Online Geocoding Services

A benchmarking analysis to some European cities

Gianfranco Di Pietro, Fabio Rinnone
Geofunction Srls
Via Luigi Sturzo, 71, Niscemi, Italy
dipietro@geofunction.it, fabiorinnone@geofunction.it

***Abstract*—Geocoding is the computational process that allows to transforming an address of a place into a location on the Earth's surface using geographic coordinates. The quality of geocoding have a relevant importance in spatial analysis. Furthermore, in recent years we seen an enormous spread of Web technologies that involved a growing number of users. At the same time have been born many Online Geocoding Service that support the increasingly WebGIS consumer applications.**

For these reasons, we present GFgeocoder, a tool based on a methodology capable to benchmark efficiently Online Geocoding Services. GFgeocoder analyzes geocoding result obtained by some Online Geocoding Services. The tool is written in Java language and it allow to make parallel requests to the Online Geocoding Services, passing in input the files containing strings of addresses. We used as input results of geocoding strings of addresses, published by some European local governments in their open data portals.

Keywords— geographic information systems, geospatial analysis, benchmarking testing

I. INTRODUCTION

The geocoding process consists of translating an address entry, searching for the address, and delivering the best candidate as a point feature on a map [1]. Today many Online Geocoding Services offer the possibility to convert addresses strings into geographic coordinates (latitude, longitude). The quality of these services has important implications in the analysis of spatial data [2].

This research is based on the following assumptions:

1) the precise location of house numbers provided by Local Governments is the most accurate result achievable, because it is manually performed via a surveying on the field executed by an human operator (census and survey initiatives on a local basis) and it is an official dataset;
2) it is not possible expecting that the accuracy of geolocation cannot exceed bounds of 2-3 meters around every point, both for limits of equipment used for the surveying, and for the real position of the point, which it can be difficult to be obtained (e.g. car accesses, shared private building accesses, etc.).

II. METHODS

Given the above assumptions, we have analyzed the results of geocoding string of addresses published by local governments of Cagliari, Florence, Trento (Italy) and Kristiine, a district of Tallinn (Estonia), using three different Online Geocoding Services, both commercial and open source: Google Geocoder, MapQuest and OpenRouteService. The first one is the geocoder used by the commercial service of map offered by Google, Inc., named Google Maps. The second one is the geocoder used by MapQuest, the map service offered by AOL. The last one uses OpenStreetMap data, the collaborative project that allow to create a free editable by users map of the world [3]. Similar comparisons were carried out with data of some American [4] and Brazilian cities [5], and often they were made respect only to commercial geocoding systems [6]. In other cases, comparisons were useful to epidemiologic studies to map residences with geographic information systems [7] or to public health research [8]. Moreover, it has also been proposed efficient algorithms to dynamically determine which Geocoding Service is more accurate [9].

The procedure performed for benchmarking include a pre-processing of data downloaded from open data portals of some local governments:

1) from the dataset of dati.toscana.it we pre-process data from Road graph version 1.7.10 supplied by Iter.Net project and made available on the portal of open data of the Tuscan region (Italy) [10];
2) from the dataset of comune.cagliari.it we pre-process data from database of house numbers of Cagliari;
3) from the dataset of comune.trento.it we pre-process data from database of first and second level house numbers of Trento;
4) from the dataset of xgis.maaamet.ee/adsavalik we pre-process data from buildings of Tallinn and we extract 22515 addresses of Kristiine district.

From databases listed above, we extract tables containing the addresses of strings and geographic location (Latitude and Longitude, North and East) provided by local governments.

978-1-5090-6040-5/17/$31.00 ©2017 IEEE

Italian cities uses WGS84 system and Tallinn uses EPSG:25884-TMBaltic93.

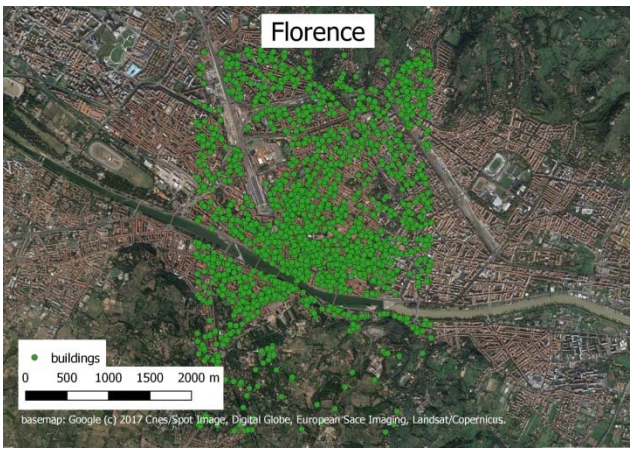

Fig. 1. Map of dataset used for the city of Florence.

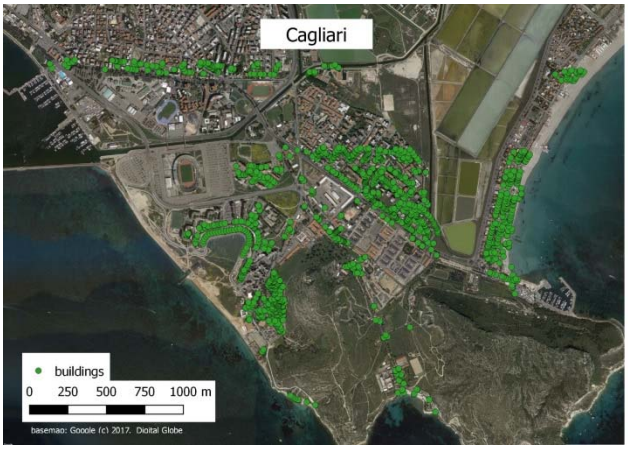

Fig. 2. Map of dataset used for the city of Cagliari.

