UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Exatas
Programa de Pós-Graduação em Ciência da Computação

Rodrigo Smarzaro

Spatial Data Integration from Heterogeneous Sources for Urban Computing

Belo Horizonte
2023

Rodrigo Smarzaro

Spatial Data Integration from Heterogeneous Sources for Urban Computing

Final Version

Dissertation presented to the Graduate Program in Computer Science of the Federal University of Minas Gerais in partial fulfillment of the requirements for the degree of Doctor in Computer Science.

Advisor: Clodoveu Augusto Davis Jr.

Belo Horizonte
2023

Silva, Rodrigo Smarzaro da.

S586s Spatial data integration from heterogeneous sources for urban computing [recurso eletrônico] / Rodrigo Smarzaro da Silva – 2023.
1 recurso online (139 f. il, color.) : pdf.

Orientador: Clodoveu Augusto Davis Júnior.
Tese (Doutorado) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Ciência da Computação.
Referências: f. 119 -139.

1. Computação – Teses. 2. Integração de dados espaciais – Teses. 3. Computação urbana. – Teses. I. Davis Júnior,Clodoveu Augusto. II. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Ciência da Computação. III. Título.

CDU 519.6*04(043)

Ficha catalográfica elaborada pela bibliotecária Belkiz Inez Rezende Costa
CRB 6/1510 Universidade Federal de Minas Gerais - ICEx

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

SPATIAL DATA INTEGRATION FROM HETEROGENEOUS SOURCES FOR URBAN COMPUTING

RODRIGO SMARZARO DA SILVA

Tese defendida e aprovada pela banca examinadora constituída pelos Senhores(a):

Prof. Clodoveu Augusto Davis Júnior - Orientador
Departamento de Ciência da Computação - UFMG

Profa. Ana Paula Couto da Silva
Departamento de Ciência da Computação - UFMG

Prof. José Alberto Quintanilha
Departamento de Engenharia de Transportes - USP

Profa. Mariana Abrantes Giannotti
Escola Politécnica - USP

Profa. Cristina Dutra de Aguiar
Instituto de Ciências Matemáticas e de Computação - USP

Prof. Anisio Mendes Lacerda
Departamento de Ciência da Computação - UFMG

Belo Horizonte, 26 de abril de 2023.

Documento assinado eletronicamente por **Clodoveu Augusto Davis Junior**, **Professor do Magistério Superior**, em 28/04/2023, às 11:19, conforme horário oficial de Brasília, com fundamento no art. 5° do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Ana Paula Couto da Silva**, **Professora do Magistério Superior**, em 08/05/2023, às 16:00, conforme horário oficial de Brasília, com fundamento no art. 5° do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Anisio Mendes Lacerda**, **Professor do Magistério Superior**, em 09/05/2023, às 10:26, conforme horário oficial de Brasília, com fundamento no art. 5° do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **José Alberto Quintanilha**, **Usuário Externo**, em 10/05/2023, às 14:48, conforme horário oficial de Brasília, com fundamento no art. 5° do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Cristina Dutra de Aguiar**, **Usuária Externa**, em 07/06/2023, às 10:27, conforme horário oficial de Brasília, com fundamento no art. 5° do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Mariana Abrantes Giannotti**, **Usuária Externa**, em 20/06/2023, às 11:38, conforme horário oficial de Brasília, com fundamento no art. 5° do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **2252542** e o código CRC **A894E8D6**.

To my wife, my daughters, my parents, and my brothers.

Acknowledgments

Developing research and producing a thesis's text may seem like an individual effort, but this is far from the truth. The work is the result of many people's collaboration, direct and indirect.

In no particular order, I would like to thank people and entities that have helped me in some way along the way.

I thank the Federal University of Viçosa. The institution welcomed me as a student and teacher, which allowed me to continue my studies at the doctoral level.

I thank the Federal University of Minas Gerais, specifically the Computer Science Department, for providing the environment and structure for developing this thesis.

I thank the friends in Belo Horizonte that I disturbed at various times and always made themselves available to offer me a place to stay. Juliano, Thaís, Gustavo, Giordano, Alessandra, thank you. I take this opportunity to remember the other friends from "La Squadra", Anderson, Humberto, Rafael, and Maycow (Mico). A team formed to participate in a sporting event became a group of good friends. There go about ten years of conversation and relaxation that helped me keep my mental sanity, especially in recent years.

I thank Prof. José Alberto Quintanilha. His words and our conversations were critical in turbulent times.

I thank my advisor, Clodoveu Augusto Davis Júnior. There are no words to thank enough for how much his attitude, understanding, and experience were essential for developing this work. I learned a lot and grew as a person and as a researcher. Thank you very much.

I thank my parents, Alcides and Maria, and my brothers, Ricardo and Renan. Although we are far apart geographically, we have always been close, and your support has been fundamental.

I thank my wife, Guanaeli, and my daughters, Júlia and Beatriz. Your existence is enough for gratitude. I know I was mentally distant in many moments, even though I was physically present, and you were always there to help me.

“*A computer which can calculate the Question to the Ultimate Answer is a computer of such infinite and subtle complexity that organic life itself shall form part of its operational matrix. And you yourselves shall take on new forms and go down into the computer to navigate its ten-million-year program! Yes! I shall design this computer for you. And I shall name it also unto you. And it shall be called . . . The Earth*”

(Douglas Adams)

Resumo

A urbanização global tem resultado em cidades cada vez mais populosas, aumentando a necessidade de prestação eficiente de serviços essenciais que afetam diretamente a qualidade de vida da população. Entre esses serviços, destaca-se o transporte público. Órgãos governamentais coletam grandes volumes de dados assim como usuários de redes sociais, por meio de seus smartphones, podem complementar essas fontes oficiais com uma gama de informações que vão desde dados objetivos até opiniões e sentimentos pessoais. Entretanto, a integração de fontes tão diversas e heterogêneas apresenta desafios significativos. Este trabalho tem como objetivo propor, desenvolver e validar métodos e técnicas para integrar múltiplas fontes de dados urbanos heterogêneos dentro da estrutura conceitual da Computação Urbana. Para tanto, os métodos desenvolvidos foram aplicados em um estudo de caso que consistiu na construção de uma rede multimodal de transporte para Belo Horizonte. Para validar os resultados, um conjunto de rotas foi determinado comparando a rede de transporte multimodal e o Google Maps, obtendo-se resultados próximos em termos de tempo e distância. Além disso, foi criado um estudo de caso para determinar um índice de qualidade de vida urbana a partir de dados integrados de diferentes fontes, o que demonstrou a possibilidade de utilização da rede de transporte multimodal. Os modelos de dados e os métodos desenvolvidos neste trabalho podem ser utilizados para obter informações relevantes sobre a cidade e subsidiar a análise e tomada de decisões em diversas disciplinas que lidam com problemas urbanos.

Palavras-chave: Integração de dados espaciais. Computação urbana.

Abstract

Global urbanization is creating increasingly populous cities, and their services must become more efficient. Public transport is one of those essential services that directly affect the quality of living among the population. Today, various government and transportation agencies generate large volumes of data. At the same time, users of social networks, using smartphones, can enrich official sources with a range of information, from objective data to personal opinions and sentiments. There is an essential challenge in integrating such diverse and heterogeneous data sources. This work aims to propose, develop, and validate methods and techniques for integrating multiple heterogeneous urban data sources within the conceptual framework of Urban Computing. The methods developed were used in a case study to build a multimodal transportation network for Belo Horizonte. To test the results, a set of routes were determined using the multimodal transport network created and Google Maps, obtaining results close to time and distance. A case study was created to determine the urban quality of life indexes from integrated data from different sources to demonstrate the possibility of using the multimodal transport network. The data model and methods developed in this work can be used to obtain relevant information about the city and to subsidize analysis and decision-making in the various disciplines that deal with urban problems.

Keywords: Spatial data integration. Urban computing

List of Figures

2.1 Spatial data matching taxonomy 29
2.2 Equivalence among the similarity measures classifications 32
2.3 Hausdorff vs. Fréchet distance 33
2.4 Strategies classification to handle inconsistent data 36
2.5 GIS-T Enterprise Logical Data Model 45
2.6 MDLRS conceptual view in UML notation 47
2.7 Crowdsourcing and crowdsensing taxonomy based on user collaboration 48
2.8 Relationship diagram for GTFS files 50

3.1 Diagram of research design 60
3.2 Conceptual schema for the MUTN in UML notation 61
3.3 An overview of the steps for building the multimodal network. 70
3.4 Cardinalities of matching pairs. 72

4.1 Case study overview 82
4.2 Examples of administrative boundaries for Belo Horizonte 83
4.3 Conceptual schema of data from BH's open data portal 83
4.4 Data available from BH SDI 84
4.5 Exclusive pedestrian segments from HVTL integrated with OSM dataset . . . 91
4.6 MUTN routing examples for WALK and TRANSIT transport modes 97
4.7 MUTN routing examples for DRIVE and DRIVE–WALK transport modes . . 98
4.8 The difference in routes generated by Google Maps and MUTN 100

5.1 Steps to calculate the IQVU index 103
5.2 The result of the data integration tasks for POIs data 109
5.3 Comparison of results for IOL Indicator *Hyper and supermarket* 114

List of Tables

2.1 MDLRS domains and their elements 46

3.1 Different values for key *highway* in OSM to represent streets and respective preferred modes of transport (cars, pedestrian, bicycle) 66

3.2 Combinations of key-value pairs used for public transportation stops in OSM. 66

4.1 Top ten values for the key *highway* from OSM for Belo Horizonte, by total length in meters 85

4.2 Number of points and lines from datasets before and after schema matching procedures. 86

4.3 Time (in seconds) and average number of candidates for matching found per segment varying the buffer size (in meters) 88

4.4 Number and length(meters) of segments matched between OSM and HVTL datasets. 89

4.5 Value mapping for attribute '*level*' in the segments of the MUTN data model. 90

4.6 Number and length (meters) of segments matched between DSA$_d$ and TC datasets. 92

4.7 Comparison between MUTN and Google Maps 99

5.1 Correspondence between OpenStreetMap key-value pairs and IQVU indicators 108

5.2 Quantity of places (points) of each step of data source cleaning process. 109

5.3 IQVU Indicators from each source to build the MUTN integrated version . . . 110

5.4 Accuracy (acc.), Precision (prec.) and Recall (rec.) of indicators calculated from Foursquare, Google Places, Yelp, and the integrated dataset using the MUTN data model. The Precision and Recall values are the mean of the five IQVU classes 112

Contents

1 Introduction **17**

1.1 Objectives . . . 19

1.2 Motivation . . . 20

2 Literature Review **23**

2.1 Spatial Data Integration . . . 24

2.1.1 Schema Matching . . . 27

2.1.2 Spatial Data Matching . . . 28

2.1.2.1 Similarity Measures . . . 31

2.1.2.2 Evaluation Metrics . . . 35

2.1.3 Spatial Data Fusion . . . 35

2.1.4 Discussion . . . 37

2.2 GIS for Transportation . . . 38

2.2.1 Multimodal Transport Networks . . . 39

2.2.2 Multimodal Transport Data Models . . . 42

2.2.2.1 Geographic Data File . . . 42

2.2.2.2 ESRI Transportation Data Model . . . 43

2.2.2.3 GIS-T Enterprise Data Model . . . 43

2.2.2.4 MDLRS . . . 44

2.3 Spatial Data Sources . . . 45

2.3.1 General Transit Feed Specification . . . 49

2.3.2 OpenStreetMap . . . 52

2.3.3 Location Based Social Networks . . . 54

2.3.4 Discussion . . . 56

2.4 Quality Metrics for Spatial Data . . . 57

3 Methodology **59**

3.1 Multimodal Urban Transportation Network Data Model . . . 60

3.2 Data Selection . . . 64

3.2.1 OpenStreetMap . . . 65

3.3 Data Gathering . . . 67

3.3.1 Location-Based Social Networks . . . 67

3.4 Data Cleaning . . . 68

3.5 Building the Multimodal Urban Transportation Network 68
3.5.1 Initial Definitions 69
3.5.2 Schema Matching for the Reference and Complementary Datasets . 69
3.5.3 Data Matching for Network Data 71
3.5.3.1 Building the Set of Candidates for Matching 72
3.5.3.2 Finding Matching Pairs 74
3.5.3.3 Selection of Exclusive Features from the Complementary Dataset 75
3.5.4 Data Fusion for Network Data 75
3.5.5 Creation of the Collective Transportation Network 76
3.5.6 GeoFeatures Matching and Duplicates Removal 77
3.5.7 Selection of ConnectionNodes 78
3.5.8 Creation of Transfers between Transport Modes 78
3.5.9 Linking GeoFeatures to the Multimodal Urban Transportation Network 78
3.6 Discussion 79

4 Creation of a MUTN for BH 81
4.1 Official Datasets 82
4.2 Alternative Datasets 84
4.3 Schema-Matching Procedures 86
4.4 Data Matching and Fusion between OSM and HVTL 87
4.5 Data Matching and Fusion between DSA and TC 91
4.6 Quantitative Evaluation of Data Matching Results 92
4.7 Creation of the Collective Transportation Network from GTFS 93
4.8 Integration of *ConnectionNodes* into the MUTN 94
4.9 Cost Assignment to Segments 94
4.10 Multimodal Routes Using the MUTN 95
4.11 Results and Discussion 96

5 MUTN and the Quality of Urban Life 101
5.1 Quality of Urban Life Index 103
5.2 Methods 105
5.2.1 Official Data Sources 105
5.2.2 Crowdsourced Data Sources 105
5.2.3 Data Collection 106
5.2.4 Data Cleaning and Characterization 107
5.2.5 Data Integration 107
5.2.6 Calculation of the Local Availability Index Indicators 107
5.3 Results and Discussion 111

5.3.1 Limitations 113
5.3.2 Final Remarks 115

6 Conclusion **117**

References **121**

Chapter 1

Introduction

The world's population is increasingly urban. In 1950 the urban population was approximately 740 million people. By 2014 it was 3.9 billion and is expected to surpass six billion by 2045 [233]. The urban population is now more prominent than the rural. This phenomenon has no precedent in human history and brings tremendous challenges. Making cities work more efficiently and sustainably is a priority, and the use of technology becomes more critical [176, 71, 135, 218, 74, 223]. One of the initiatives to provide solutions in this task is called *urban computing*.

Urban computing applications must collect and process a large amount of heterogeneous data and are frequently required to help solve near real-time problems, such as traffic conditions. Methods to organize and integrate these data are essential to discover new knowledge which would remain unknown considering the data sources individually [267].

Urban computing is closely related to other concepts such as *smart cities*, *wired cities*, *cyber cities*, *digital cities* and *sentient cities* [120]. While these terms highlight some aspects of a "smart" city, the broader meaning is twofold. One view implies having cities with high use of pervasive and ubiquitous computing. The second involves using technology to empower people to innovate so that government and society can improve their decision-making processes, leading to higher city efficiency. Both meanings share the need to produce, integrate, and maintain large volumes of heterogeneous data from multiple sources with varied formats and spatiotemporal granularity.

Two concepts widely used in urban computing are mobility and accessibility, which have greatly impact the quality of life for individuals and communities. Mobility refers to the ability to move and travel, enabling individuals to access education, employment, healthcare, social activities, and other essential services. It encompasses the physical movement of people, vehicles, and goods, and is influenced by transportation infrastructure, available modes of transport, and personal circumstances [131]. On the other hand, accessibility is about ensuring equal and inclusive access to goods, services, and opportunities. It focuses on eliminating barriers, both physical and socio-economic, and ensuring that everyone can reach and utilize resources without limitations [76, 131, 211]. By enhancing mobility and accessibility, we empower individuals to fully participate in society, promote social inclusion, foster economic growth, and create sustainable and livable communities. Both concepts are vital for improving the quality of life by enabling individuals to connect, engage, and thrive in their surroundings [193, 144]. Despite its importance, urban mobility-related data is not always easily obtained. For example, road network maps are traditionally produced by governments or commercial agencies aiming at high accuracy and using expensive, labor-intensive means. Still, the final product has limited use due to licensing terms [113]. Also, road networks are encoded for car navigation systems and lack data on other transportation modes, such as public transportation [136]. The expensive production process causes data to be often outdated. It is necessary to have detailed information in an integrated multimodal transport network to create helpful mobility information for all dwellers, not only those using private vehicles.

When dealing with multiple data sources, there are usually three approaches to making them interoperable [260]. First, the datasets are stored separately, but their contents are linked. Second, one dataset is entirely integrated into the other. Third, the two datasets are merged into a new one. This work focuses on the third approach using conceptual data modeling to achieve spatiotemporal data integration. Using a proper and unified schema may prevent errors and quality problems [180]. Although there are some alternatives for the modeling of transit networks, they were not designed to consider the integration of data from unstructured, heterogeneous data sources, and their main focus is to represent the static components of the transport network, such as the road network, bus stops, and footways. Besides these static components, urban mobility is also affected by many dynamic factors that must be considered. Weather, vehicular flow, accidents, authoritative alerts, street maintenance, and data from Location-Based Social Networks (LBSN) like Foursquare, Twitter, and Flickr can enrich the dataset and provide valuable information to users. For example, pedestrian users can decide on walking routes considering the alerts of robberies posted on Twitter. A commuter can take the metro based on real-time information about traffic delays caused by accidents. A jogger may prefer a more pleasant (beautiful, quiet, shadowy) route based on online photos and comments. People's movement is highly associated with their online connections and activities [87], and transit data models that do not consider this may not be taken to their full potential. Moreover, online data can be used as an alternative or complement official data sources at a fraction of the cost [111].

One possibility is to use Volunteered Geographic Information (VGI). VGI [81] is a type of User Generated Content (UGC) in which the geographic factor plays a central role in the information. In this context, *produsers*, users as data producers [31, 45, 93], using the combination of Web 2.0, portable devices equipped with Global Positioning System (GPS), and Internet access to produce and consume geospatial data. OpenStreetMap (OSM) may be the best example of this concept [29, 63, 160]. OSM has over three million registered users, and it's open data policy and flexible data model encouraged many companies, like Foursquare, Flickr, MapQuest, Wikipedia, and Craiglist, to start using it as a map provider instead of resorting to proprietary ones. Much research has been made to assess the quality of VGI data and to understand its community and factors behind volunteer motivations [149]. While most of the research on VGI is focused on developed countries, more attention is needed in developing countries, like Brazil, to verify the status of VGI data and its suitability to replace and/or complement official sources [263, 205].

1.1 Objectives

The research hypothesis of this work is that multiple heterogeneous urban data sources, such as UGC and official data producers, can be integrated to combine the best characteristics of each source. While official sources tend to be static and precise but outdated, crowdsourced data can be more timely, lacking coverage and accuracy. If stable data can be adequately combined with up-to-the-minute information provided by human sensors, a completely novel set of analyses and results can be obtained, directly impacting decision-making in all aspects of urban life.

Therefore, this work aims to propose, develop, and validate methods and techniques for integrating multiple heterogeneous sources of urban data within the conceptual framework of Urban Computing. Such integrated data can then obtain relevant information about the city and subsidize analysis and decision-making in various disciplines dealing with urban problems. Our approach concentrates on problems in multimodal urban transportation since this is a subject of broad interest, for which data sources are varied, voluminous, and cover both static and dynamic aspects of reality.

Specific objectives include:

- to build a multimodal transport network data model;
- to select data sources and verify their suitability to transport information;
- to create methods for the integration of static data on the urban transport networks, including transportation modes, urban maps, transportation and street network elements, spatial data infrastructures, and others;
- to create methods for integrating dynamic and typically unstructured data from sources such as location-based social networks and crowdsourcing to the multimodal transport network;

- to validate the integration of multiple sources through novel analyses and applications.

1.2 Motivation

Even though much spatial data is produced, integrating data from different sources is often difficult. Sources usually focus on a thematic subset and are developed to fulfill the needs of particular applications. Therefore, several datasets cover the same geographic area with potentially complementary data. Being able to integrate these multiple spatial data sources and enrich or complement them with UGC serves several purposes:

- Spatial data integrated from multiple sources can be used to build better datasets. Better in this context means a more complete, accurate, and useful dataset. For example, one dataset may be highly accurate but has insufficient associated data, while another may have complete attributes but faulty geometry. So the integration of these datasets may be much more useful to users, and the comparison among different datasets may improve the overall data quality [83, 192, 32, 1, 58, 29, 225, 205]

- Integration methods can also be used to keep the dataset up-to-date. Multiple data sources can be analyzed to identify when the original data have to be updated or new data have to be incorporated [132, 251].

- An integrated dataset on static and dynamic transit behavior impacts how people interact with the city [224, 135, 218, 107].

- The integration of UGC brings several new opportunities for use and analysis. UGC can enrich the spatial dataset with user behavior, preferences and other kinds of data [84, 15, 117].

This work is organized as follows. Chapter 2 presents the literature review about spatial data integration (Section 2.1) and its applications on transportation problems within the scope of urban computing (Section 2.2). Chapter 3 presents the research design and proposed methods to fulfill the objectives of this work. A case study scenario to build a multimodal urban transportation network for the city of Belo Horizonte[1] is presented in Chapter 4). Chapter 5 uses the results from Chapter 4 to estimate a quality

[1] Belo Horizonte is the capital of the Brazilian state of Minas Gerais.

of urban life indicator for the city of Belo Horizonte. Chapter 6 presents conclusions and future work directions.

Chapter 2

Literature Review

Since the 19th century, the world has experienced high population migration rates from rural to urban areas. At the beginning of the 20th century, only about 10% of the world's population lived in urban areas. By 2007, however, the urban population surpassed those living in rural areas. In 2014, 53% of the world's population was living in cities, and the projections for 2050 indicate that the urban population will reach 64% in developing countries and 86% in developed ones.

Urban population growth demands a lot of effort from local governments to improve services and infrastructure. The capacity to detect problems or anticipate and avoid them is crucial to making the city administration more dynamic. To achieve this, city managers must collect, integrate, and process a large amount of heterogeneous data. That is the essence of the *Urban Computing* concept. Zheng [267] defines it as the "process of acquisition, integration, and analysis of big and heterogeneous data" that can be generated from a large variety of sources like sensors, vehicles, and smartphones, to help improve common city problems such as energy consumption, air pollution, traffic flow, and others.

Urban computing is an interdisciplinary field with computer science as its "engine". To act in the urban context, it needs other disciplines such as transportation, sociology, economy, biology, ecology, architecture, and civil engineering [229] act in the urban context. Many exciting and innovative works have been proposed and presented in recent years. For example, obtaining detailed air pollution data across the city is difficult and expensive. The government depends on monitoring stations to collect data, and regions without one have no reliable information. Zheng [268] infer air quality information for regions without monitoring stations using historical and real-time air quality data from available stations, combined with other data sources such as meteorology, traffic flow, human mobility, road networks, and points of interest (POIs). Noise pollution is another problem in major urban areas that affects people's health and is correlated with traffic conditions, varying throughout the day [183]. Similar to air quality data, it is also costly and time-consuming to gather reliable data. Some work uses smartphones as noise sensors to measure noise pollution across the city [138, 239, 146] or methods to infer such data from different sources, such as online social media, formally complain reports, road networks, POI locations [243].

Other works focus on discovering mobility patterns and assessing city dynamics using time series of mobile phone locations [19, 35, 34, 256, 257] and handoff[1] statistics [49]. GPS tracklogs from taxicabs are used to analyze patterns and efficiency in refueling operations at gas stations [259] and to infer urban traffic flow [126, 177]. Using the same kind of data, [167] identify traffic anomalies when driver behavior differs significantly from the usual, then search social media (WeiBo, a Twitter-like online social network in China) for relevant terms to infer what caused the anomaly. Ibeas *et al* [100] use traffic flow, transit timetables, and socio-demographic data to obtain the optimal spacing of bus stop locations to minimize the overall social costs. Bike sharing has become a vital transportation alternative in many cities worldwide. He [96] use crowdsourced data on bike station locations, usage, and cost to suggest a better placement configuration. Duan [60] seek to increase user satisfaction at bicycle stations, avoiding the unavailability of bicycles by recruiting workers to move them from stations with higher availability to smaller ones.

Some inferences can be made about urban computing applications from the works mentioned. First, dealing with data from unstructured and heterogeneous sources is often necessary. Second, traffic and transit issues are critical in urban centers and directly influence everyday life and health. Third, data from online sources, such as Location-Based Social Networks and VGI, must be considered.

This work relates to public and private urban transportation issues. Specifically, we want to integrate spatial data from multiple and heterogeneous sources. The methods and techniques proposed will be used to build and maintain a multimodal transport network enriched with data from various sources, such as official transit authorities and crowdsourced content.

The remainder of this chapter is organized as follows. Section 2.1 presents spatial data integration concepts and methods. Section 2.2 presents the use of Geographical Information Systems (GIS) technology in transportation. Section 2.2.1 presents concepts related to multimodal transport networks. Section 2.2.2 presents the available multimodal transport data models, and Section 2.3 presents examples of transit-related data sources.

2.1 Spatial Data Integration

With the advancement and dissemination of spatial technology, the volume and diversity of spatial data being produced increase daily. Although it is positive to have more

[1] Handoff or handover is the process of transferring an ongoing connection (voice or data) from one channel to another.

data, some concerns are also created. Many agencies, applications, and users produce and maintain spatial data. Each one has specific needs and therefore uses its policy regarding the data, which causes a high level of inconsistency and heterogeneity among available spatial data sources [155].

Spatial data plays a vital role in the decision-making process. Early works estimate that about 80% of all information used in the decision-making process have spatial properties [189, 122] and its correct use implies better decisions [232]. The need for quality spatial data and the high production cost motivate the sharing and integration of available sources [173].

Due to its increasing necessity, spatial data integration has received much attention from researchers. There are different definitions and approaches. Uitermark [230, 231] define spatial data integration as establishing relationships between corresponding elements in different spatial datasets. Usery [234] consider different datasets to be integrated when there are geometrical and topological matches so that the spatial relationships between different versions of the dataset objects and the real world are the same, and their attributes correspond. Samadzadegan [196] uses a broader scope of data integration by including activities to collect, process, and combine data from various datasets. Kolahdouzan [123] consider spatial data integration when various datasets are integrated to create a single composite dataset from the integrated elements. Longley[139] describe it as combining geographic information to retain accurate data, minimize redundancy, and reconcile data conflicts. The Open Geospatial Consortium (OGC) defines it as "the process of unifying two or more separate datasets, which share certain characteristics into one integrated all-encompassing result" [163, p. 3].

For [244], spatial data integration is the process of combining the spatial data of different sources, providing users with a unified view of these data while maintaining the integrity and reliability of the data. The definition proposed by [1] also includes this unified view of the integration results to provide processing, modeling, and visualization but, like [196], includes the process of collecting data within the scope of spatial data integration.

From these definitions, we can summarize by establishing that spatial data integration uses spatial data from several sources to create a new product better than the originally distinct datasets, given a certain purpose. There are various approaches to integrating spatial data sources, such as Federated Databases, mediation, and ontology-driven integration [156].

Federated Databases (FDB) [208] are virtual databases used to integrate data from different data sources. FDBs are virtual because they do not store the actual data but define a common schema used among the data sources. Data are materialized upon user requests to the FDB. Thus, FDB provides users with centralized and transparent data access, so they do not have to worry about where the actual data is located. Each separate data source can define a subset of its shared data with the FDB. [51] create a process to integrate two different database schemas representing road network data from the same region. The two schemas are analyzed to find inter-schema correspondences and then merged into a federated schema.

Mediation-based data integration approaches are based on providing a uniform interface to access data using a common global data model [21]. Two main components are involved: a *mediator* and *wrappers*. The mediator is responsible for making semantic translations of user requests to the common global data model and then using it to query different data sources. Each data source must have a wrapper to provide an interface between the mediator and the specific data source's query language [46].

Ontologies are explicit and formalized specifications of conceptualizations [85]. An *Ontology-driven* data integration approach relies upon providing a common understanding of the semantics of data objects. To accomplish this, spatial data integration is shifted to integrating the ontologies corresponding to the different data sources [30]. Three main tasks are used to create a unifying ontology. Given two different ontologies, *ontology merging* creates a new one, *ontology mapping* relates the similar elements found on both ontologies using an equivalence relation, and *ontology integration* obtains the missing parts of one ontology from another. Ontology merging and mapping do not modify the original ontologies, but one of the two original ontologies is modified in ontology integration.

All these different data integration approaches can be organized into three main tasks [42]. The first task is *Schema Matching*, identifying and corresponding the semantics of datasets objects. Once the semantics is resolved, the second task, *Data Matching*, identifies corresponding dataset objects. The third and last task, called *Data Fusion*, involves the resolution of schematic differences among the matched objects to produce a single and consistent representation of them. Some works do not consider schema matching and data matching as independent tasks and suggest viewing them as components of a larger task within the data integration process [164, 26, 264, 250].

