ORIGINAL BREVE

COMPARACIÓN DE LAS APLICACIONES DE GOOGLE Y YAHOO PARA LA GEOCODIFICACIÓN DE DIRECCIONES POSTALES CON FINES EPIDEMIOLÓGICOS

JoseAntonio Quesada, Andreu Nolasco y Joaquin Moncho.

Unidad de Investigación de Análisis de la Mortalidad y Estadísticas Sanitarias. Departamento de Enfermería Comunitaria, Medicina Preventiva y Salud Pública e Historia de la Ciencia. Universidad de Alicante.

No existen conflictos de intereses.

RESUMEN

Fundamentos: Geocodificar es asignar coordenadas geográficas a puntos del espacio, frecuentemente direcciones postales. El error cometido al aplicar este proceso puede introducir un sesgo en las estimaciones de modelos espacio-temporales en estudios epidemiológicos. No se han encontrando estudios que midan este error en ciudades españolas. El objetivo es evaluar los errores en magnitud y direccionalidad de dos recursos gratuitos (Google y Yahoo) respecto a GPS en dos ciudades de España.

Método: Se geocodificaron 30 direcciones aleatorias con los dos recursos citados y con GPS en Santa Pola (Alicante) y en Alicante. Se calculó la mediana y su IC95% del error en metros entre los recursos y GPS, para el total y por el status reportado. Se evaluó la direccionalidad del error calculando el cuadrante de localización y aplicando un test Chi-Cuadrado. Se evaluó el error del GPS midiendo 11 direcciones dos veces en un intervalo de 4 días.

Resultados: La mediana del error total desde Google-GPS fue de 23,2 metros (16,0-32,2) para Santa Pola y 21,4 metros (14,9-31,1) en Alicante. Para Yahoo fue de 136,0 (19,2-318,5) para Santa Pola y 23,8 (13,6-29,2) para Alicante. Por status, se geocodificó entre un 73% y 90% como 'exactas o interpoladas' (menor error), tanto Google como Yahoo tuvieron una mediana del error de entre 19 y 22 metros en las dos ciudades. El error del GPS fue de 13,8 (6,7-17,8) metros. No se detectó direccionalidad.

Conclusiones: El error de Google es asumible y estable en las dos ciudades, siendo un recurso fiable para geocodificar direcciones postales en España en estudios epidemiológicos.

Palabras clave: Geocodificar. Análisis espacial. Métodos epidemiológicos. Sesgos. Mapeo geográfico.

Correspondencia
Jose Antonio Quesada Rico.
Departamento de Enfermería Comunitaria, Medicina Preventiva y Salud Pública e Historia de la Ciencia.
Universidad de Alicante.
Campus de San Vicente s/n.
ja.quesada@ua.es

ABSTRACT

Comparison of Google and Yahoo Applications for Geocoding of Postal Addresses in Epidemiological Studies

Background: Geocoding is the assignment of geographic coordinates to spatial points, which often are postal addresses. The error made in applying this process can introduce bias in estimates of spatiotemporal models in epidemiological studies. No studies have been found to measure the error made in applying this process in Spanish cities. The objective is to evaluate the errors in magnitude and direction from two free sources (Google and Yahoo) with regard to a GPS in two Spanish cities.

Method: 30 addresses were geocoded with those two sources and the GPS in Santa Pola (Alicante) and Alicante city. The distances were calculated in metres (median, CI95%) between the sources and the GPS, globally and according to the status reported by each source. The directionality of the error was evaluated by calculating the location quadrant and applying a Chi-Square test. The GPS error was evaluated by geocoding 11 addresses twice at 4 days interval.

Results: The overall median in Google-GPS was 23,2 metres (16,0-32,1) for Santa Pola, and 21,4 meters (14,9-31,1) for Alicante. The overall median in Yahoo was 136,0 meters (19,2-318,5) for Santa Pola, and 23,8 meters (13,6-29,2) for Alicante. Between the 73% and 90% were geocoded by status as "exact or interpolated" (minor error), where Goggle andYahoo had a median error between 19 and 23 metres in the two cities. The GPS had a median error of 13.8 meters (6,7-17,8). No error directionality was detected.

Conclusions: Google error is acceptable and stable in the two cities, so that it is a reliable source for geocoding addresses in Spain in epidemiological studies.

Keywords: Geocoding. Spatial analysis. Epidemiological methods. Bias. Geographic mapping.

INTRODUCCIÓN

Geocodificar consiste en asignar coordenadas geográficas (latitud, longitud) a puntos del espacio, siendo frecuente en estudios en salud (análisis de la mortalidad, vigilancia epidemiológica, encuestas de salud, etc.) que los puntos se correspondan con direcciones postales.

