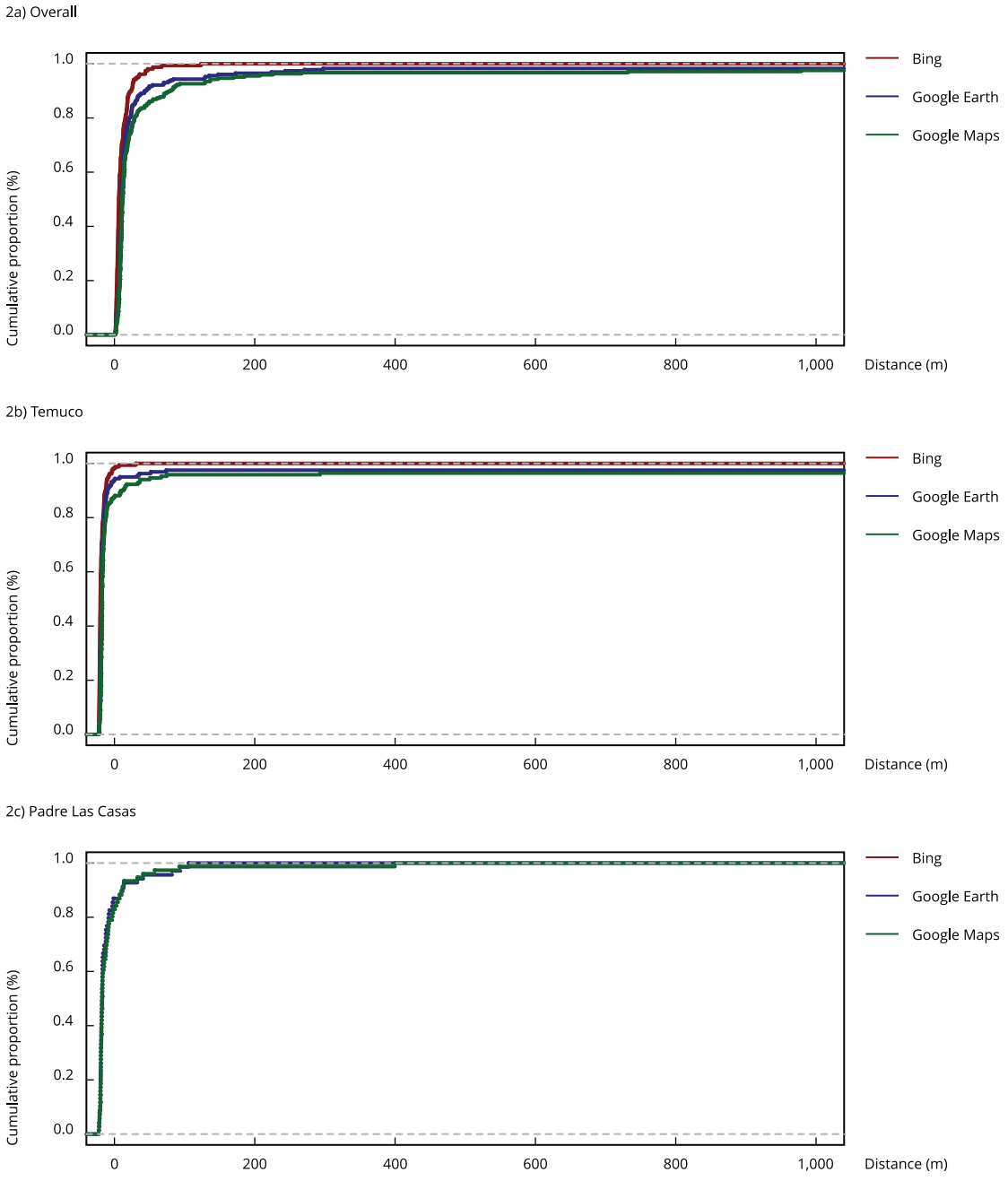


Table 8

Positional error summary statistics. Municipalities of Temuco and Padre Las Casas, Chile, 2009-2015.

Method	n	Mean (SD)	Minimum	p5	p95	p98	Maximum
Padre Las Casas							
Google Earth	69	32.5 (58.2)	1.0	4.0	140.0	261.0	298.0
Google Maps	76	42.1 (117.3)	1.0	3.0	133.0	226.0	979.0
Temuco							
Bing	155	11.1 (13.9)	2.0	2.0	32.0	48.0	124.0
Google Earth	160	100.3 (531.4)	2.0	4.0	73.0	2,298.0	4,255.0
Google Maps	168	101.7 (431)	1.0	4.0	194.0	1,436.0	3,184.0
All							
Bing	155	11.1 (13.9)	1.5	2.0	32.1	47.5	123.6
Google Earth	229	79.9 (446)	1.2	3.6	129.1	282.5	4,254.6
Google Maps	244	83.2 (364.2)	1.0	3.6	166.5	1,145.2	3,184.5

p: percentile; SD: standard deviation.