The purpose of spatial data integration techniques in this work is related to urban transport data. Up-to-date transportation data is a need for the population, as well as for businesses and institutions. The road infrastructure and different transport mode networks, land use information, the behavior of people, weather, and other data that impacts everyday urban life must be considered in decision-making. Such elements must be presented in an integrated fashion to be helpful to a broader spectrum of users.

The remainder of this Section presents works on spatial data integration, focusing on urban transportation. We use the three-tasks view of the data integration process from [42] to present their respective approaches and techniques. Section 2.1.1 presents schema-matching concepts. Spatial data matching is discussed in Section 2.1.2. Section 2.1.3 describes spatial data fusion techniques.

2.1.1 Schema Matching

Consider the task of integrating two completely unknown spatial datasets. Without knowing the database schemas and what the features represent, it makes little sense to match them based only on their geometries. For example, one could match two line features from the two datasets, but one may represent a road while the other is a river. Lack of information or proper data formats for disseminating spatial data makes schema matching a challenging task [61].

Schema matching is finding semantic correspondences between elements from different schemas [53]. Since spatial databases can easily become large, and as the number of object classes grows, schema matching must be executed as the first step in the spatial data integration process to reduce the computational work by filtering the original data objects into valid classes [143]. Schema-matching techniques rely on schema information such as data types, element names, and structural properties. Characteristics of object instances can also be used to help with schema matching. Some other data-matching approaches use external information, such as ontologies and dictionaries.

Many works have been developed on schema matching involving transport-related data. [240, 241] created a mapping across the different representations a feature can have in multiple spatial datasets of street data. In each relation, information on how consistent the mapping is concerning geometry, topology, or thematic properties is described. This method is called Multi-Representational Relations in his work. [16] uses an iterative algorithm to match two road network datasets. One contains traffic information provided by the government of Korea. The other is a road network produced by navigation companies. The integration process aimed to enable proper visualization of both road and traffic data. [4] worked on Ordnance Survey and OpenStreetMap datasets. They combined structural, semantic, geometric, and attribute in a weighted similarity measure to resolve possible matches. However, they did not obtain good results and suggested using more directed ontologies to improve spatial data integration.

Guan [86] used ontologies to match Geographic Markup Language (GML) schemas and tested the proposal over highway and road data (among other kinds of data such as states, cities, rivers, and lakes) from Canada and USA. Vaccari [235] develop a structure-preserving semantic matching used on interactions of web services exchanging spatial data. They also used ontologies to solve semantic heterogeneity between the different implementations of web services used to provide and request spatial data. Du [59] works used ontologies to integrate authoritative (Ordnance Survey) and crowdsourced (OpenStreetMap). They used to integrate road vector data [59] and polygon data [57].

Prudhomme [175] applied a semantic interpretation process to infer an ontology from a dataset schema without prior knowledge. The produced ontology is then used for schema matching through ontology matching techniques. Their approach to semantic interpretation is based on geocoding and natural language processing.

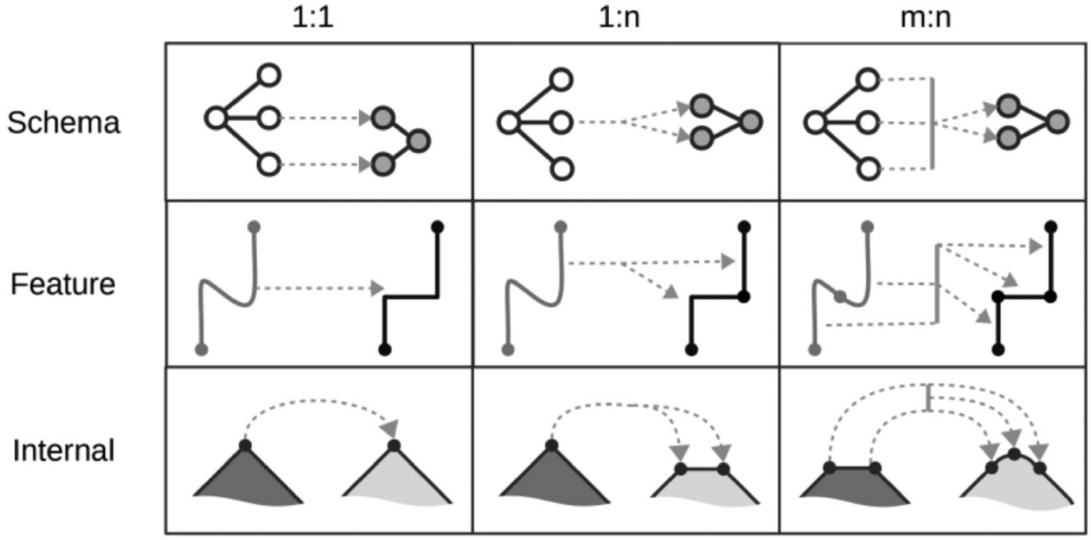
2.1.2 Spatial Data Matching

Spatial data matching has been a target of much research over the last years. It can be defined as the correct correspondence between different geospatial dataset objects [58] and is a requirement for integration, management, and quality evaluation of spatial datasets [250]. It can also be called *linking* [59], *alignment* [210] or *reconciliation* [194]. Researchers have different categories for spatial data matching. Devogele [52] and Yuan [255] propose, as data matching categories, semantic/attribute, topologic and geometric. [241] and [55] propose a classification based on primitive geometric types: point, line, and area-based data matching. Quddus [178] classify the matching methods based on geometry, topology, probability, and advanced techniques, but focus only on algorithms to match trajectory data (e.g., GPS data) with the corresponding road networks. Xavier [250] propose a novel classification using two criteria: level and case of correspondence. The *level* refers to where the matching will occur in the data model hierarchy. Three values are possible for level: *schema*, *feature*, and *internal*. The *case of correspondence* regards the cardinality of the match and can be one-to-one (1:1), one-to-many (1:N), and many-to-many (M:N). Figure 2.1 shows this proposed taxonomy for spatial data matching. It is important to establish that these categories are not closed, and it is common for a method to use techniques from more than one category. Schema-level Matching was already described in Section 2.1.1.

Feature Level Matching

Feature-level matching methods consider that schema matching is already complete. The focus now is to find correspondences among features using one or more similarity measures. Volz [241] uses an iterative matching method to match street network data from the ATKIS database to GDF files. The method first reduces the geometric deviation by using rubber sheet transformations and then splits the features topologically by inserting new nodes to make it easy to find 1:1 node matches. The next step selects the seed nodes with a high likelihood of correspondence to start a process in which multiple

Figure 2.1: Spatial data matching taxonomy



Source: Adapted from [250].

iterations try to recognize matches with different cardinalities.

Buffer growing is a popular technique many works use. Zhang [260] uses an unsymmetrical buffer growing method to match road networks. Chen [39] extend the method to establish possible matchings between nodes and edges, while the original method only matches edges. Kim [119] combine buffer growing, Voronoi diagrams, triangulation, and other geometric measures to create a context similarity-based matching process. Ying [254] also use it to find candidate objects to match and then apply probabilistic match based on multiple measures to select the correspondent object with maximum likelihood.

Another common method to feature level matches is the *Nearest Neighbor* (NN). This method works by searching, for each object in a dataset, the closest object from another dataset. This approach has known drawbacks since it is possible for one object in a database to be the nearest neighbor to more than one object in another dataset. Despite that, many researchers use it in data-matching tasks. Beeri [20] define the Mutually-Nearest Method, which establishes a symmetric property to the match operation. Given two objects, $a \in A$ and $b \in B$, they are mutually nearest if a is the nearest neighbor of b and b is the nearest neighbor of a. They also develop two other methods: Probabilistic and Normalized-Weights. Their results show that when the overlap between datasets is small, mutually nearest and normalized weights get better results. The normalized weights method performs better for medium-size overlaps; for significant overlaps, the probabilistic approach outperforms the other two. Based on [20], Safra [192, 191] use the mutually nearest method to match road networks by isolating the endpoints of polylines. Tong [228], and Ying [254] extend the probabilistic method to include multiple similarity measures instead of only one, as used by [20].

Anand [7] use geometric similarity measures (distance and angle) to match Ordnance Survey data with OSM. This method can be used only in linear features. Ludwig [141] compare OSM road data with a Navteq dataset. They segment the OSM data using buffer operations to level the number of features on both datasets and facilitate finding 1:1 correspondences. Then the length, category, and name of features were used as similarity measures to find the best candidates to match.

Koukoletsos [127] assess the completeness of VGI data (OpenStreetMap) using a multi-stage approach, combining geometric and attribute constraints. The method splits the dataset into smaller areas and uses distance, direction, length of roads, and also road names and road types in the comparison. Characteristics of VGI data, such as topological inconsistencies and abbreviations used on feature names, may impact the method's performance.

Luan [140] apply skeleton extraction to road networks and resolve the road matching problem using the maximum common subgraph algorithm. One advantage of this method is that it can be applied to datasets of different coordinate systems.

Yang [252] develop a heuristic probabilistic relaxation method to match road networks from OSM and authoritative datasets. The method starts with a probabilistic matrix built from similarity measures on feature shapes and then incorporates compatibility coefficients of neighboring candidates until the probabilistic matrix gets globally consistent. The method finds 1:1 matching pairs and then expands to find M:N matchings.

Fan [65] use a polygon-based approach to match road networks. The initial step uses the urban block polygons so that road lines are assigned to the edges of the urban blocks. The matching process starts by finding matches on the block polygons, then on the edges, and finally on the road lines.

Abdolmajidi [2] compare segment-based and node-based approaches for road (network) matching. They chose the node-based approach due to the reduced computational cost and improved it to handle topological relationships and other network components. The resulting method is used to assess the completeness of OSM data concerning the Swedish National Road Database.

Internal Level Matching

Internal level matching regards the comparison of parts of geometry features. It plays an essential role in the quality assessment of shape features [66, 112, 188].

Huh [99] propose a method to detect conjugate-point pairs to align two polygons matching their contour. Their method starts by identifying the corresponding polygon pairs, then creates an approximation of the shapes using virtual corner vertices, and finally detects conjugate-point pairs with contour matching using the algorithm Vertices-Attribute-String-Matching (VASM). This algorithm matches the contour of polygons calculating the minimum-cost edit sequence necessary to convert one polygon vertex string into the other.

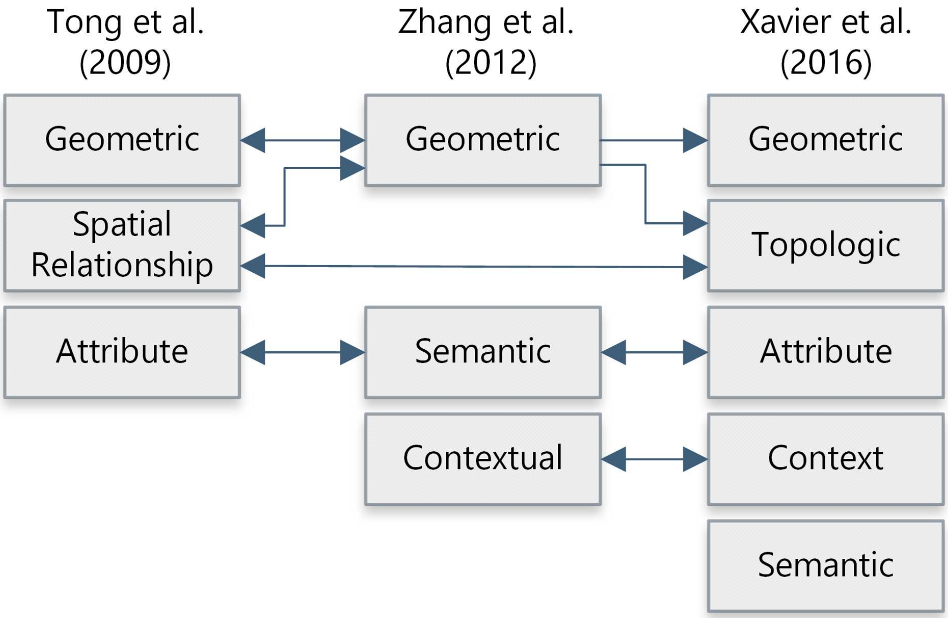
Fan [66] use internal matching as part of the process to assess the quality of OSM 2D buildings data. When a correspondence pair is found, its geometry is simplified using the Douglas–Peucker algorithm to avoid problems with different LoD between the datasets. The minimum bound rectangle (MBR) is determined for both polygons. Any edge of the polygon that touches the MBR is marked. The MBR from OSM is shifted to the center of the MBR from the authoritative dataset (ATKIS). Then, if the marked edges are at the exact location or very close, their ending points are considered identical points of the two polygons.

Ruiz-Lendinez [188] use the turning function [11] to find homologous points on polygons. These homologous points are then used to assess the positional accuracy of the features.

2.1.2.1 Similarity Measures

The core of any data matching method is the similarity measure used. Every method needs criteria to compare entities along the matching process. Tong [228] classify the measures into *geometric*, based on *spatial relationships* and based on *attributes*. Geometric measures relate to the shape of the object or its location. Spatial-relationship-based measures use distance, topology, or orientation of features. Attribute-based measures compare attributes like name, type, or other information. Zhang [262] use similar classifications for such measures: geometric, semantic, and contextual. The "attribute-based" similarity measures from Tong and "semantic" from Zhang are equivalent. Geometric and Spatial Relationship from Tong is combined into Zhang's "geometric" category. The contextual measure has no equivalence in Tong's classification. It is based on analyzing the neighborhood of an object in the hopes that if two features are similar, their vicinities are similar too. Based on both works, [250] uses five categories of similarity measures: *geometric*, *topological*, *attribute-based*, *context-based* and *semantic*. The first four categories can be related to the ones from Tong and Zhang. The last, "semantic," has no equivalent. It relates to the distance in the meaning of the feature and uses some common ontology or taxonomy to be implemented. Figure 2.2 shows the equivalence among the three classifications presented. To prevent confusion, this work adopts the classification from [250], as it represents a superset of Tong and Zhang classifications and presents additional concepts. Each class is discussed in more detail, briefly describing the main similarity measures.

Figure 2.2: Equivalence among the similarity measures classifications proposed by [228] (left), [262] (middle) and [250] (right)



Source: Made by the author.

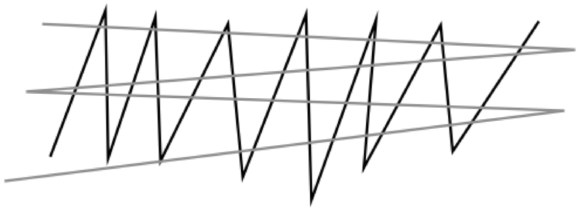
Geometric Measures relate to the position, length, perimeter, area, shape, or angle of features and their components [216]. Position measures are based on distance, while shape can be related to the length or area of the feature.

The most common distance measure used is the *Euclidean Distance*, which represents the straight-line distance between two points. Euclidean distance is mostly used in point matching tasks [255, 32, 157, 151] but is widely used as a component in other more complex similarity measures [261, 227, 148].

Hausdorff Distance [186] measures the distance between two sets of features. The intuition is that, given two sets of features, the Hausdorff Distance is the largest distance someone can find between one arbitrary point from the first dataset and another from the second dataset. It is a popular measure to match lines and polygons [255, 241, 99, 134, 72, 2], but it has some problems determining shapes (see Figure 2.3) [5].

Fréchet Distance calculates the distance between two curves, like Hausdorff, but considers the ordering of the points along the curves. The intuition can be explained by imagining a person walking along a curve and a dog at another curve and both connected by a leash and also with varying velocities. The Fréchet distance is the minimum length of the leash to complete the walk. It can be a better similarity measure than Hausdorff distance when the feature geometry has many sinuous curves. Mascret [147] used it to develop a matching process to coastlines. Chen [38] used it to match GPS data in the road network. Fréchet distance can overcome some problems Hausdorff distance can have when the features have many curves or are not convex, but it has a higher computational cost to calculate [5].

Figure 2.3: Example of two features with small Hausdorff distance and large Fréchet distance.



Source: [5].

One common shape measure uses the *area overlap* between two areal features. It is used primarily with polygon features but can be used with the line features using a buffer operation [82, 204, 29] or the minimum bounding box (MBB) [187]. The calculation is the simple ratio of the overlapped area between two polygons over their average area values.

Some shape-based methods inspect the angle direction of lines that connect to a point. The *spider function* [190] divides the angle area into 2^{n} sectors, and then the presence of lines in each sector are coded to form a spider code that represents all the sectors that have lines. Although classified as a geometric similarity measure by [250], the spider function is classified as a topological similarity measure by [228] (under the spatial relationship category) and [260].

Another popular approach to shape similarity is the use of *turning functions*. Arkin [11] propose a turning function $\theta_A(s)$ which "measures the angle of the counter-clockwise tangent as a function of the arc length s, measured from some reference point 0 on A's boundary" [11, p. 209]. Each change in the angle along the geometry is recorded, increasing the value of the function in left-hand turns and decreasing otherwise. The result is a shape signature that can be compared to other features. Ruiz-Lendinez [188] use turning functions to develop a method to assess the positional accuracy of spatial data. Fan [66] use to determine shape similarity by comparing OpenStreetMap data with ATKIS (German Authority Topographic-Cartographic Information System) data for Munich, Germany.

Topological measures analyze the spatial relationships among features and is mainly used to match network (node-arc) structures. Many metrics from the networks study field are used, such as *node degree* [191, 118], *centrality* [88, 29], *betweenness* [47, 88, 78] and *closeness* [88, 78].

Attribute-based measures are used to compare features based on their nonspatial data. This category of similarity measures relies on traditional data types operations. For example, using arithmetic or logic operations, the similarity between two integers, floats, or boolean values is straightforward. String types cannot be compared so directly. It is usual in spatial databases and critical in VGI data to find many spellings to refer to the same place or location. Two or more strings can be different but may relate to the same real-world entity. For example, the city of Belo Horizonte can have its name attribute as "Belo Horizonte", "BH", "B. Horizonte". Some text measures can be used to overcome misspellings and minor differences in strings. *Levenshtein distance* computes the difference between two strings by counting the minimum number of edit operations necessary to transform one string to the other. Olteanu [165] and Samal [197] used it to compare feature names. *Hamming distance* computes the difference between two strings of the same size and was used by [134] to compare feature names. To use strings of different lengths, they divided them using their average size. McKenzie [151] used an interesting approach applying phonetic algorithms to string comparison. They tested *Soundex* [92] and *Double Metaphone* [172] to measure the feature names' similarity. Double Metaphone had better results in their work. Phonetic algorithms try to compare strings by their pronunciation, so they have the drawback of being language-dependent.

Context-based measures use the geographic context of the features to help determine their similarity. Geographic context means analyzing how a feature relates to other reference features. For example, [197] used landmarks to build a proximity graph to compare the similarity between features. This proximity graph represents a directed weighted graph, and the total vector offset of the candidate feature matchings on both datasets determines the similarity measure. Using a similar approach, [119] and [261] used Voronoi diagrams and Delaunay triangulation, respectively. The difference from using the proximity graph is that the area intersection ratio now determines the similarity measure. One drawback of these approaches is that selecting the reference landmarks *a priori* is necessary. With the landmarks list done, context measures can help find correct matches when the datasets have little information.

Semantic measures try to determine the distance between concepts of the features, where the concepts can be classes, methods, or attributes [250]. Applying such methods is difficult because it usually requires some formal knowledge representation, like an ontology or taxonomic tree. Hastings [95] use the Least Common Superconcept (LCS) in a taxonomic tree to evaluate the similarity between gazetteer terms. The LCS uses the number of steps (possibly weighted) on the tree necessary to navigate from one concept to another.

2.1.2.2 Evaluation Metrics

The most common evaluation metrics in the literature to assess the spatial data matching process are precision, recall, and f-measure. They can be calculated from simple concepts. Consider reference and test datasets on a spatial data matching task. Suppose a feature from the test dataset correctly matches the correspondent feature from the reference dataset. In that case, it is a true positive (TP), but if it is incorrectly matched, it is a true negative (TN). If a feature from the test dataset does not find an existing match on reference data, it is called a false negative (FN) case. In contrast, if the feature doesn't exist on reference data, it is a true negative (TN) situation. Precision, recall, and F-measure can be calculated using Equations 2.1, 2.2, and 2.3, respectively.

$$precision = \frac{TP}{TP + FP} \tag{2.1}$$

$$recall = \frac{TP}{TP + FN} \tag{2.2}$$

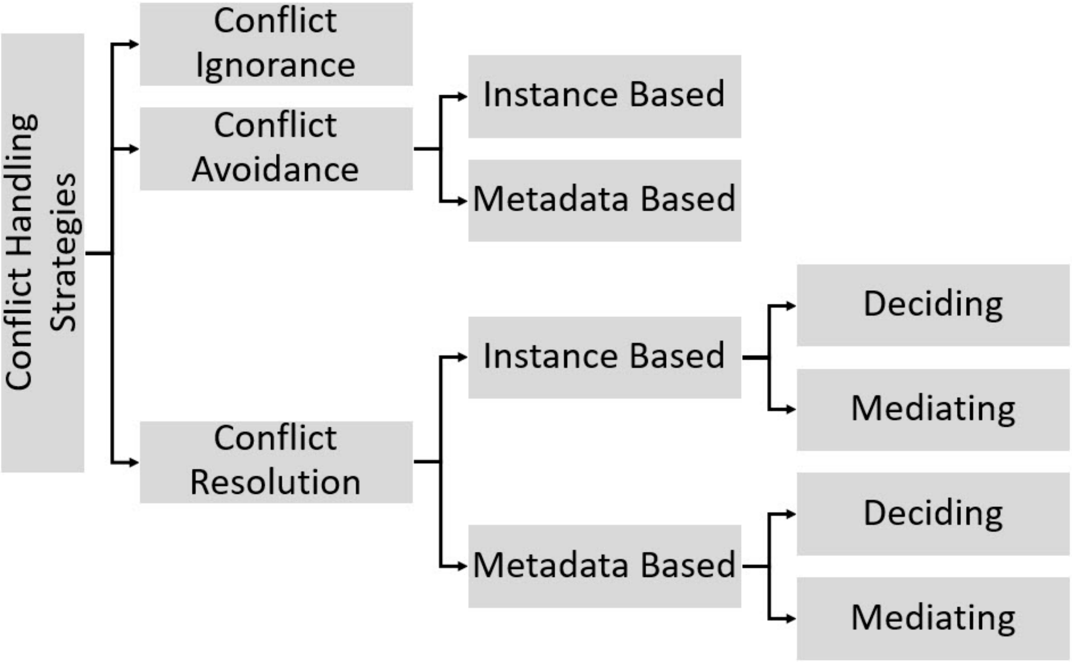
$$F = 2 \times \frac{precision \times recall}{precision + recall} \tag{2.3}$$

Precision and recall are the most popular evaluation metrics in spatial data matching works. F-measures appear mostly in works using ontology-driven matching methods. Unfortunately, not every work about spatial data matching indicates their results using precision, recall, or F-measure [250]. Some work [241, 133, 134] inform the results using a correct match ratio and do not give details if the other cases were false positive or false negative values.

2.1.3 Spatial Data Fusion

The Spatial Data Matching task results in matched pairs of elements from the integrated datasets. The next step is to use the multiple versions found of the elements across the databases to produce a complete and consistent one. This is the responsibility of *Spatial Data Fusion* step on Spatial Data Integration process [37, 24, 128, 152]. To illustrate the problem spatial data fusion faces, consider the following situation. Two road objects, $\mathcal{R}_a$ and $\mathcal{R}_b$, from different spatial data sets with slightly different geometries and values for the name attribute. What object represent better the geometry of the real-world road? Which one has the better representation of the name of the road? The desired output is a road object with better real-world correspondence for every attribute. Bleiholder [24] describe a taxonomy for the strategies to deal with data conflicts (Figure 2.4).

Figure 2.4: Strategies classification to handle inconsistent data.



Source: Adapted from [24].

Conflict Ignorance is a class of strategies that do not resolve or even care about data conflicts. There is no concern about the detection of conflicts. The strategies in this category are Pass it On and Consider all Possibilities. The former takes all attribute values and passes them to the next authority (user or another application) without modifications. The latter involves considering all combinations of possible attribute values and giving the user a choice.

Conflict Avoidance is the class of strategies that make a quick decision on how to deal with inconsistencies globally without analyzing specific conflicting values. It has a predefined choice of what attribute value to keep. The strategies of this category can be divided into Instance-based and Metadata based. Take the Information and No Gossiping are instance-based strategies while Trust your Friends are metadata based. Take the Information always prefer some value instead of Null values. No Gossiping ignores the inconsistencies and preserves only the values without conflict. Trust your Friends defines some criteria to give trust to a specific data source. The criteria may be user preference or determined by choosing the cheapest, largest, most complex, or other characteristics. The source that earns the trust has all its values carried out, with or without conflict.

Conflict Resolution analyses the data and metadata before deciding what to do. Thus, it has a higher computational cost than previous categories. This category adds another level of classification considering *deciding* and *mediating* strategies. Instance-based deciding strategies are Cry with the Wolves and Roll the Dice. The former uses some criteria to make the correct values prevail over incorrect ones (for example, using the mode among conflicting data). The latter pick one value at random from the available values. Better Bend than Break is an example of an instance-based mediating strategy that, instead of choosing one of the available values, tries to invent a value as close as possible to all present values (for example, calculating the average).

Metadata-based deciding strategies for conflict resolution are Keep up to Date, Ignorance is a Bliss, and Oops, I did it Again. The first keeps the most recent value. The second chooses a "known value" (a value that appears elsewhere, possibly in another table) among conflicting values. The third chooses values that were already chosen before and proven successful. A mediating strategy based on metadata example is Better Bend than Break II which is similar to Better Bend than Break but applied to metadata. For example, we can use the lowest common ancestor as a possible value if there is a conflict.

The characteristics of the data sources, the available data, and the desired output must be analyzed to choose an appropriate strategy. For example, some can only be applied if there are metadata available. Others cannot be used if the desired output does not support the insertion of values not present in any of the sources.

2.1.4 Discussion

The importance of spatial data in decision-making, especially in the urban context, makes using spatial data integration methods almost mandatory to take appropriate advantage of the availability of spatial data from different sources. The revised spatial data integration methods provide the tools to integrate urban data and build a multimodal transport network enriched with data representing complementary information to enable a more complex urban context analysis.

2.2 GIS for Transportation

Previously, the amount of data related to transportation and urban problems was quite limited[154]. Nowadays, the accelerated process of urbanization and the availability of connected technologies (internet, smartphones, sensors...) allow the creation of large volumes of data (Big Data). Among this data, it is estimated that about 80% have a spatial component [137]. In the initial stage, technology and conceptual advances made it easier for governmental organizations and private companies to adopt GIS-T. On the technology side, the consolidation of relational databases with spatial extensions and low-cost hardware has been decisive, while on the conceptual side, the development of algorithms to analyze optimal path [166], routing procedures, and a systematic approach to transportation planning [220] has evolved continuously since the early days of GIS. Since then, the evolution of technology has made it even easier to get, store, and process spatial data. Devices with GPS receivers and Internet access, like smartphones enable applications to gather and distribute spatial data in near real-time [137].

These kinds of devices helped to expand the concept of online social networks to Location-Based Social Networks (LBSN), as user interactions can now be located in space and time. These interactions can affect how dwellers interact and move across the city. For example, if someone is informed beforehand of a traffic jam, he can take another route, or if a friend is known to be nearby, one can decide to join him for happy hour. If dwellers change their movement behavior, the transportation system can be affected. To discover to what extent these online interactions affect the actual (or physical) world, we need to be able to gather, store and analyze a large volume of data that are typically heterogeneous and nonstructured. Government agencies, and social media, like Twitter and Facebook, produce data regarding traffic flow, the movement of people, bikes, buses, and weather reports that can be easily gathered but not easily integrated. Providing ways to collect, store, manage, integrate, analyze, and visualize spatial data from such heterogeneous sources is necessary to make the urban transportation system work efficiently and effectively [154, 253].

Urban transportation has become increasingly complex. The larger the city, the greater the need for multiple transportation modes to move efficiently. However, the availability of various modes of transport is not a sufficient condition for efficient transport. All transport modes should work in an integrated way. Providing affordable and efficient mobility to the population is crucial to promote quality of urban life [182, 158, 221, 245]. Spatial technology allows a change of paradigm in urban transportation policies. Due to difficulties in collecting spatial data, the results focused on the road network for vehicles, leaving aside pedestrians and cyclists. Today, it is possible to collect spatial data from individual entities such as cars, pedestrians, and buses in near real-time, bringing new possibilities to analyze the transportation system in an integrated way. In addition, data produced by online social networks and other Internet services can enrich the database and help achieve better application results.