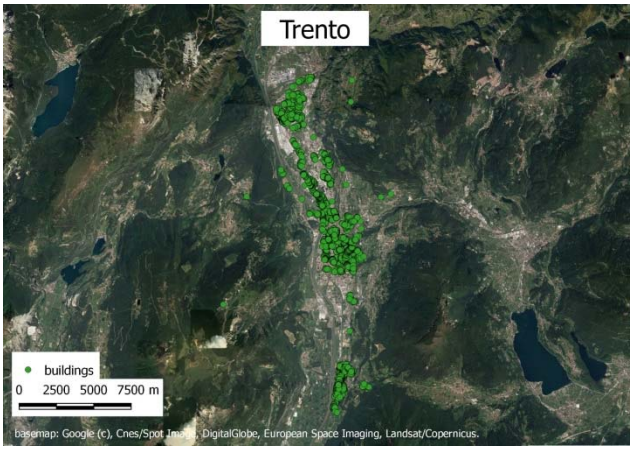

Fig. 3. Map of dataset used for the city of Trento.

From these tables we randomly extracted address to be passed to the geocoder: the total amounts to about 30000 strings. All files generated are stored in CSV format, and the localization of these is showed in Fig. 1, Fig 2, Fig. 3, and Fig. 4.

We propose a benchmarking analysis performed using a parameter named "Geocoder Approx 10 m". We define this parameter, referred to as GA_{10}, as the " ratio (in percentage) of geocoded results in a distance less than 10 meters from position provided by official datasets".

Let δ the distance from geocoded position and official dataset position for each point, we can introduce the following parameter:

$$\delta_{10} = \begin{cases} \delta_i & \text{if } \delta_i \leq 10\ m \text{ (meters)} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

The distance is measured on the same reference system of source data. If coordinates are provided in WGS84 spatial reference (latitude, longitude) we assume a local approximation of the surface of the Earth as a sphere, using average radii of curvature of the ellipsoid (geometric mean of radius in prime vertical and radius in meridian) [11-13].

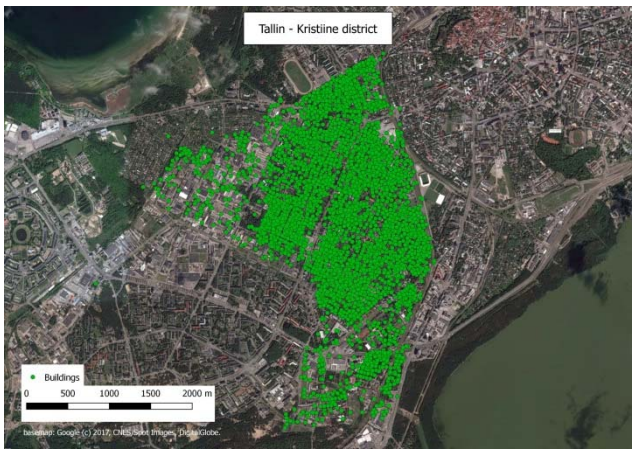

Fig. 4. Map of dataset used for the district of Kristiine (Tallinn).

Now, we can define GA_{10} as follow:

$$GA_{10} = \sum_i^n \theta(\delta_{10}) \quad (2)$$

The parameter GA_{10} represent a unique value that classify whole addresses investigated. This parameter is useful for describe the performance of a geocoding service globally for a location, e.g. a city, a district or a generic buildings dataset.

III. Implementation

We developed an automated tool named GFgeocoder, written in Java language, that allow to make parallel requests to the Online Geocoding Services, passing in input the files containing strings of addresses collected. Obtained results are stored in tabular mode, in CSV format.

