Geospatial Data Enrichment through Address Geocoding: Challenges and Solutions

Emil Hristov [1], Dessislava Petrova-Antonova [1], Flavio De Paoli [2], Iva Krasteva [1], Michele Ciavotta [2], Roberto Avogadro [2]
[1] Sofia University, GATE Institute, 1164 Sofia, Bulgaria – emil.hristov@gate-ai.eu, dessislava.petrova@gate-ai.eu, iva.krasteva@gate-ai.eu
[2] Dipartimento di Informatica, Sistemistica e Comunicazione, Universit`a di Milano–Bicocca, Milan, Italy – flavio.depaoli@unimib.it, michele.ciavotta@unimib.it, roberto.avogadro@unimib.it

Keywords: Geospatial Data Enrichment, Address Geocoding, SemTUI Framework, QGIS Platform, Accessibility Analysis.

Abstract

Enriching geospatial data, particularly through address geocoding, is fundamental to many geospatial analytical solutions, facilitating resource allocation, accessibility assessment, and urban development planning. This paper explores the address geocoding approaches from traditional GIS-based methods to advanced data-driven solutions, emphasising their accuracy, efficiency, and complementarity. Geospatial data enrichment challenges are explored using the SemTUI framework and QGIS platform. SemTUI employs a data-driven approach that facilitates interactive semantic enrichment and address reconciliation. QGIS is utilised for address geocoding, using the MMQGIS plugin with two geocoding web services provided by OSM and Google. The comparison between SemTUI and QGIS workflows highlights their complementary strengths. SemTUI offers intuitive interfaces and an automated solution, while QGIS provides advanced geospatial visualisation and analysis capabilities. Practical applications in supply-demand analysis of social amenities and walkability index calculation demonstrate the significance of geocoding in urban planning. Further research efforts are focused on enhancing the integration between SemTUI and QGIS to improve the precision and performance of address geocoding and further spatial data analysis.

1. Introduction

Data enrichment methods and tools play a significant role in data processing by delivering meaningful data to implement analytical solutions. However, they face limitations such as a lack of comprehensive support for the entire data enrichment process, including data discovery, understanding, transforming, cleaning, linking and classifying (Ciavotta et al., 2022). Still, comprehensive and user-friendly approaches are needed to address geocoding challenges, particularly in urban planning contexts, where high-quality spatial data is crucial for informed decision-making. Additionally, existing solutions often suffer from the absence of essential Humans-In-The-Loop (HITL) assistance necessary to navigate and resolve uncertainties encountered throughout the data enrichment process, scalability and repeatability remain critical considerations in addressing the evolving demands of geospatial data processing.

Enriching geospatial data is fundamental to many analytical solutions related to resource allocation, accessibility assessment, and urban development planning. In the context of address geocoding, it brings challenges due to the requirement for precise geographic coordinates and the complex nature of address data, which lacks common representation and standardisation (Davis and Fonseca, 2007, Koumarelas et al., 2018). Traditional methods for address geocoding have evolved significantly with advancements in data-driven approaches, probabilistic models, and machine learning techniques enhancing accuracy and efficiency.

This paper explores the possibilities for geocoding addresses by evaluating two distinguished approaches as follows. First, a data-driven approach is applied using the SemTUI framework, which supports interactive semantic data enrichment and address reconciliation to produce high-quality geospatial datasets. Second, QGIS is utilised for the same task by using the MMQGIS plugin with two geocoding web services provided by OSM and Google. A real-world scenario concerning municipal kindergartens and residential buildings in Sofia, Bulgaria, is developed to assess both approaches and obtain insights about their applicability. The practical application of address geocoding is demonstrated through two use cases: supply-demand analysis of social amenities and walkability assessment of residential buildings.

The rest of the paper is organised as follows. Section 2 outlines the related work. Section 3 presents the research methodology, including a description of the SemTUI and QGIS workflows, while Section 4 summarises the results. Section 5 is dedicated to the use cases, and Section 6 concludes the paper, giving directions for future work.

2. Related Work

Geocoding, the process of associating addresses with precise geographic coordinates, has evolved significantly from traditional rule-based methods. In this chapter, we explore the evolution of geocoding from traditional approaches to advanced, data-driven methods. The advancements enhance accuracy, integrate external data, and refine georeferencing techniques, which are crucial for location-based services and spatial data analysis.

2.1 Traditional Rule-Based Geocoding Approaches

Traditional geocoding was primarily accomplished using rule-based systems and databases, often relying on heuristic methods to match textual addresses to their corresponding geographic coordinates. These systems, while effective to a certain extent, suffered from limitations in terms of accuracy, especially when dealing with complex addressing structures, multilingual variations, and user-generated content (Rahul Bakshi et al., 2004). This leads to the need for more sophisticated techniques to improve the precision of geocoding systems. Numerous algorithms and methods have been proposed in the literature to enhance geocoding accuracy based on probabilistic and machine learning methods. The probabilistic geocoding approaches use statistical models to estimate the likelihood of an address being associated with specific geographic coordinates. Approaches such as Hidden Markov Models, Bayesian Networks, and Conditional Random Fields have improved geocoding accuracy, particularly when dealing with ambiguous or incomplete addresses (Christen et al., 2004). Machine learning techniques, including supervised and unsupervised learning algorithms, have

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLVIII-4-2024-239-2024

also shown promising results in geocoding. Support Vector Machines (SVM), Random Forest, and deep learning models like Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) have been employed to learn complex relationships between textual addresses and geographic locations (Lee et al., 2020, Zhang et al. 2022). These approaches often incorporate vast datasets to train models that can handle variations in addressing conventions and formats.