This scenario also changes GIS-T technology from a static to a dynamic concept. The typical approach to GIS-T applications in the last decades has focused on organizing spatial data in map layers so that each layer represents a snapshot of some transportation-related data [206]. For example, the street network can be visualized as a map layer, and any changes to data entities lead to an update of the current layer or the creation of a new one. Each layer is independent of the other, and the primary resource is to visualize multiple layers simultaneously in map overlay operations. This static approach no longer supports city planners' and transit users' needs. City planners must integrate, analyze and visualize data from several sources to improve transportation infrastructure and organization decisions. Transit users need route information considering traffic conditions, alternative transportation modes, security, and other concerns. To provide these services, GIS-T has to evolve to include dynamic concepts and to be integrated into static data to reflect the dynamic behavior of the urban transportation system.

2.2.1 Multimodal Transport Networks

People daily move around the city to work, study, sports, and other entertainment activities. In urban environments, combining more than one mode of transport is often necessary to reach a destination. Modes of transport can be private, using cars, cycling, walking, or public, using a bus, metro, train, or ferry. Each mode has its characteristics. For example, private cars are usually a fast mode of transport and can get you to almost any place in a city. However, traffic delays can affect cars and have a high maintenance cost (gas, insurance, parking, taxes). On the other hand, cycling and walking are cheap but limited in range, physical effort, or loading capacity.

Most large cities have traffic problems that force dwellers to spend hours on the commute. This has a direct impact on the economy and quality of life. Brazil spent around 2.7% of its Gross Domestic Product (GDP) on traffic jams [70]. In São Paulo metropolitan area, the total traffic jam reached 300 km/day in 2013, which cost R$ 69.4 billion (approximately 7.8% of the urban GDP) and is expected to reach 357 km/day in 2022 with a cost of R$ 120 billion. This constitutes a handicap, especially for the poor population, who usually live far from work or study and must spend over two hours a day on commute [170].

It is urgent to promote better alternatives to transport for the population. It is necessary to know the infrastructure and to analyze it to identify the bottlenecks and opportunities for improvement. However, it is difficult to consider the multimodal transport system due to the lack of integrated data. It is relatively easy to find data about the road network from commercial and open sources that can be used to find car routes. Still, it is not straightforward to find pedestrian [113, 114] or public transport data and integrate everything [102].

A Multimodal Urban Transport Network (MUTN), enriched with other sources of information, enables various analyses of the city, citizens, and their interactions. Waddell [242] use it to assess the regional and local accessibility of a neighborhood. Martens [145] assesses the effects of including more bike facilities on the transport system performance, especially regarding using bike-and-ride to access public transport. Zuidgeest [273] also model bike and bus integration and measure accessibility based on travel time, access based on the catchment area, and estimate the number of people who would access a bus stop or transfer station. Boyac [27] studied the dependency between characteristics of the multimodal network and the throughput of passengers and vehicles. Gil [78] builds a multimodal urban network to evaluate urban areas' mobility potential and performance.

Some researchers [10, 110, 246, 68, 97] use transit network data to map food deserts (regions without easy access to healthy food) across the city. Using a similar concept, other researchers [109, 226, 108, 6, 237, 129] also use the transit network to find "transit deserts", which can be defined as areas where the supply of public transit services is not sufficient to fulfill user demand. Badland [15] found a positive correlation between living far from transit stops (bus, tram, or train) and poorer self-rated health due to longer commuting, which increases the amount of overall sitting time. Blanchard [23] developed measures of transit accessibility considering transit schedules from GTFS files and pedestrian networks derived from OpenStreetMap. Salonen [195] analysed the differences in accessibility comparing travel times by different travel modes: cars and public transport. For each one they considered three approaches varying if congestion, parking and schedule (for public transport) times were taking into account.

Smarzaro [213, 214] explore the use of VGI data to calculate a Local Offering Index (IOL[2]) for indicators such as green area, gas station, health centers, bank agency, and other. IOL is a component used to calculate the Quality of Urban Life Index (IQVU[3]) [158] for the city of Belo Horizonte. To get the IQVU value, it is necessary to apply an accessibility index on IOL to measure how people from one area can access services in another. The accessibility index is based on the trip time among the unities of planning[4] using public mass transport. A multimodal transport network plays an essential role in this task. It can be used to calculate other distance metrics using other transport modes,

[2] *Índice de Oferta Local* in Portuguese.

[3] *Índice de Qualidade de Vida Urbana* in Portuguese.

[4] Unities of Planning are aggregated from census sectors. Belo Horizonte has 80.

like bike or car, or changing the criteria to cost of the trip instead of time.

As a MUTN is an important component when dealing with urban analysis, it is important to define it. Nes [161] defines multimodal transport as one in which at least two modes are used, and the traveler has to transfer from one mode to another. Zuidgeest [273] see it as a set of subsystems where each one represents a transport mode. The connections among systems are implemented as exchange points at nodes or terminals, but in this view people can only change to a different transport mode using a terminal. Madloi [142] and Chen [40] characterize multimodal transport as whenever the movement of people or goods involves at least two modes of transport from origin to destination. Considering only these definitions, if a person leaves home, walks to the bus stop, takes the bus, and then walks to the final destination, the trip is classified as multimodal because two modes were used: walking and bus. Some authors classify trips like this as unimodal because walking and cycling are considered only as access and egress modes [273] or consider only mechanized modes for multimodal transport [98]. Considering a complete origin-destination route, the access part is from the origin to the first transportation node (a bus stop, metro station, etc.), and the egress part is from the last transportation node to the final destination.

In the urban environment walking is the "glue" that ties all other transport modes and can function as a mode of its own, especially in developing countries [238]. Using this interpretation, any route that involves walking as one of the transport modes can be considered to be multimodal.

Schoemaker [203] organize a three-layer framework to analyze transportation systems: *Activities*, *Transport Services*, and *Traffic Services*. Each layer demands from and provides services to the others. For example, the Activities layer provides "people" and can demand quality, prices, routes, and timetables from the Transport layer. Van Binsbergen [236] extend this transportation system layer view, comparing it with the Open Systems Interconnection (OSI) networking model [272]. They use seven layers: user, agent, integrator, carrier, traffic management, infrastructure, and corridor & nodes. Compared with [203], the user layer corresponds to the Activities; the agent, integrator, and carrier correspond to the Transport services, and the remaining layers correspond to the Traffic services. Their work was focused on the transportation of goods instead of people, and they evolved this networking view to become more suitable to their purpose. Nonetheless, the layered approach has many advantages that can be applied to MUTN management.

The MUTN has to store data about infrastructure (streets, rail, bus stations) and transport services of each mode (car, walking, cycling, bus, metro). It needs to store data about all the possible transfers from one mode to another, and it is essential to capture the static and dynamic properties of the system [40]. Integrating and incorporating data from users and other heterogeneous data sources is also important to better understand urban transportation patterns. The first step is to choose a data model to represent all the entities and their relationships.

2.2.2 Multimodal Transport Data Models

Creating transit data models that can be used widely has been the focus of many researchers from academia and industry. Proposals vary in some aspects, like the location referencing method, spatial data structures, or topological representation. The most used data models are Geographic Data File, ESRI Trans Model, MDLRS, and GIS-T Enterprise Data Model.

2.2.2.1 Geographic Data File

The Geographic Data File (GDF) is an ISO international standard to model, store and transfer geographic data, focusing on data for in-car navigation systems [202]. Its history began with concerns of the producers of road maps about a common standard to exchange data by the late 1980s. The European Committee for Standardization (CEN - *Comité Européen de Normalisation*) developed the first versions, which evolved to the ISO GDF, described on the ISO/TR 14825:1996 standard as “a system for the interchange of digital road-related geographic information. It considers all the application requirements in the road transport and traffic telematics (RTTT) field” [103]. The standard was revised in 2004 [105], and the current version is ISO GDF 5.0, or ISO 14825:2011 [106]. The changes in GDF 5.0 from previous versions include “UML model migration and refinements, harmonization with linear referencing and geospatial web standards, support for 3-D content and time coordinates, comprehensive character set and phonetic representations, and new XML- and SQL-based delivery formats.” [105].

The adoption of UML for modeling made it easier to maintain the entire model's consistency and provide different physical realizations for data exchange. GDF 5.0 provides three: ASCII flat file, XML schema specifications, and SQL encoding specifications.

GDF uses a three-layer view of the transport model. Level 0 is for the physical infrastructure. Level 1 models the transport infrastructure as a graph for routing purposes. Level 2 simplifies the graph structure of Level 1 to provide directions to drivers.

2.2.2.2 ESRI Transportation Data Model

The ESRI Transportation Data Model (ETDM) is an updated version of the Unified NEtwork-TRANSportation data model (UNETRANS) project [64]. It started as an initiative by ESRI, a consortium of organizations, software developers, and university faculty to develop a "template" data model for transportation. It is an object-oriented data model [181], using Unified Modeling Language (UML) as model notation, and is based on the work of [33]. ETDM closely follows ESRI's geodatabase model, so ESRI software ArcGIS can easily use it. Its main objective is to provide an initial model that transportation agencies can customize to their needs and then facilitate data sharing among agencies and other entities.

ETDM models the transportation system using three layers: reference network, route features, and events [169]. The reference network layer contains the infrastructure data, like road and rail networks. It generally stores the centerline of the transportation networks using a linear spatial data representation. Connectivity and adjacency information is also stored in this layer. The route features layer uses the data on the reference layer to build more complex features, like bus routes, carriageways with direction information, or streets. For example, a street may be represented on the routes layer as a sequence of connected linear features on the reference network. The events layer can store any information relevant to the data model user that needs to be located on the reference or the routes layer. This layer can represent any asset or occurrence. Traffic signals, bus stops, accidents, moving objects' locations, and road sections on maintenance are examples of events that can be stored. Each is directly related to an object of the reference layer and uses Linear Reference System and Dynamic Segmentation to be located and represented.

Besides the layers, ETDM also organizes its elements using packages to group the feature classes that implement some data model functions. There are six packages: (1) reference network package, (2) route and location referencing package, (3) asset package, (4) activities package, (5) incidents package, and (6) mobile objects package.

2.2.2.3 GIS-T Enterprise Data Model

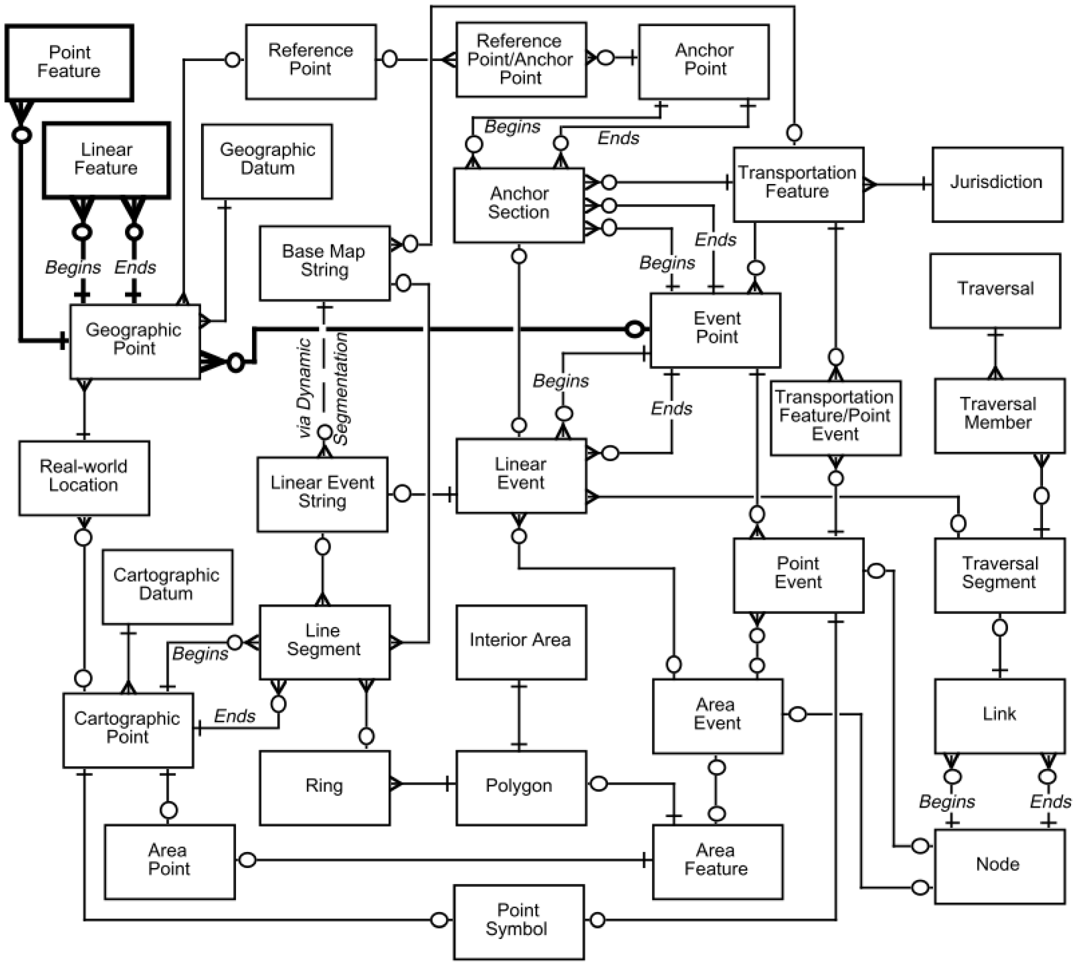
The GIS-T Enterprise Data Model was developed by [62]. It aims to facilitate sharing of road map databases among transportation organizations and supports linear and nonlinear referenced data. It consists of five logical units that group entities responsible for a specific set of functions: (1) basic model, (2) topology, (3) cartography, (4) linear Datum, and (5) non-transportation features [153]. The *basic model* (Seven entities: Jurisdiction, Transportation Feature, Event Point, Transportation Feature/Point Event, Point Event, Area Event, and Linear Event) is responsible for the physical aspects of the road network. The *topology* module (Five entities: Traversal, Traversal Member, Traversal Segment, Link, Node) enables the definition of paths (e.g., bus routes) through the physical transportation network. The *cartography* module (Eleven entities: Area Feature, Point Symbol, Polygon, Interior Area, Ring, Area Point, Cartographic Point, Cartographic Datum, Line Segment, Linear Event String, Base Map String) provides geometric representations to entities. The *linear Datum* module (Seven entities: Anchor Point, Anchor Section, Reference Point/Anchor Point, Reference Point, Geographic Datum, Geographic Point, Real-World Location) defines anchor points that are related to the transportation features and events. Finally, the *non-transportation features* module (Two entities: Linear Feature and Point Feature) adds the relationship between the Linear LRS and a nonlinear one, so any entity with geometric representation can be located directly on the surface of the Earth without using its relative position to a transportation feature. Figure 2.5 shows the complete Logical Data Model for GIS-T Enterprise with colors representing entities added to each development step.

2.2.2.4 Multi-Dimensional Location Referencing System

The Multi-Dimensional Location Referencing System (MDLRS) was developed by the USA National Cooperative Highway Research Program (NCHRP) from the NCHRP 20-27 models. It intends to provide a data model that can be used by the entire GIS-T community (agencies, software developers, researchers, and transportation agencies) [3].

MDLRS divides the data model using four domains: aspatial, spatial, temporal, and causal. The aspatial domain regards objects that help to answer “what” is at a location. The spatial domain is interested in “where” an object is. The temporal domain answers “when” an object is at a location. Finally, the causal domain answers “how” an object changed and “what caused” the changes. Table 2.1 lists the components of each domain [125, 124]. Figure 2.6 presents the conceptual view of MDLRS using UML notation.

Figure 2.5: GIS-T Enterprise Logical Data Model.



Source: Adapted from [62].

2.3 Spatial Data Sources

When the urban environment becomes larger and more complex, the need to use data from different sources to help decision-making also increases. Many types of data can help us understand city dynamics, such as geographical data, human mobility data, online social network data, environmental data, etc.

Geographical data is the basis for urban computing applications. The road network, transportation network, points of interest, and land use data represent this category. The *road network data* are usually represented as a graph of road intersections and segments as nodes and edges, respectively. Each intersection and segment can have many associated properties besides their latitude and longitude. For example, segments can have a speed limit, type of road, number of lanes, mode of transport allowed, and other related data. The *transportation network data* can be understood as an information layer using the

Table 2.1: MDLRS domains and their elements [124]

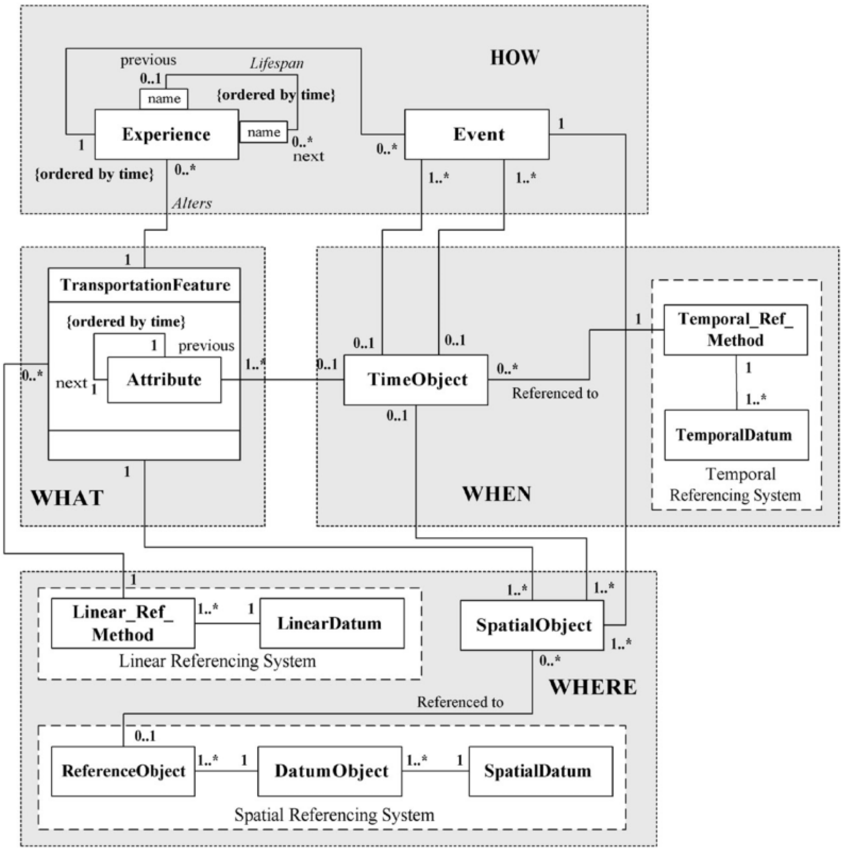
Aspatial	- Transportation Features - Transportation Complex - Conveyances
Spatial	- Spatial Referencing Systems - Spatial Objects - Topological/Geometric Relationship - Geometric Objects - Coordinate Object - Error Propagation - Topologic Objects
Temporal	- Temporal representations - Temporal Referencing Systems - Temporal Relationship Operators
Causal	- Event and Experience Objects - Metadata

road network to record the transit routes and stop facilities of the bus, metro, bike sharing, and other transit modes. Transportation network data should also include schedule information about the expected time for each vehicle to reach each stop facility on their routes. *Points of Interest* are locations, usually represented as single points, representing facilities such as hospitals, schools, supermarkets, tourist attractions, etc. *Land use data* represents the many activities where people use land and its resources.

Human mobility data help us understand how people interact with each other and their surroundings. Zheng [265] categorizes human mobility data into traffic, commuting, mobile phone, and geo-tagged social media data. *Traffic data* are generated by monitoring taxis, buses, metros, trains, and other traffic vehicles with embedded sensors (for example, GPS navigation systems). Other forms to collect traffic-related data evolve using road sensors and cameras. Road sensors can be placed on roads to measure the time interval a vehicle takes to get across them, and camera images can be used to assess traffic conditions visually. However, road sensors usually lack citywide coverage, and extracting information automatically from camera images can be difficult. *Commuting data* represents data gathered from passenger cards to use the transit system and usually includes the time getting in and out of transportation vehicles, the type of transportation, and the fares. It is valuable data to help improve public transportation systems [258]. *Mobile Phone data* are collected from the recorded interactions between mobile phones and the telecom infrastructure (cell stations).

Online Social Networks data relates to any post on social media that can be georeferenced. This type of data contributes to understanding people's behavior because social posts' contents hint at what activity they are doing. Also, the links among users allow us to analyze how people interact with each other, revealing important information

Figure 2.6: MDLRS conceptual view in UML notation.



Source: Adapted from [125].

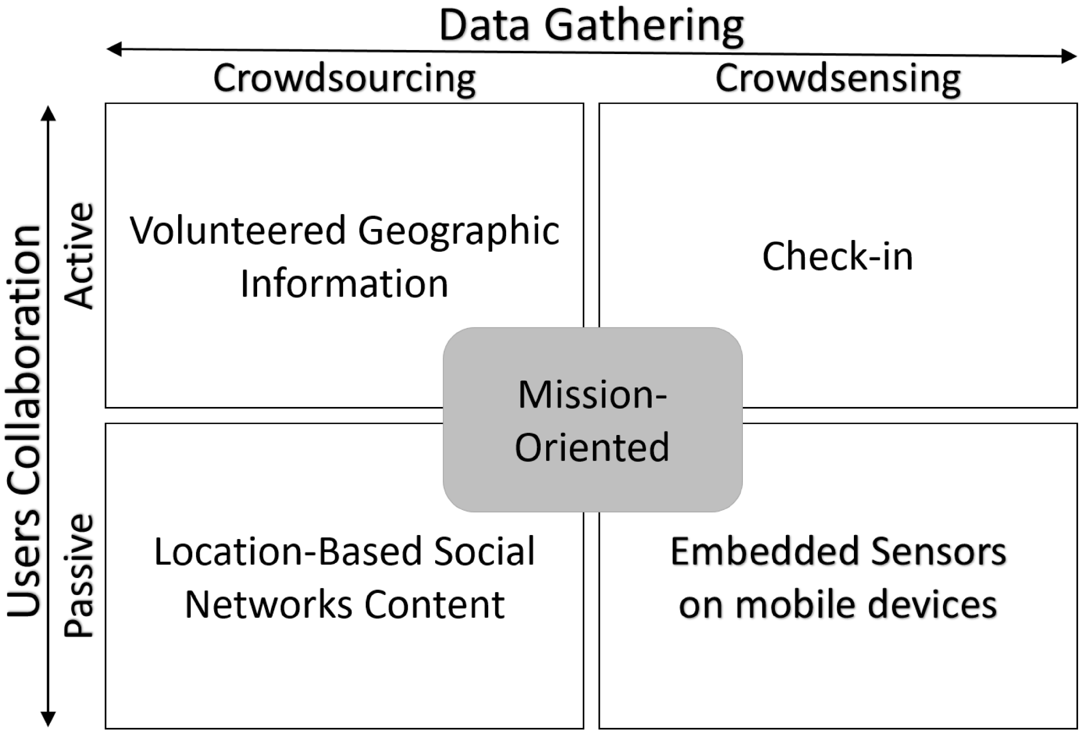
about communities. When geo-referenced data is central to the social media service, it is called a Location-Based Social Network (LBSN).

Environmental data is another important type of data on the scope of urban computing as it enables us to make better decisions regarding the use of natural resources and measure the impact of the city on the environment. Temperature, humidity, rain conditions, air quality, water quality, electricity, and water consumption are examples of environmental data.

These data can be gathered from official or unofficial sources. *Official data sources* are data provided by governmental entities or agencies and are considered to have "high quality". The process of producing official data is usually made by professionals, which must comply with some national or international standard. Also, the cost of gathering official data is high and is not updated frequently. A good example of official data is the National Census of Brazil [101]. The last Census was made in 2010 and is theoretically repeated every ten years. It used 240 thousand people in the process, consumed R$ 1.4 billion in resources, and produced detailed social, economic, and spatial data about 67.569.688 domiciles and their residents. The census data is available only in aggregated form for each census sector. Brazil was divided into 314.018 census sectors in 2010.

Unofficial data sources are those governmental agencies that do not disclose. Online Social Networks are a great source of unofficial data. There are plenty of works using it as a complement [199, 200] or as a proxy when official data are not available or up-to-date [213]. Unofficial sources are easily accessible (using API) and, for some cases, can be as complete and accurate as official ones [66] but unfortunately have many quality concerns that must be evaluated before appropriate use [56, 36]. Unofficial data sources from online social media can be classified using two dimensions: how the data are collected and how the user collaborates on the data creation [149]. The former is classified as crowdsourcing and crowdsensing, and the latter as active and passive (Figure 2.7).

Figure 2.7: Crowndsourcing and crowdsensing taxonomy based on user collaboration.


Data Gathering
Crowdsourcing
Crowdsensing
Users Collaboration
Active
Passive
Volunteered Geographic Information
Check-in
Mission-Oriented
Location-Based Social Networks Content
Embedded Sensors on mobile devices

Source: Adapted from [149].

Active crowdsourcing occurs when the user provides the information consciously and knows how the data will be used. OpenStreetMap is an example of active crowdsourcing. In *passive crowdsourcing*, the user is unaware of how or whether the data provided by her is used. Twitter posts (tweets) and geo-tagged photos from photo-sharing services like Flickr and Panoramio are examples in *passive crowdsensing*. In *active crowdsensing*, the user can employ embedded sensors from her devices to create spatial content. For example, when the user records her presence at a place (a check-in) in services like Foursquare or Yelp, the GPS sensor of the smartphone is used to register the user's location. In *passive crowdsensing*, the user uses an application that records data from the sensors on the device. Still, there is no involvement of the user in the recording process. The routing application Waze is an example. When used while driving, the application continuously records position and speed data to improve the road network map, and the navigation by other users. It is important to note that this classification is not rigid. For instance, Waze can also be used as an active crowdsensing application when the user informs about an accident along her route.

The remaining section presents some common data sources used in urban computing and transport-related problems, such as General Transit Feed Specification (GTFS), OpenStreetMap (OSM), and Location-Based Social Networks (LBSN). It finishes presenting the available spatial data quality metrics used to access data from unofficial sources.

2.3.1 General Transit Feed Specification

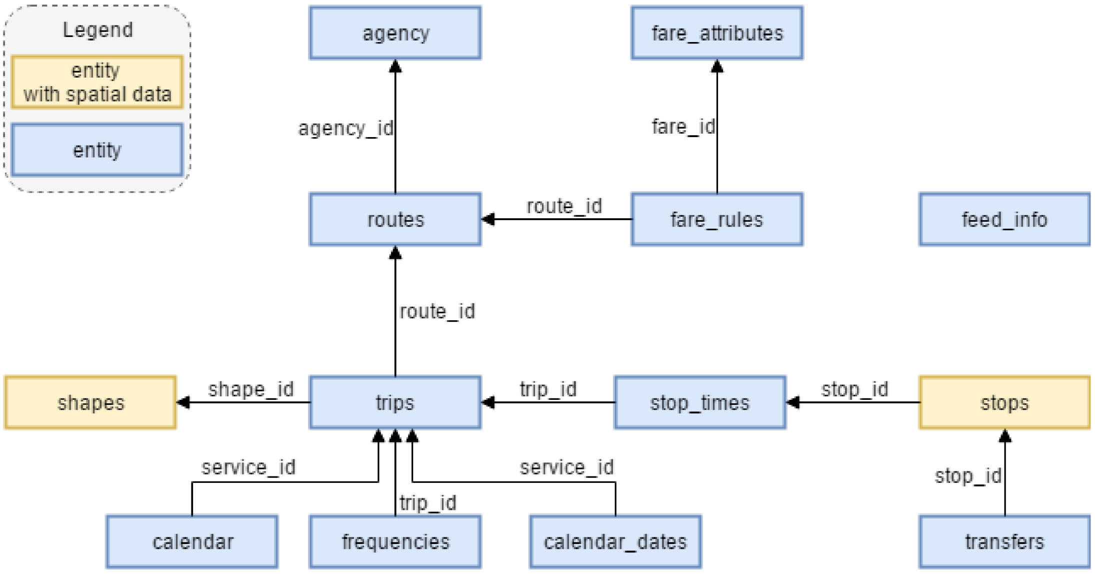
The creation of the *General Transit Feed Specification* (GTFS) began in the summer of 2005, when Tim and Bibiana McHugh, IT managers at TriMet, the transit agency for Portland (Oregon, USA), got frustrated when they couldn't find transit information on mapping programs while traveling. For them, it should be as easy to plan a trip using public transportation as it is by driving. They asked Mapquest, Yahoo!, and Google if they had plans to include transit information in their mapping platforms. Only Google replied.