978-1-5090-6040-5/17/$31.00 ©2017 IEEE

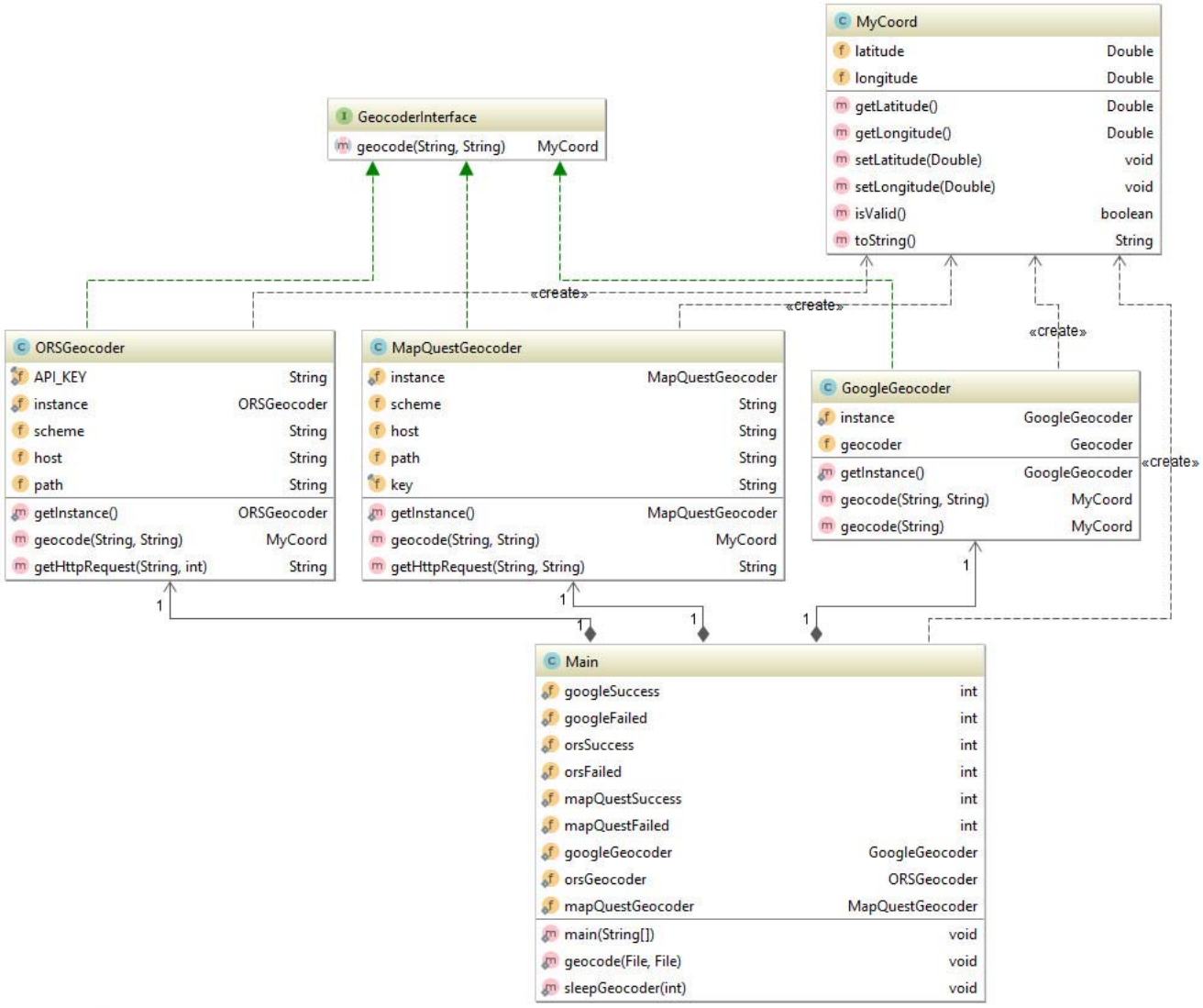

Fig. 5. Class diagram of GFgeocoder

GFgeocoder is a multiplatform tool executable from command line and can be run as a background daemon. GFgeocoder allows loading of input CSV file that contains two columns: first column, named “address”, must contains addresses and house numbers, and second column, named “city”, must contains city name. Output CSV file contains more than two columns: the columns named “address” and “city” are identical to input file columns; the other columns contain respectively pairs of coordinates (latitude and longitude) obtained as results from Online Geocoders Services. GFgeocoder implements geocoding tasks using third party APIs made available to developers. In particular, we used three sets of public APIs:

1) Geocoder Java APIs: a set of unofficial Google Java APIs that is interfaced with official JavaScript Google Maps API v3 [14], via HTTP requests: these APIs provide a set of native calls that make remote requests to Google's official servers, which can be integrated directly into Java application sources.
2) MapQuest RESTful APIs: the official set of geocoding libraries offered by MapQuest. MapQuest offers to developers a RESTful web service that provides a JSON [15] or XML [16] representation of the output requested via a HTTP request. The user provides input parameters, such as address and city name, and builds an HTTP GET request that will be sent to server and waits for its response. Response can contain some detailed information, such as latitude and longitude of point, and some other useful data.
3) OpenRouteService [17] is an online route planning application based on open source software, open data

978-1-5090-6040-5/17/$31.00 ©2017 IEEE

and open standards. It offers many features, such as route services, geocoder and reverse geocoder, accessibility analysis services and emergency route services. It also provides a set of public RESTful APIs that allows user to send geocoding requests. A request for geocoding can be done by an HTTP GET call that returns the response in XML format.

GFgeocoder implements three Java classes that extends a common interface called GeocoderInterface, which have the role to connect the client to these three Online Geocoding Services. Obtained results are passed to a Main class that is responsible for format them and save them in a tabular format to an output CSV file. In Fig. 5 is showed the complete class diagram of GFgeocoder.

IV. RESULTS

Encoded addresses of Italian cities were exactly 7935: 4624 of Trento, 2000 of Florence, and 1311 of Cagliari. For district of Kristiine in Tallinn were 22515. At the end of the first phase of geocoding, we have done appropriate analyzes based on geometric comparison of the distances between the strings of the coordinates obtained from Online Geocoding Services and the coordinates provided by Municipalities.

TABLE I. GA_{10} RESULTS

Geocoding services benchmark - GA_{10}				
City	***Google Maps***	***MapQuest***	***OpenRoute***	***Addresses***
Trento (IT)	0,04 %	0,00 %	0,00 %	4624
Firenze (IT)	90,05 %	0,00 %	6,50 %	2000
Cagliari (IT)	54,51 %	49,12 %	21,97 %	1311
Kristiine (ES)	0.00 %	0.02 %	0.00 %	22515

To analyze the distance between two geographical coordinates (latitude and longitude) we operated an approximation of WGS84 ellipsoid to a local sphere [18]: errors from this approximation are compatible with the basic assumptions. The radius of local sphere is calculated for each of the three locations of analyzed data. Many examples of ranking criteria for benchmarking geocoding services are proposed in literature [19]. To analyze the distance in Baltic93 system we pre-convert results given by GFgeocoder with libproj4 [20] library. After this step, we simply calculate distance on East/North coordinate system given by projected reference system Baltic93. Some preliminary results for cities of Trento, Cagliari and Florence are explained in [21]. Result values of GA_{10} obtained by input data analysis are showed in Fig. 9, Fig. 10, Fig. 11, Fig. 12 and Table I.

For Google Maps Geocoder the best performance of GA_{10} is for Florence addresses, with 90,05 % of string addresses correctly geocoded into 10 meters. The worst is only Kristiine district of Tallinn with no one address correctly geocoded into 10 meters.

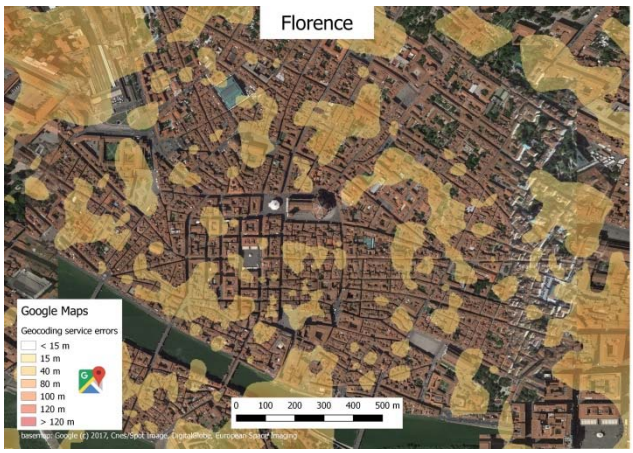

Fig. 6. Google Maps geocoding errors IDW spread of Florence old-town area.