2.2 Integration of External Data Sources

Integrating external data sources, such as street maps, building footprints, and land-use information, has become increasingly popular. By incorporating these datasets, geocoding systems can leverage additional context to enhance accuracy and reduce ambiguity. Research in this area focuses on data fusion techniques and matching algorithms that combine textual information with spatial data (Lane et al., 2023, Zandbergen et al., 2008). Address matching and geocoding systems must deal with errors and ambiguities in textual addresses. Research in this area has led to advancements in error analysis, quality assessment, and methods for resolving inconsistencies. Algorithms for address standardisation, data cleansing, and error correction play a vital role in improving the overall performance of geocoding systems (Lane et al., 2023, Schootman et al., 2007, Kravets et al., 2007).

2.3 Semantic Data Enrichment

A systematic literature review on recent approaches for the automatic identification of addresses revealed the importance of semantics with regard to methods and algorithms for address geocoding (Cruz et al., 2022). The SemTUI framework facilitates the data enrichment process by assisting users in enriching tabular data by integrating data linking and extension services. It supports data discovery and linking by employing advanced algorithms to integrate data from various reference sources while leveraging open-linked data and private knowledge graphs. Furthermore, SemTUI provides interactive reconciliation and extension, involving users in the identification and revision of uncertain links (Ripamonti et al., 2022, Cara et al., 2023, Udroiu et al., 2023). It has a backend serving as an advance gateway for access to a variety of services, including built-in (e.g., Alligator), third-party (e.g., Atoka and services compatible with OpenRefine), or a combination of both (e.g., the built-in GEO service relying on HERE and OSM services) (Cutrona et a., 2019, Avogadro et al., 2022).

2.4 Geographic Information Systems

Geographic Information Systems (GISs) provide a powerful functionality that covers the whole data lifecycle, such as capturing, storing, enriching, analysing, visualising, etc. There is much software that supports the use of GIS technologies in different fields like urban planning and management, health, services, etc. (Chang, 2019, Mennecke et al., 1996). Some of the most widely used are QGIS and ArcGIS Pro. QGIS is an open-source GIS for creating, editing, visualising, analysing, and publishing geospatial information across various operating systems. To geocode a bulk amount of addresses effectively, it's essential to identify a plugin tailored to the specific needs of the task. One of the most promising ones is the MMQGIS plugin (Minn, 2021). It can work with 5 different web services, namely Google with API key, OpenStreetMap/Nominatim, US Census Bureau, ESRI server and NetToolKit. Each of these services offers varying levels of geocoding accuracy, depending on the geographic area of the addresses being processed and their format. ArcGIS Pro is a GIS application supporting data visualization, data analysis, and maintenance, boasting various instruments and capabilities. The geocoding of addresses requires using of ArcGIS World Geocoding Service (ArcGIS). The service can geocode addresses that are stored in a single field, divided into multiple fields, or stored in a single field and a country field. Some locators support multiple input address fields, where the address component can be separated into multiple fields, and the address fields are concatenated at the time of geocoding.

2.5 Internet-based Mapping Services

In the past few years, the democratisation of internet-based mapping services has empowered non-GIS users to use geocoding services such as Google Maps or MapQuest (Wu et al., 2005; Roongpiboonsopit and Karimi, 2010). The OpenStreetMap (OSM) Nominatim API supports geocoding and reverse geocoding, providing a straightforward translation between geographic coordinates and place names (Geoglify, 2024). NetToolKit consults several data sources to deliver better results when attempting to geocode an address. These data sources have overlapping coverages of the US (i.e., an address might be included in one source but not the others). For non-US addresses and landmarks, NetToolKit relies on Nominatim. It also queries PostGIS for US addresses, as it offers a different method of parsing input (NetToolKit). In the context of geocoding, PostGIS enhances accuracy and efficiency by parsing and standardising address data, utilising spatial indexing for faster geocoding, and integrating with external datasets for richer context. It supports multi-field address parsing, error correction, and spatial querying, particularly beneficial for US addresses, making it a valuable tool in the geocoding process within the PostgreSQL database environment.

3. Methodology

This section outlines the methodology employed to assess geocoding approaches using both the SemTUI framework and QGIS. Their efficacy in handling the geospatial data enrichment challenges is evaluated in the context of urban planning scenarios, particularly in Sofia, Bulgaria.

3.1 Study Area and Data

The study area covers Sofia, the capital of Bulgaria, where historical planning variations have caused insufficient coverage of essential services like schools, medical centres, and, especially, kindergartens. Sofia presents an ideal case study for evaluating geocoding approaches in urban planning contexts due to its diverse spatial characteristics and ongoing development projects. To evaluate the accessibility of kindergartens within walking distance, a spatial analysis should be performed based on the residential address of the children. The dataset, comprising over 23,621 data points sourced from the municipal Directorate of Education, encompasses addresses of children aged 1 to 6 years, along with an extensive list of over 300 childcare amenities.

A dataset is prepared to fulfil the specific requirements of the SemTUI framework and QGIS. Subsequently, a subset of 1028 addresses, representing diverse children enrolled across 12 kindergartens, is selected to validate the SemTUI service's functionality. To explore QGIS, 200 addresses are selected from the dataset prepared for the SemTUI framework.

3.2 SemTUI workflow

The SemTUI framework is used for interactive semantic enrichment, including the following steps:

- Preprocessing, excluding lacking addresses and preschool classes and removing non-relevant columns.
- Address reconciliation, comprising the identification and utilisation of a linking service from those available in SemTUI. At this step, geo-coordinates are assigned to each address, leveraging the HERE Geocoder as the geocoding service.
- Geocoding extension, including further augmentation of the reconciliation process by supplementary information. This optional step adds route length, walking duration, and a well-formed Latin transliteration of the addresses.
- Review, engaging the user to find potential inaccuracies caused by the geocoding task.
- Resolve reconciliation discrepancies, involving correction actions to deal with inaccuracies identified during the review process.
- Enrich for downstream analytics, including extending the geocoded dataset with supplementary data.