Google's software engineer, Chris Harrelson, struggled to incorporate transit data into Google Maps. He joined Google after finishing his Ph.D. in routing problems in public transportation systems and, using the "twenty percent flexible project time" company policy to dedicate to side projects; he got in contact with the McHugh couple. They work together to elaborate a format to export TriMet's transit data to be included in Google's geospatial database. By the end of 2005, Google made the bus and light rail schedules available for Portland on Google Maps. Following the successful experience in Portland, Google offered its trip planner for free to any transit agency, using the same format to export their data. In 2006, Seattle (WA), Eugene (OR), Honolulu (HI), Pittsburgh (PA), and Tampa (FL) joined the group of cities with transit data available. In 2007, the data format was published as *Google Transit Feed Specification* [150].

The adoption of GTFS quickly spread to many cities around the world. Several applications emerged consuming data in this format. In parallel, the Open Government Data initiative gained momentum, and the pressure to provide public data grew. These factors and their simplicity contributed to adopting GTFS as a *de-facto* standard. By 2009, there were so many agencies and applications using GTFS that Joe Hughes, a software engineer working on Google Transit, suggested a name change to *General Transit Feed Specification* and transferred the responsibility of the standard development to a larger community of agencies and developers [185].

GTFS specifies the contents of a maximum of thirteen CSV files. Figure 2.8 shows the relationship between the files and the attributes that connect them. There are some alternatives to represent the transit information. For example, the “stop_times.txt” file can contain all the possible trips, which means that if a specific trip occurs multiple times along the day, each one of these has to be represented on the stop_times file. Depending on the number of repetitions of the trips over the day, the file size can become huge. The alternative is to use the optional file “frequencies.txt” to inform the frequency of a specific trip. For example, if a trip is represented by a sequence of 50 stops on stop_times file and is repeated 40 times along the day, it would result in 2000 rows (50 × 40) on stop_times file. However, using the frequencies file, only one copy of the trip needs to be on the stop_times file, so it is possible to retrieve the offset time between stops, and the start time of each trip repetition is derived from the frequencies file data.

Figure 2.8: Relationship diagram for GTFS files



Source: Made by the author.

The wide adoption of the GTFS standard changed how people and transit agencies interact. The common approach in the past was for transit agencies to develop applications to provide people with transit information. Now, the agencies need only to make the data available; anyone can create and consume the transit data. This latter approach works more dynamically as the availability of applications increases and becomes more customer-focused [150]. Applications such as multimodal trip planning, transit timetable, real-time transit information, and others [9] can be generically developed and reused anywhere GTFS data is available.

Many researchers have also used GTFS. Tran [116] develop a framework to synchronize the bus stop information on GTFS with OpenStreetMap data so the transit agencies can upload official data to OSM and retrieve back the contributions and corrections made by OSM users. Wong [249] used GTFS data for transit analysis by calculating several indicators from the report "Transit Capacity and Quality of Service Manual" [121]. This document describes several metrics for transit service availability, but it lacks in providing ways to get the necessary data. Tao [224] combined smart card travel data and GTFS files to reconstruct bus passengers' trajectories and created flow-comaps to visualize temporal changes in the flow patterns of bus users. Flow-comaps combine two techniques: flow mapping and conditional plotting. Flow mapping is used to spatially visualize movement patterns. The conditional plot uses the classification of raw data into subsets plotted individually based on certain conditions.

Brosi [28] developed a tool called TRAVIC[5] (Transit Visualization Client), to visualize the movement of public transit all around the world. It uses the static GTFS data to interpolate the position of the vehicles along the routes contained on `shapes.txt` file, or the vehicle real-time positions when the GTFS-realtime is available.

Farber [67] used GTFS data as the basis to calculate the time necessary to reach the nearest supermarket, at different times of the day, from every census block using transit transportation to investigate food deserts, areas which have access to healthy food is harder. Researchers in Nairobi, Africa, are looking to make transit data available to the population, collect bus route information manually by riding the buses, and use a GPS device to log the trip. These data were then used to build the GTFS files, but they extended them by providing a new type of field called "continuous stop". A route with this property enabled allows one to board or debark the bus at any point on the route path. So any route of this kind must have the path information with GTFS, usually as shapefile [248].

Perrine [171] proposed methods to match data from GTFS and OSM regarding stops and routes as a basis for building a multimodal network that can be used in transit analysis. Steiner [219] analyzed the quality of GTFS-realtime available for the Netherlands and found that it was unsatisfactory due to the lack of a large part of vehicle positions and trip delay data. Without these data, it was not possible to make a good analysis of the benefits of real-time data for trip planning.

Farber [68] used GTFS to build data cubes with the origin, destination, and time of the trip, considering the start time every minute of the day. The centroid of each region was considered as the origin or destination. The shortest path for each pair was computed based on GTFS file data. As the region analyzed had 1326 blocks, approximately 2.5 billion shortest paths were calculated (1326 × 1326 × 1440 minutes daily). These data cubes were then combined into a transportation survey to seek mismatches in transit demand and supply based on socioeconomic aspects. Similarly, Boisjoly [25] also used GTFS data to compute a transit travel time matrix to calculate different accessibility

[5] TRAVIC is available at http://tracker.geops.de/, accessed on April 09, 2022

measures to job sites.

The GTFS files need a road network dataset to provide transit information to its users correctly. OpenStreetMap is one of the most used and freely available sources of the street network, land use, and points of interest.

2.3.2 OpenStreetMap

In the early 2000s, digital maps were mostly proprietary. Inspired by the success of Wikipedia and the advent of cheap GPS devices, Steve Coast created OpenStreetMap (OSM), in 2004. The purpose was simple: let the users create the map collaboratively. The users can use GPS tracklogs, aerial images, and any other free source of geographical data to feed the OSM database. All content is free to anyone to use under the Open Database License (ODbL)[6]. It is possible to download a file containing all data from the entire world from Planet OSM[7]. It is updated weekly but has a huge size. In April 2022, the uncompressed file, in XML format, has 114 GB, and the compressed, in PBF[8] (Protocolbuffer Binary Format) format, has 63 GB. It is possible to download smaller files with only the changes in metadata, which make it easier to update a local copy. It is also possible to download data using the Overpass API[9] which provides more options to control which data will be retrieved. There are also some services that provide organized extracts of OSM data. The Geofabrik[10] service has extracts of OSM data organized by continents, countries, and regions in PBF and shapefile formats.

OSM's data model uses three types of basic elements: nodes, ways, and relations. A *node* represents a geographical point. It has at least an id number and geographical coordinates. A *way* represents linear features (streets, rivers) or area boundaries (buildings, forests, lakes) and is formed by an ordered list of between 2 and 2,000 nodes. When the way represents an area boundary, the first and last nodes must coincide spatially. The area can be solid (e.g., a building) or not (e.g., a roundabout), and the tags associated with the way must be examined to define its type. *Relations* represent a relationship between two or more OSM elements (nodes, ways, or other relations). For example, an area boundary with a hole can represent a relation between two ways representing areas. A turn restriction can be mapped as a relation among two ways representing streets and a node connecting them, and a bus route can be a relation containing a list of ways. As a

[6] http://opendatacommons.org/licenses/odbl/, Accessed on April 09, 2022

[7] http://planet.openstreetmap.org/. Accessed on April 12, 2022.

[8] http://wiki.openstreetmap.org/wiki/PBF_Format. Accessed on April 12, 2022.

[9] http://wiki.openstreetmap.org/wiki/Overpass_API. Accessed on April 12, 2022.

[10] http://download.geofabrik.de/openstreetmap/. Accessed on April 12, 2022.

way is an ordered list of nodes, a relation is an ordered list of the objects it contains, called the relation's members. The typical elements in GIS data are points, lines, and polygons. The correspondence between nodes and points; ways, and lines is straightforward. In OSM, a polygon is a closed way with a tag denoting it represents an area feature (e.g. building=yes). OSM tagging system is flexible but makes it harder to query and manipulate data [77], and the lack of a formal standard to enforce the use of a predetermined set of tags can generate inconsistencies. Although there is no formal standard, there is an informal one that emerges from the user community of map editors. It is important to analyze the usage and verify how the tags are being used to identify the relevant ones.

Based on the content created by its users, OSM can be classified as a volunteered geographic information (VGI) application [81]. Kazemi [115] classifies OSM as a self-incentivized crowdsourcing platform where users contribute voluntarily without a defined reward. Mateveli [149] classifies OSM as an active crowdsourcing platform where users contribute consciously and voluntarily and know how the data will be used. Some authors [31, 45, 93] use the term *produsers* to designate those users that can "choose to receive, appropriate, creatively use, share, and/or produce geospatial information independently or in collaboration with others" [31, p. 155]. Users can use different editors to edit data on OSM. The first, called iD, is suitable for beginners and is the default choice. The second, called JOSM (Java OpenStreetMap Editor) which allows them to download data from OSM and edit locally before uploading the changes in batch back to OSM. Since OSM is an open platform, there are many other editors available[11] to edit OSM data.

Although the use of the editors facilitates the insertion and editing of the data, the collaborative nature of the OSM allows the input of inconsistencies. This brings many concerns about data quality, and many researchers have sought to analyze it. Haklay [90] was the first work to systematically analyze OSM data's quality. He used a reference dataset (Ordnance Survey Meridian 2) to represent the ground truth and compared the positional accuracy of OSM data in England, using a buffer-zone method technique [82]. The roads had a buffer set to 20m and 1m on the reference and OSM datasets, respectively, then the overlap area was calculated. The results indicated a positional accuracy of over 80% for roads classified as primary. He also tested the *completeness*. England region was divided into a grid of one km resolution (discarded all in the coastline, i.e., an area less than one km^2) then the road length of both datasets was compared in each cell.

Complementing [90], Ather [14] used another high-quality dataset (OS Mastermap) to compare with OSM data. He extended the analysis to include lower-level roads beyond primary ones. The results were similar for positional accuracy (over 80%) and also found a positive correlation between attribute completeness (considering road name) and the number of users per area.

Zielstra [271] used the completeness metric to evaluate the quality of OSM data in

[11] https://wiki.openstreetmap.org/wiki/Editors, Accessed on April 09, 2022.

Germany, the state of Florida, and other US cities [270]. The reference dataset used was TeleAtlas MultiNet. They found that OSM data has a high level of detail in urban areas, but as it gets away from the city center, the detail level drops significantly. Ciepluch [43] found similar results analyzing Ireland OSM data using 5 km grid cells to calculate quality data metrics such as completeness, accuracy, the density of users' contributions, the spatial density of points and polygons, types of tags and dominant contributors to a particular area or geographic feature type. Girres [79] analyzed OSM data in France using eight metrics to assess OSM data quality: geometric accuracy, attribute accuracy, completeness, logical consistency, semantic consistency, temporal accuracy, lineage, and usage. The reference dataset used was BD TOPO Large Scale Referential. They highlighted the responsiveness and flexibility as a strength of OSM data and the heterogeneity (different data sources, data capture methods, and user profiles) as a problem that can limit its use.

DeLeeuw [48] found that road classification made by OSM contributors with local knowledge had an average accuracy of 92%, which is better than professional surveyors without local knowledge [263].

Neis [159] also used German OSM data to compare with a proprietary dataset and found that, by 2011, OSM only missed 9% of car navigation data compared to a commercial dataset (TomTom) but provided 27% more data regarding the total street network and pedestrian information. Zielstra [269] compared free and proprietary street data sources, calculating the shortest paths for pedestrian navigation. They extracted, from each dataset, only the streets that allow pedestrian use and randomly selected 1,000 pairs of origin-destination locations. It was shown that OpenStreetMap provided a complete source of free data for pedestrians. It also stated the importance of integrating available datasets from different sources to build and use a complete one in transportation applications.

2.3.3 Location Based Social Networks

Online Social Networks (OSN), like Facebook and Twitter, enable users to create a network of friends (a following/follower network, in the case of Twitter) and share any content with all or a group of its contacts. Smartphones and other portable devices equipped with Global Positioning System (GPS) and connected to the Internet motivated the creation of Location Based Social Networks (LBSN) [174]. In LBSNs, geographical location can be shared along with the regular content of interest.

Zheng [266] uses location's role on LBSNs to classify them into Geo-tagged-media-based, point-location-driven, and trajectory-centric. *Geo-tagged-media-based* LBSNs add the location to the shared media. It is important to note that the location can be added to the media after posting it. Geo-tagged photos, videos, and tweets are examples of geo-referenced media. Flickr and Twitter can be considered LBSNs in this category. *Point-location-driven* LBSNs encourage users to share their location in real-time. For example, a user can arrive at a shopping mall and share his location. His contacts are informed, and maybe they can meet at the place. The user also can share his opinion about the place. Foursquare, Yelp, Google Places, and Facebook Places are examples of point-location-driven LBSN. *Trajectory-centric* LBSNs enable users to share a route, which is a sequential connection of point locations. Information like distance, speed, duration, altimetry, and others the user can provide (such as tags, photos, and opinions) are also shared. Sports logging services like Garmin Connect, Nike+, and Strava represent this category.

Foursquare was launched in 2009. First, it was available in 100 cities worldwide, but in January 2010, it became global. In 2011, Foursquare reached about 7 million users. There are currently more than 50 million users [73]. Initially, the user could only share her location with friends by making a "check-in" at a venue. Later, more features were added, such as a user "to-do list", which allows saving locations to be notified when nearby. Foursquare issues rewards based on the user's check-in history. For example, users who make 50 check-ins earn a "superstar" badge in their profiles. Another reward is venue mayorships. The user who makes more check-ins on a venue is considered its "mayor". Many venue owners offer advantages to Foursquare mayors, like discounts or free parking. Foursquare lets users post opinions about venues. People searching for a restaurant nearby can see how Foursquare users reviewed that place. Good or bad reviews can have a significant impact on business performance. In 2014, Foursquare was divided into two applications: Foursquare and Swarm. Social aspects, like sharing locations with friends and finding the location of friends, migrated to Swarm, while Foursquare concentrates on venue recommendation.

Yelp was launched in 2004 to help people find good jobs and places. Like Foursquare, Yelp users can check in at venues and share their location with friends. However, Yelp focuses on user reviews of business venues. Users can rate other user reviews so that Yelp can make better place recommendations. In 2022, Yelp had over 38 million unique users on its website and 59 million using the service through a mobile app [217].

Google Places is the service Google provides to store any POI for users. Users use Google Places data whenever they search for a place in Maps or other Google location services. Users can record a missing venue by providing its required attributes name, location, category, optional days/time open, phone, website, and pictures. Data can be retrieved using Google Places API[12] with a limitation of 2,500 requests per day for free.

[12] https://developers.google.com/places/, accessed on April 09, 2022

All venues' categories must be selected from a pre-existing list[13], but there is no explicit hierarchy on it.

Facebook Places works very similarly to the others. It is also organized as a place listing, where users can record new places and edit and review existing ones. Facebook also has an API that can retrieve venues by a search on a region.

2.3.4 Discussion

Location Based Social Networks and other VGI sources are valuable sources of data that can be collected at low cost and that cannot be ignored by governments and organizations interested in the interactions among people, their online presence, and the impacts on real-world [207]. For example, [18] integrated census data, social media, online traffic data, water, and electricity consumption to seek relations between all these data and land and residential prices. Sarwat [201] uses the location of users and their contacts to improve recommendation systems for places. Garcia-Marti [75] uses a dataset of tick bite records, in the Netherlands, and enriched it with data on temperature, precipitation, and vegetation to understand when and why tick bites are happening, so proper health interventions can occur in order to reduce Lyme disease cases. Slope information for roads is valuable data to be considered when one is planning a route by foot or bike. As commercial or authoritative datasets are more focused on cars they do not pay much attention to topographic data. John [111] used GPS traces from users to derive the slope of road segments with satisfactory accuracy showing that VGI data has the potential as a low-cost data source alternative.

As long as technology continues improving, the world is becoming more and more connected and also urbaner. Methods to gather and integrate different sources of heterogeneous data are necessary to understand city dynamics and help to solve problems in this scenario with the increasing demand for services and infrastructure at urban spaces [229].

[13] https://developers.google.com/places/supported_types, Accessed on April 09, 2022

2.4 Quality Metrics for Spatial Data

Official datasets are usually made by professionals. This produces high-quality datasets but it has a high cost associated and is not updated frequently [44, 54]. On the other side, VGI data are produced by users who present great skill variability to deal with spatial data [91]. The problem of spatial data quality has been always a concern but the change in the data production paradigm and the popularity and massive use of crowdsourced and VGI data boosted research in this field [209].

Spatial data quality metrics can be classified as internal and external [13]. External quality metrics verify the *Fitness of Use* [50] which means the suitability of the dataset according to a specific purpose. Internal quality metrics "reflects the data production specifications, which recognizes errors in data" [13, p. 41].

The internal metrics are defined by [104] and consist of Completeness, Positional Accuracy, Attribute Accuracy, Logical Consistency, Semantic Accuracy, Temporal Accuracy, and Lineage.

Completeness regards the coverage of geographic data in a dataset in relation to ground truth. A reference dataset is needed to represent the ground truth and this is one of the difficulties to calculate the metric. When some reference dataset is not available some intrinsic measures can be applied as a proxy for completeness. [17] suggests that if the number of additions to a VGI dataset is decreasing for a feature type it might indicate high completeness. Completeness can be a measure used to assess fitness-for-use for particular applications [12].

Completeness can also be viewed from different perspectives. When the term completeness is used alone it usually means the ratio in which the features from the reference dataset are represented in the test dataset. It can also be used to measure if all types of features are represented (*class completeness*) or how many features have valid values for an attribute (*attribute completeness*). For example, on a road network dataset, completeness could be calculated by the ratio between the summed lengths of all roads from the reference and the tested dataset [204]. Class completeness would evaluate if all types of roads of the reference dataset are present on the tested dataset, while attribute completeness could evaluate how many of the roads from the tested dataset have values for the name attribute.

Positional Accuracy measures how well the geographical coordinates of a feature represent its real location [63, 90]. This metric is affected by the way data was collected. GPS sensor imprecision and miss displaced or low-resolution aerial images used as the basis to insert data are common causes of problems in the positional accuracy of geographical datasets [66, 17].

Attribute accuracy evaluates how correct is the non-spatial attributes of a feature

[89, 79]. The method to compare values to measure attribute accuracy depends on the data type. For example, string values can be compared using exact match or some string similarity metric like Levenshtein distance or Soundex (See Section 2.1.2.1).

ISO standard [104] defines *logical consistency* as “degree of adherence to logical rules of data structure, attribution, and relationships” and subdivides it into four elements: conceptual consistency, domain consistency, format consistency, and topological consistency. Conceptual consistency verifies if the tested data respects the conceptual data schema. Domain consistency verifies if attribute values are within the valid range of a specific domain. Format consistency verifies if tested data respects the physical data structure. Topological consistency verifies the correctness of topological characteristics. One-way roads finishing in a dead end and non-closed areas are examples of topological errors that affect topological consistency [94].

Semantic Accuracy verifies if the features on the tested dataset have compatible types as its correspondence features on the reference dataset [79, 66, 12]. For example, on a road dataset, the semantic accuracy can measure the ratio of how much of the road classified as primary on the reference dataset was correctly classified as the primary of equivalent type on the tested dataset. Semantic Accuracy can be considered as an attribute accuracy metric if the feature class is considered a nonspatial attribute.

Temporal Accuracy evaluates the rate of updates and the validity of the changes in a dataset. [79] defines it as the “actuality of the database relative to changes in the real world”. Methods to evaluate this metric usually rely on time since the last update [36, 159]. Camboim [36] use the number of editors to determine temporal accuracy considering that the more editors working in a given region, the higher the likelihood of accurate mapping over time.

Lineage regard to the history of the features of the dataset [79]. The collection method, the software used, the source of the data, and how the feature evolved are characteristics that affect the lineage [90].

External quality metrics measures to which degree the spatial database addresses the needs and requirements of the users (fitness-for-use). External quality represents the user perspective while internal quality represents the perspective of the data producer [8].

As the users’ needs can vary significantly among applications there is no absolute metric to measure external quality. It is usually a combination of internal quality metrics that can be used to define if a dataset has “enough quality” for a specific use. For example, a road dataset with only major roads may not be of much use for an urban driver, but it can be detailed enough for a driver on a long trip.

Chapter 3

Methodology[1]

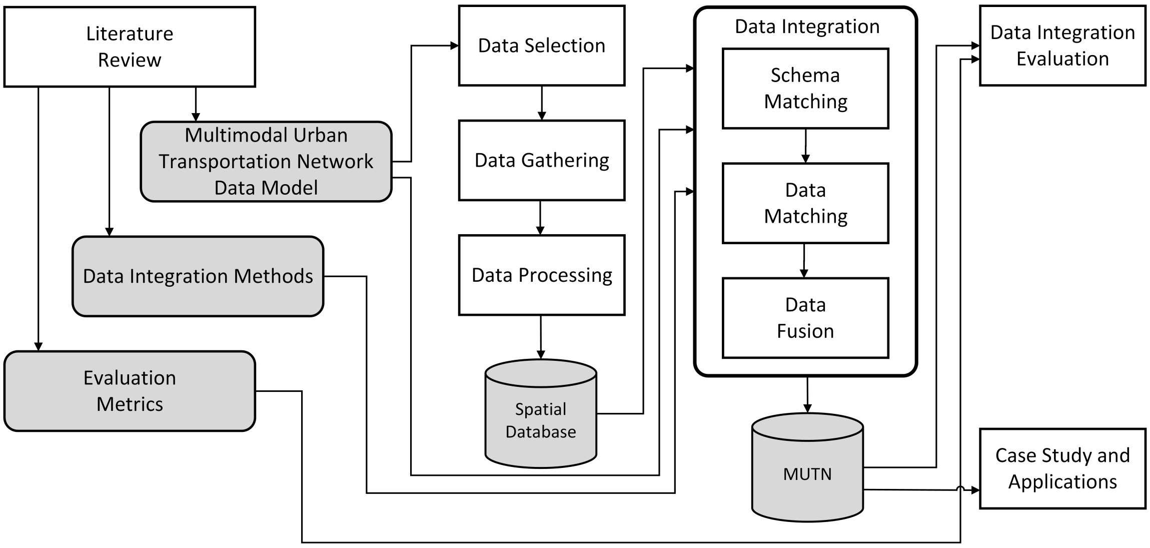
The focus of this work is the development of methods and techniques to integrate spatial data from heterogeneous and, possibly, unstructured data sources with an emphasis on multimodal urban transport networks (MUTN). The study proposes a MUTN data model and then builds it by integrating different data sources. The transport network plays an important role in city dynamics and people's life, hence it is essential to have detailed and updated data about its structure. It is also important to be able to enrich it with thematic data from heterogeneous sources, enabling analyses that would be hard or impossible to make without it.

Figure 3.1 shows the structure of this research. The literature review has three main outputs: a multimodal urban transport data model; identification of relevant data integration methods, with a focus on spatial data; and identification of evaluation metrics used to assess the quality of spatial data and the results of data integration methods. The data selection task relates to the selection of data sources relevant to build and enriching the multimodal transport data model. Data Gathering involves the acquisition of the selected data and the Data Processing task involves cleaning and preparation of data to be stored in a spatial database. Once data is appropriately selected and stored, data integration tasks take place. The Schema Matching task consists in integrating all data using the same schema. Data Matching identifies correspondences among the objects of different data sources, and the Data Fusion task consolidates the multiple versions of data in a multimodal urban transport network. The results of data integration tasks are evaluated and then applied to implement the case study, which is analyzed and evaluated.

Although Figure 3.1 implies a linear sequence of tasks, the process is, in fact, iterative and some tasks can be revised based on facts learned in any phase. For instance, the data processing and data integration tasks may cause a revision of the initial multimodal urban transport data model to tackle some newly discovered requirements.

[1] Parts of this chapter are based on and extend the works in [212]

Figure 3.1: Diagram of the research design. White boxes are tasks. Grey boxes are outputs.



Source: Made by the author.

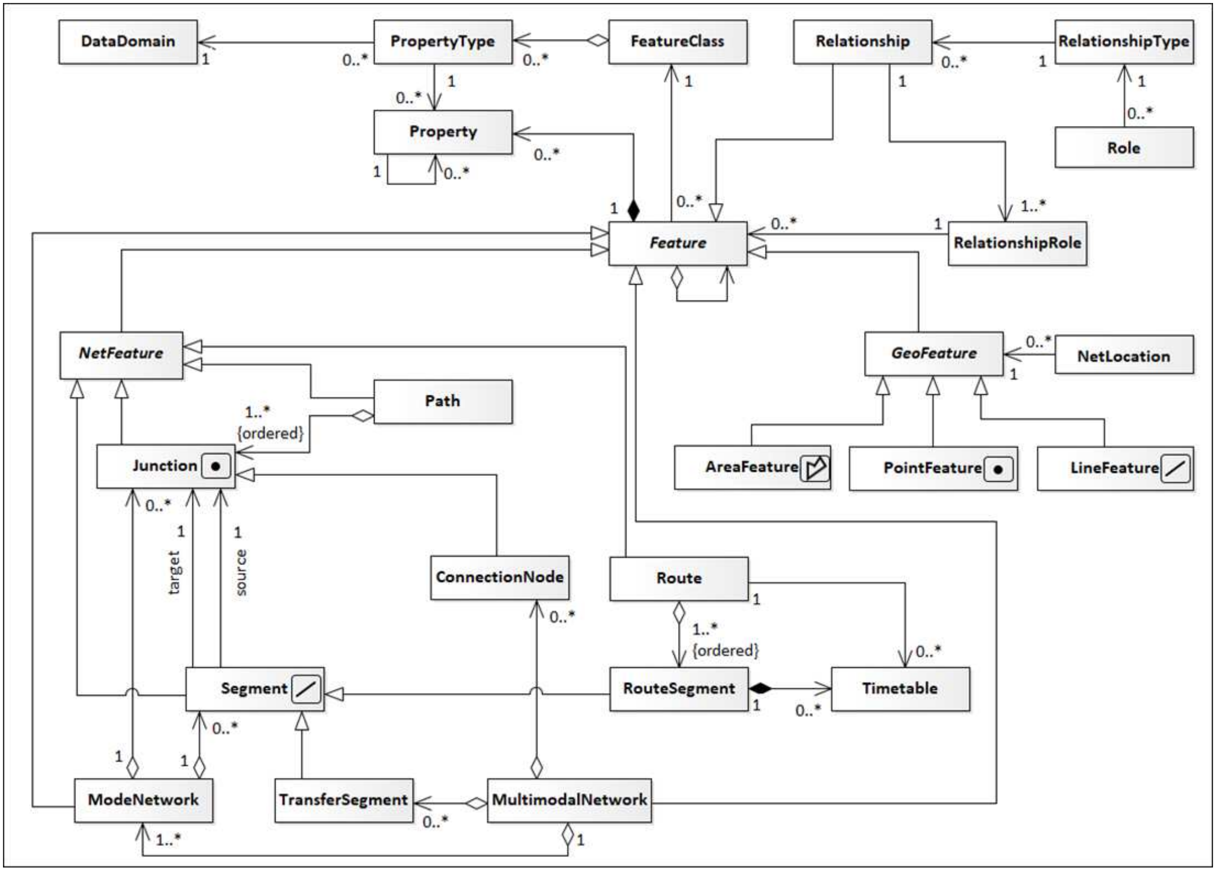
3.1 Multimodal Urban Transportation Network Data Model

The Multimodal Urban Transportation Network (MUTN) model represents the integrated infrastructure of urban transport, considering individual and collective transportation modes. The individual mode comprises the infrastructure for private or shared vehicles (including taxis, rentals, car sharing, bicycles, and others) and pedestrians, while the collective transportation mode is responsible for public transit such as bus and metro systems. The difference between them is that the transit system typically follows a pre-established structure where routes, stops, and schedules are defined. Multiple agencies may be responsible for the management of public transit alternatives. The network for each mode of transport is represented geographically, using geospatial coordinates, and topologically, using directed graphs.

We introduce a conceptual schema (Figure 3.2) to be used as the basis for data integration, including schema matching, data matching, and data fusion. All source datasets must be matched and transformed as needed to fit the proposed schema. Next, we describe the proposed schema in detail.

The Property class stores attribute for each feature using a key-value schema, where the key is an instance of the PropertyType class that has a name and a domain, given through the DataDomain class. In turn, the DataDomain class has a name, a data type, and a unit (e.g., km/h, meters, seconds, and other measurement units) for the

Figure 3.2: Conceptual schema for the Multimodal Urban Transportation Network in UML notation. Attributes were omitted for readability.