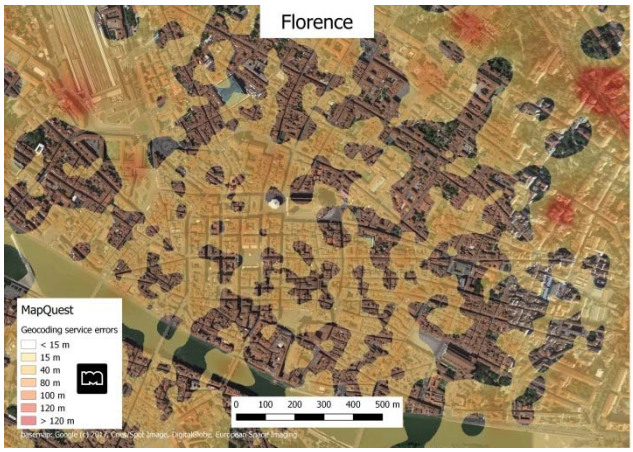

Fig. 7. MapQuest geocoding errors IDW spread of Florence old-town area.

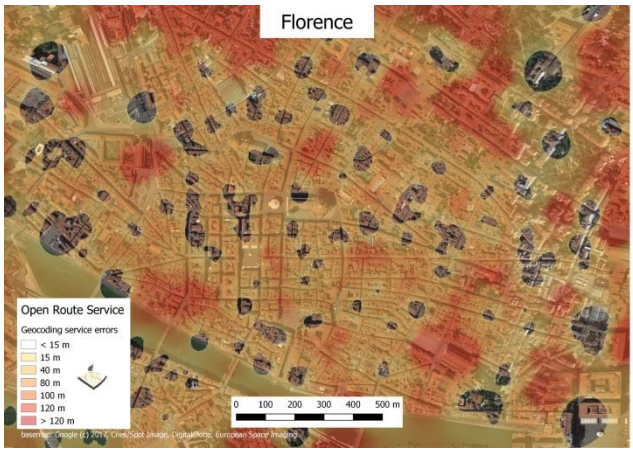

Fig. 8. ORS geocoding errors IDW spread of Florence old-town area.

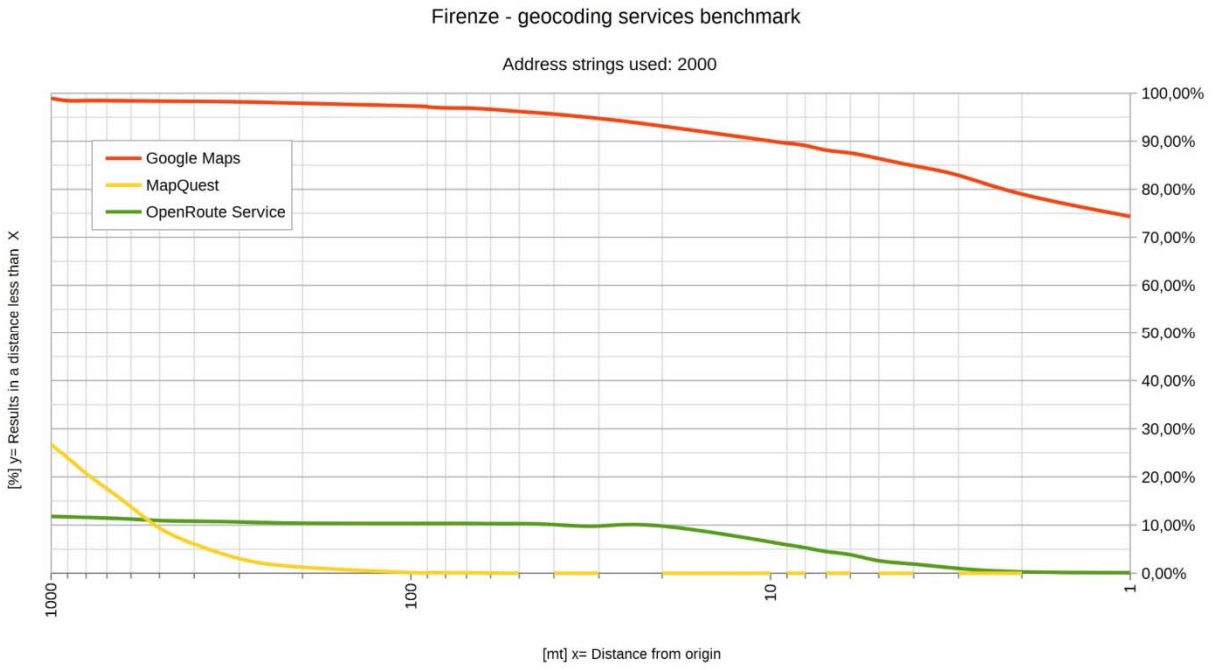

Fig. 9. Benchmark results of the addresses of Florence.