Table 9

Distribution of positional error for the different methods according to socioeconomic status. Municipalities of Temuco and Padre Las Casas, Chile, 2009-2015.

Method/Distance	Low		Medium		High		All	
	n	%	n	%	n	%	n	%
Bing								
< 5m	38	34	13	32	0	0	51	33
≥ 5m and < 10m	43	38	11	28	1	50	55	35
≥ 10m and < 20m	23	20	8	20	0	0	31	20
≥ 20m and < 50m	8	7	6	15	1	50	15	10
≥ 50m and < 100m	1	1	1	2	0	0	2	1
≥ 100m	0	0	1	2	0	0	1	1
Google Earth								
< 5m	17	11	4	6	0	0	21	9
≥ 5m and < 10m	61	39	18	26	1	50	80	35
≥ 10m and < 20m	53	34	25	36	0	0	78	34
≥ 20m and < 50m	17	11	13	19	1	50	31	14
≥ 50m and < 100m	4	3	2	3	0	0	6	3
≥ 100m	6	4	7	10	0	0	13	6
Google Maps								
< 5m	14	8	5	6	0	0	19	8
≥ 5m and < 10m	56	34	22	29	1	50	79	32
≥ 10m and < 20m	52	32	25	32	0	0	77	32
≥ 20m and < 50m	21	13	13	17	1	50	35	14
≥ 50m and < 100m	13	8	3	4	0	0	16	7
≥ 100m	9	5	9	12	0	0	18	7

Table 10

Pearson's chi-squared test of the distance rate by socioeconomic status. Municipalities of Temuco and Padre Las Casas, Chile, 2009-2015.

Label	Statistic	Parameter	p-value
Compare methods			
All socioeconomic status	28.1	10	**0.01**
Low socioeconomic status only	53.1	10	**< 0.01**
Medium socioeconomic status only	23.7	10	**0.01**
High socioeconomic status only	0.0	2	0.99
Compare by socioeconomic status			
Bing only	11.0	10	0.36
Google Earth only	12.4	10	0.26
Google Maps only	8.5	10	0.58

Note: the bold p-values mean statistical significance.

Regarding limitations, we recognize that this subsample was created based on a population of pregnant women seeking services at a public hospital, and, therefore, it is likely that women of higher socioeconomic status are less represented. We can speculate that these women from higher socioeconomic status are likely to live in well-established, higher socioeconomic status neighborhoods; and, concerning the commercial interest of the engines, it is possible that the quality of the geocoding of their addresses might not differ much from the one reported here. Another limitation is that the reference method was derived from two complementary methods: one that could be called a "gold standard", i.e., GPS and Google Street View. These methods could have different inherent errors [7], but we emphasize that the supervised process should keep errors relatively small and thus, should allow estimating whether the automated geocoding methods are falling in the identification of address coordinates within a smaller (below 20m) or larger (more than 100m) range of errors.

As strengths of this study, we can mention the fact that this is the first of its kind in Chile and the second in South America. Also it is linked to a real-world health study that uses handwritten addresses, a challenge that many research teams face. Furthermore it was performed in Temuco, a mid-sized regional capital, and not in Santiago, the capital city, where approximately 40% of Chile's population resides. A study conducted in Santiago would likely have yielded results unlikely to be compared to small, or medium-size cities in Chile.

Overall, methods showed an acceptable, but heterogeneous performance, corroborating with other international studies. If the methods are used combined in a tiered protocol, the geocoding results may present adequate quality to perform health studies in Temuco and Padre Las Casas. The heterogeneity of the performance warns against using the automatic methods without assessing quality in other cities in Chile and Latin America.

Contributors

M. E. Quinteros and C. Blazquez contributed to the study design, data acquisition, analysis and interpretation, and writing. F. Rosas, X. M. O. García, J. M. Delgado-Saborit, R. M. Harrison, and S. Ayala contributed to the data analysis and interpretation. P. Ruiz-Rudolph and K. Yohannessen contributed to the data analysis and interpretation of data and writing. All authors approved the final version of the manuscript.

Additional informations

ORCID: Maria Elisa Quinteros (0000-0001-8815-1513); Carola Blazquez (0000-0003-4760-885X); Felipe Rosas (0000-0001-7321-2491); Salvador Ayala (0000-0002-5277-4177); Ximena Marcela Ossa García (0000-0003-2626-0946); Juana Maria Delgado-Saborit (0000-0002-7096-9744); Roy M. Harrison (0000-0002-2684-5226); Pablo Ruiz-Rudolph (0000-0001-7872-2546); Karla Yohannessen (0000-0003-4248-0121).