Source: Made by the author.

interpretation of values associated with the domain.

The main building block of the MUTN model is the abstract Feature class. A feature represents a real-world object or a relationship among features. It must have a unique identification (fid) and belong to a FeatureClass. Features may have a set of properties. The FeatureClass contains all possible feature types the data model can use and stores information about the properties of each feature class. A Feature can be specialized as a Relationship, a GeoFeature, a NetFeature, a ModeNetwork, or a MultimodalNetwork.

In many situations in modeling, we need to establish relationships between several features so that each one can play a role in a relationship with others. The Relationship, RelatioshipRole, RelationshipType, and Role classes are used in these situations. RelationshipType and Role define relationships for each one of the possible roles a feature can assume. For example, consider a forbidden conversion constraint between s_1 and s_2 segments passing through road junction j_1. This constraint can be modeled as follows: There must be a 'no_turn' RelationshipType associated with 'from', 'via', and 'to' roles, a new instance of the Relationship class with type 'no_turn', and three new instances of the RelationshipRole class are created for the segment s_1, junction j_1, and segment s_2 in the roles ‘from’, ‘via’, and ‘to’, respectively.

The abstract class NetFeature represents features that relate to others in topological structures to form networks. A NetFeature can be a Junction, a Segment, a Path, or a Route. A Junction corresponds to a network node, but with a geographic representation. The Path class is used to represent a path through the transportation network using an ordered sequence of Junctions. The Route class is used to represent a collective transportation service with a fixed schedule, for example, a bus or subway line. Junction and Segment classes are the basis for establishing network structures as the ModeNetwork class. In the proposed data model, the networks are modeled as directed graphs. From graph theory, a directed graph is defined as an ordered pair $G = (V, E)$, where V is a set of vertices, and E is a set of edges defined as ordered pairs of vertices. In the MUTN data model, a ModeNetwork represents the network for one mode of transport as a directed graph in which the vertices and edges are Junctions and Segments, respectively.

Every Segment starts and ends at a Junction whose identifiers are stored in the segment as its ’source’ and ’target’ attributes. The direction of the flow through the segment is always from source to target. There are other mandatory attributes for segments besides source and target, such as length, orientation, and cost. The length represents the size of the segment geometry in meters. The orientation attribute is the direction angle of the segment, considering East as 0, North as 90, West as 180, and South as 270 degrees. The cost attribute is used for routing calculations. The default value is to store the time in seconds to traverse the segment. A segment can be specialized as TransferSegment or RouteSegment. The former is used to represent segments representing intra- and inter-modal transfers. The latter is used to represent routes in collective transportation networks where there are defined departure and arrival times for a given service. The geometry attribute for TransferSegment and RouteSegment class segments may not precisely represent the real-world path. For instance, sometimes the exact path taken by a bus is not known, but it is possible to determine the sequence, position, and interval between its stops on a route (a common situation in General Transit Feed Specification (GTFS) files, as the path, is optional). In this case, a RouteSegment represents the link between each stop on the route and has an associated timetable that stores information about the arrival and departure time of each transport service that uses the segment. An isRealGeometry attribute can be checked to determine if the RouteSegment’s geometry represents the real path or just the transition between the stops.

Each Junction has a point geometry. A Junction represents an intersection between segments in the network. However, a ConnectionNode represents a point where it is possible to transition between different transportation networks or between different services within the same network, for example, a connection between different bus lines. A Junction can be of the type of intersection, station, or transfer. A ConnectionNode can be of the busStop, subwayStation, lightrailStation, railwayStation, parkingLot, parkAndRide, airport, intercityBusStation type. The origin and destination Junction types of a segment determine its type. For example, suppose both the source and target junctions are of the intersection type. In that case, the segment will be of the default textitSegment type. If one is of the textitintersection type and the other is of the transfer or station type, it denotes a segment of the OuterTransfer type, indicating that there will be a change in the mode of transport. Segments between two junctions of station type can be either RouteSegment or InterTransfer; that is, the bus user when arriving at a station, can continue on the same bus line or change to another line.

To represent elements not necessarily associated directly with the transportation network, the classes PointFeature, LineFeature, and AreaFeature can be used. For example, a city boundary or a lake can be AreaFeature instances. A river can be modeled as a LineFeature. Trees, lamp posts, traffic lights, and accidents can be represented as a PointFeature. Although they do not necessarily need to be connected to the transport network, it is often necessary to assign a network location to some GeoFeature. For example, the geometry assigned to record a traffic accident may not match a Junction or Segment. In this case, GeoFeatures may have a NetLocation attribute that assigns to them a location on the transport network based on its elements. The position can be related to a Junction or a Segment. In the case of the Junction, the location coincides with the position of the junction since the representation is a point. In the case of a Segment, the assigned location can be either a point or a line. If the NetLocation value references a Segment of the network, a start position and, optionally, an end position must be provided. This location is recorded as a position along the Segment line, using a value between 0 (start position) and 1 (end position). For example, on a segment with 100 m, a start position with a value of 0.1 and an end position of 0.5 indicates that the GeoFeature is located from 10m until 50m, measured from the segment origin, along its line geometry. If no end position is informed, it is assumed the location is a point along the segment given by the start position.

Finally, the MultimodalNetwork class is used to combine several ModeNetworks, using TransferSegments and ConnectionNodes to integrate all modes into a single network. Each ModeNetwork stores the data for one mode of transport. A transition between modes of transport occurs at a ConnectionNode, which is linked to ModeNetwork via TransferSegment. Each ConnectionNode contains incoming (fromMode) and outgoing (toMode) transport mode information. A ConnectionNode has an associated cost for transport mode transition. In this way, one can assign the cost of an intra- or intermodal switch. For example, a driver (ModeNetwork; mode = DRIVE) can leave their car in a parking lot (ConnectionNode;fromMode = DRIVE;toMode = WALK) and walk (ModeNetwork; mode = WALK) the rest of the way. The average time to park the car can be considered a cost of changing the mode of transport.

3.2 Data Selection

The *Data selection* tasks select candidate datasets from various data sources that can be used to build and enrich the MUTN. It is desirable that the selected databases be publicly available through direct download or API requests, as in the case of crowdsourced data. Using open data facilitates the reproducibility of this work by other researchers and for other geographical regions.

The data sources must cover various aspects that can impact urban transportation or other urban-related activities. Among these aspects are transportation infrastructure, transportation services, land use, administration, socioeconomic aspects, and environmental conditions. Transport infrastructure aspects concern using the transport network through different modes (car, bike, pedestrian, bus, metro) and their combinations, and related information such as bus stops, metro stations, timetables, and routes used by public transport. Land use aspects are important for the analysis of the built environment and activities in which people engage in a certain area. Administrative aspects include the political boundaries and administrative divisions (municipal boundaries, neighborhoods, units of planning, census sectors). Socioeconomic aspects can often be assessed from census data (gender, income, education, population) and other sources. Finally, environmental aspects involve temperature records, rainfall, wind, air quality, or other climatic phenomena.

Data to cover these aspects can usually be found in existing Spatial Data Infrastructures (SDI), Open Data portals, city websites, transit agency websites, statistical and Census organizations, crowdsourcing applications, and Location Based Social Networks. For this work, besides the available official data from government agencies, we used LBSNs (Section 2.3.3) and OpenStreetMap (Section 2.3.2). LBSNs such as Foursquare, Facebook Places, Google Places, and Yelp have many similarities in how they work but their data are not the same. Integration of data among them can lead to a representative dataset of points of interest related to different themes, like health, transport, food supply, education, and sports. OSM can provide a large amount of transport infrastructure data, and also data to cover aspects, such as land use and administrative limits. The next section details relevant aspects of OSM data to build the MUTN.

3.2.1 OpenStreetMap

The OSM data is beneficial for the construction of the MUTN road network. Its global coverage and the open nature of its data make it easy to use for any location on the planet. Questions arise regarding the quality of the data. In urban areas, this issue is not so problematic as OSM data is often more complete and up-to-date than available official data [69].

The main tag involved in the structure of the road network is *highway.* In OSM you can find different values for highways that can determine whether the feature represents a street, bus stop, or subway station. Table 3.1 shows values found in OSM data using the highway as a key to represent streets and indicates the preferred mode of transport associated with the feature. Other tags can modify this behavior by adding or restricting the use of the modes. For example, in *living_street* pedestrians and cyclists have the preference of use, but vehicles at a very low speed can be allowed.

The metro system uses a different set of tags. The key *railway* is similar to *highway* but has fewer possible values.

Transit stops appear in OSM as nodes. There is a feature, approved on April, 20 2010[2] that consists in the use of the key *public_transport* and the values: stop_position, platform, station, and stop_area. Table 3.2 shows combinations of tags for public transportation in OSM. Each line of the table represents a combination. For example, the third line means that there are objects using *highway=bus_stop* and *public_transport=platform*, simultaneously.

Besides this feature, the most popular use of the tag to identify bus stops is *highway=bus_stop.* There are objects with tags using keys *highway* and *public_transportation* simultaneously, and again the *highway=bus_stop* is the most popular. There are two possible reasons for this. First, the use of it is more traditional, so users may not be aware of the new tags. Second, the new tags of the proposal are not rendered on the main page of OSM, hence the redundancy of the tag attribution. The same observations apply to those elements with the tag *subway=yes* and *public_transportation=stop_position.*

Bus routes are mapped to relations in OSM. The route relation is identified by a tag *type=route*, and another tag specifies the mode. For bus routes, the tag is *route=bus*, and for the metro is *route=subway.* Both keys, type, and route are mandatory. There are several other keys to describe the route, such as the operator (the company that operates the route), distance (the official distance of the route), duration (official duration of the route), to (destination station), from (origin station), and others. Unfortunately, OSM has rules against the insertion of time-dependent data such as timetable information

[2] https://wiki.openstreetmap.org/wiki/Proposed_features/Public_Transport. Accessed on April 13, 2022.

Table 3.1: Different values for key *highway* in OSM to represent streets and respective preferred modes of transport (cars, pedestrian, bicycle)

Values for highway	Mode of Transport		
	Car	Ped.	Bic.
residential	X		
tertiary	X		
secondary	X		
service	X		
unclassified	X		
trunk	X		
primary	X		
motorway	X		
footway		X	
track	X		
trunk_link	X		
motorway_link	X		
cycleway			X
living_street	X	X	
secondary_link	X		
tertiary_link	X		
primary_link	X		
pedestrian		X	
path			
steps		X	
construction	X		
services	X		

Table 3.2: Combinations of key-value pairs used for public transportation stops in OSM.

highway	public_transport	railway	subway
bus_stop			
	stop_position		
bus_stop	platform		
		buffer_stop	
		level_crossing	
	stop_position		yes
bus_stop	stop_position		
		station	
		switch	
	stop_position	level_crossing	
	station	station	yes

OSM. The data for public transportation stops and routes are only used for basic routing operations in OSM. To overcome this flaw, one could look if there are GTFS files available for the region to find complete and official data from transit agencies.

3.3 Data Gathering

The *Data gathering* task involves the acquisition of data. In this work, the data will be gathered using direct download or using API requests and should be freely available. One of the primary data sources that can be used is OpenStreetMap (See Section 2.3.2). Its scope and variety allow data collection that can be used to compose the main components for the MUTN, such as the road network and other points of interest such as parking lots, subway stations, and bus stops. Besides the OSM, different data sources can be used to compose the MUTN, such as city open data portals, transit agencies, Census data, and Location-Based Social Networks (e.g., Foursquare, Yelp, Google Places, Facebook Places).

3.3.1 Location-Based Social Networks

Location-Based Social Networks (LBSN), such as Yelp, Facebook Places, Google Places, and Foursquare, have many similarities in how they work but their data are not the same. Integration of data among them can lead to a representative dataset of points of interest related to different themes, like health, transport, food supply, education, and sports.

The process used to gather data from LBSN work is as follows. First, a regular grid of points separated by 25 meters is generated for the area of interest. For each point, a call to the LBSN's API is made using the point's coordinates and a radius of 25 meters as parameters. The radius size of 25 meters was considered small enough to get all the available places for each social network since a larger radius might cause the APIs to return only a subset of the available places instead of the full list due to limitations of the maximum number of features for each LBSN API call. Collecting the data in this way allows an overlap in the area considered in each API call.

For this work, we implemented crawlers for each LBSN using Python programming language. Usually, there are some limits to API usage and the crawler must be aware of them. For example, Yelp's API allows 25,000 search requests per day and Foursquare allows 5,000 requests per hour. The API returns JSON data that were stored and then processed to extract, for each location retrieved, its ID, latitude, longitude, and category.

Due to the collection strategy, overlap occurred and many duplicate entries is recorded for each LBSN. They can be easily detected and deleted using the *id* attribute as a key to finding duplicates.

3.4 Data Cleaning

The *Data processing* task is responsible for data cleaning and transformation. The cleaning task is used to eliminate noise and redundancies in data. After cleaning, data is transformed (if needed) to be stored using spatial database technology, and appropriate geometry types of points, lines, and polygons are reprojected using a standard Coordinate Reference System (CRS) for all data. For this work, we use PostGIS, a spatial extension of the PostgreSQL database management system to store all the MUTN data. This choice is justified by its technical features and advanced tools, as well as the fact that it is open-source software, its compliance with the standards defined by the Open Geospatial Consortium (OGC), a large user community, and is natively supported by all the GIS tools used in this work (e.g. QGIS, osmosis).

3.5 Building the Multimodal Urban Transportation Network

The first step to building a MUTN is the creation of a street network, which is used by pedestrians, bicycles, and vehicles. This network is also where the components of collective transportation infrastructure are connected, and other GeoFeatures can be located. Our approach is to build the street network using data from different sources to get a more complete and up-to-date dataset, to use as the basis to integrate data from public transport and other Geofeatures. An overview of the steps for building the multimodal network is shown in Figure 3.3. The remainder of this section presents each process in detail.

3.5.1 Initial Definitions

The following definitions are used in the description of the process:

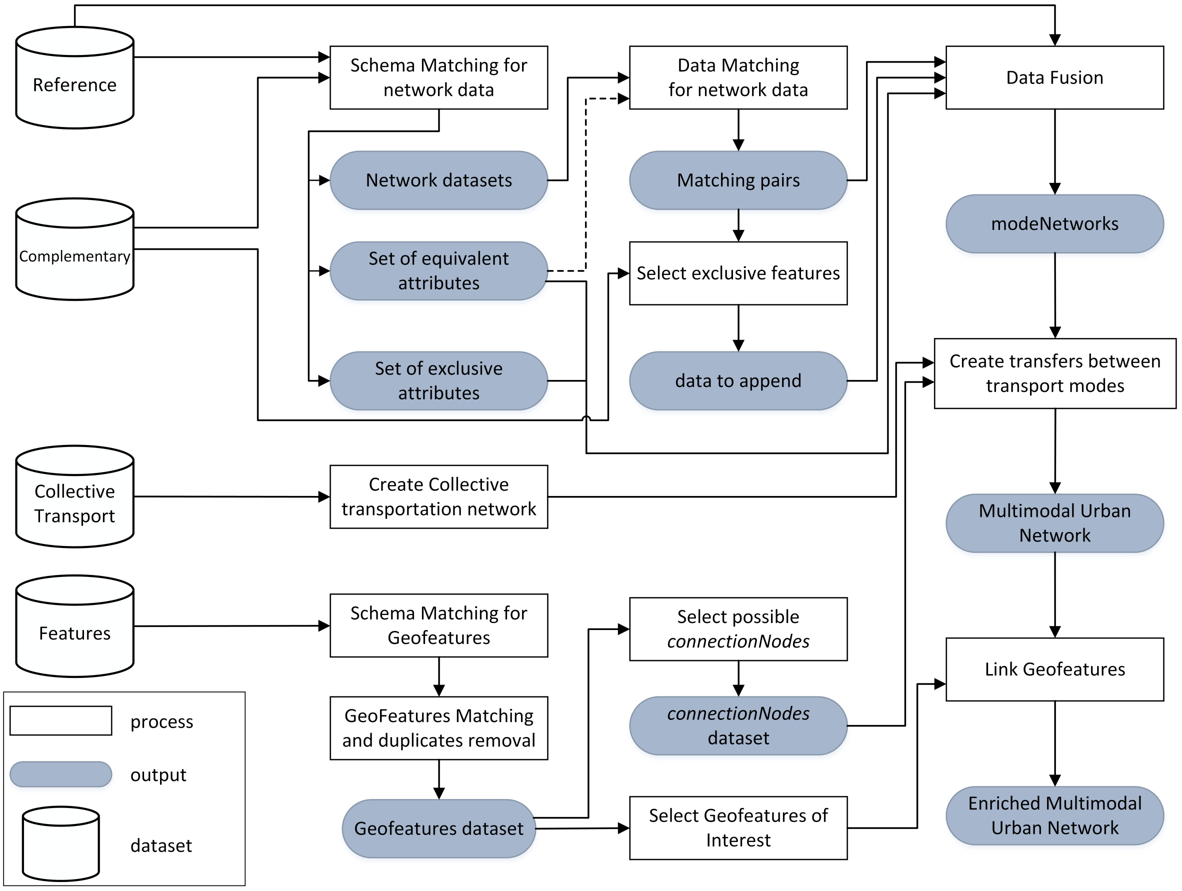
- Reference Dataset: This dataset is the basis for the integration process and construction of the multimodal transportation network. It follows the proposed conceptual schema and is the dataset whose data will be given preference when resolving data conflicts in data integration. Usually, but not necessarily, it should be an authoritative dataset.
- Complementary Dataset: contains data that can complement, expand, correct, or update the Reference Dataset.
- Collective Transportation Data: data related to routes, stops, and schedules of collective transportation infrastructure available at the same region of the Reference and Complementary datasets. The most common sources are GTFS files.
- Features Dataset: various datasets that can be used to enrich the resulting multimodal transportation network to enable its use in urban computing applications. This dataset provides features that are related to transportation mode transfer, such as parking lots or car-sharing points, to enable multimodal routing.

3.5.2 Schema Matching for the Reference and Complementary Datasets

The MUTN schema proposed in this work establishes that a transportation network is represented as a directed graph. The first step of the work is to transform the reference and complementary datasets into a uniform graph representation, following the proposed schema. In the resulting network, each segment must begin and end in a junction. There must be a junction at every segment intersection if the transition from one segment to the other is possible. For example, in a street network, a road intersection must be a junction, but the point where a road (segment) intersects a tunnel or a bridge cannot be a junction since the transition is not possible.

Every junction must represent an intersection or a dead-end to match the MUTN schema. A cleanup operation should identify useless junctions, i.e., pass-through nodes

Figure 3.3: An overview of the steps for building the multimodal network.


Reference
Complementary
Collective Transport
Features
Schema Matching for network data
Network datasets
Set of equivalent attributes
Set of exclusive attributes
Data Matching for network data
Matching pairs
Select exclusive features
data to append
Data Fusion
modeNetworks
Create transfers between transport modes
Multimodal Urban Network
Create Collective transportation network
Schema Matching for Geofeatures
GeoFeatures Matching and duplicates removal
Geofeatures dataset
Select possible connectionNodes
connectionNodes dataset
Select Geofeatures of Interest
Link Geofeatures
Enriched Multimodal Urban Network
process
output
dataset

Source: Made by the author.

that can be removed without altering the network's topology. When such nodes are eliminated, the neighboring segments are geometrically merged. This operation can only be performed if the attributes of the neighboring segments are compatible. A set of attributes is considered compatible if it differs only in the values that relate to the geometry of the edge (e.g., length).

After the simplification process, two new properties are added (or updated) to the datasets, the length and the orientation of each segment. The length is the size of the segment's geometry, in meters. The orientation of the segment is the angle, in degrees, from the source junction to the target junction considering east = 0, north = 90, west = 180, and south = 270 degrees.

It must be possible to identify the mode (or modes) of transport for the segments in all datasets. Usually, this information is stored as an attribute, or else the entire dataset relates to a single transport mode.

Each dataset can have an arbitrary number of attributes for both segments and junctions. We opted to make manual matches in the case study, but existing semantic schema matching techniques can be used [168, 80, 162].

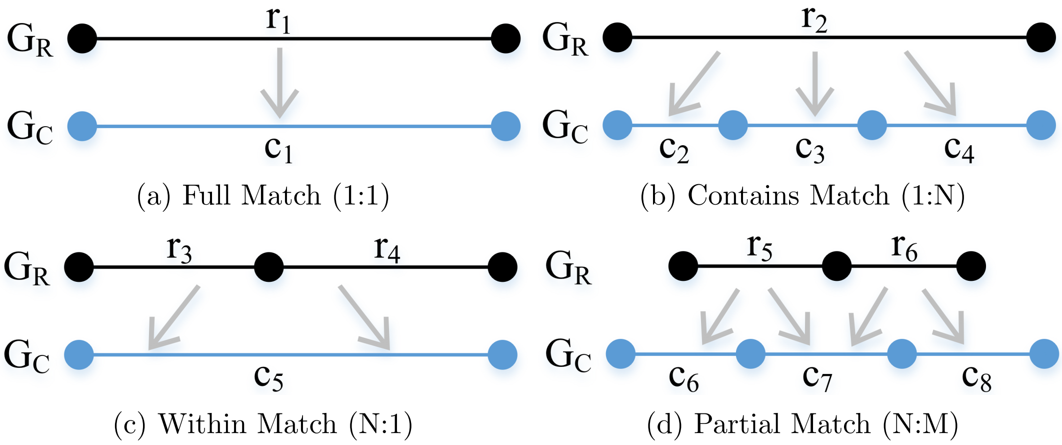
Finally, the last step is to transform all geometries to use the same coordinate reference system (CRS). The result of the schema matching is the graphs, G_R, and G_C, representing the reference and complementary datasets, respectively, with their attributes mapped to properties from the MUTN data model. The exclusive attributes from the complementary dataset are kept to be used, if necessary, in the data fusion process. The common attributes can be used in the data-matching step to improve matching results by confirming or rejecting matching pairs based on available semantic information.

3.5.3 Data Matching for Network Data

The data-matching process works by finding matching pairs with increasing cardinality. We defined four cardinalities for the matching pairs: full, contains, within, and partial. Figure 3.4 shows in a simplified way the possible cardinalities for matching pairs. A fifth category, called null (one-to-zero cardinality), is used for features that have no match in the other dataset. This category is of fundamental importance for complementary data fusion, allowing one dataset to expand on the contents of the other to improve the completeness of the result. A full match (one-to-one cardinality), occurs when one segment from G_R has an exact counterpart in G_C and vice versa, which means that the source and target junctions of both segments are closer than a threshold and both geometries are similar. In Figure 3.4a, the segment r_1 from G_R has a full match with segment c_1 from G_C. A contains match occurs when one segment of G_C has the projections of its source and target junctions located at the same segment in G_R. In Figure 3.4b, the segment r_2 from G_R has a contains match (one-to-many cardinality) with segments c_2, c_3 e c_4. A within match (many-to-one cardinality) is symmetrical to the contains match. It occurs when one segment in G_R has the projection of its source and target junctions located in the same segment in G_C. Figure 3.4c shows that segment r_3 and r_4 from G_R have within match with c_5 from G_C. A partial match (many-to-many cardinality) happens when the source and target junctions of a segment from G_R have its projections in different segments in G_C and correspondent segments in G_C also cannot be related to one single segment in G_R. In Figure 3.4d segment, r_5 has a partial match with segments c_6 and c_7 from G_C.

The data matching process starts with a list of all possible candidate matching pairs (L_{MP}) from G_R and G_C. Next, L_{MP} is analyzed to find full matching pairs, then contains and within matches, and finally, the remaining non-matched edges are tested to find partial matching. If semantic information is available, an additional procedure can be triggered to check the reliability of the matching pairs found and to seek other possible matches in the non-matched edges.

Figure 3.4: Cardinalities of matching pairs.



(a) Full Match (1:1)

(b) Contains Match (1:N)

(c) Within Match (N:1)

(d) Partial Match (N:M)

Source: Made by the author.

3.5.3.1 Building the Set of Candidates for Matching

The first step in the process to find the list of all candidates for matching (L_{PM}) is to build an R-Tree-based spatial index to accelerate the process. The index is created for the segments in G_C. Then, we search for nearby segments in G_R. Each segment in G_R is buffered and used to search the index for segments in G_C that intersect the segment's buffer. The size of the buffer depends on the road density of the region considered. Also, the larger the buffer size, the higher the computational cost of the process because more candidates are expected to be found. Typically values up to 20 meters achieve a good balance for creating the list of all candidates and the computational cost.

All segments from G_C that intersect the buffer are inserted in L_{PM} along with the counterparts in G_R as candidate matching pairs, with the following metrics: the difference between the segment orientations (in degrees) (D_b), the distance between the source junction of both segments (D_{uu}), the distance between the target junctions of both segments (D_{vv}), the distance between the source junction from G_R and the target segment from G_C (D_{uv}), the distance between the target junction from G_R and the source junction from G_C (D_{vu}), a flag indicating if the buffer of the segment from G_R contains the candidate segment from G_C (B_{GT}) and the length difference ratio (L_{dr}). All segments for which no candidate matching is found (a null matching) are marked as exclusive to the particular dataset and are not considered in the next matching steps, but they can be used in the data fusion process.

In the next steps, some metrics are calculated to guide the matching process. They are Node Proximity, Length Similarity, and Angle Similarity.

Node Proximity The node proximity is used to verify if the source and destination junctions of a segment r are close enough to the source and destination junctions of a segment c, considering distance tolerance, t_d. It is defined as:

$$P_{sim}(n_1, n_2) = \frac{dist(n_1, n_2)}{t_d} \tag{3.1}$$

where n_1 and n_2 are junctions in the transportation network; $dist(n_1, n_2)$ is a function to calculate the distance between the two junctions, for example, Euclidean distance, and t_d is the maximum distance to consider the two junctions as a possible match. There is no fixed value for t_d, as it depends on both datasets' positional accuracy. For example, if both datasets have high positional accuracy, a threshold of 5 or 10 m can be used to determine if a junction is close enough to the other. If the accuracy is low, it may be necessary to use a higher tolerance.

Length Similarity The similarity by length considers that merely defining a tolerance based on a ratio of the difference in lengths is not appropriate. For example, if a segment r_1 is 20 m long and a segment c_1 is 16 m long they may match, even with a 20% difference in length between them. However, if r_1 is 1000 m long and c_1 is 800 m long, possibly a 200 m difference is too high to consider them a match. The same principle applies if we only consider an absolute value for the difference. Suppose a difference of up to 40 m is used to consider two segments similar in length. In this case, a r_1 edge with 10 m and a c_1 edge with 50 m would be considered a match, which is not desirable. This way, lower and higher absolute limits for the difference in length are defined, while intermediate values depend on the length difference ratio between the segments. The length similarity is defined as:

$$L_{sim} = \frac{|l_r - l_c|}{\min(t_{lmax}, \max(t_{med}, t_{lmin}))} \tag{3.2}$$

where

$$t_{med} = \max(l_r, l_c) \times t_{ratio} \tag{3.3}$$

and l_r and l_c are the lengths of segments r and c, respectively. The t_{lmax} and t_{lmin} are the maximum and minimum absolute distance tolerance value, respectively; and t_{ratio} is the tolerance value, in terms of the ratio between l_r and l_c.

Angle Similarity The angle similarity establishes if the difference of orientation angle of segments r and c is smaller than a threshold. It is defined as:

$$B_{sim} = \frac{|D_b|}{t_{bmax}} \tag{3.4}$$

where D_b is the angle between segments r and c, and t_{bmax} is the threshold difference (in degrees) to consider the orientation angle of both segments to be similar. For example, a t_{bmax} of 15 degrees means that segments with angle differences up to 15 degrees are considered similar in orientation angle.

3.5.3.2 Finding Matching Pairs

The process of finding matching pairs works iteratively, searching for matches according to their cardinality. First, full matches are searched, then the contains and within matches, and finally the partial matches. Matching results are stored in hash lists keyed by the segment or junction ID for efficient retrieval.

A full matching occurs when one segment r in G_R, with r_s and r_t as the source and target junctions, respectively, corresponds to exactly one segment c in G_C, with c_s and c_t as the source and target junctions, respectively. The candidates list L_{PM} is used to find full matching pairs, which are identified by checking if the values for length, angle similarity, and node proximity, $P_{sim}(r_s, c_s)$ and $P_{sim}(r_t, c_t)$, are all less than or equal to one. The segments that satisfy this criterion are marked as full matching. If a segment r has more than one candidate segment in G_C for full matching, the one with the largest name similarity is chosen. In the case of a new tie, the candidate segment with the shortest distance is chosen. The candidate segments not chosen are available for new matching.