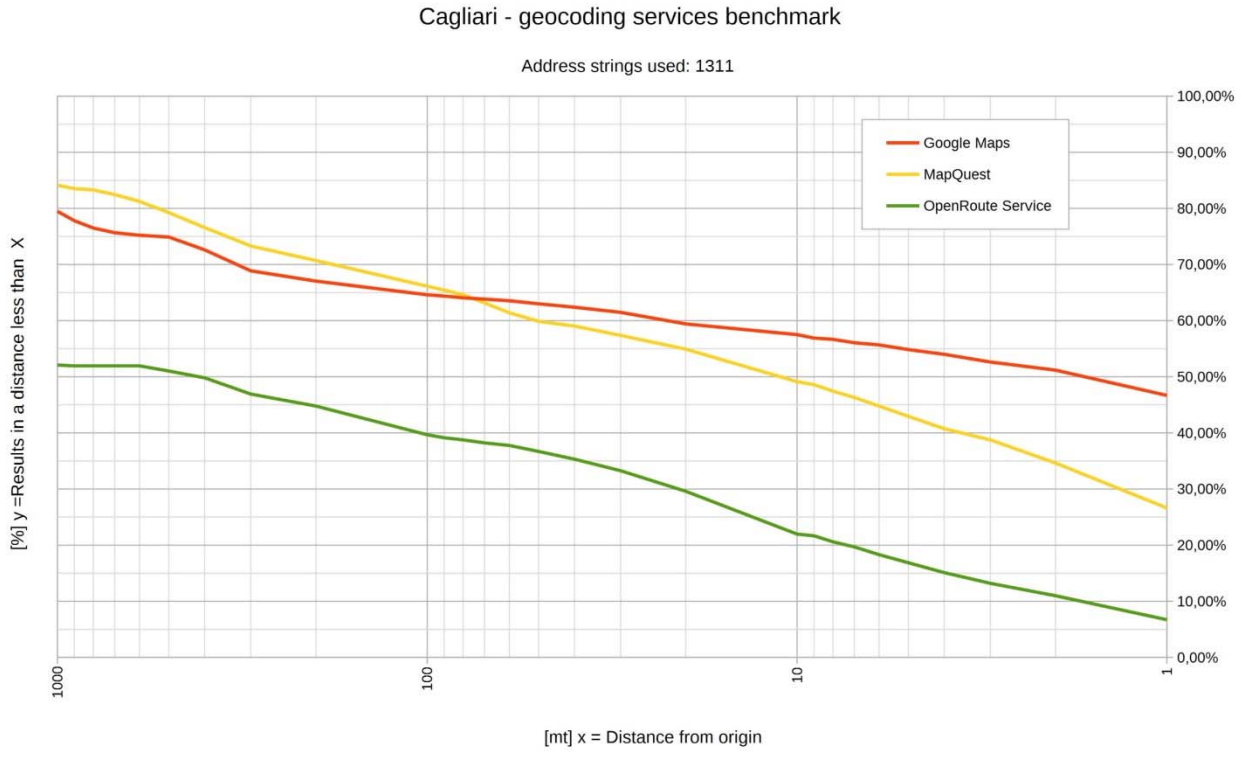

Fig. 10. Benchmark results of the addresses of Cagliari.

978-1-5090-6040-5/17/$31.00 ©2017 IEEE

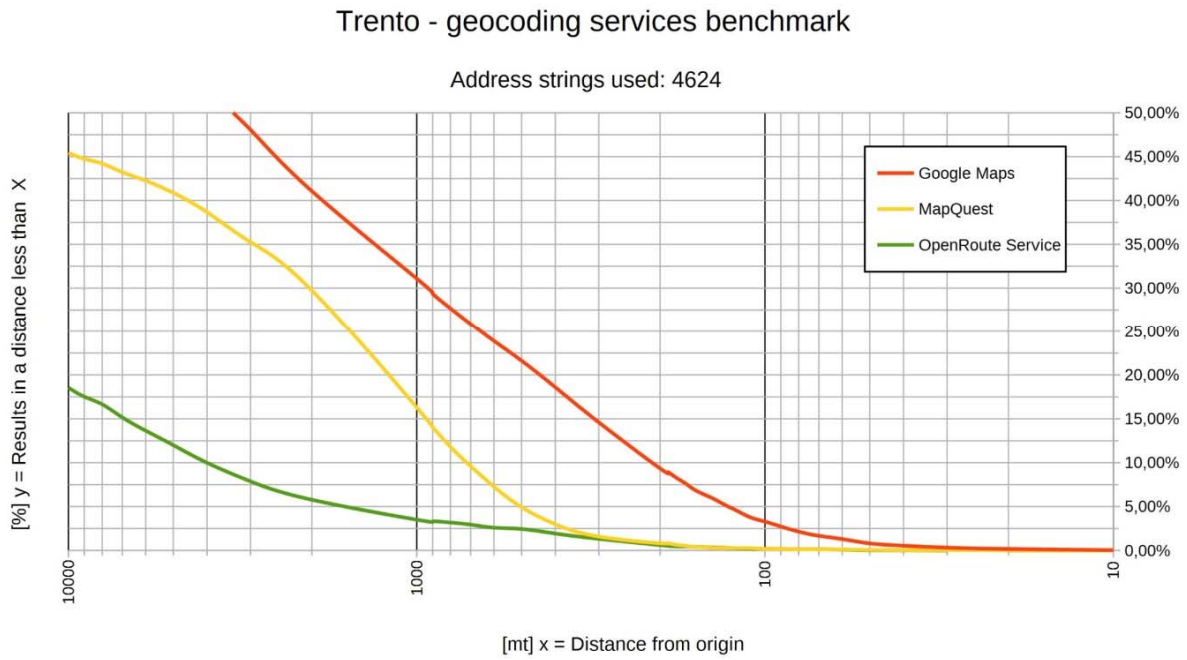

Fig. 11. Benchmark results of the addresses of Trento.