Sobre todo en EEUU, existen compañías especializadas que, previo pago, geocodifican direcciones postales[1]. También existen webs gratuitas que realizan el proceso para direcciones aisladas pero no utilizan geocodificación masiva de ficheros[2].

La Junta deAndalucía ha desarrollado una aplicación sólo para territorio Andaluz[3]. Existen programas comerciales que realizan este proceso, como el conocido ArcGis de Esri[4]. El programa libre GvSig prevé la geocodificación masiva de direcciones para su próxima versión 2[5]. Es conocido que la geocodificación tiene asociado un error o sesgo que puede introducir un ruido significativo en los análisis epidemiológicos espaciotemporales[6- 7] y afectar a los resultados[8]. Estos errores pueden deberse: 1)A la mala calidad del fichero de las direcciones (errores tipográficos, símbolos no reconocibles, etc.), por lo que el recurso para geocodificar no puede reconocer la dirección, no la codifica y devuelve un error. 2) Debidos al propio recurso que no identifica la calle y asigna unas coordenadas que no son las correctas, por ejemplo, localizándo otra dirección o un centro geométrico de la ciudad. 3) Errores menores al interpolar el número de policía entre dos conocidos por no encontrar el que se solicita.

Dependiendo de la magnitud y/o direccionalidad de estos errores, los resultados se pueden ver afectados al relacionar espacialmente factores con indicadores en salud, detectando asociaciones o agrupaciones cuando no las hay o no detectándolas cuando sí que existen, pudiendo producirse sesgos en las estimaciones[9]. Si bien hay trabajos que han estudiado estos errores con recursos de EEUU[6-8], no se han encontrado trabajos que evalúen este error en municipios de España.

El objetivo de este estudio es evaluar la magnitud y direccionalidad de los errores tipo 1) y 2) al geocodificar direcciones postales de dos recursos gratuitos y accesibles vía internet (Google geocoding API - (Application Programming Interface)- versión 3 y Yahoo Place Finder, Google y Yahoo en adelante)[10,11], comparadas con coordenadas obtenidas por GPS (Global Positioning System). También se pretende comparar este error en dos municipios de España: Santa Pola (Alicante) con menos de 30.000 habitantes y Alicante capital con más de 330.000.

MATERIAL Y MÉTODOS

Se determinó el tamaño muestral necesario para detectar una diferencia mínima de 20 metros en la media del error de geocodificación entre las dos ciudades, con nivel de confianza del 95%, potencia de 80% y desviación típica de 27 metros, estimada mediante una muestra piloto previa. El tamaño final estimado fue de 30 direcciones en cada ciudad.

El listado de las 30 direcciones postales de cada ciudad (incluyendo tipo de la vía, nombre de la calle y número) fue obtenido por muestreo aleatorio simple sobre la población de residentes con número de identificación SIP (Sistema de Información Poblacional) de la Conselleria de Sanidad de la Comunidad Valenciana.

Se realizó una depuración del fichero de direcciones eliminando errores tipo-

gráficos. Se obtuvieron las coordenadas de estas direcciones, tanto por los dos recursos como por el dispositivo GPS (Samsung modelo SII).

Cada recurso reporta un indicador de la calidad de la geocodificación (status). Los status de cada recurso no tienen porque coincidir y se construyen según criterios propios. Los status pueden ser: exacto cuando el recurso encuentra la calle y el número; interpolado cuando encuentra la calle, pero no el número, el cual es interpolado entre otros dos; aproximado y centro geométrico cuando no encuentra ni calle ni número. Se clasificaron las direcciones según el status como 'correctas' (menor error) cuando el status fue exacto o interpolado, y 'no correctas' (mayor error) cuando fue aproximado o centro geométrico.

Para medir el error de cada recurso respecto del GPS se calcularon las distancias en metros mediante la fórmula del Haverseno[12], tanto para el total como por status de cada recurso y para cada ciudad.

Para evaluar la direccionalidad del error, en cada ciudad se etiquetó el cuadrante donde se posicionó cada coordenada respecto a las de GPS . Así, cuando la coordenada tuvo mayor latitud y longitud que la de GPS, se posicionó en el primer cuadrante; menor latitud y mayor longitud en el segundo cuadrante, etc. Se aplicó un test Chi-Cuadrado para comparar las proporciones de cada cuadrante.

También se evaluó la exactitud del GPS midiendo el error cometido en metros en 11 coordenadas aleatorias dos veces con un intervalo de 4 días.

Se calculó el mínimo, máximo, media, deviación típica y mediana del error en metros de cada recurso al GPS en cada ciudad y el IC al 95% para la mediana, debido al bajo número de efectivos en algunos grupos, tanto para el total como en cada status.