3.3 GIS-based Tools Workflow

MMQGIS plugin is used for address geocoding in QGIS. The steps performed for address geocoding are as follows:

- Preprocessing, including splitting the address information based on address parts and removing the unnecessary address components to streamline the geocoding process.
- Translating the addresses from Cyrillic to Latin to enable compatibility with the plugin.
- Geocoding, involving the specification of the fields that will be used, determination of how to deal with duplicates and storage location for the output and the list with unmatched addresses.
- Export, including a selection of data format for the export, usually as a shapefile with a geometry point for each address.
- Visualisation of the geocoded dataset on 2D maps and further visual exploration of the geocoding issues that arise.

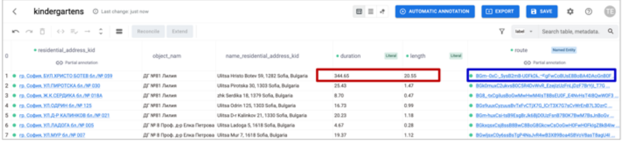
4. Results

This section presents results obtained from the SemTUI framework and GIS, specifically QGIS in this case.

4.1 SemTUI Results

The SemTUI framework automatically enriched the source table by adding geocoordinates of the Cyrillic addresses and successfully converted them into their Latin counterparts. This opens opportunities for enhanced cross-platform compatibility and data interoperability. The address geocoding demonstrated an accuracy rate of approximately 95% for residential addresses. Problems with the remaining 5% were identified using the visual support provided by SemTUI to identify and correct anomalies if they appear during the reconciliation to enforce a human-in-the-loop approach.

An example is illustrated in Figure 1, where anomalies above a certain threshold can be detected by numbers (enclosed in a red rectangle) and verified by clicking on the polyline (enclosed in a blue rectangle), which opens the view in Figure 2. This view clearly indicates that the discovered address is located outside the city; therefore, this anomaly appears to be an error. In this case, it was caused by a misinterpretation of the abbreviation for ‘boulevard’ in the address.



0	residential_address_kid	object_nam	name_residential_address_kid	duration	length	route
1	гр. София, БУЛ.ХРИСТО БОТЕВ бл./№ 059	ДГ №81 Лилия	Ulitsa Hristo Botev 59, 1282 Sofia, Bulgaria	344.65	20.55	BGm-0xC-_SysB2mB-U0FkDk-~FgFwCoBUsE8BoBA4DAoGnB0F
2	гр. София, УЛ.ПИРОТСКА бл./№ 030	ДГ №81 Лилия	Ulitsa Pirotska 30, 1303 Sofia, Bulgaria	25.43	1.47	BGk0muxC2ukvsB0C5R4DvWvR_EzejlzUzFnLjDzF7BrYjL_T7G ...
3	гр. София, Ж.К.СЕРДИКА бл./№ 018А	ДГ №81 Лилия	zhk Serdika 18, 1379 Sofia, Bulgaria	8.70	0.47	BG8_-txCgilusBoGwMwHwM4IsTBBsEU0F_E4NvHsT4I8QwW0F3 ...
4	гр. София, УЛ.ОДРИН бл./№ 125	ДГ №81 Лилия	Ulitsa Odrin 125, 1303 Sofia, Bulgaria	16.73	0.99	BGs9uuxCyzuusBvTxFvCTjX7G_lCrT3X7G7sCvWrEnB7L3DzrC ...
5	гр. София, УЛ.Д-Р КАЛИНКОВ бл./№ 021	ДГ №81 Лилия	Ulitsa D-r Kalinkov 21, 1330 Sofia, Bulgaria	20.23	1.18	BGm-huxCsi-tsB9EsgBrJk6BjD0UzFsnB7B0K7BwM7BsJnBoGv ...
6	гр. София, УЛ.ЛАДОГА бл./№ 005	ДГ № 8 Проф. д-р Елка Петрова	Ulitsa Ladoga 5, 1618 Sofia, Bulgaria	4.67	0.28	BGkxqsxCsj8ss888wC8BcG8GkcwCsOoGwH0FwH0FklgZ8kB4w ...
7	гр. София, УЛ.МУР бл./№ 007	ДГ № 8 Проф. д-р Елка Петрова	Ulitsa Mur 7, 1618 Sofia, Bulgaria	19.37	1.12	BGwljsxC0y6ssBsTgP4NsJvR4wB3X89Boa4S8VzV8asT8agU4I ...

Figure 1. Visual detection of anomalies in discovered addresses by examining the added columns.

Anomalies can be detected and possibly fixed by domain experts. Thus, various forms of user input can be integrated, allowing domain adaptation through user feedback. Consequently, fine-tuning of the provided services can be achieved for a specific domain of the processed data.

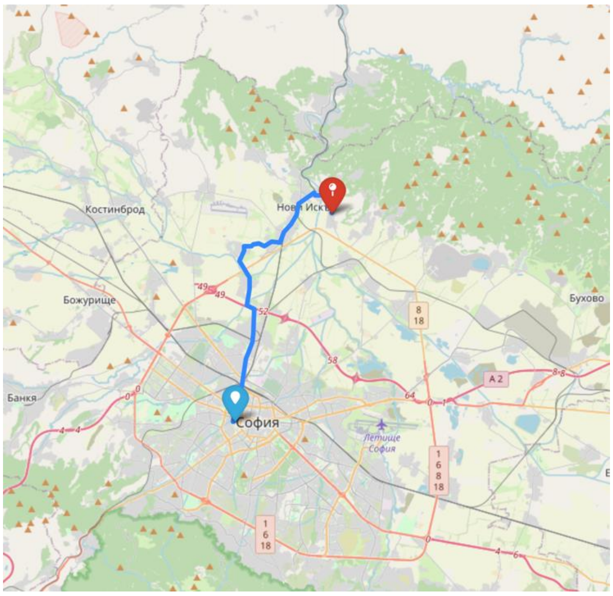
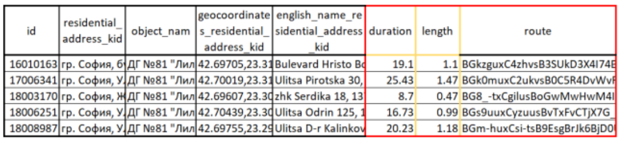


Figure 2. Visualisation of a polyline in SemTUI to check anomalies: the discovered address is outside the city.