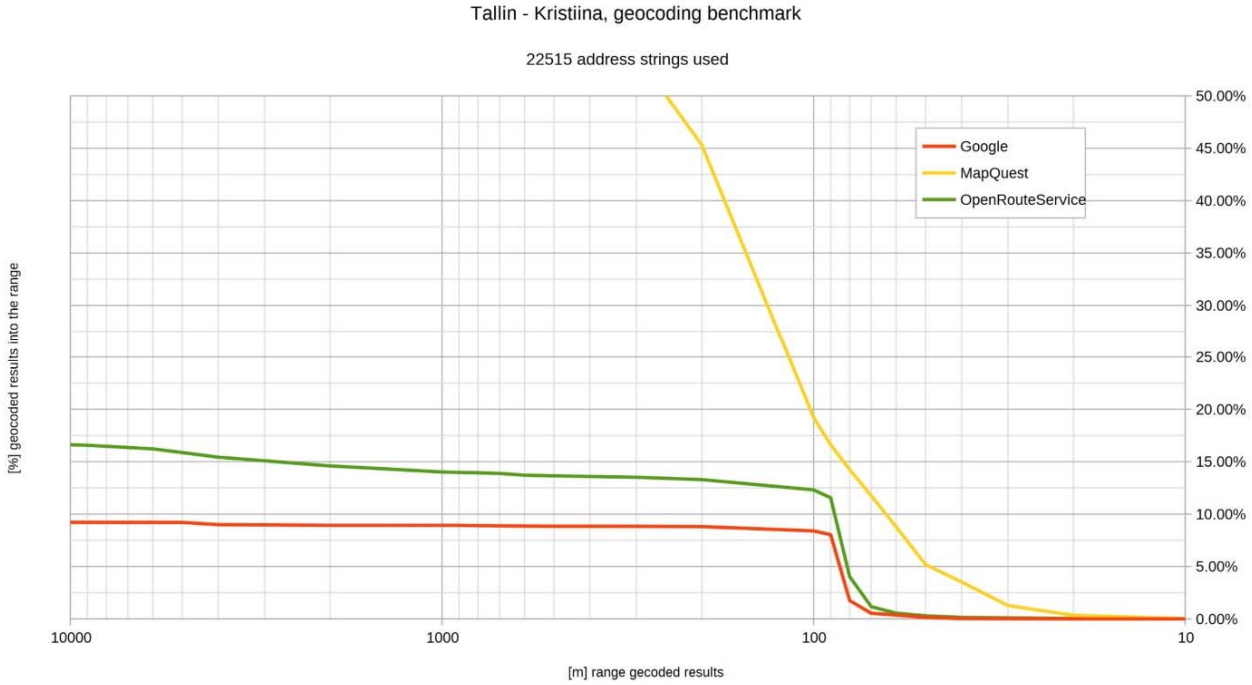

Fig. 12. Benchmark results of the addresses of Kristiine district (Tallin).

978-1-5090-6040-5/17/$31.00 ©2017 IEEE

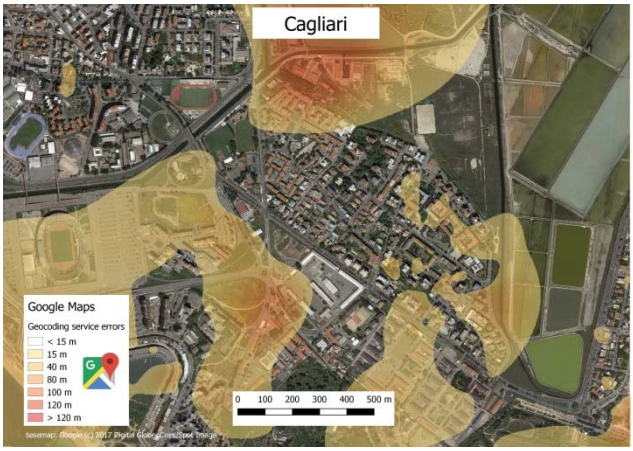

Fig. 13. Google Maps geocoding errors IDW spread of Cagliari old-town area.

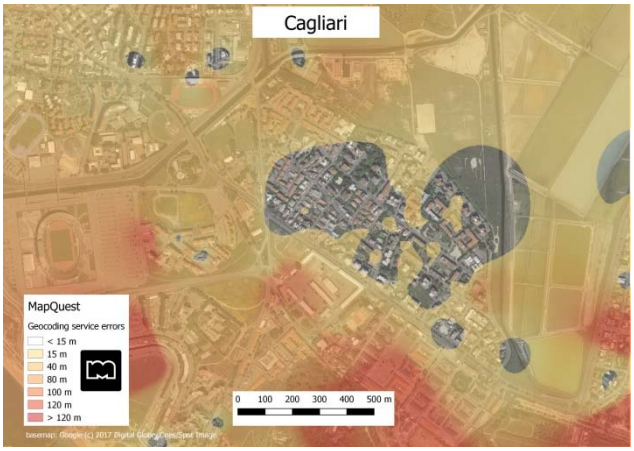

Fig. 14. MapQuest geocoding errors IDW spread of Cagliari old-town area.

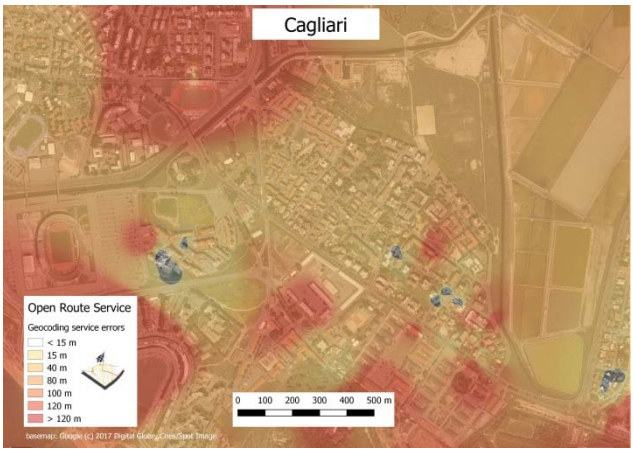

Fig. 15. ORS geocoding errors IDW spread of Cagliari old-town area.

For MapQuest Geocoder the best performance of GA_{10} is for Cagliari city; with 49,12 % of string addresses correctly geocoded into 10 meters. The worst is both Trento city and Florence with less than 0.005 %.

For OpenRouteService the best performance of GA_{10} is for Cagliari city; with 21,7 % of string addresses correctly geocoded into 10 meters. The worst is Kristiine district of Tallinn with only four addresses on 22515 are in a range less than 10 meters.

If we want investigate the spread of these geocoding errors we need use GIS analysis. Using QGIS [22] open source software is possible to realize thematic maps of error spread. To generate a map of errors we used IDW geo-algorithm (Inverse Distance Weighted): it allows to generate GRID data of 2.5×2.5 m (10 x 10 m for Kristiine) cell resolution. For the map plot we used OpenStreetMap as base layer.

In Fig. 6-8 and 13-18 are showed the results of geocoding errors IDW spread for Florence, Cagliari and Trento old-town areas. In Fig. 19-21 are showed the results of geocoding errors IDW spread for Kristiine (Tallin).