If a candidate pair fails the full matching test, the verification for the contains and within matching types occurs. A contains matching is established when one segment r from G_R corresponds to one or multiple segments from G_C, and these segments in G_C entirely fit the geometry of r, so we can say that r contains the segments from G_C. A segment pair (r, c), where $r \in G_R$ and $c \in G_C$, is a contains match if r strictly contains c, and the segment in r that corresponds to c (the projection of c in r) has L_{sim} and B_{sim} less than or equal to one. We defined that segment r strictly contains c if c_s and c_t have a valid projection in r, and, if the projection of c_s in r is equal to r_s, then $P(r_s, c_s)$ must be less than one, and, if the projection of c_t in r is equal to r_t, then $P(r_t, c_t)$ must less than one. An edge can have a within relation with only one other segment. When multiple candidates appear, the pair with the smallest distance is selected.

To find partial matches, we check if only one of the junctions of a segment c, c_s, or c_t, has a projection inside segment r. Considering r' as the part of r representing the projection of c in r, and c' the part in c representing the projection of r in c, if r' and c' have L_{sim} and B_{sim} less than than one, then r and c partially match each other.

3.5.3.3 Selection of Exclusive Features from the Complementary Dataset

After the matching process, the G_C features that had no match (null matching) in G_R are analyzed for a possible data fusion operation with G_R. This operation is also called conflation in the literature [197, 227, 247, 250, 69]. Merging one dataset's exclusive data into another allows for complementing the data in the reference dataset and improving its coverage and completeness. The next section details the fusion process.

3.5.4 Data Fusion for Network Data

In this stage, the data fusion occurs in two ways: redundant and complementary. In redundant data fusion, the matched features can have their attribute values updated. For example, if two road segments are matched, the value for a name attribute of one feature can be used to update the other. One problem that arises is how to define the attribute value of the feature resulting from the fusion of features. There is no single strategy, and cases may vary depending on the characteristics of the data sources and the purpose of the data fusion. When dealing with authoritative and crowdsourced data, the default strategy is to use the trust your friends' technique from the conflict resolution category (see the data conflict taxonomy in [24]) to give preference to the authoritative data source. If the value is not present in the authoritative data, a take-the-information strategy from the conflict avoidance category can be used to take the value available from other sources, when available.

The complementary data fusion techniques are used to complement a dataset with features from other datasets without correspondence (null matches). In this case, a fully automated process is complex and may be subject to errors that must be verified by humans. In this work, the complementary fusion at this stage is used in two situations: missing driving directions information and the inclusion of connected segments for which no match was found.

To detect missing driving directions, the road segments for which there are matching candidates that could not be matched are analyzed. If, for instance, there is a mismatch due to the angle similarity metric, and the angle difference is close to 180 degrees, then the segment is considered an erroneous driving direction, and a new segment is inserted.

The data fusion process to include sets of connected segments that did not match checks if there are any connections of previously matched segments to any segment from the set. If connections exist, they are inserted in the reference dataset and connected. Otherwise, the junctions in the set closer than a distance tolerance (t_d as default) from a junction or segment in the reference dataset are connected. The new segments created to connect the sets of segments receive a flag 'needs_review' to indicate they need further validation.

3.5.5 Creation of the Collective Transportation Network

The creation of a collective transportation network dataset has particularities that must be taken into account. First, unlike individual transportation networks, collective transportation routes are defined with a specific schedule. Second, the actual physical path taken by a vehicle in collective transportation is not always available; however, it is possible to collect data regarding the lines and their sequence of stops. A currently adopted standard for collective transportation information dissemination is GTFS files.

The proposed data model allows the building of a public transport network with pre-defined routes through ConnectionNode, Route, RouteSegment, and Timetable classes by mapping GTFS data to the proposed schema. A Route stands for a path through a sequence of collective transportation stops. Each stop is represented as a ConnectionNode as they allow a change in the mode of transport (WALK→BUS). The GTFS file allows the grouping of stops in stations. When building the collective transportation network for the proposed data model, the same station‘s stops are unified in the same ConnectionNode represented as a station. When leaving a route, the user can change the transport mode (BUS→WALK) or make a connection to another route (BUS→BUS). To enable inter and intra-modal routing, each ConnectionNode used by several routes is duplicated (one for each possible route), and TransferSegments of type InterTransfer are created to enable the assignment of a cost when a collective transportation user makes the connection. The connection of ConnectionNode to the individual transportation network is made according to the possible mode of transport. Generally, the collective transportation network will be connected to the pedestrian (street) network through TransferSegments of type OuterTransfer. For each RouteSegment, the corresponding Timetable is created containing the information of days and times of arrivals and departures of a vehicle traveling along a certain route.

3.5.6 GeoFeatures Matching and Duplicates Removal

Geofeatures can appear in the MUTN data model as points, lines, or polygons. The task of consolidating data from different sources for the features is complex. For example, Geofeatures represented as points have no geometric attributes that can identify duplicates beyond their position. Therefore, the use of semantics in the matching process is always necessary. Even so, the task remains hard to be fully automated because features have different sets of attributes, attributes that represent the same information appear with different names or data types, and attribute values may be in different languages, among other challenges related to automated schema matching.

To identify duplicates, the strategy is to compare PointFeatures close to each other at an arbitrary tolerance distance and with similar names (all GeoFeatures must have a value for the name property, null values are not allowed). The Levenshtein distance is a widely used similarity metric to compare names. However, its results are sensitive to the order in which the words appear in the strings, to punctuation, and to lowercase or uppercase letters. For example, a place p_1 named "Capitólio Estacionamento" and another p_2 named "Estacionamento Capitolio" has a normalized Levenshtein similarity of 0.58. Crowdsourced data has great variability in the attributes whose values the user can provide freely. To minimize this variability and improve the matching results, we preprocess the names before using the Levenshtein distance. First, the names are converted to lowercase characters, and the punctuation is eliminated by tokenizing the strings. The tokens are then sorted alphabetically and concatenated. Then the Levenshtein distance is calculated and normalized. The name similarity, N_{sim}, can be expressed as Equation (3.5):

$$N_{sim}(p_1, p_2) = \frac{(length(p_1.name') + length(p_2.name')) - levenshtein(p_1.name', p_2.name')}{length(p_1.name') + length(p_2.name')} \tag{3.5}$$

where $length$ is a function to return the number of characters of the string representing the name of the PointFeature, and $name'$ represents the processed $name$ of the feature after conversion to lowercase characters, tokenization, sorting, and concatenation. Applying N_{sim} to the previous example of p_1 and p_2 results in a value of 0.98.

PointFeatures p_1 and p_2, with the same FeatureClass that are close enough to each other and have similar names according to a given tolerance (t_{name}), are considered to be duplicated. If $P_{sim}(p_1, p_2)$ (Equation (3.1)) is less than than one, p_2 is automatically considered a duplicate. If not, those points that names have $N_{sim}(p_1, p_2)$ with a value of t_{name} or more are considered duplicates until a distance up to d_m (distance multiplier) times the t_d (distance tolerance). Formally, the $isDuplicate(p_1, p_2)$ function is defined as:

$$isDuplicate(p_1, p_2) = \begin{cases} true, & \text{if } (P_{sim}(p_1, p_2) \leq 1) \text{ or} \\ & (P_{sim}(p_1, p_2) \leq d_m \text{ and } N_{sim}(p_1, p_2) \geq t_{name}) \\ false, & \text{otherwise.} \end{cases} \tag{3.6}$$

3.5.7 Selection of ConnectionNodes

ConnectionNodes are selected from the GeoFeatures. To create the multimodal transportation network, we select GeoFeatures types that can be used to change the mode of transport. For example, parking lots can be used in the transition from car to rail transport mode, and vice versa.

3.5.8 Creation of Transfers between Transport Modes

The points to be used as ConnectionNodes are classified according to the mode of transport from which a transition in and out can occur. For each set, connections are created by looking for the Junction closest to the position of the ConnectionNode and creating a TransferSegment of type OuterTransfer. For example, a set of ConnectionNodes that will be used to transition from DRIVE to WALK will be connected to the DRIVE network via an incoming OuterTransfer and to the WALK network via an outgoing OuterTransfer segment.

3.5.9 Linking GeoFeatures to the Multimodal Urban Transportation Network

The MUTN data model allows us to store GeoFeatures for different applications. For those that are not directly related to routing, it is not necessary to create Junctions for them. Instead, the GeoFeatures are created, and the class NetLocation is used to store where in the transportation network a GeoFeature can be reached. This way, the MUTN data model is kept stable without excessive partitioning of the segments to create links to GeoFeatures.

3.6 Discussion

This chapter presented spatial data integration methods and a data model to store the results. The spatial data integration method is composed of schema-matching steps, data matching, and data fusion. In the schema-matching stage, datasets with different schemas and levels of detail are made compatible with the proposed data model. In the data matching stage, matching pairs are found in the datasets, with different cardinalities, full (one-to-one), contains and within (one-to-many), and partial (many-to-many). The segments that have no matching candidate are identified and marked as null matches. In the data fusion stage, such null matches can be incorporated into the integrated database, and attributes can be transferred and consolidated. Once the datasets were integrated, information regarding collective transportation and transitions between modes of transport were incorporated, also using data integration methods.

In Chapter 4, a study case was performed to test the methods on real-world data for the city of Belo Horizonte. Data from authoritative and crowdsourced datasets were integrated into a multimodal dataset containing information that allows performing multimodal routing and analysis in the urban environment.

Chapter 4

Creation of a Multimodal Urban Transportation Network for Belo Horizonte[1]

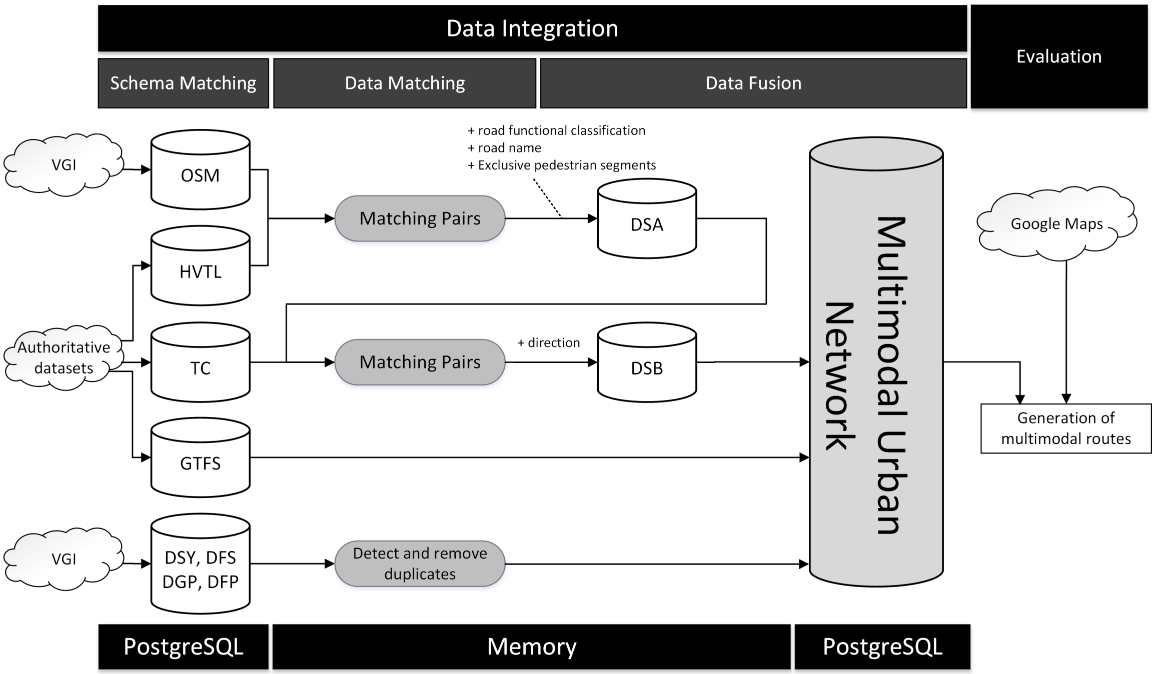
To test the framework's validity, a multimodal urban transportation network for the Brazilian city of Belo Horizonte was built. Data from different sources were used and integrated to create multimodal routes. Official (reference) and alternative (complementary) datasets were used. Datasets were considered official if their provider was an agency connected to the public administration. Otherwise, they were considered to be alternatives. First, the datasets' schemas were mapped to the MUTN proposed schema. Second, datasets were integrated using data matching and fusion techniques to build the individual transportation network dataset. Then, GTFS files were used to build the collective transportation network dataset, which in Belo Horizonte includes the bus and subway systems. Finally, data from additional and heterogeneous sources were integrated to establish *ConnectionNodes* between modes of transport.

The resulting multimodal urban transportation network was used to find routes among eighty points using *DRIVE*, *WALK*, and *TRANSIT* transportation modes. Each point represents the centroid of an aggregation of census sectors comprising a planning unit, as defined by the city's administration. The routes created were then compared against the equivalent Google Maps routes. The experiments were conducted on a laptop computer with an Intel Core i5-9300H processor, 1 TB hard disk, 20 GB RAM, PostgreSQL 11.7 (64-bit) with extensions PostGIS (3.0.1) and hstore (1.5) enabled. All the methods were implemented using the Python (3.8.5) language. Figure 4.1 shows an overview of the procedures executed in the case study. Acronyms used in the figure and a process description are given in the next subsections.

The following Sections describes official and alternative datasets used in the case study and explains how they were integrated to build the MUTN and Section 4.11 discusses the results.

[1] Parts of this chapter are based on and extend the works in [212]

Figure 4.1: Case study overview



Source: Made by the author.

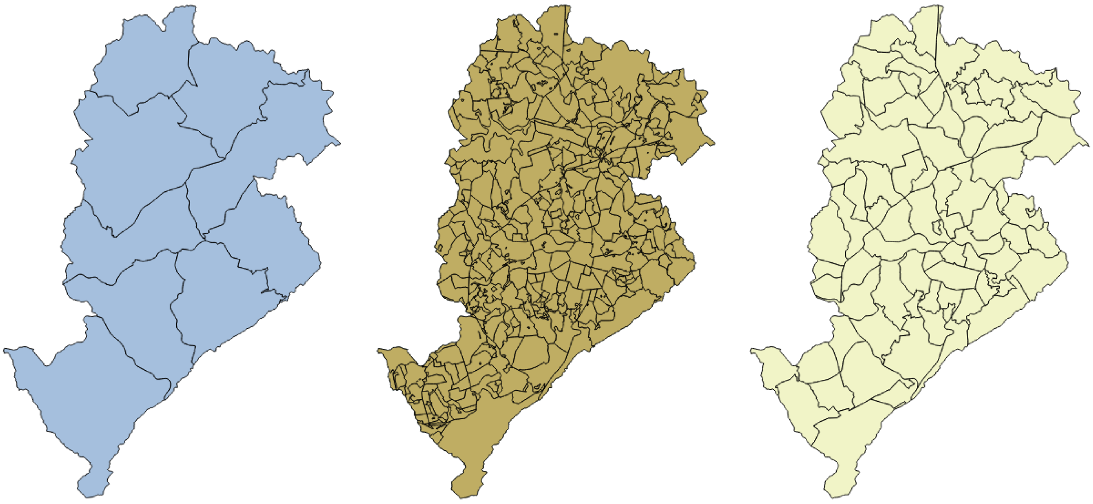
4.1 Official Datasets

BH's Open data portal content covers many aspects needed in this work, such as infrastructure, land use, and administration. For example, on the administrative aspect, BH is divided into nine administrative regions (Figure 4.2a), 487 popular neighborhoods (Figure 4.2b), and 80 units of planning (Figure 4.2c). The planning units are aggregations of Census sectors, so estimating Census data at that level of aggregation is straightforward. Some data are clearly outdated. For example, the shapefile representing the coverage area of health centers has attributes for the population from the 2000 Census, while the most recent one is from 2010.

The Infrastructure aspect can use streets, intersections, and address locations. Figure 4.3 shows the conceptual schema of the available data for transportation. It contains 231,535 streets with a total length of 9,044,490m. Figure 4.4 shows an overview of the streets and intersections of the entire city area and a zoomed view at Raul Soares Square, the central region of BH. The conversion rules are represented as link arcs (red ones in Figure 4.4c). If two streets are linked, it means the conversion is allowed. The data also includes 527,643 geolocated addresses (blue squares on Figure 4.4c) across the entire city.

Four official datasets were used. The first dataset, called "Classificação Viária" (HV), stores data about functional classification for each road segment. The second dataset, "Trecho Logradouro" (TL), contains the name of each road segment. These two

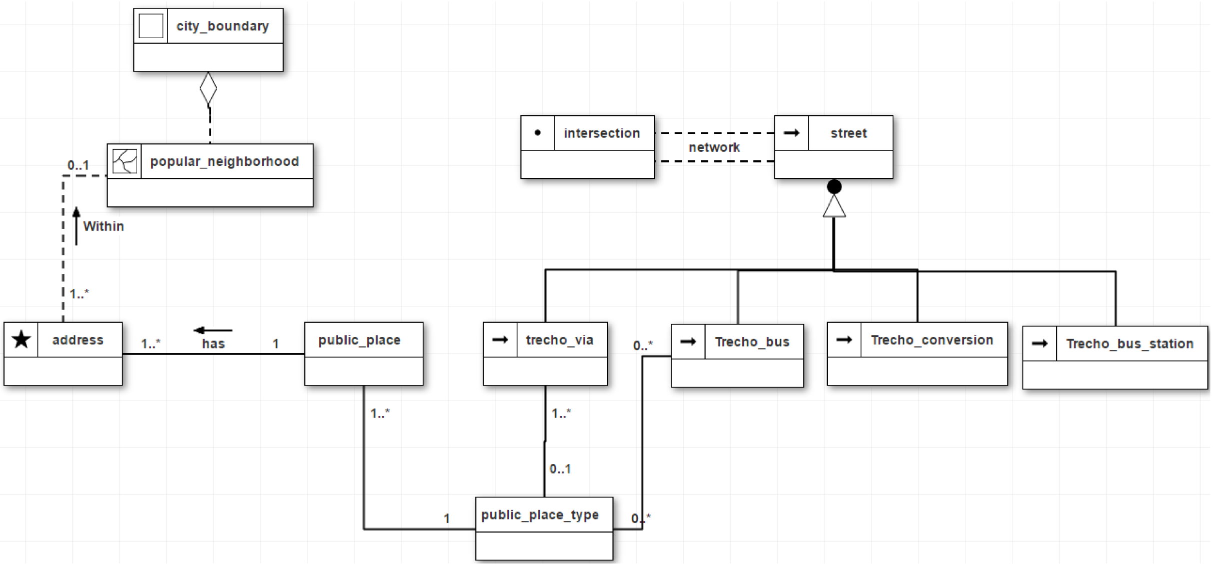
Figure 4.2: Examples of administrative boundaries for Belo Horizonte



(a) Administrative Regions (b) Popular Neighborhoods (c) Units of Planning

Source: Made by the author.

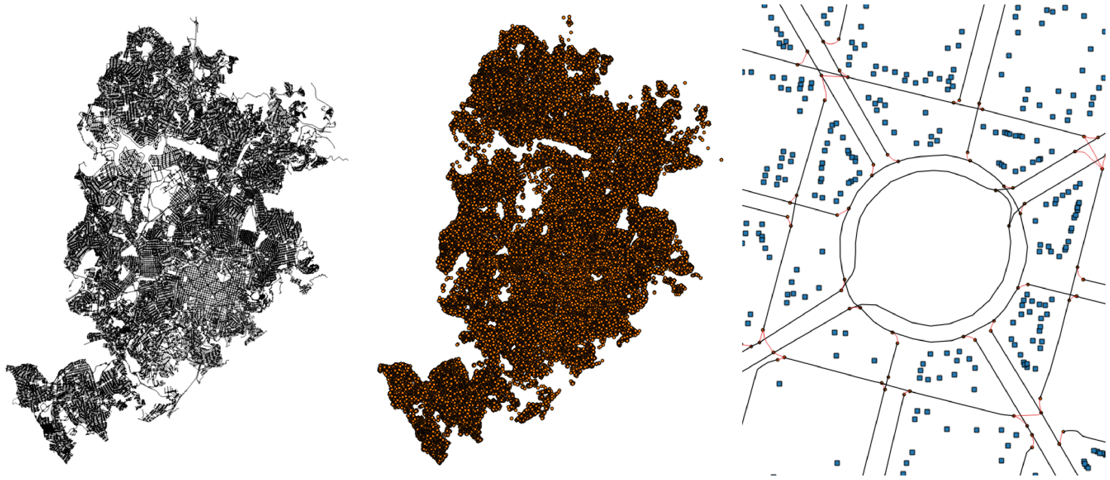
Figure 4.3: Conceptual schema (OMT-G notation) of administrative and infrastructure aspects available at BH's open data portal.



Source: Made by the author.

datasets have relational integrity constraints defined. Thus, it is straightforward to join the information of both datasets using relational database operations ($HVTL$). The third dataset, called "Circulação Viária" (TC), has data about the city street network. Each segment is related to an origin and a destination node. Street data corresponds to a directed graph using two edges to represent two-way streets, which causes many duplicate nodes at intersections used to represent turn permissions. There is no way to link a segment in TC to a segment in $HVTL$ only using attribute values, so it is necessary to

Figure 4.4: Data available from BH SDI. Streets (a) and intersections (b). Close-up image of Raul Soares Square, including points representing door addresses in (c).



(a) Streets (b) Intersections (c) Door address points

Source: Made by the author.

use spatial data matching operations to integrate the data from both datasets. All three datasets are part of Belo Horizonte's Spatial Data Infrastructure [2], created and managed by the city's administration.

The fourth dataset is the set of GTFS files provided by the city's traffic department, BHTrans[3]. The data used is from 29 July 2020, with 9,328 stops, 643 routes, 56,771 trips, and 3,202,454 timetable entries for all trips and stops. To get data about subway trips, we used GTFS files from 23 January 2019, the last issue to include subway information. It has 37 stops, 1 route, and 629 trips.

4.2 Alternative Datasets

We are considering alternative data as those derived from non-official sources. For this work, we are looking for datasets with data that can complement the road network and add information for other modes of transport (pedestrians, bicycles, ...) usually not covered by official data sources. In addition to the road network, we are looking for data from points of interest that can enrich the MUTN and provide new analyses about the city's transportation dynamics. Furthermore, the data sources must be easily obtained through direct download or API use. To complement the road network, OpenStreetMap

[2] http://bhmap.pbh.gov.br/ Accessed on 9 July 2020.

[3] https://dados.pbh.gov.br/dataset/gtfs-estatico-do-sistema-convencional. Accessed on 3 July 2020

was selected for its worldwide coverage and high data availability, especially for urban areas. For the points of interest, we also used data from OpenStreetMap plus crowdsource services such as Foursquare, Google Places, Facebook Places, and Yelp.

Data from open sources such as OpenStreetMap, Foursquare, Yelp, Google Places, and Facebook Places can play a crucial role in complementing official datasets when addressing urban problems. OpenStreetMap provides detailed and crowdsourced geospatial information, including roads, buildings, and points of interest, which can enhance official maps and contribute to urban planning initiatives. Foursquare, Google Places, Yelp, and Facebook Places offer a wealth of data on businesses, venues, and user-generated content, providing valuable insights into local activities and trends. By integrating these open data sources with official datasets, urban planners and policymakers can gain a more comprehensive understanding of urban dynamics, such as traffic patterns, commercial activities, and social interactions. This combination of datasets can aid in identifying areas of congestion, planning transportation infrastructure, optimizing urban services, and making informed decisions to improve the overall livability and sustainability of cities.

OSM data was downloaded from Geofabrik[4], a service that hosts OSM extracts for several regions. Data used in this case study represent a snapshot from 1 July 2020. The data should be clipped to include only the objects inside the polygon representing Belo Horizonte's city boundary. However, looking at the collective transportation data, several points along bus routes fall outside the official city limits. Because of this, it was necessary to expand the original polygon using a 1200m buffer to clip the original data. The resulting OSM dataset representing the road network includes 33,772 road segments totaling 7,121,282 m. Table 4.1 shows the top ten types of roads inside the Belo Horizonte boundary. The residential type is the most common, representing 64.9% of the total length of all roads.

Table 4.1: Top ten values for the key *highway* from OSM for Belo Horizonte, by total length in meters

highway	Total Length (m)
residential	4,625,083
tertiary	671,352
service	375,959
secondary	370,964
unclassified	192,825
primary	151,402
trunk	142,712
footway	138,897
motorway	97,466
track	86,999

[4] https://download.geofabrik.de/south-america/brazil/sudeste.html. Acessed on July 15, 2020.

Facebook Places (DFP), Google Places (DGP), Yelp (DYP), Foursquare (DFS), and OSM (DOP) were used as sources for points of interest. All services provide APIs for data querying. However, there are limitations on the volume of queries that can be executed at a given time. To respect the limits of each service, the collection took place from June 1, 2020, to June 25, 2020. The collection was assembled by querying reference points 25m away from each other and distributed as a grid across the area of interest. The total number of reference points for BH was 530,044. Data for each service were cleaned to eliminate duplicates (see Section 3.5.6) and stored. For instance, the number of points representing parking lots was initially 1,613. After cleaning and eliminating duplicates, the total count dropped to 1,238 (Table 4.2). Facebook Places and Yelp contributed a relatively small amount of data. However, some of their places were unique, so we chose not to remove them from the data integration process to get a complete result.

Table 4.2 shows an overview of the number of point and line objects gathered from official and alternative datasets and the results after schema matching procedures.

Table 4.2: Number of points and lines from datasets before and after schema matching procedures.

Dataset	**Before**		**After**	
	Points	**Lines**	**Points**	**Lines**
TC	146,542	231,112	145,625	125,554
HVTL	—	54,354	40,287	111,740
GTFS	9,365	—	42,452	472,839
OSM	123,308	260,265	47,458	127,656
DOP (OSM - parking lots)	317	—	49	—
DGP (Google Places)	918	—	857	—
DFP (Facebook Places)	6	—	5	—
DSY (Yelp)	52	—	31	—
DSF (Foursquare)	320	—	296	—

4.3 Schema-Matching Procedures

The schema-matching process starts by creating a directed graph representation of the datasets to match the proposed MUTN schema. The *TC* dataset is already in the proper format since it has a segment for each direction, and each segment has a source and a destination junction. However, the *TC* dataset has segments that do not follow the physical counterpart in the real world, representing the allowed turns between segments. These segments were used to build the *TC* network but were not considered in the data-matching process. After schema matching, the TC dataset had 145,625 nodes (junctions) and 125,554 lines (segments) (Table 4.2).

A network structure representing $HVTL$ had to be built since only the segments' geometry was available. A junction was created for each segment intersection, and the respective segments received the attributes for their source and target junctions. The original data included no information to infer the traffic flow in $HVTL$. This dataset was used primarily to transfer information about road functional classification, road names, and segments exclusively dedicated to pedestrians to the MUTN data model. After schema matching, the $HVTL$ dataset had 40,287 points (junctions) and 111,740 lines (segments) (see Table 4.2).

OSM data required some transformations to match the MUTN data model. Road segments representing two-way streets in OSM were duplicated and inverted to create two one-way segments. An OSM way feature was considered unidirectional if it had a tag oneway with any values: yes, true, 1, or -1. In the case of value -1, the direction of the segment's geometry was reversed. Source and target junctions are not readily available in the OSM dataset. Each way in OSM has a nodes attribute, an ordered list of all node codes that compose the way's geometry. OSM graph is first constructed using all nodes and then simplified to eliminate intermediate nodes, following the procedures described in Section 3.5.2. After schema matching, the OSM dataset had 47,458 points (junctions) and 127,656 lines (segments) (see Table 4.2).

POI data from Foursquare, Google Places, Facebook Places, OSM, and Yelp were selected from their respective datasets, filtering only those corresponding to parking locations. We manually identified the attribute values needed to filter the data in each dataset correctly. For example, the data in OSM was filtered using the tag value *amenity = parking*. The resulting number of points from each dataset is shown in Table 4.2.

4.4 Data Matching and Fusion between OSM and HVTL

The data matching procedure finds corresponding pairs of segments in the datasets. First, the matching between OSM and HVTL datasets follows the procedures presented in Section 3.5.3. To determine the buffer size for creating the list of candidates for matching (L_{pm}), we ran a series of tests varying the buffer size. The time (in seconds) and the average number of possible candidates found for each segment given a specific buffer size (in meters) can be seen in Table 4.3. The further away, the lower the chance of finding a segment that will match, so it is not interesting to increase the candidate list due to the computational cost and the little benefit to be obtained. We used a buffer size of 15 meters, a good balance between the processing time and the extent needed to find the best candidates for matching.