A sample illustrating the results from the enrichment process is presented in Figure 3.



id	residential_address_kid	object_nam	geocoordinates_residential_address_kid	english_name_residential_address_kid	duration	length	route
16010163	гр. София, б	ДГ №81 "Лил	42.69705,23.31	Bulevard Hristo Bo	19.1	1.1	BGkzguxC4zhvsB3SUkD3X4I74E
17006341	гр. София, У	ДГ №81 "Лил	42.70019,23.31	Ulitsa Pirotska 30,	25.43	1.47	BGk0muxC2ukvsB0C5R4DvWvF
18003170	гр. София, Ж	ДГ №81 "Лил	42.69607,23.30	zhk Serdika 18, 13	8.7	0.47	BG8_-txCgilusBoGwMwHwM4I
18006251	гр. София, У	ДГ №81 "Лил	42.70439,23.30	Ulitsa Odrin 125, 1	16.73	0.99	BGs9uuxCyzuusBvTxFvCTjX7G_
18008987	гр. София, У	ДГ №81 "Лил	42.69755,23.29	Ulitsa D-r Kalinkov	20.23	1.18	BGm-huxCsi-tsB9EsgBrJk6BjD0

Figure 3. Enrichment outcome: new columns added to the source table with duration, length, and polyline of routes.

The enriched dataset includes a dedicated column with encoded polylines representing routes. A specialised Python script was developed and applied to convert the routes from polylines to arrays of coordinates, rendering them suitable for visualisation in GIS, specifically QGIS. Moreover, two supplementary columns, route duration and length, have been added. As a result, the quality of the dataset is enhanced, offering insights into the routes' temporal and spatial dimensions and thereby improving the understanding of the data.

This exploration of results from the SemTUI framework shows minor issues. Incorrect or unconnected routes were discovered

(Figure 4). For instance, some residential addresses are accurately geocoded, and their corresponding kindergartens were identified but remained unconnected.

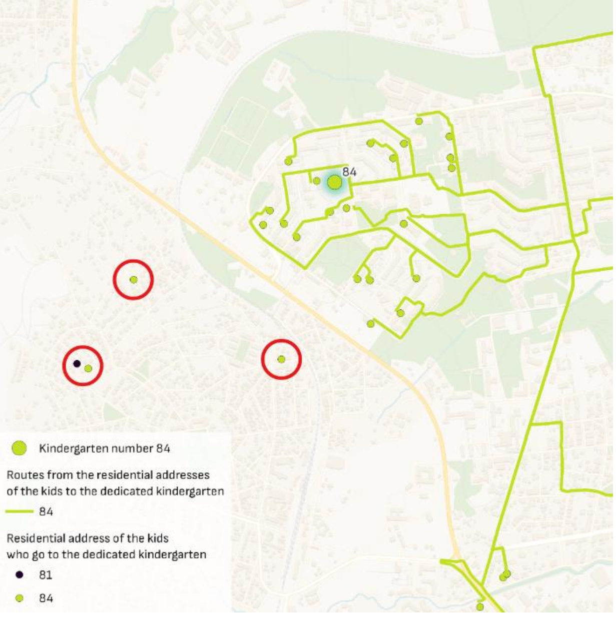
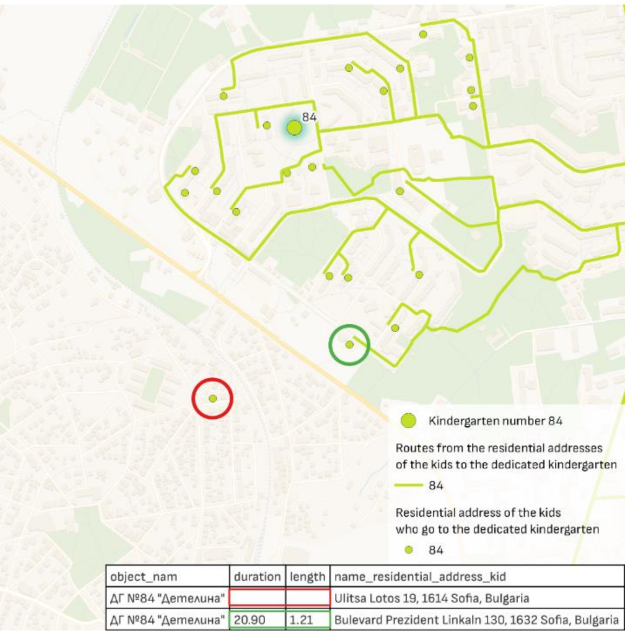

84
Kindergarten number 84
Routes from the residential addresses of the kids to the dedicated kindergarten
84
Residential address of the kids who go to the dedicated kindergarten
81
84

Figure 4. Unconnected residential buildings and kindergartens.

The lack of information on the route length and duration also leads to a missing connection between a residential building and its corresponding kindergarten (Figure 5).