RESULTADOS

La tabla 1 muestra los resultados descriptivos del error entre las coordenadas de GPS y los dos recursos en cada ciudad, para el total y por el status.

La mediana global desde Google a la GPS fue de 23,2 metros para Santa Pola (IC95%: 16,0-32,2) y 21,4 metros en Alicante (IC95%: 14,9-31,1). Para Yahoo fue de 136,0 metros para Santa Pola (IC95%:19,2-318,5) y 23,8 metros para Alicante (IC95%:13,6-29,2). El error cometido con el recurso de Yahoo fue puntualmente mayor que el de Google en Santa Pola y similar en Alicante.

Por status, Google geocodificó en Santa Pola como correctas 26 (86,7%) direcciones y en Alicante 27 (90%). Para Yahoo, 22 (73,3%) direcciones fueron devueltas como correctas en Santa Pola y 26 (86,7%) en Alicante. Estos porcentajes fueron similares para los dos recursos en las dos ciudades.

Para las direcciones correctas, la mediana del error de Google al GPS de las 26 direcciones fue de 22,5 metros (IC95%: 12,6-32,1) en Santa Pola y 21,1 (IC95%: 14,0-28,4) en Alicante. Para Yahoo, en Santa Pola fue de 22,3 metros (IC95%: 14,8-183,7) y en Alicante de 19,3 (IC95%: 11,4-28,0). Los dos recursos aportan errores puntualmente similares para las dos ciudades, si bien Yahoo presenta mayor variabilidad en Santa Pola que Google.

No se detectó direccionalidad en las coordenadas Google y Yahoo respecto a las GPS (test Chi-Cuadrado, tabla 2). La mediana del error del GPS fue de 13,8 metros (IC95% 6,7-17,8).

Tabla 1
Estadísticos descriptivos del error de cada recurso respecto al GPS en cada ciudad, para el total y por status reportado

ciudad	recurso	estatus	n	mínimo	máximo	media	desviación típica	mediana	IC 95%
SANTA POLA	GOOGLE	Interpolado	24	3,7	114,4	31,4	28,3	23,2	(12,6-32,2)
		Exacto	2	4,2	20,5	12,3	11,5	12,3	-
		Centro geométrico	2	15,2	61,0	38,1	32,3	38,1	-
		Aproximado	2	183,4	660,4	421,9	337,2	421,9	-
		Total	**30**	**3,7**	**660,4**	**56,6**	**120,4**	**23,2**	**16,0-32,2**
	YAHOO	Interpolado	0	-	-			-	-
		Exacto	22	4,3	2102,4	210,6	462,1	22,3	14,8-183,7
		Centro geométrico	0	-	768,2			-	-
		Aproximado	8	306,4		578,3	200,9	667,2	306,4-748,1
		Total	**30**	**4,3**	**2102,4**	**308,7**	**437,9**	**136,0**	**19,2-318,5**
ALICANTE	GOOGLE	Interpolado	9	4,8	24,1	14,6	6,9	14,8	5,7-21,1
		Exacto	18	1,9	178,2	35,9	39,2	27,5	14,2-32,3
		Centro geométrico	0	-	-			-	-
		Aproximado	3	242,9	1091,2	542,5	475,8	293,4	-
		Total	**30**	**1,9**	**1091,2**	**80,2**	**202,9**	**21,4**	**14,9-31,1**
	YAHOO	Interpolado	0	-	-			-	-
		Exacto	26	6,4	182,4	29,4	35,1	19,3	11,4-28,0
		Centro geométrico	0	-	-			-	-
		Aproximado	4	288,9	1234,0	571,1	444,7	380,7	-
		Total	**30**	**6,4**	**1234,0**	**101,6**	**237,9**	**23,8**	**13,6-29,2**

Tabla 2
Direccionalidad: número, porcentaje y valor p del test Chi-Cuadrado en cada ciudad

	Cuadrante	n (%) Google		n (%) Yahoo	
SANTA POLA	I	8 (26,7%)	p = 0,343	8 (26,7%)	p = 0,276
	II	4 (13,3%)		10 (33,3%)	
	III	11 (36,7%)		9 (30,0%)	
	IV	7 (23,3%)		3 (10,0%)	
	Total	30 (100%)		30 (100%)	
ALICANTE	I	6 (20,0%)	p = 0,753	9 (30,0%)	p = 0,276
	II	7 (23,3%)		10 (33,3%)	
	III	10 (33,3%)		3 (10,0%)	
	IV	7 (23,3%)		8 (26,7%)	
	Total	30 (100%)		30 (100%)	

DISCUSIÓN

Todas las direcciones fueron geocodificadas satisfactoriamente tanto por Google como por Yahoo. Esto es debido al proceso previo de depuración de errores tipográficos en el fichero de direcciones. Este proceso de estandarización se hace imprescindible para evitar este tipo de errores (errores de tipo 1) mencionados anteriormente.