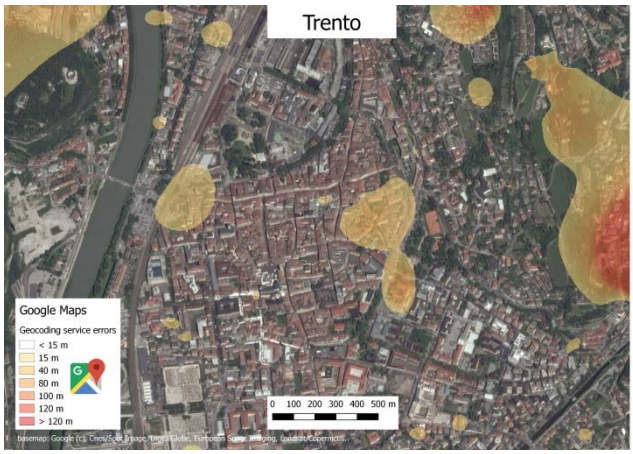

Fig. 16. Google Maps geocoding errors IDW spread of Trento old-town area.

V. CONCLUSIONS

In conclusion, the analysis of the local spread of geocoding errors, allows us to assert that into old-town areas

of the three Italian cities the geocoding errors are least than the total. In particular case of Trento, if we compare the result of whole addresses string geocoded (Fig. 11), to the map of geocoding errors (Fig. 16-21), there is a seeming contradiction due to a high accuracy of all geocoding service for urbanized area (for example old-town area), the geocoding errors for suburban and outskirts area penalize the overall performance result for a city, given by GA10 parameter.

More in-depth analysis showed that the different performance of geocoding services are mainly due to two factors: correct interpretation of the addresses strings, a presence of a direct or indirect survey carried out by the company or organization that manages the service, and other.

978-1-5090-6040-5/17/$31.00 ©2017 IEEE

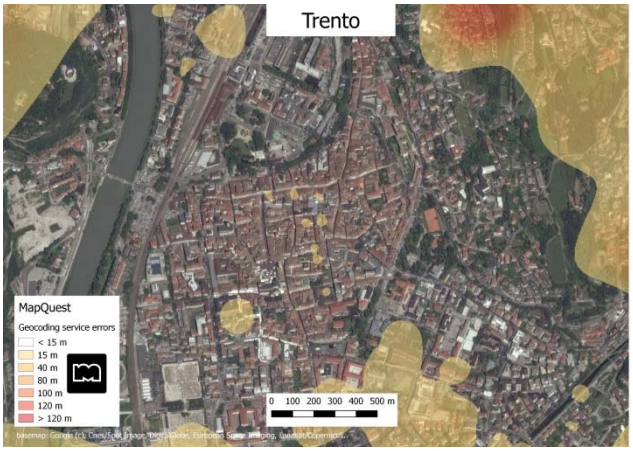

Fig. 17. MapQuest geocoding errors IDW spread of Trento old-town area.

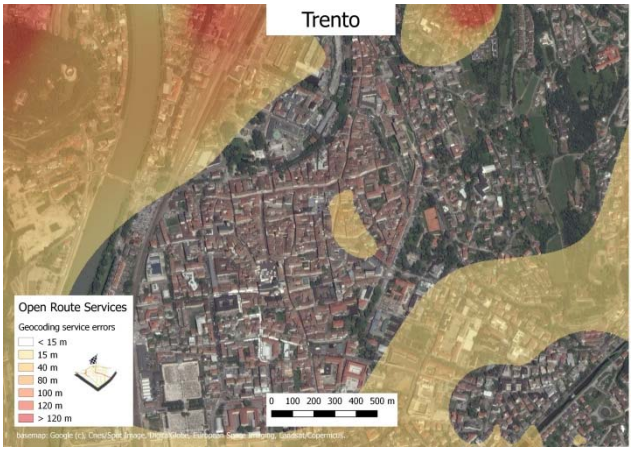

Fig. 18. ORS geocoding errors IDW spread of Trento old-town area.

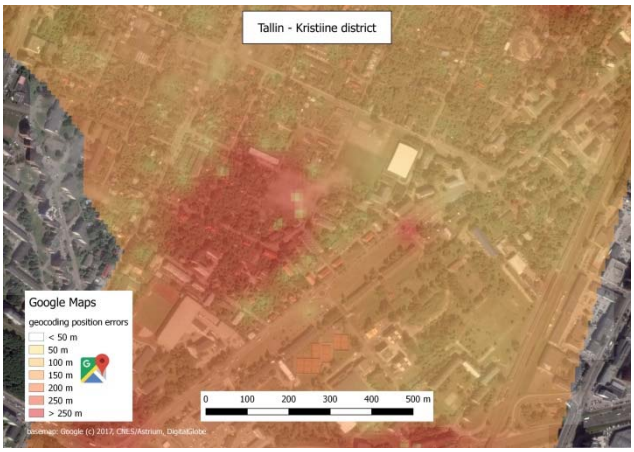

Fig. 19. Google Maps geocoding errors IDW spread of Kristiine (Tallin).

Further investigations will be conducted to investigate these aspects. Also, new comparisons will be made using results obtainable via Nominatim [23], the open source search and geocoding engine that consumes OpenStreetMap data [24]. The methodology explained in this paper is still underway with data of other international locations.

Furthermore, it is part of a larger project that will try to offer a global performance overview of most important Online Geocoding Services.

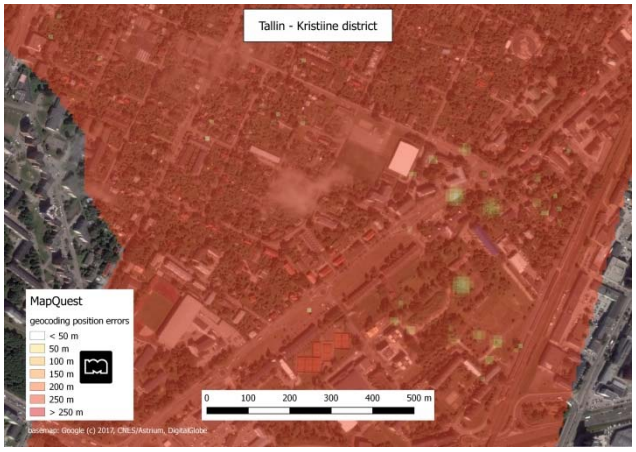

Fig. 20. MapQuest geocoding errors IDW spread of Kristiine (Tallin).