84
Kindergarten number 84
Routes from the residential addresses of the kids to the dedicated kindergarten
84
Residential address of the kids who go to the dedicated kindergarten
84

object_nam	duration	length	name_residential_address_kid
ДГ №84 "Детелина"			Ulitsa Lotos 19, 1614 Sofia, Bulgaria
ДГ №84 "Детелина"	20.90	1.21	Bulevard Prezident Linkaln 130, 1632 Sofia, Bulgaria

Figure 5. Connected (in green circle) kindergarten and unconnected (in red circle) kindergarten due to missing route duration and length.

4.2 QGIS Results

Geocoding with QGIS is tested using the MMQGIS plugin. Two of the five available geocoding web services, provided by OSM and Google, were used. Each of them gives a different degree of success regarding the number of geocoded addresses. Note that the way and order the addresses are written affect the results. For some addresses, OSM gives a better result, and for others - Google. In addition, the same address geocoded with a different web service might receive a different location. In some scenarios, MMQGIS cannot process addresses with any web service.

The results from address geocoding performed on a dataset with 200 residential addresses are shown in Figure 6. Notably, Google web service successfully geocoded 195 addresses, while OSM web service managed only 28. However, it's essential to recognise the limitations of Google web service, as it imposes restrictions on the number of free geocoding requests before payment is required. In contrast, OSM offers a free-for-use web service but demands careful preprocessing of address data.

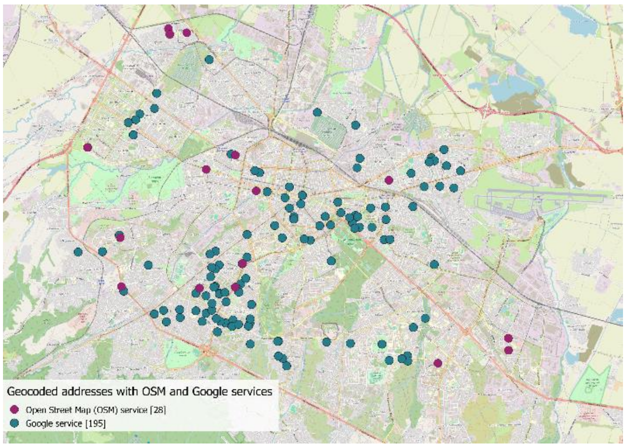
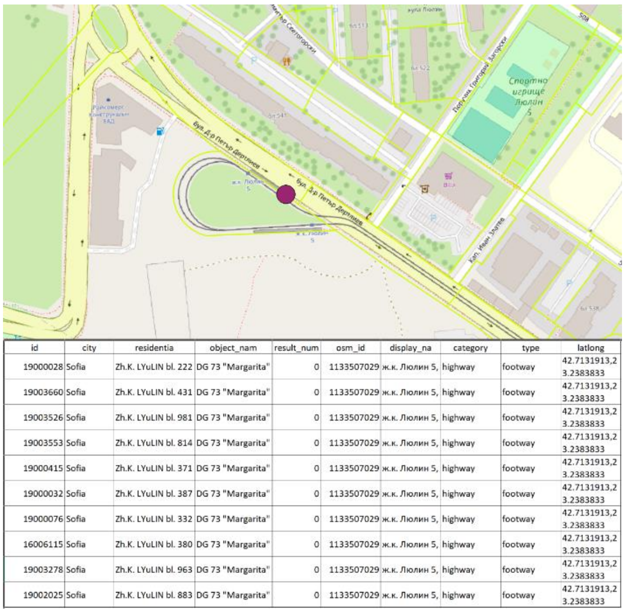

Geocoded addresses with OSM and Google services
Open Street Map (OSM) service [28]
Google service [195]

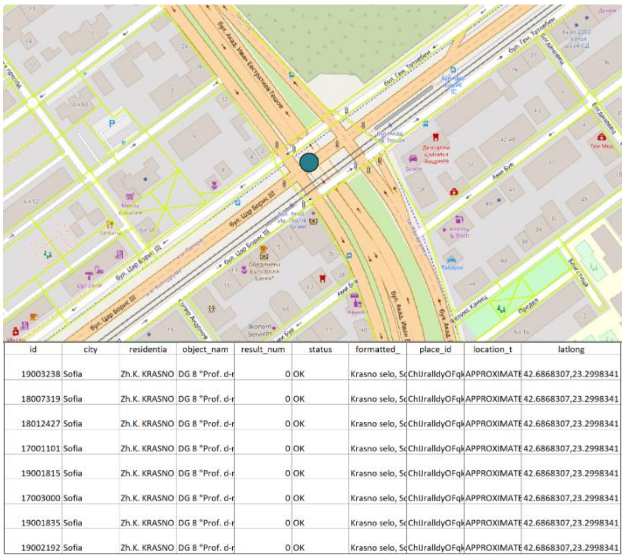
Figure 6. Geocoding results, obtained with Google and OSM web services.

Despite the quantity of successful geocoded addresses, both web services demonstrated errors in quality. Figures 7 and 8 illustrate a common issue where multiple addresses are incorrectly geocoded to a single point, often located in a road infrastructure zone outside of a residential area.