En Santa Pola hay un mayor error promedio en Yahoo que en Google. Una de las causas puede deberse a que en el cálculo mediante Google se puede utilizar una ventana de coordenadas para la ciudad de estudio, esto es, un input de coordenadas que delimitan un rectángulo donde se pretende obtener las coordenadas de resultados en la ciudad de estudio. Esta ventana no existe en Yahoo, pudiendo haber resultados muy inexactos por ubicarlo, por ejemplo, en una pedanía cercana. Pudo ser el caso de la distancia máxima de 2102 metros de una dirección en Santa Pola.

El porcentaje de geocodificación de direcciones correctas es similar tanto en Google como en Yahoo, entre el 73% y 90%. Estos porcentajes se aproximan a los obtenidos en un estudio[14] en el que un software GIS geocodificó el 97% de las direcciones. En otro trabajo este porcentaje con el programaArcGis fue de 94%[13].

Google muestra valores similares en la mediana del error para las direcciones correctas de unos 22 metros en las dos ciudades. Para Yahoo sí se observan diferencias y, aunque la mediana del error es similar, 22 metros para Santa Pola y 19 para Alicante, el error medio es de 210,6 y 29,4 metros respectivamente, con mayor variabilidad en Santa Pola que en Alicante. Estas diferencias pueden volver a deberse al uso de la ventana de acotación. Los resultados indican que Google es bastante fiable y proporciona un error constante tanto en un municipio pequeño como en una ciudad grande. Este error de 22 metros (25,9 en promedio) es perfectamente asumible en análisis epidemiológicos de ciudades, ya que una parte podría deberse al error del GPS, que fue de 13 metros. Yahoo parece que proporciona errores similares a Google en ciudades grandes y mayor error en ciudades pequeñas, pero habría que realizar más estudios para contrastarlo.

Los resultados obtenidos por Google son parecidos a los obtenidos por Vieira et al. donde comparaban los resultados de dos recursos con GPS en direcciones de EEUU[7], resultando unos errores medios de 39 metros para ArcView y 188 para TeleAtlas.

En el trabajo de Ward MH et al. se evalúa también el error cometido por el programa ArcView y una empresa especializada respecto a GPS[6], siendo 62 y 61metros de error medio cometido respectivamente. No se ha encontrado direccionalidad en los errores en las dos ciudades, coincidiendo con Zandbergen et al[8].

Podemos concluir que la geocodificación realizada por Google es sensiblemente más exacta que la realizada por Yahoo en municipios pequeños, debido posiblemente a la ausencia de ventana acotadora en Yahoo. Se hace imprescindible una depuración previa del fichero de direcciones. La magnitud del error detectada por Google es asumible en estudios epidemiológicos de ciudades de España.

BIBLIOGRAFÍA

1. Unxos GmbH, Switzerland. The GeoNames geographical database. Disponible en: http://www.geonames.org.

2. BatchGeo LLC. BatchGeo database. Disponible en: http://www.batchgeo.com.

3. IDEAndalucia. Infraestructura de Datos Espaciales de Andalucía. Disponible en: http://www.ideandalucia.es/index.php/es/geocodificar.

4. Esri España Programa ArcGis. Disponible en: http://www.arcgis.com

5. Asociación GvSig. Programa GvSig. Disponible en: http://www.gvsig.com

6.WardMH, Nuckols JR, Giglierano J, et al. Positional accuracy of two methods of geocoding. Epidemiology 2005, 16:542-47.

7. Vieria V, Howard GJ, Gallagher LG, et al. Geocoding rural addresses in a community contaminated by PFOA: a comparison of methods. Environ Health 2010. 9:8.

8.Zandbergern 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.

9.Zimmarman D, Sun P. Estimating Spatial Intensity andVariation in Risk from locations subject to geocoding errors. 2006. Disponible en: http://www.stat.uiowa.edu/sites/default/files/techrep/tr363.pdf

10. API de codificación geográfica de Google. Disponible en: https://developers.google.com/maps/documentation/ geocoding/.

11. API decodificación geográfica Yahoo Place-Finder. Disponible en: http://developer.yahoo.com/boss/geo/docs/

12.Sinnott RW, Virtudes de la Haversine. Cielo y 68 Telescopio. 1984; 159

13.Duncan DT, Castro MC, Blossom JC, et al. Evaluation of the positional difference between two common geocoding methods. Geospat Health 5, 2011.pp 265-73.

14.McElroy JA, Remington PL, Trentham-Dietz A, et al. Geocoding addresses from a large population-based study: lessons learned. Epidemiol. 2003; 14:399-407.