Table 4.3: Time (in seconds) and average number of candidates for matching found per segment varying the buffer size (in meters)

Buffer Size (m)	**Time (s)**	**Cand/Seg (average)**
5	138.63	8.33
10	137.27	10.37
15	138.34	11.06
20	143.81	11.64
25	147.94	12.14
30	145.14	12.62
35	141.86	13.13
40	156.67	13.72
45	151.93	14.39
50	147.08	15.07
55	147.02	15.86
60	153.10	16.75
65	154.01	17.78
70	162.76	19.26
75	162.66	21.65
80	166.26	23.46
85	166.53	25.07
90	171.92	27.22
95	181.46	29.47
100	179.08	30.96
150	221.74	52.85
200	269.53	80.12
250	324.57	111.48
300	393.02	150.75
350	485.25	193.30
400	584.73	241.41
450	692.84	295.45
500	798.13	354.05

The resulting matching pairs are used to merge data between the datasets. The OSM dataset contributed information about the mode of transport allowed in each segment (derived from the tags). The HVTL dataset was used to check the segments' street names and functional classification information. It was also used as a source of additional pedestrian segments.

Table 4.4 shows the number and total length (in meters) of segments in each dataset that were matched and discriminated by the matching type. This information can characterize the potential of each dataset to contain complementary or redundant data relative to the other. Still, it does not show whether the matches are correct (see Section 4.6). Approximately 69% of the segments and total length of OSM and 86% of the segments and 91% of the total length of segments in HVTL were matched. The high percentage of segments and length matched in HVTL indicates that it will contribute mostly as redundant or confirmation data in the data integration process. At the same time, OSM has more complementary information to contribute.

Table 4.4: Number and length(meters) of segments matched between OSM and HVTL datasets.

	OSM				**HVTL**			
	Count	**%**	**Length**	**%**	**Count**	**%**	**Length**	**%**
full	49,010	38	4,620,917	37	49,010	44	4,624,037	49
contains	10,055	08	1,565,083	12	7805	07	1,004,634	11
within	9877	08	740,788	06	18,113	16	1,000,159	11
partial	18,678	15	1,776,123	14	21,495	19	1,939,919	21
Matched	87,620	69	8,702,913	69	96,398	86	8,568,751	92

Once the matching pairs have been established, the fusion procedure for the datasets takes place. Three attributes were used in the fusion process: *width*, *level*, and *name*, which represent the width, functional classification, and the segment name, respectively. OSM dataset had few segments with *width* value (115). In this case, the fusion strategy was to rely on data from the HVTL dataset. In case of a value difference, if the same segment is involved in more than one matching pair, the new value for *width* is calculated by averaging the values found. A total of 87,532 segments had their *width* value assigned or updated.

During the schema matching phase, each dataset's attributes representing the *level* value in the MUTN data model were mapped to corresponding values. Table 4.5 shows the correspondences in the values. In the OSM dataset, the values in the table represent the contents of the '*highway*' tag for the segments. In the HVTL dataset, the values represent the contents of the '*desc_class*' attribute. In case of disparity, the fusion strategy adopted was to consider the lowest level to prevail over the most restrictive classification in terms of speed allowed in the segment. At the end of the process, 2553 segments had their level values updated.

The OSM dataset has 3438 segments with no value for the name attribute among those with a corresponding pair. When merging the name attribute, a strategy was adopted to update the values only when the corresponding pair's value had a similarity below 80% (Equation (3.5)). In this case, the name value of the dataset HVTL was preferred since it is an official source (*Trust your friend* fusion strategy). The HVTL dataset values were considered only when more than 50% of the segment length was matched for partial or contains matching. At the end of the process, 2813 segments had new values for the name attribute, and 10,599 segments were updated.

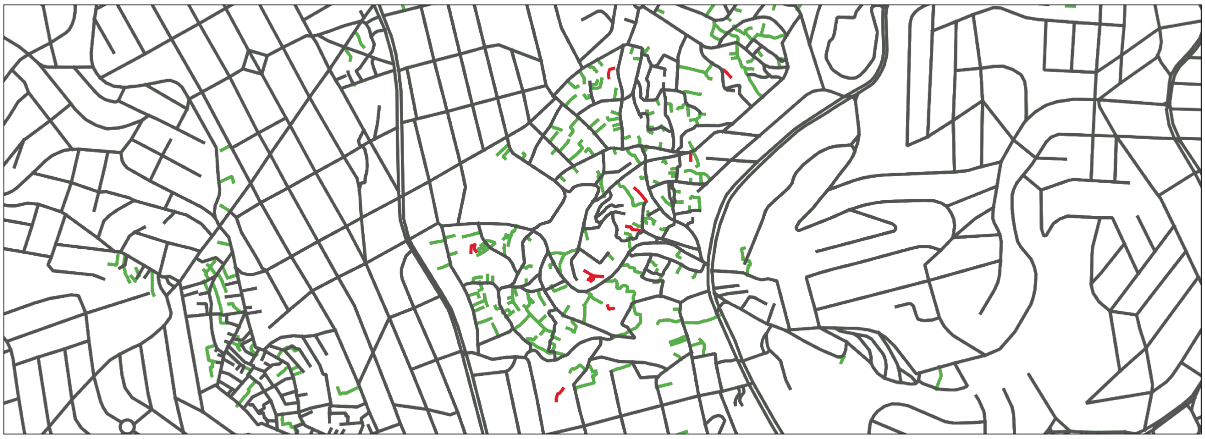
Table 4.5: Value mapping for attribute '*level*' in the segments of the MUTN data model.

MUTN **Level**	**OSM** **Highway**	**HVTL** **Desc_Class**
1	residential, unclassified, service, services, construction, corridor, crossing, cycleway, disused, dummy, footway, industrial, living_street, path, pedestrian, steps, track	local
2	secondary, secondary_link, tertiary, tertiary_link	coletora
3	primary, primary_link	arterial
4	motorway, motorway_link, trunk, trunk_link	ligação regional

The last procedure in the fusion between OSM and HVTL datasets was inserting exclusive pedestrian segments from HVTL. The segments were identified by the attribute values 'tipo_lograd' equal to 'VIA DE PEDESTRE' (walkway), 'BECO' (alley), or 'TRAVESSA' (a narrow cross-street). Even if some of these segments could be used for motor vehicles, they were considered only for pedestrian use. There was not enough information in the HVTL dataset to guarantee, for example, whether or not a segment could be used by cars and which would be its correct driving direction.

In the fusion strategy, the exclusive pedestrian segments in HVTL that did not match another in OSM were grouped into connected components. For each connected component, Junctions within a tolerance distance of some segment of the OSM dataset were detected. The respective segments are connected if they exist, and the entire group is integrated. If not, all the connected component is disregarded. Figure 4.5 shows segments from the OSM dataset (in black) and the segments from HVTL that were successfully integrated (in green), and the ones that were dismissed (in red). In this process, 5591 HVTL segments were grouped into 1219 connected components. The resulting dataset from the fusion was named DSA and had 136,675 segments, 51,179 junctions, and a total length of 12,694,791 m.

Figure 4.5: Exclusive pedestrian segments from HVTL integrated with OSM dataset. Segments in black represent the OSM original dataset. Segments in green represent the exclusive pedestrian segments in HVTL that were integrated with OSM. The ones in red were not integrated as they are far from any segment in OSM. The bounding box of this area is (606757.440, 7793281.679, 610038.036, 7794457.742) with EPSG = 31983.



Source: Made by the author.

4.5 Data Matching and Fusion between DSA and TC

The integration between DSA and TC follows the same procedures used in the fusion between OSM and HVTL. First, the matching pairs are found. At this stage, only the DSA segments (DSA_d) that allow motor vehicles were considered since, in the dataset TC, there is only this street type.

Table 4.6 shows the number and total length (in meters) of segments that were matched between the DSA_d (only segments for motor vehicles) and TC datasets, discriminated by the type of matching. Approximately 64% of the segments of DSA_d and 96% of the segments in HVTL were matched. DSA_d had most of the matched segments of type contains, which is in line with the high rate of within matches in the TC dataset. The results indicate a more significant fragmentation of TC segments. Still, the high overall rate of matched segments suggests it is a redundant data source to the data integration process. These matching results show that correspondences were found among the segments of the dataset but do not confirm if the match was correct or not. Section 4.6 presents a quantitative evaluation of the matching process to assess the quality of the matchings.

Once the matching pairs were found, they were used to check the driving direction in DSA_d. For this purpose, we analyzed all the segments in DSA_d that had a match, but their equivalent in the opposite direction did not. Using this approach, 1300 segments with the incorrect direction were found and removed from DSA$_d$. The DSA dataset, with the removal of the segments in the wrong driving direction, was named DSB.

Table 4.6: Number and length (meters) of segments matched between DSA$_d$ and TC datasets.

	DSA$_d$				**TC**			
	Count	**%**	**Length**	**%**	**Count**	**%**	**Length**	**%**
full	4346	4	294,352	2	4346	3	273,233	4
contains	57,320	47	6,375,217	53	2697	2	396,296	5
within	3836	3	209,065	2	64,461	51	5,132,231	68
partial	12,131	10	1,051,726	9	49,777	40	1,528,043	20
Matched	77,633	64	7,930,362	66	121,272	96	7,329,804	97

4.6 Quantitative Evaluation of Data Matching Results

To quantitatively evaluate the data matching process, we manually matched a random sample of 400 features for each process, OSM-HVTL, and DSA$_d$-TC, and compared the respective results. This sample size gives us a 95% confidence interval with less than 5% margin of error. The samples were selected in QGIS using the random selection tool. Then, each selected feature was manually matched by visual inspection. The results were then compared with the data matching processes for OSM-HVTL and DSA$_d$-TC. Two evaluation metrics were used, precision and recall, defined by Equations (2.1) and (2.2):

True Positive (TP) is the number of segment pairs corrected matched. False Positive (FP) is the number of segment pairs wrongly matched. False Negative (FN) is the number of segment pairs missed by the data-matching process. The intuition is that precision relates to the correctness and recall to the completeness of matching.

The data matching process between OSM and HTVL had a precision of 97.7% and recall of 96.7%. The results for the matching between DSA and TC were 98.2% and 97.7%, for precision and recall, respectively.

4.7 Creation of the Collective Transportation Network from GTFS

Although OSM can represent the geography of collective transportation, it lacks information to be effectively used for route planning. For example, OSM data has only 1457 bus stops (nodes with tag highway = bus_stop), while GTFS data for Belo Horizonte has 9328 stops. Furthermore, although there are some proposals to store timetable data in OSM, it is not clear if the community will embrace it since it violates some principles of not including temporal and seasonal features. Hence, we rely on GTFS data to build the collective transportation transport network dataset to be integrated into MUTN.

The processing of GTFS files for Belo Horizonte follows the steps described in Section 3.5.5. For each stop-route combination, a Junction is created. Then, TransferSegments are created to connect each Junction, which represents the same stop. This way, we can create TransferSegments between the routes.

For each segment of a route, a transition between two stops, a RouteSegment, is created. Each RouteSegment has an associated timetable object with all departure times assigned to that route between the two stops (in GTFS, this is represented as a trip). It is common that the GTFS files do not have the complete departure time data for each stop since it is only mandatory for them to be present at the first and last stop. In this case, each stop's estimated departure time was interpolated using the total time spent on the route by the number of stops. Then, each timetable object of each TransferSegment between ConnectionNodes representing collective transportation stops is fulfilled with all departure times from one stop to another and the traversal time (in seconds).

A collective transportation stop is where a change in transportation mode can occur, which means it is a ConnectionNode in the MUTN. This way, each stop is connected to the closest segment that allows the pedestrian transportation mode. TransferSegments are created, both inbound and outbound, for each ConnectionNode and its nearest pedestrian segment. The numbers resulting from creating the collective transportation network for the MUTN were shown in Table 4.2. The total number of ConnectionNodes was 35,250, and TransferSegments was 322,122.

4.8 Integration of *ConnectionNodes* into the MUTN

The creation of ConnectionNodes used data from five datasets: OSM (DOP), Yelp (DSY), Facebook Places (DFP), Google Places (DGP), and Foursquare (DSF). The points from all datasets identified as parking lots were selected. In the case of DOP, it is possible to find parking lots also represented using area features. For them, a point inside the area is automatically generated to represent the parking lot as a ConnectionNode.

The integration of the points uses two criteria (as seen in Section 3.5.6): the node proximity and the name similarity. For the case study, we used a distance tolerance (t_d) of 5 m. So, if two points were less than or equal to 5 m from each other, they were considered to be duplicates. Else, up to 20 m ($d_m = 4$, which means four times the tolerance), the name similarity (N_{sim}) is executed, and any two points with a similarity of 0.8 (t_{name}) or more are considered to be duplicates. The values for t_d, d_m, t_{name} were determined empirically.

After processing, 1238 instances of ConnectionNode were created. The integration of these points into the MUTN was made using the information on the possibility of a change of transport mode at the TranferJunctions. We considered the parking lots as a local to change from DRIVE (motorized vehicles) to WALK (pedestrian). For each one, we identify the nearest segment with transport mode DRIVE and connect them with a TransferSegment of type OuterTransfer (DRIVE → ConnectionNode). Similarly, we find the nearest segment with transport mode WALK and connect to the ConnectionNode (ConnectionNode → WALK).

After this integration step, the MUTN is almost complete, and it is necessary to associate traversal costs for the segments to enable the calculation of multimodal routes. Our approach was to use the time in seconds to traverse the segment as the default cost.

4.9 Cost Assignment to Segments

The maximum speed and segment length are required for the cost calculation. Only 5,72% of the segments have a value for the maximum speed assigned. For segments that do not have an assigned value, a default value is derived from the segment's functional classification (*level*). The Brazilian traffic code[5] establishes four classifications for urban

[5] https://tinyurl.com/46v77eze Accessed on 7 August 2020.

roads: fast, arterial, collector, and local traffic. Each of them has a maximum speed of 80 km/h, 60 km/h, 40 km/h, and 30 km/h, respectively, if no sign indicates otherwise. The lowest value is used if there is already a speed indication for the segment. However, a vehicle does not always move at the maximum allowed road speed, and many variables affect its speed, such as type of vehicle, time of day, weather conditions, and school hours. We adopted a value of 65% of the maximum speed for the cost calculations. This value is an estimation based on radar data from BHTrans[6].

For pedestrians, an average walking speed of 4.8 km/h was used [23]. The segments for collective transportation already have the time in seconds of transition between their points defined in the original GTFS files. These values were used as the cost of the segments. For segments that represent transfer between routes in collective transportation (*InterTransfers*), a cost of half of the interval between departures on the destination route was used. For the study case, only parking lots were used as possible points to change the mode of transport between *DRIVE* and *WALK* modes (*OuterTransfers*). The time spent to park a car varies widely depending on location and time of day, and it is difficult to estimate it accurately [22, 41]. For the case study, we empirically set a cost of 300s when a transition happens.

4.10 Multimodal Routes Using the MUTN

After the segment costs were defined, the MUTN had all the necessary information to generate routes using different transport modes. In the case study, the possible transitions between transport modes are from walking to collective transportation (and vice versa) and private vehicles to walking. The first is the typical situation of a collective transportation user who walks to a station or stops, takes a bus, and possibly changes lines until the end of their journey. The second case considers a driver who needs an appropriate place to park their vehicle near the destination.

The possibility of stopping the vehicle on the streets was not considered, only in specific parking lots. We consider that parking on the streets is already contemplated by the transportation mode, considering only the private vehicle (although a time penalty may be applied according to the expected time to find a parking spot near the destination). Therefore, the MUTN for Belo Horizonte supports routing for DRIVE, WALK, TRANSIT, and D-W (drive and walk) for the modes of transport of private vehicle, pedestrian, collective transportation, and private vehicle with the need for parking and walking to the destination, respectively.

[6] https://tinyurl.com/bdefnv66. Accessed on 7 August 2020.

Dijkstra's algorithm was used to determine optimal MUTN routes based on segment traversal time costs. Figures 4.6 and 4.7 show examples of routes created in the MUTN network considering DRIVE, WALK, TRANSIT, and D-W modes between the same points (Origin: (614840.946, 7794233.024); Destination: (611644.780, 7806446.191); EPSG:31983).

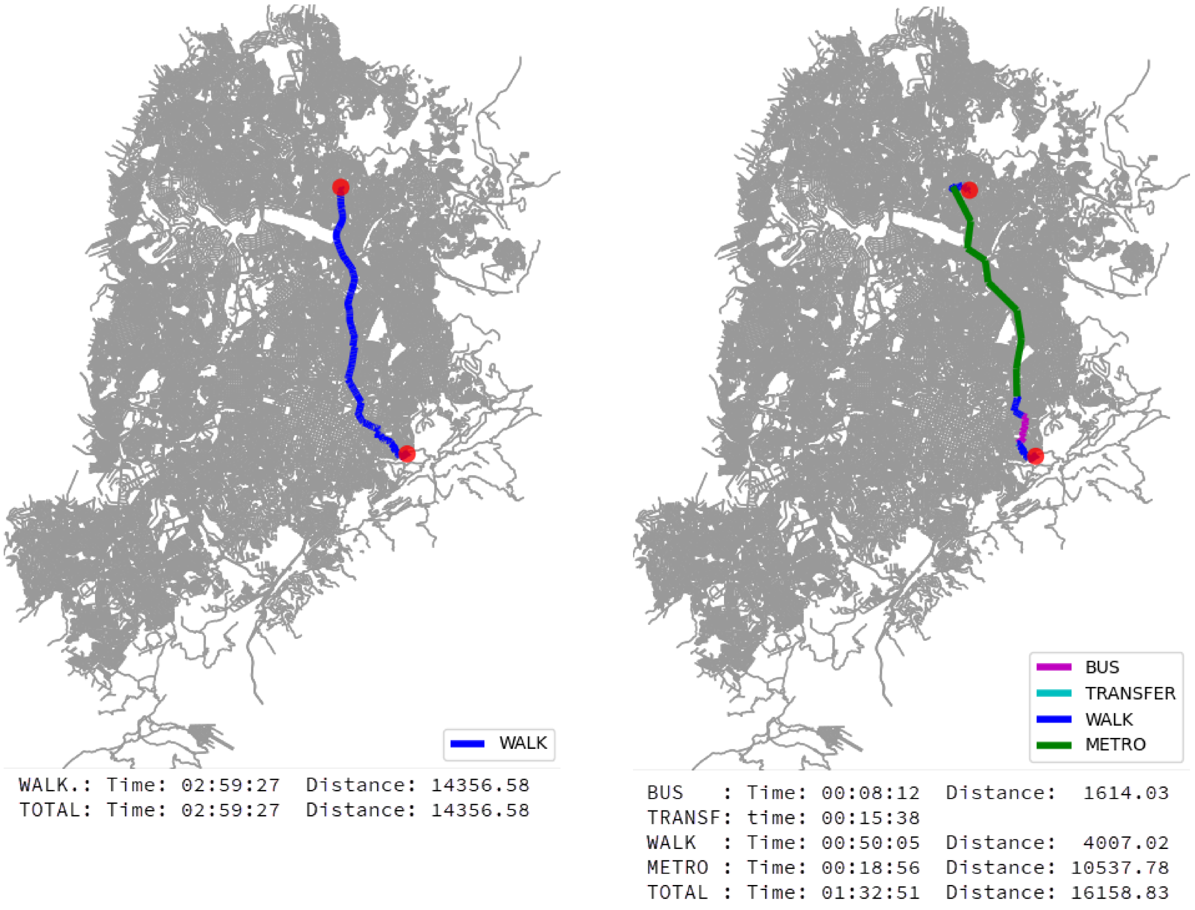
As expected, the WALK path (Figure 4.6a) was the shortest in the distance (14,356.58 meters) because it has no pedestrian turn restrictions on the streets, and therefore it is possible to create a path without detours. An exception would be if the path crossed some main road where pedestrian traffic is not allowed, in which case there would be a detour to a safe path, such as a pedestrian overpass, which would cause an increase in the final distance. The DRIVE route (Figure 4.7a) was second in the distance (15,510.97 meters) with only an 80% difference from the WALK route due to the street turning restrictions. The TRANSIT route (16,158.83 meters, Figure 4.6b) was 12% longer than the WALK route and 4% longer than the DRIVE route. The vehicles used on TRANSIT have a defined route, so it is expected that they will make more considerable detours on the direct path between origin and destination. Finally, the D-W path (17,123.01 meters) had the longest distance. This mode of transportation is susceptible to the availability of parking lots near the destination. It can be observed that the DRIVE stretch (16,316.99 meters, Figure 4.7b) on the D-W route was longer than the direct path to DRIVE, which indicates that it was necessary to go beyond the destination in search of a parking lot and then return on foot (WALK).

Concerning the time spent, there was more variation among the modes of transport than in the distance. The fastest one was DRIVE taking 1938 seconds for the commute, followed by D-W with 3011 seconds (+19% compared to DRIVE), TRANSIT with 5571 seconds (+187% compared to DRIVE and +85% compared to D-W), and WALK with 10,767 seconds (+455% compared to DRIVE, +257% compared to D-W and 93% compared to TRANSIT).

4.11 Results and Discussion

To compare the results obtained by modeling and integrating the data, we created routes between 80 points spread throughout the municipality area. Each point represents a location at the MUTN closest to the centroid of each of Belo Horizonte's units of planning. Units of planning are territories formed by the aggregation of census sectors, used by the public administration in various situations, such as calculating socioeconomic indicators (e.g., urban life quality, social vulnerability), and distributing participatory

Figure 4.6: MUTN routing examples for WALK and TRANSIT transport modes (distance in meters).



```
WALK.: Time: 02:59:27  Distance: 14356.58
TOTAL: Time: 02:59:27  Distance: 14356.58
```

```
BUS   : Time: 00:08:12  Distance:  1614.03
TRANSF: time: 00:15:38
WALK  : Time: 00:50:05  Distance:  4007.02
METRO : Time: 00:18:56  Distance: 10537.78
TOTAL : Time: 01:32:51  Distance: 16158.83
```

(a) WALK routing

(b) TRANSIT Routing

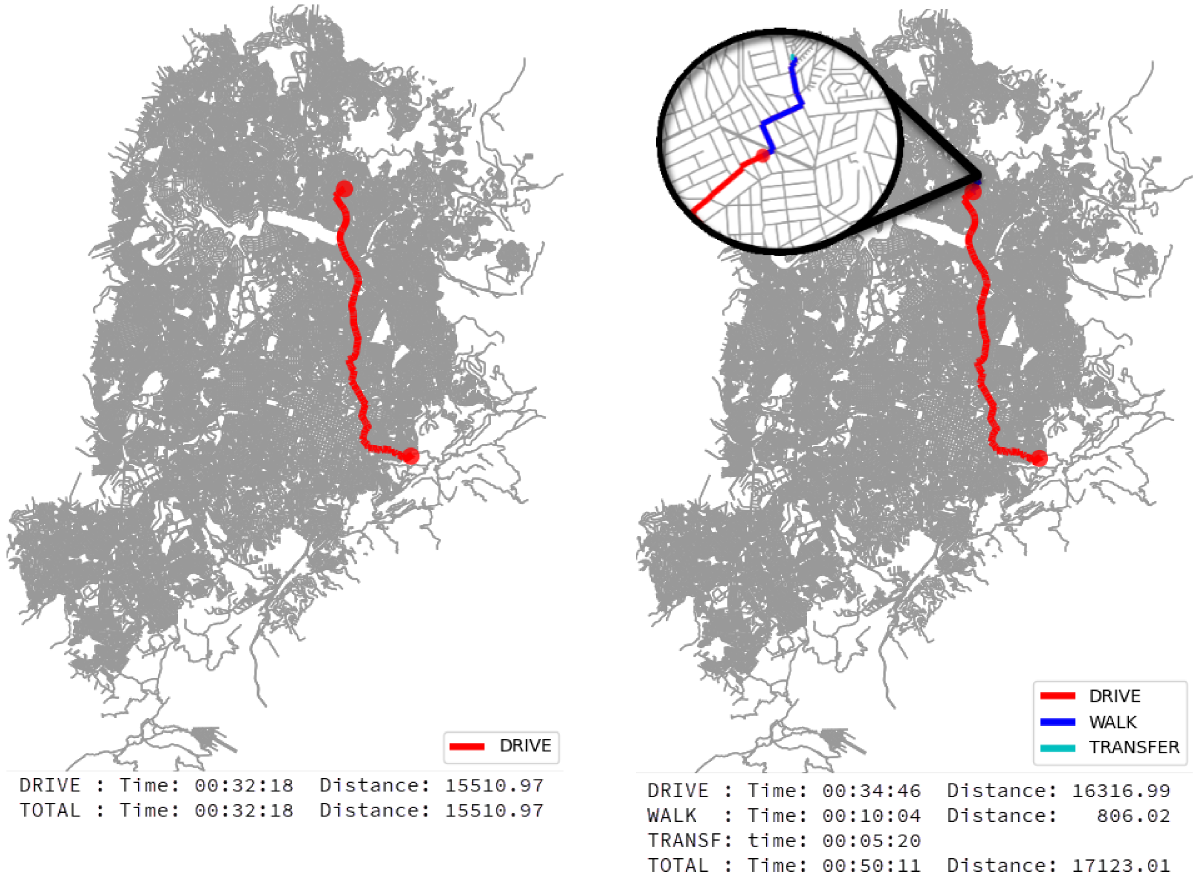
Source: Made by the author.

budget resources. As there are 80 units of planning in Belo Horizonte, 6400 routes were calculated considering the round trip between each pair of points.

Routes between all pairs of points were calculated for the WALK, DRIVE, TRANSIT transport modes using MUTN, and Google Maps. Google Maps does not have an option for car routes looking for parking near the destination, so it was impossible to compare it with the D-W routing option.

For each route, the time and the distance were calculated using MUTN and Google Maps. Then, the differences between distances and times were calculated, and finally, the ratio between the differences and the respective values was obtained by MUTN. Table 4.7 shows a comparison of the results. The table's values represent the average of the absolute values of the ratios for time and distance. The time difference between the routes created through MUTN and Google Maps was 9.5%, 9.9%, and 15.5% for WALK, DRIVE, and TRANSIT modes of transport, respectively. Simultaneously, the distance difference among the routes was 4.7%, 7.3%, and 18.4%.

Figure 4.7: MUTN routing examples for DRIVE and DRIVE–WALK transport modes (distance in meters).



```
DRIVE : Time: 00:32:18  Distance: 15510.97
TOTAL : Time: 00:32:18  Distance: 15510.97
```

```
DRIVE : Time: 00:34:46  Distance: 16316.99
WALK  : Time: 00:10:04  Distance:   806.02
TRANSF: time: 00:05:20
TOTAL : Time: 00:50:11  Distance: 17123.01
```

(a) DRIVE Routing

(b) D-W Routing

Source: Made by the author.

To investigate if the distance between the points has any significant effect on the difference among the routes, we divided the results into ten distinct groups, each one with 617 ± 4 routes, and calculated the respective time and distance difference averages. Table 4.7 shows that the smallest differences are found in the groups of routes with the largest distances between the origin and destination. In contrast, the most significant differences occurred in the group with smaller distances for each mode of transport. A possible explanation for this situation is that any difference in routes considering a small distance will significantly impact the difference between them, while for longer distances, small variations in routes do not have a significant impact.