id	city	residentia	object_nam	result_num	osm_id	display_na	category	type	latlong
19000028	Sofia	Zh.K. LYuLIN bl. 222	DG 73 "Margarita"	0	1133507029	ж.к. Люлин 5,	highway	footway	42.7131913,23.2383833
19003660	Sofia	Zh.K. LYuLIN bl. 431	DG 73 "Margarita"	0	1133507029	ж.к. Люлин 5,	highway	footway	42.7131913,23.2383833
19003526	Sofia	Zh.K. LYuLIN bl. 981	DG 73 "Margarita"	0	1133507029	ж.к. Люлин 5,	highway	footway	42.7131913,23.2383833
19003553	Sofia	Zh.K. LYuLIN bl. 814	DG 73 "Margarita"	0	1133507029	ж.к. Люлин 5,	highway	footway	42.7131913,23.2383833
19000415	Sofia	Zh.K. LYuLIN bl. 371	DG 73 "Margarita"	0	1133507029	ж.к. Люлин 5,	highway	footway	42.7131913,23.2383833
19000032	Sofia	Zh.K. LYuLIN bl. 387	DG 73 "Margarita"	0	1133507029	ж.к. Люлин 5,	highway	footway	42.7131913,23.2383833
19000076	Sofia	Zh.K. LYuLIN bl. 332	DG 73 "Margarita"	0	1133507029	ж.к. Люлин 5,	highway	footway	42.7131913,23.2383833
16006115	Sofia	Zh.K. LYuLIN bl. 380	DG 73 "Margarita"	0	1133507029	ж.к. Люлин 5,	highway	footway	42.7131913,23.2383833
19003278	Sofia	Zh.K. LYuLIN bl. 963	DG 73 "Margarita"	0	1133507029	ж.к. Люлин 5,	highway	footway	42.7131913,23.2383833
19002025	Sofia	Zh.K. LYuLIN bl. 883	DG 73 "Margarita"	0	1133507029	ж.к. Люлин 5,	highway	footway	42.7131913,23.2383833

Figure 7. Common geocoding error: multiple addresses mapped to a road infrastructure zone (OSM).



id	city	residentia	object_nam	result_num	status	formatted_	place_id	location_t	latlong
19003238	Sofia	Zh.K. KRASNO	DG 8 "Prof. d-r	0	OK	Krasno selo, So	ChIJralldyOFqk	APPROXIMATE	42.6868307,23.2998341
18007319	Sofia	Zh.K. KRASNO	DG 8 "Prof. d-r	0	OK	Krasno selo, So	ChIJralldyOFqk	APPROXIMATE	42.6868307,23.2998341
18012427	Sofia	Zh.K. KRASNO	DG 8 "Prof. d-r	0	OK	Krasno selo, So	ChIJralldyOFqk	APPROXIMATE	42.6868307,23.2998341
17001101	Sofia	Zh.K. KRASNO	DG 8 "Prof. d-r	0	OK	Krasno selo, So	ChIJralldyOFqk	APPROXIMATE	42.6868307,23.2998341
19001815	Sofia	Zh.K. KRASNO	DG 8 "Prof. d-r	0	OK	Krasno selo, So	ChIJralldyOFqk	APPROXIMATE	42.6868307,23.2998341
17003000	Sofia	Zh.K. KRASNO	DG 8 "Prof. d-r	0	OK	Krasno selo, So	ChIJralldyOFqk	APPROXIMATE	42.6868307,23.2998341
19001835	Sofia	Zh.K. KRASNO	DG 8 "Prof. d-r	0	OK	Krasno selo, So	ChIJralldyOFqk	APPROXIMATE	42.6868307,23.2998341
19002192	Sofia	Zh.K. KRASNO	DG 8 "Prof. d-r	0	OK	Krasno selo, So	ChIJralldyOFqk	APPROXIMATE	42.6868307,23.2998341

Figure 8. Common geocoding error: multiple addresses mapped to a road infrastructure zone (Google).

The generation of the shortest path in QGIS has not been explicitly performed. The software offers various functionalities provided by both external plugins like ORS and Traveling Salesman analysis, and inbuilt network analysis tools. Figure 8 shows a successful generation of the shortest path between two points using the Network analysis tool and, more specifically, the Shortest path function. It should be noted that the execution time has taken around 30 seconds, which raises the question of the applicability of the tool on large datasets.

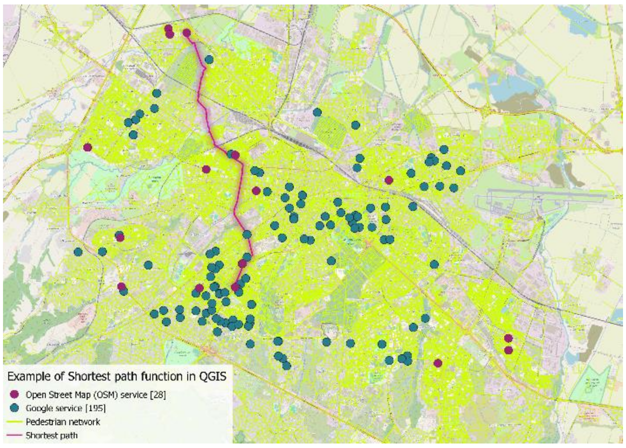

Example of Shortest path function in QGIS
● Open Street Map (OSM) service [28]
● Google service [195]
Pedestrian network
Shortest path

Figure 9. Shortest path between two points, generated with Network analysis tool in QGIS.

5. Use Cases

The geocoding of addresses is successfully applied in two use cases: supply-demand analysis of social infrastructure and walkability index of residential buildings based on walking access to different points of interest (POIs).

A unified solution for supply-demand analysis of public amenities considering a 15-minute walking distance. The solution is validated through a real case study related to the kindergartens and nurseries in Sofia Municipality, where a shortage of these amenities affects around 12,000 children. It is based on geospatial data for the public facilities, residential buildings, and pedestrian network; demographic data for estimating the demand capacity of the residential building; and regulations used to calculate the supply capacity of the kindergartens and nurseries. The proposed solution finds the best possible distribution of the children living in the residential buildings among the public facilities by achieving optimal coverage. A sample result from the supply-demand analysis of Sofia city centre is shown in Figure 10.