Acknowledgments

Impact of Wood Burning Air Pollution on Preeclampsia and other Pregnancy Outcomes in Temuco, Chile (DPI20140093) was granted by CONICYT and Research Councils UK. J. M. Delgado-Saborit is supported by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement n. 750531. M. E. Quinteros and S. Ayala were supported by a doctoral scholarship by CONICYT Chile Beca Doctorado Nacional n. 21150801 and n. 21191111, respectively.

Conflict of interests

The authors declare no conflict of interests.

References

1. Waller LA. Geospatial data for environmental health. In: Frumkin H, editor. Environmental health: from global to local. 3rd Ed. San Francisco: Jossey-Bass; 2016. p. 111-22.
2. Ganguly R, Batterman S, Isakov V, Snyder M, Breen M, Brakefield-Caldwell W. Effect of geocoding errors on traffic-related air pollutant exposure and concentration estimates. J Expo Sci Environ Epidemiol 2015; 25:490-8.
3. Nuckols JR, Ward MH, Jarup L. Using geographic information systems for exposure assessment in environmental epidemiology studies. Environ Health Perspect 2004; 112:1007-15.
4. Zandbergen PA. Influence of geocoding quality on environmental exposure assessment of children living near high traffic roads. BMC Public Health 2007; 7:37.
5. Akseer N, Bhatti Z, Mashal T, Soofi S, Moineddin R, Black RE, et al. Geospatial inequalities and determinants of nutritional status among women and children in Afghanistan: an observational study. Lancet Glob Health 2018; 6:447-59.
6. Drewnowski A, Aggarwal A, Cook A, Stewart O, Moudon AV. Geographic disparities in Healthy Eating Index scores (HEI-2005 and 2010) by residential property values: findings from Seattle Obesity Study (SOS). Prev Med (Baltim) 2016; 83:46-55.
7. Ribeiro AI, Olhero A, Teixeira H, Magalhães A, Pina MF. Tools for address georeferencing – limitations and opportunities every public health professional should be aware of. PLoS One 2014; 9:e114130.
8. Baldovin T, Zangrando D, Casale P, Ferrarese F, Bertoncello C, Buja A, et al. 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:E88-94.
9. Faure E, Danjou AMN, Clavel-Chapelon F, Boutron-Ruault MC, Dossus L, Fervers B. Accuracy of two geocoding methods for geographic information system-based exposure assessment in epidemiological studies. Environ Health 2017; 16:15.
10. Shah TI, Bell S, Wilson K. Geocoding for public health research: empirical comparison of two geocoding services applied to Canadian cities. Can Geogr 2014; 58:400-17.
11. Longley PA, Goodchild MF, Maguire DJ, Rhind DW. Geographical information systems and science. 2nd Ed. Chichester: John Wiley & Sons; 2005.
12. Singh SK. Evaluating two freely available geocoding tools for geographical inconsistencies and geocoding errors. Open Geospatial Data, Software and Standards 2017; 2:11.
13. Lamprecht A-L, Tiziana M. Process design for natural scientists. Berlin: Springer; 2014. (Series Communications in Computer and Information Science).