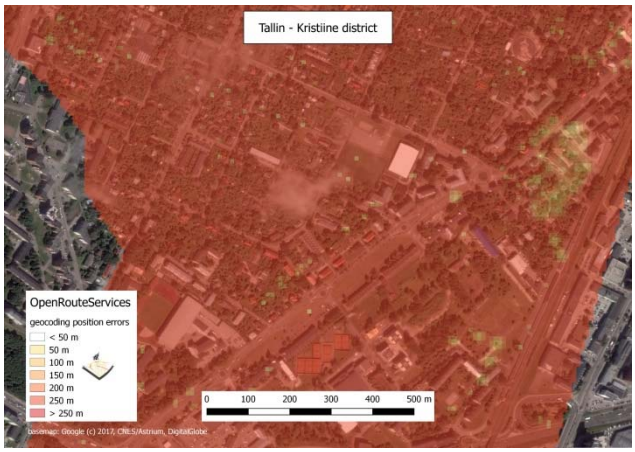

Fig. 21. ORS geocoding errors IDW spread of Kristiine (Tallin).

Sources of last release of GFgeocoder [25] are released under Apache License 2.0 and are currently available for download.

References

[1] D.W. Goldberg, J.P. Wilson, and C.A. Knoblock, "From text to geographic coordinates: The current state of geocoding", URISA Journal, 19(1), pp. 33-46, 2007.

[2] P.A. Zandbergen, "Geocoding quality and implications for spatial analysis", Geography Compass, 3(2), pp. 647-680, 2009, DOI: 10.1111/j.1749-8198.2008.00205.x.

[3] M. Haklay and P. Weber, "OpenStreetMap: User-generated street maps", IEEE Pervasive Computing, 7(4), pp. 12-18, 2008, DOI: 10.1109/MPRV.2008.80.

[4] D. Roongpiboonsopit and H.A. Karimi, "Comparative evaluation and analysis of online geocoding services", International Journal of Geographical Information Science, 24(7), pp. 1081–1100, 2010, DOI: 10.1080/13658810903289478.

[5] C.A. Davis Jr. and R.O. de Alencar, "Evaluation of the quality of an online geocoding resource in the context of a large Brazilian city", Transactions in GIS, 15(6), pp. 851-868, 2011, DOI: 10.1111/j.1467-9671.2011.01288.x.

[6] D.T. Duncan, M.C. Castro, J.C. Blossom, G.G. Bennett, and S.L. Gortmaker, "Evaluation of the positional difference between two common geocoding methods", Geospatial Health, 5(2), pp. 265-273, 2011, DOI: 10.4081/gh.2011.179.

[7] Ward, Mary H., et al. "Positional accuracy of two methods of geocoding", Epidemiology 16.4, pp. 542-547, 2005, DOI: 10.1097/01.ede.0000165364.54925.f3.

[8] T.I. Shah, S. Bell, and K. Wilson. "Geocoding for public health research: Empirical comparison of two geocoding services applied to Canadian cities", The Canadian Geographer, 58(4), pp. 400-417, 2014, DOI: 10.1111/cag.12091.

[9] H.A. Karimi, M.H. Sharker, and D. Roongpiboonsopit, "Geocoding recommender: an algorithm to recommend optimal online geocoding services for applications", Transactions in GIS, 15(6), pp. 869-886, 2011, DOI: 10.1111/j.1467-9671.2011.01293.x.

[10] M. Trevisani, "Gli Open Geodata e la regione Toscana", GEOmedia, 18(6), 2015.

[11] J.R. Clynch, "Radius of the Earth – Radii Used in Geodesy", http://clynchg3c.com/Technote/geodesy/radiigeo.pdf.

[12] R.H. Rapp, "Geometric Geodesy, Part I", Ohio State University Department of Geodetic Science and Surveying, 1991.

[13] W.D Lambert, "The distance between two widely separated points on the surface of the earth", J. Washington Academy of Sciences, 32(5), pp. 125–130, 1942.

[14] G. Svennerberg, "Beginning Google Maps API 3", Apress, 2010.

[15] T. Bray, "The JavaScript Object Notation (JSON) Data Interchange Format", RFC 7159, March 2014.

[16] T. Bray, J. Paoli, C.M. Sperberg-McQueen, E. Maler, and F. Yergeau, "Extensible markup language (XML) 1.0", World Wide Web Consortium Recommendation, 1998, http://www.w3.org/TR/1998/REC- xml-19980210.

[17] P. Neis and A. Zipf, "OpenRouteService.org is three times "Open": Combining OpenSource, OpenLS and openStreetMaps", Proceedings of the GISRUK 2008 conference, Manchester, UNIGIS UK, April 2008.

[18] C. Bernasconi, "Rappresentazioni parziali e totali dell'ellissoide di rotazione sulla sfera", Pure and Applied Geophysics, 33(1), pp. 1-8, 1956, DOI: 10.1007/BF02629941.

[19] D. Teske, "Geocoder accuracy ranking", Process Design for Natural Scientists, Springer Berlin Heidelberg, 2014, DOI: 10.1007/978-3-662-45006-2_13.

[20] Gerald I. Evenden, "A Comprehensive Library of Cartographic Projection Functions (Preliminary Draft)", Falmouth, MA, USA, 2008.

[21] G. Di Pietro and F. Rinnone, "Servizi di geocoding on-line: un'analisi di benchmarking per alcune città italiane", Proocedings of VIII Convegno Nazionale AIT, June 22-23-24, 2016, Palermo, DOI: 10.13140/RG.2.1.3475.2240.

[22] M. Hugentobler, "Quantum GIS", Encyclopedia of GIS, Springer US, 2008.

[23] OpenStreetMap Wiki contributors, "Nominatim", OpenStreetMap Wiki, 2017, http://wiki.openstreetmap.org/wiki/Nominatim.

[24] K. Clemens, "Geocoding with OpenStreetMap Data", GEOProcessing, 2015.M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[25] G. Di Pietro and F. Rinnone, GFgeocoder: v0.2.6 [Data set], Zenodo, 2017, DOI: 10.5281/zenodo. 375794.

978-1-5090-6040-5/17/$31.00 ©2017 IEEE