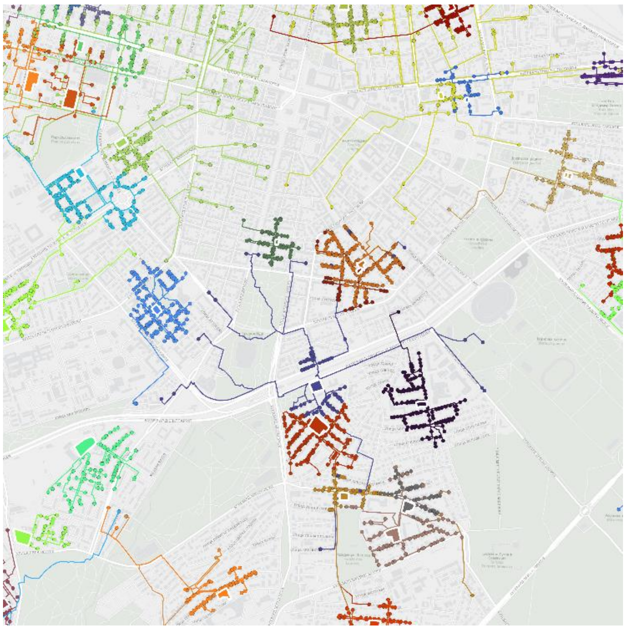


Figure 10. Sample result from the supply-demand analysis of Sofia city centre.

The primary goal of the supply-demand analysis is to ensure maximum utilisation of the capacity of public facilities while minimising the overall distance children need to travel to access these facilities. The distance between public facilities and residential buildings is determined based on the shortest possible path within the pedestrian network. As a result, underserved areas and areas with an over-supply are determined.

Although the coverage of kindergartens and nurseries is considered for validation, the solution is developed in a unified way that allows its application to other public facilities such as schools, health centres, public transport stops, etc. This motivates the second use case related to the walkability index of residential buildings in District Lozenets”, Sofia. The study evaluates the walkable access to the available POIs, considering perceptions of convenience, how people spend their valuable time and their proximity. The POIs are classified into primary and secondary groups and assigned weights based on their significance, permanence, and public or private status. The walkability index for a single residential building is calculated as follows: first, a score for each primary group is calculated by summing up the weights of accessible POIs in secondary groups and multiplying the result by the weight of the primary group, second, a final score is calculated by sum up all calculated scores for each primary group. A result from the calculation of the walkability index is shown in Figure 11.

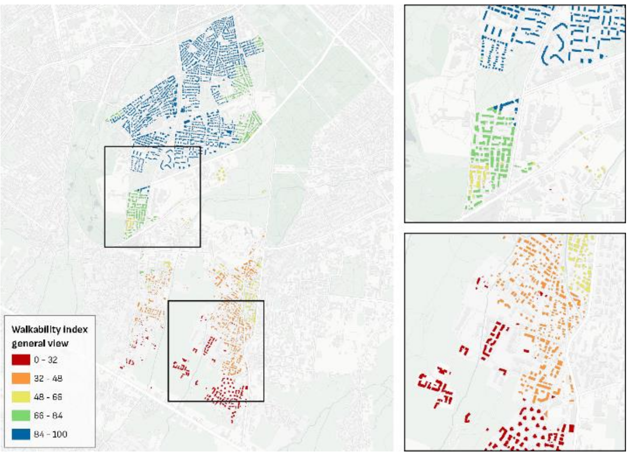


Figure 11. Walkability index calculated for District “Lozenets”.

6. Conclusion

This study demonstrated the important role of data enrichment methods in facilitating analytical solutions in the geospatial domain, particularly in addressing geocoding. The SemTUI framework and QGIS are used to assess the applicability of both data-driven and GIS-based approaches. SemTUI offers an intuitive interface and fully automated solution, making it practicable even for users with limited technical proficiency. However, it lacks valuable visualisation capabilities. In contrast, QGIS requires the usage of plugins or external tools for geocoding but provides inbuilt visualisation and geospatial analysis functionalities.

The comparison between SemTUI and QGIS workflows reveals complementary strengths. SemTUI excels in efficiently processing datasets to generate geocoded points and optimal routes. On the other hand, QGIS provides a platform for advanced visualisation, analysis and publishing of geospatial data on different platforms and devices. Such complementarity provides an opportunity to combine or integrate them into a common data pipeline that can enhance performance, execution time, and ease of use.

Furthermore, the presented use cases, including the supply-demand analysis of social amenities and the calculation of walkability indices, highlight the significance of geocoding in urban planning and design. They illustrate how geocoding, facilitated by tools like SemTUI and QGIS, can provide valuable insights into resource allocation, accessibility analysis, and urban development strategies.

Further research and development efforts should focus on enhancing the integration between SemTUI and QGIS, as well as providing additional, rich functionalities for the end users, such as improved visualisation tools and automated error detection mechanisms. This integrated approach could lead to improved efficiency and effectiveness in the address geocoding and spatial data analysis, ultimately benefiting various industries in the geospatial domains.

The integration of SemTUI and QGIS presents a promising opportunity to enhance the precision, performance, and efficiency of address geocoding and spatial data analysis. By leveraging the complementary strengths of both tools, users can benefit from a comprehensive data pipeline that combines data enrichment capabilities offered by SemTUI with advanced visualisation and analysis functionalities provided by QGIS. Integrating SemTUI for data preprocessing and enrichment, followed by seamless data transfer to QGIS for visualisation and analysis, can streamline workflows and improve overall productivity. This integrated approach holds the potential for addressing the evolving demands of geospatial data processing and facilitating more informed decision-making in urban planning and analysis contexts.

Acknowledgements

This research was supported by the GATE project funded by the Horizon 2020 WIDESPREAD-2018-2020 TEAMING Phase 2 programme under agreement no. 857155, the enrRichMyData project, funded by Horizon Europe research and innovation programme under agreement No. 101070284, and the FLEdge project, funded under the Driving Urban Transitions (DUT) Partnership program, under agreement no. KP-06-D002/5.

References

Ciavotta, M., Cutrona, V., De Paoli, F., Nikolov, N., Palmonari, M., Roman, D., 2022. Supporting semantic data enrichment at scale, *Technologies and Applications for Big Data Value*, no. 732590, Springer, 19–39.