14. Davis Jr. CA, Alencar RO. Evaluation of the quality of an online geocoding resource in the context of a large Brazilian city. Trans GIS 2011; 15:851-68.
15. Krieger N, Waterman P, Lemieux K, Zierler S, Hogan JW. On the wrong side of the tracts? Evaluating the accuracy of geocoding in public health research. Am J Public Health 2001; 91:1114-6.
16. Diez Roux AV, Slesinski SC, Alazraqui M, Caiaffa WT, Frenz P, Jordán Fuchs R, et al. A novel international partnership for actionable evidence on urban health in Latin America: LAC-Urban Health and SALURBAL. Glob Challenges 2019; 3:1800013.
17. Fry D, Mooney SJ, Rodríguez DA, Caiaffa WT, Lovasi GS. Assessing Google Street View image availability in Latin American cities. J Urban Health 2020; 97:552-60.
18. Instituto Nacional de Estadística. Compendio estadístico región de la Araucanía. Santiago: Instituto Nacional de Estadística; 2017.
19. Universidad Mayor. Estudio de actualización diagnóstico territorial de Temuco para modificación al plan regulador. Santiago: Universidade Mayor; 2015.
20. Observatório Social, Ministerio de Desarrollo Social. Situación de pobreza: sintesis de resultados. Casen 2017. http://observatorio.ministeriodesarrollosocial.gob.cl/storage/docs/casen/2017/Resultados_pobreza_Casen_2017.pdf (accessed on 20/Nov/2020).
21. Quesada JA, Nolasco A, Moncho J. Comparación de las aplicaciones de Google y Yahoo para la geocodificación de direcciones postales con fines epidemiológicos. Rev Esp Salud Pública 2013; 87:201-6.
22. Vasquez C, Blazquez C. Simple spatial data implementation and error analysis for transportation applications. In: XV Congreso Panamericano de Ingeniería de Tránsito y Transporte. Cartagena de Indias: Universidad del Norte; 2008. p. 009.
23. Zhan FB, Brender JD, De Lima I, Suarez L, Langlois PH. Match rate and positional accuracy of two geocoding methods for epidemiologic research. Ann Epidemiol 2006; 16:842-9.
24. Roongpiboonsopit D, Karimi HA. Comparative evaluation and analysis of online geocoding services. Int J Geogr Inf Sci 2010; 24:1081-100.
25. Goldberg D, Goldberg M, Ballard J, Boyd N, Mullan C, Garfield D, et al. An evaluation framework for comparing geocoding systems. Int J Health Geogr 2013; 12:50.
26. Adimark: Investigaciones de Mercado y Opinión Pública. Mapa socioeconómico de Chile. https://docplayer.es/1497857-Mapa-socioeconomico-de-chile-nivel-socioeconomico-de-los-hogares-del-pais-basado-en-datos-del-censo.html (accessed on 20/Nov/2020).
27. Ward MH, Nuckols JR, Giglierano J, Bonner MR, Wolter C, Airola M, et al. Positional accuracy of two methods of geocoding. Epidemiology 2005; 16:542-7.
28. Bonner MR, Han D, Nie J, Rogerson P, Vena JE, Freudenheim JL. Positional accuracy of geocoded addresses in epidemiologic research. Epidemiology 2003; 14:408-12.
29. Cayo MR, Talbot TO. Positional error in automated geocoding of residential addresses. Int J Health Geogr 2003; 2:10.

Resumen

Los métodos automáticos de geocodificación se han convertido en algo popular durante los últimos años para facilitar el estudio de la asociación entre resultados de salud y lugar para vivir. No obstante, más bien pocos estudios han evaluado la calidad de la geocodificación, siendo realizados la mayoría de ellos en EE.UU. y Europa. El objetivo de este artículo es comparar la calidad de tres herramientas automáticas de geocodificación en línea frente a un método de referencia. La submuestra de 300 direcciones escritas a mano, procedentes del registro hospitalario, se geocodificaron usando Bing, Google Earth y Google Maps. Los porcentajes de coincidencia fueron mayores (> 80%) en el caso de Google Maps y Google Earth comparados con Bing. Sin embargo, la precisión de las direcciones fue mejor con Bing, en una proporción más grande (> 70%) de direcciones que tenían errores de posición por debajo de 20m. En general, el rendimiento no varió en cada método para diferentes niveles estatus socioeconómico. En general, los métodos mostraron un rendimiento aceptable, pero heterogéneo. Esto previene contra el uso de métodos automáticos sin evaluar la calidad en otras ciudades, particularmente en Chile y Latinoamérica.

Mapeo Geográfico; Características de la Residencia; Análisis Espacial

Resumo

Os métodos de geocodificação automática se tornaram populares nos últimos anos para facilitar o estudo da associação entre desfechos de saúde e lugar de residência. Entretanto, poucos estudos avaliaram a qualidade da geocodificação, e a maioria dos estudos existentes foi realizada nos Estados Unidos e Europa. O estudo teve como objetivo comparar a qualidade de três ferramentas de geocodificação eletrônica automática em relação a um método de referência. Foi geocodificada uma subamostra de 300 endereços anotados à mão em prontuários hospitalares, usando Bing, Google Earth e Google Maps. As taxas de correspondência dos registros foram mais altas (> 80%) com Google Maps e Google Earth, comparado com Bing. Entretanto, a acurácia dos endereços foi melhor com Bing, com uma proporção maior (> 70%) de endereços com erros de localização menores que 20 metros. Em geral, o desempeno não variou para cada método de acordo com condição socioeconômica. Os métodos apresentaram desempenho geral aceitável, porém heterogêneo. Os resultados servem de alerta contra o uso de métodos automáticos sem avaliar a qualidade em outras cidades, particularmente no Chile e no resto da América Latina.

Mapeamento Geográfico; Características de Residência; Análise Espacial

Submitted on 22/Oct/2020
Final version resubmitted on 10/Feb/2021
Approved on 25/Mar/2021