Davis, C. A., Fonseca, F. T., Assessing the certainty of locations produced by an address geocoding system, *Geoinformatica* 11 (2007) 103–129.

Koumarelas, I. Kroschk, A., Mosley, C., Naumann, F., 2018: Experience: Enhancing address matching with geocoding and similarity measure selection. *Data and Information Quality (JDIQ)* 10 (2), 1–16.

Bakshi, R., Knoblock, C. A., Thakkar, S., 2004. Exploiting online sources to accurately geocode addresses. *The 12th annual ACM international workshop on Geographic Information Systems (GIS '04)*. Association for Computing Machinery, New York, NY, USA, 194–203. doi.org/10.1145/1032222.1032251

Christen, P., Churches, T., & Willmore, A., 2004. A Probabilistic Geocoding System based on a National Address File. https://api.semanticscholar.org/CorpusID:9169323.

Lee K., Claridades ARC, Lee J., 2020: Improving a Street-Based Geocoding Algorithm Using Machine Learning Techniques. *Applied Sciences*. 2020; 10(16):5628. doi.org/10.3390/app10165628

Zhang, C., Guo, R., Ma, X., Kuai, X., and He, B., 2022: W-TextCNN: A TextCNN model with weighted word embeddings for Chinese address pattern classification. *Computers, Environment and Urban Systems*, vol. 95, 2022. doi:10.1016/j.compenvurbsys.2022.101819.

Lane, K., Scammell, M., Levy, J., Fuller, C., Parambi, Zamore, R., Mwamburi, W. M., Brugge, D., 2013: Positional error and time-activity patterns in near-highway proximity studies: An exposure misclassification analysis. *Environmental health* 12, 75. doi:10.1186/1476-069X-12-75.

Zandbergen, P. A., 2008: A comparison of address point, parcel and street geocoding techniques. *Computers, Environment and Urban Systems* 32 (3), 214–232, discrete Global Grids. doi.org/10.1016/j.compenvurbsys.2007.11.006.

Schootman, M., Sterling, D. A., Struthers, J., Yan, Y. Laboube, T., Emo, B., Higgs G., 2007: Positional accuracy and geographic bias of four methods of geocoding in epidemiologic research. *Annals of Epidemiology* 17 (6) 464–470. doi.org/10.1016/j.annepidem.2006.10.015.

Kravets, N., Hadden, W. C., 2007: The accuracy of address coding and the effects of coding errors. *Health & Place* 13 (1) 293–298, part Special Issue: Environmental Justice, Population Health, Critical Theory and GIS. doi.org/10.1016/j.healthplace.2005.08.006.

Chang, K.-T. 2024: Geographic Information System. *International Encyclopedia of Geography* (eds D. Richardson, N. Castree, M.F. Goodchild, A. Kobayashi, W. Liu and R.A. Marston). doi.org/10.1002/9781118786352.wbieg0152.pub2.

Mennecke, B. E. Crossland M. D., 1996: Geographic information systems: applications and research opportunities for information systems researchers, *The 29th Hawaii International Conference on System Sciences*, Wailea, HI, USA, 1996, pp. 537-546 vol.3, doi.org/10.1109/HICSS.1996.493249.

Minn, M., 2021. MMGGISS.
https://michaelminn.com/linux/mmqgis/ (1 March 2024)

ArcGIS, Geocode Addresses, https://pro.arcgis.com/en/pro-app/latest/tool-reference/geocoding/geocode-addresses.htm (1 March 2024)

Cara, C., Udroiu, C., Ciavotta, M., 2023. D1.2 enRichMyData architecture, Tech. rep., enRichMyData: Enabling Data Enrichment Pipelines for AI-driven Business Products and Services, grant No. 101070284, Horizon Europe.

Udroiu, C., Cara, C., Nicola, L., 2023. D3.1 enRichMyData integrated toolbox v1, Tech. rep., enRichMyData: Enabling Data Enrichment Pipelines for AI-driven Business Products and Services, grant No. 101070284, Horizon Europe.

Ripamonti, M., De Paoli, F., Palmonari, M., 2022. Semtui: a framework for the interactive semantic enrichment of tabular data, *ArXiv*. https://arxiv.org/abs/2203.09521

Cutrona, V., Ciavotta, M., De Paoli, F., Palmonari, M., 2019. Others, ASIA: A tool for assisted semantic interpretation and annotation of tabular data, *International Workshop on the SemanticWeb*, Vol. 2456, CEUR-WS, 209–212.

Avogadro, R., Cremaschi, M., D’Adda, F., De Paoli, F., Palmonari M., 2022. Others, LamAPI: a Comprehensive Tool for String-based Entity Retrieval with Type-base Filters, i17th ISWC workshop on ontology matching (OM).

Cruz, P., Vanneschi, L., Painho, M., Rita, P., 2022: Automatic Identification of Addresses: A Systematic Literature Review. *ISPRS Int. J. Geo-Inf.*, 11, 11. doi.org/10.3390/ijgi11010011

Wu, J., Funk, T. H., Lurmann, F. W., Winer, A. M., 2005. Improving spatial accuracy of roadway networks and geocoded addresses. *Transactions in GIS* 9 (4):585-601.

Roongpiboonsopit, D., Karimi, H. A., 2010: Comparative evaluation and analysis of online geocoding services. *International Journal of Geographical Information Science* 24 (7):1081-1100.

Geoglify, 2024. OpenStreetMap Nominatim API: Mastering Geocoding. https://www.geoglify.com/blog/openstreetmap-nominatim-api-mastery/ (1 March 2024)

NetToolKit,
https://www.nettoolkit.com/docs/geo/overview/geocoding (1 March 2024)