While the routes for DRIVE and WALK had a difference of less than 10% in both time and distance, TRANSIT results obtained higher values, of 18.4% and 15.5% of difference for distance and time, respectively. The reasons for this difference in data can be summarized in two hypotheses: Difference in data and difference in cost parameters for routing. To check for the difference in data, we selected routes with the most significant

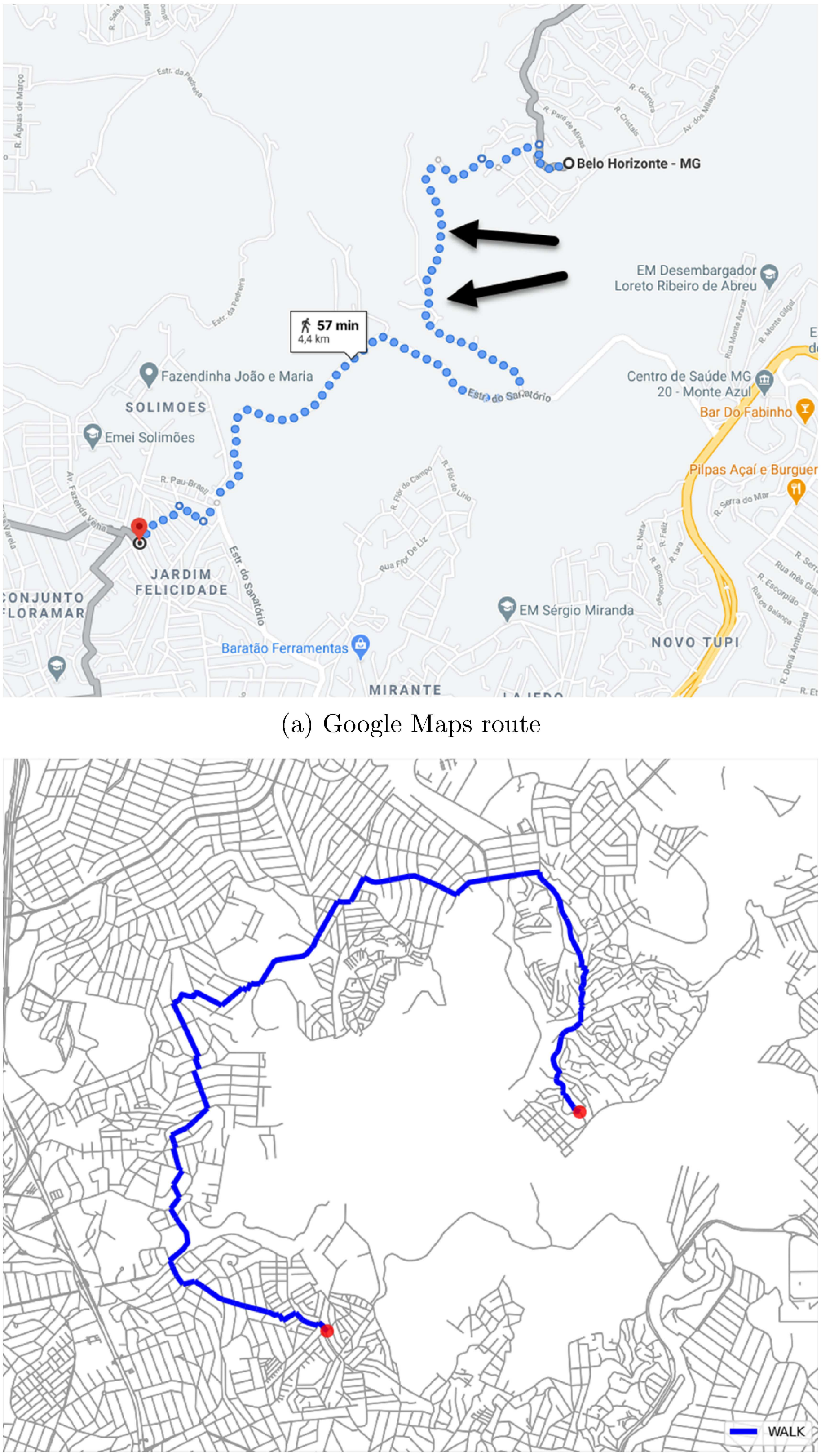
Table 4.7: Comparison by time and distance, between the routes created using the Multimodal Urban Network and Google Maps. Values represent the average of the absolute ratio difference (e.g. 0.127 means 12.7%).

Distance Range (m)	*WALK*		*DRIVE*		*TRANSIT*	
	time_diff	*dist_diff*	*time_diff*	*dist_diff*	*time_diff*	*dist_diff*
(533, 3508]	0.127	0.057	0.154	0.112	0.207	0.284
(3508, 5116]	0.112	0.051	0.117	0.090	0.196	0.229
(5116, 6546]	0.102	0.050	0.105	0.078	0.165	0.209
(6546, 7810]	0.091	0.044	0.090	0.071	0.149	0.197
(7810, 9123]	0.096	0.050	0.083	0.075	0.139	0.182
(9123, 10444]	0.087	0.044	0.080	0.071	0.137	0.177
(10444, 11932]	0.090	0.046	0.079	0.069	0.143	0.156
(11932, 13717]	0.081	0.040	0.079	0.060	0.135	0.144
(13717, 16301]	0.079	0.040	0.086	0.057	0.136	0.138
(16301, 26591]	0.081	0.043	0.121	0.050	0.140	0.125
(533, 26591]	0.095	0.047	0.099	0.073	0.155	0.184

difference in distance between MUTN and Google Maps for visual inspection. We observed that, in certain situations, the MUTN network traced longer routes than Google Maps. Figure 4.8 shows the route in which the most significant difference in relative distance occurred. The Google Maps route (Figure 4.8a) uses a path in which there is no apparent connection in the map segments and returned a route length of 4394 m. Routing on the MUTN network only returns routes by connected segments (Figure 4.8b). The route length returned by MUTN was 11,183.68 m. The difference in cost parameters is more challenging to verify because the details of how Google Maps performs the routing are unknown. Thus the cost estimates used in the MUTN for TRANSIT may be a factor that increases the difference in comparison to the routes obtained using Google Maps. In addition, distance differences in the TRANSIT routes can be explained by the lack of the geometry of the transit routes in the GTFS file of Belo Horizonte. Google Maps creates a route between bus stops using a direct path through the road network. In contrast, the routes in MUTN use the travel time from the GTFS file, and a direct link between the bus stops in the lack of geometry.

The proposed data model proved adequate as a frame of reference to organize the process and to integrate the data in a structure that is suitable for the necessary processing. The final result of MUTN for Belo Horizonte and associated data took up 184 MB of disk space.

Figure 4.8: The difference in routes generated by Google Maps (**a**) and MUTN (**b**). Google Maps uses a path by apparently disconnected routes. Origin: (614140.121,7808865.508), destination: (612378.659,7807352.595), EPSG:31983.



(a) Google Maps route

(b) MUTN route

Source: Made by the author.

Chapter 5

MUTN in the Calculation of the Quality of Urban Life Index[1]

The fast urbanization process significantly impacts the economy, health, environment, and other aspects. Many challenges arise from these impacts. Local government must have indicators that summarize the evolution of citizens' quality of life, identify regions that need further attention, assist urban planning, and allocate resources to deal with current and future problems. However, many metrics have been proposed over the years, and there is no standard agreement on defining and measuring the quality of urban life. In 1990, the United Nations Development Programme created the Human Development Index (HDI) to be a more human-centric measure instead of only economic ones to show the welfare situation. It is based on life expectancy, education, and *per capita* income. The HDI is globally used to rank countries, states, and cities. Still, it fails to "reflect inequalities, poverty, human security, empowerment" [233]) The HDI uses the city, or municipality, as its smallest territorial unit.

The creation of the HDI inspired government officials to propose and use their indicators to measure the quality of life in spatial units of various sizes, varying from country to metropolitan regions and municipalities. Based on the HDI, the Social Development Index (IDS) was created in 1991, replacing per capita income with comfort and sanitation indicators. The IDS is calculated for state-size geographic areas. The Living Conditions Index (ICV) is another city-level index and was created to measure poverty and childhood conditions [70].

One problem with these indicators is that they attempt to summarize the complexity of cities into a single number. The local government uses these indexes to monitor city performance but cannot use them to identify problems in specific city regions. To address this situation, Nahas [158] describes the idea of the quality of urban life index (in Portuguese, Índice de Qualidade de Vida Urbana, IQVU), which is composed of ten dimensions: food supply, culture, education, sports, housing, urban infrastructure, environment, health, urban services, and urban security. It is based on quantifying the spatial inequality of services that are available and accessible to the population and can

[1] Parts of this chapter are based on and extend the works in [213, 214]

be used to show areas that need increased public investment. The index was created and is used in Belo Horizonte (BH). The IQVU is calculated for a geographic subdivision of the city, called the Unit of Planning (UP), and is used to support urban planning. The first component in calculating the IQVU is a Local Availability Index (Índice de Oferta Local, IOL) that measures the availability of services inside a UP. An accessibility factor then adjusts the IOL to deal with situations where people from other UPs access services from a given UP. The IOL, adjusted by accessibility, comprises the IQVU. Since its first published results in 1996, the IQVU has been used by the government as a factor to balance public investment among the UPs. IQVU relies on official data sources provided by about 16 different governmental organizations.

Although the IQVU has been conceived to be quickly and periodically recalculated, this is not what has happened since its creation. It was expected that the variety of temporal granularities among the several data sources might be a significant cause for the irregular updating of the indicator. Other factors may include difficulties obtaining some indicator components due to methodological changes implemented by the information-providing organizations throughout the years. Thus, this chapter shows the results of an investigation into using crowdsourced data as an alternative to obtaining a quality of urban life index.

Two main components are needed to calculate the IQVU: the indicator values and the accessibility matrix. The MUTN data model proposed in this thesis can be used for both cases. Data were collected from Facebook, Foursquare, Google Places, and OpenStreetMap for the indicator values. The available API of each LBSN was used to search for points of interest over the city of Belo Horizonte and then to estimate the IOL component of IQVU, following, as close as possible, the IQVU calculation methods. Then the data was processed to eliminate duplicates in the same dataset and between different datasets to create an integrated dataset for storage according to the MUTN data model.

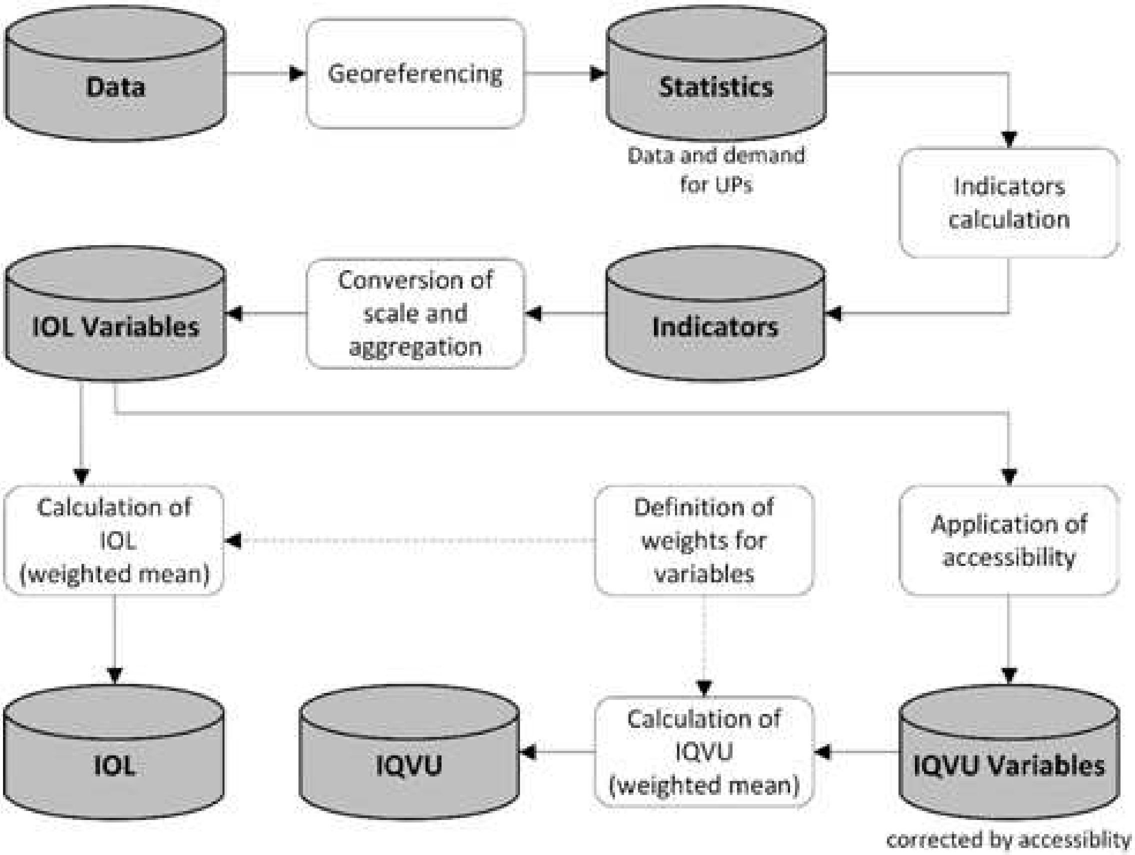
The accessibility matrix comprises the travel times by public transport between the centroid of each pair of planning units. When the work that originated IQVU was created in 1995 [130], there was no simple way to obtain the accessibility matrix. Currently, we can use services such as Google Maps, which provides an API that allows us to obtain the necessary information, but may incur costs and has the restriction of using Google's data, which may not be up to date. As seen in Section 4.10 and 4.11 the MUTN can create an accessibility matrix for different modes of transport without additional costs and with greater control of the data and parameters used in routing.

The remaining of this chapter is organized as follows. In Section 5.1 we present the IQVU index and how it is calculated. Section 5.2 offers the methods used to collect, clean, and integrate the crowdsource data to calculate IQVU indicators. Finally, Section 5.3 presents the results and discussion.

5.1 Quality of Urban Life Index

As previously described, the IQVU aims to spatially quantify the inequality of services available to the population. Georeferenced data from several government agencies calculate the indicators aggregated into components and variables. For instance, *health centers* (the number of public health facilities in a UP per thousand inhabitants) is an indicator of the *health care* component, which is part of the *health* variable. The steps to calculate the IQVU are shown in Figure 5.1.

Figure 5.1: Steps to calculate the IQVU index.


Data
Georeferencing
Statistics
Data and demand for UPs
Indicators calculation
IOL Variables
Conversion of scale and aggregation
Indicators
Calculation of IOL (weighted mean)
Definition of weights for variables
Application of accessibility
IOL
IQVU
Calculation of IQVU (weighted mean)
IQVU Variables
corrected by accessiblity

Source: Adapted from [198]

In short, after data gathering and calculation of indicators, three main results are obtained: the Local Availability Index, IQVU variables, and IQVU index.

The IOL is calculated for each dimension on each UP. First, each indicator's availability is calculated with a specific measure, generally a count per thousand inhabitants (e.g., number of hospitals/population × 1000). Then, the indicator values are normalized between 0 and 1 by Equation 5.1:

$$I_c = 1 - e^{-(f.v)} \tag{5.1}$$

where I_c is the normalized indicator value, v is the original indicator value, and f is given by Equation 5.2:

$$f = -ln(0.05)/L_{ref} \tag{5.2}$$

where L_{ref} is the reference value for the indicator, which is assumed to be the 95% percentile. This adjustment compensates for higher values, so they cause less impact on the index [130]. The IOL of a dimension is given by the simple mean of the normalized value of its indicators, and the IOL for the UP is given by the weighted average of all dimensions.

The **IQVU Variables** (sector indexes) are calculated by applying an accessibility index on the IOL. The accessibility index is based on an estimate of the travel time of a citizen using public transportation between each pair of UPs. The accessibility index can increase or decrease the variable's IOL value based on how easy it is for the population of a UP to access the services available in other UPs. For example, if a resident needs medical care, but his or her home's UP has a below-average offer of this type of service, he or she can access this service at a nearby UP that can provide the service (and to do so, it must have an offer value above the general average). The accessibility adjustment will then increase the local availability for the UP of the resident and will decrease the value for the UP where he or she used the service. When we calculate the influence of accessibility over all variables for all UPs, we can calculate the IQVU. The accessibility works differently for each variable, considering four accessibility modes: immediate, near, average, and remote. This approach can be seen as an application of a gravity potential expression [222, 76].

An immediate accessibility measure means that the services corresponding to the variable should only be accessed inside the UP. A near accessibility measure needs the displacement time to be relatively short for services available near the home. For instance, a person can go to a supermarket outside her UP, but it is less likely that she will visit a distant one. The values of the average and remote accessibility decay more slowly with distance than in the near accessibility measure. These are the cases of more complex services for which traveling a relatively long time can be tolerated; for instance, a consultation with a medical specialist across town. The calculation procedure to incorporate accessibility measures is rather long and will not be detailed further here. It can be seen in detail in [130].

The **IQVU index** for each region is given by a weighted average of IQVU variables. The weight of each dimension was established by the specialist group responsible for creating IQVU.

With IQVU results, local governments can identify which regions need more attention. It has already been used to support decision processes for public investment, but we do not have information on its current use. The index underwent methodological changes and was not published regularly. Improvements are required to use similar methods to help local governments effectively.

5.2 Methods

This section describes the materials and processes for calculating the IQVU by using and integrating data from official and crowdsourced data. First, we describe the datasets used, the collection, data cleaning, and integration process. Then the integrated data and the Multimodal Urban Transport Network are used to calculate the IQVU.

5.2.1 Official Data Sources

We used data from three official data sources: the most recent IQVU-BH dataset (2016), Brazilian Census data, and urban geographic data of Belo Horizonte.

The IQVU-BH dataset was obtained from Belo Horizonte's Web portal[2]. It is available as a spreadsheet that contains the values of indicators, components, and variables of IQVU for each UP in 1994, 2000, 2006, 2010, 2012, and 2016. The raw data and calculation formulas are not included in the file.

Almost all indicators are normalized by the population of each UP. For that, official demographic data from 2010 were used, as in previous IQVU calculations. Census data provides the population counts for each census sector, while IQVU uses Units of Planning (UP), which are spatial aggregations of census sectors. Therefore obtaining demographic data for each UP is straightforward.

5.2.2 Crowdsourced Data Sources

Facebook Places, Foursquare, Google Places, Yelp, and OpenStreetMap were used as data sources, as they maintain publicly available datasets on services and businesses, which can be used as a source for some of the IQVU indicators. They were chosen because

[2] https://tinyurl.com/27w2eywd. Accessed on March 14, 2021

they offer a public API that allows data gathering, contains a large number of users and places, and allows volunteered data contributions by anyone.

5.2.3 Data Collection

In order to collect the full set of services for Belo Horizonte, an iterative procedure was implemented. First, a regular grid of points separated by 25 meters was generated, covering the bounding box of Belo Horizonte's city limits, provided by the city's spatial data infrastructure. Then, a geometric intersection between the grid and the city limits polygon was used to select the points inside the municipal territory. This operation resulted in a set of 530,044 distinct reference points. For each point, an API call was made using the point's coordinates and a radius of 25 meters as parameters. The radius size of 25 meters was considered small enough to get all the available places for each social network since a larger radius might cause the APIs to return only a subset of the available places near the point due to the limitations of the public API.

Crawlers for Facebook Places, Foursquare, Google Places, and Yelp were implemented using Python, and data were collected in June 2020. Crawlers had to pace themselves to comply with the APIs' enforced limit on the number of daily requests. Yelp, for instance, allows 25,000 search requests per day, while Foursquare allows 5,000 requests per hour. The API returns JSON data that were initially stored as a single text file. Then the file was processed to extract, for each location, its ID, latitude, longitude, and category.

OpenStreetMap does not require a crawler, as all its data are open and available for download. The data corresponding to the southeastern region of Brazil was downloaded from the Geofabrik[3] service. Then the data were filtered using the polygon that represents the geographical limits of Belo Horizonte. Only the data inside the polygon remained. Unlike the other services, OSM can store features as points or polygons. The centroid of the polygon is used for cases where the indicator uses only feature counting. In cases where it is the area of the feature that is used in the calculation, for example, green areas, it has been preserved and stored as AreaFeature (see Section 3.1).

[3] https://download.geofabrik.de/south-america/brazil/sudeste.html. Acessed on July 15, 2020.

5.2.4 Data Cleaning and Characterization

Data collection produced datasets containing many entries for each service used. Duplicate entries were collected since intersections between areas were generated for each reference grid point. Furthermore, many places collected are irrelevant to our study since they are not used to calculate IQVU indicators. For example, the dataset has information on laundries and clothing stores, but these services are not considered in the IQVU.

The data cleaning step consisted of removing duplicate points, removing points with missing data, and mapping categories from each source to the corresponding IQVU category. First, records with duplicate IDs are removed from Foursquare, Yelp, and Facebook. OSM data did not need this step because the data collection process was different from other datasets and did not produce duplicate records. Data from Google Places had this stage postponed because Google allows each place to have a list of categories. In order to match categories, we had to duplicate each record in this situation so that one record for each category in the list was created. Second, it was necessary to match the categories of locations in the datasets with those used by the IQVU. The approach used relied on relating each category in the services with one of the IQVU indicators using manual classification. For example, categories mapping used for OSM data can be seen in Table 5.1. Table 5.2 shows the status of the datasets after each of these steps.

5.2.5 Data Integration

The integration of collected data follows the procedures described in Section 3.5.6. The integrated dataset has been stored as PointFeatures (see Section 3.1) with the properties representing their name, category, and origin. The resulting dataset has 36,690 unique points which means the data integration removed 5,770 duplicate points. Figure 5.2 shows the distribution of the points across the city of Belo Horizonte.

5.2.6 Calculation of the Local Availability Index Indicators

The IOL aggregates information from indicators that measure the availability of services (e.g., supermarkets), into IQVU variables (e.g., food supply). Then, results are

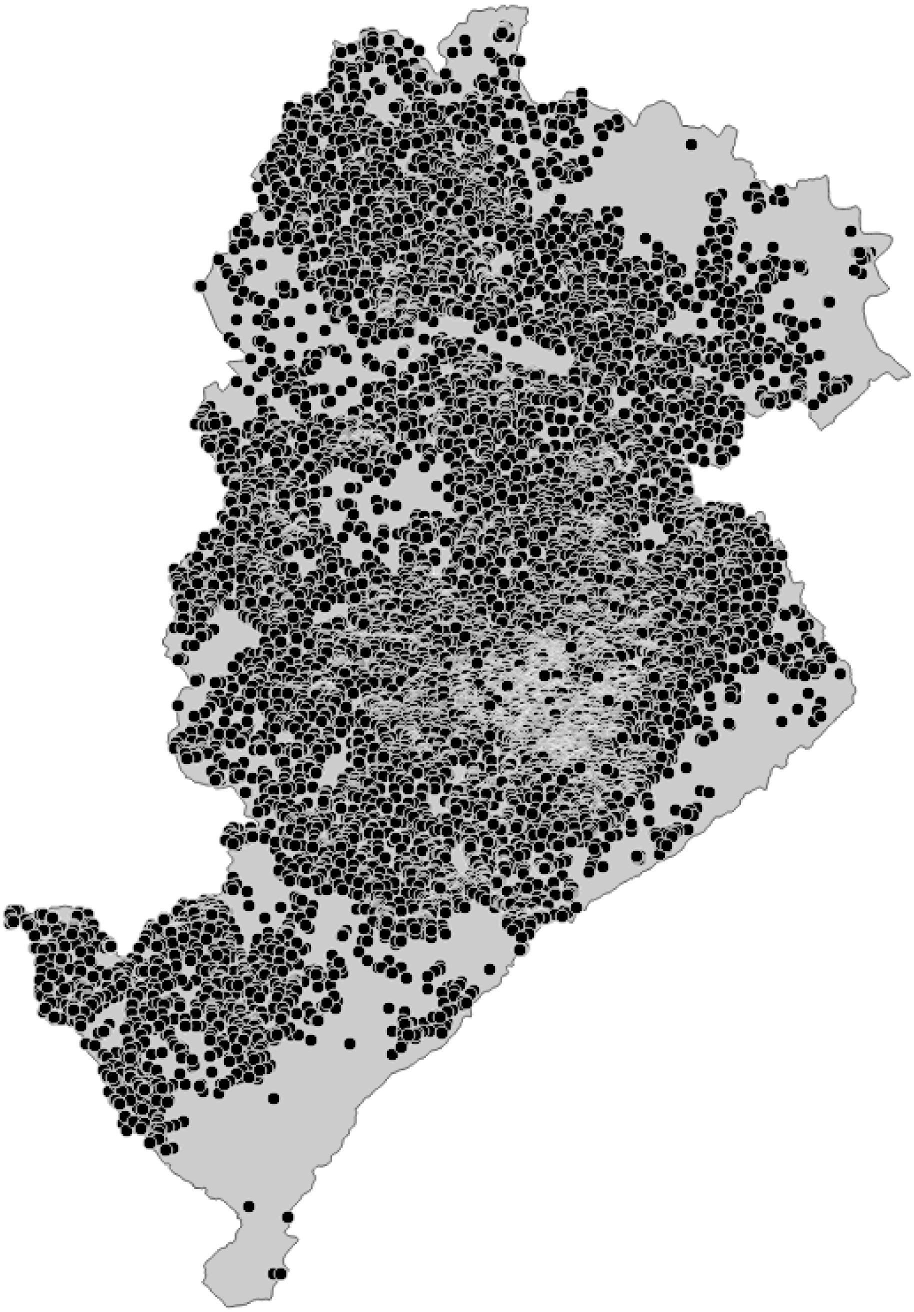
Table 5.1: Correspondence between OpenStreetMap key-value pairs and IQVU indicators

IQVU Indicator	**key**	**values**
Hyper and supermarket	shop	supermarket
Grocery store and similar	shop	convenience
Cultural equipment	amenity	library, archive, cinema, theatre
	tourism	museum
Bookstore and stationery	shop	books, stationery
Movie rental store	shop	video
Magazine stand	shop	newsagent
Sport court, field	leisure	sports_centre, fitness_centre
	amenity	community_centre
	landuse	recreation_ground
Health centers	amenity	clinic
Other health care services	amenity	hospital
Dental services	amenity	dentist
Bank agency	amenity	bank, atm
Gas station	amenity	fuel
Drugstore	amenity	pharmacy
	shop	chemist
Public Phones	amenity	telephone
Post Office	amenity	post_office
Digital Inclusion Spaces	internet_access:fee	no
Green Area	amenity	grave_yard
	landuse	allotments, cemetery, farmland, recreation_ground, meadow, orchard, village_green, vineyard, grass, greenfield,forest
	leisure	garden, golf_course, nature_reserve, park, pitch
	natural	wood, scrub, health, grassland, wetland
	tourism	camp_site

Table 5.2: Quantity of places (points) of each step of data source cleaning process.

Data Status	Facebook	Foursquare	Google Places	Yelp	OSM
Raw data	98,232	215,783	3,279,482	541,965	4,593
Duplicates removed	13,000	72,785	208,013	44,427	4,593
Categories mapped	1,226	15,617	12,528	11,301	1,788
Data integration	1,067	13,667	10,266	10,138	1,552

Figure 5.2: The result of the data integration tasks for POIs data. There are 36,690 unique points.



Source: Made by the author.

Table 5.3: IQVU Indicators from each source to build the MUTN integrated version

IQVU Indicatores	Facebook	Foursquare	GPlaces	OSM	Yelp	MUTN
Bank agency	x	x	x	x	x	x
Green area	x	x	x	x	x	x
Magazine stand	x		x		x	x
Designated Heritage Buildings			x		x	x
Health centers	x	x	x		x	x
Commerce and Cultural Services	x	x		x		x
Post office	x	x	x	x	x	x
Cultural equipment			x		x	x
Child School	x	x				x
Elementary School	x	x				x
High School		x		x		x
University education	x	x		x		x
Dental services	x	x		x	x	x
Digital Inclusion Spaces	x					x
Drugstore	x	x	x	x	x	x
Hyper and supermarket	x	x	x	x		x
Bookstore and stationery	x	x	x	x	x	x
Movie rental store	x			x	x	x
Grocery store and similar	x		x	x	x	x
Other health care services	x	x		x	x	x
Gas station	x	x	x	x		x
Sport court, field	x	x	x	x		x
Restaurants			x		x	x
Public Phones	x					x

weighted according to the importance of each IQVU variable [198].

The number of places collected inside each UP was counted for each indicator to calculate IOL variables. This number is used to get the indicator's value per thousand residents by dividing it by the UP population and multiplying it by 1,000. The exception was the indicator Green area which used the sum of the area of all polygons classified as green spaces (see Table 5.1) and then divided by population to get the raw indicator value. Results from IQVU 2016 were used as a basis to allow comparative analysis. For IOL variables calculation, the official values replaced indicators for which there is no data. Indicator values were normalized between 0 and 1 and then aggregated on IOL variables for each UP by applying a weighted average [198].

5.3 Results and Discussion

First, we evaluate IOL indicators calculated using data from Foursquare, Google Places, Yelp, and the integrated dataset using the MUTN data model. Data from Facebook and OpenStreetMap were used to enrich the integrated dataset. Still, no individual results were produced for these datasets due to the small number of points collected compared to the others. For the data collected to calculate the IQVU indicators, they must have some information regarding their function. In the case of OpenStreetMap, it is common for the maps to have a high level of completeness for the elements that can be mapped by satellite images (streets, green areas, rivers, buildings). However, semantic information is not always attributed to these elements. For example, it is known for a polygon representing a building to be marked only with the *building=yes* tag, which makes it very hard to use them to calculate the IQVU indicators. In the case of Facebook, a possible explanation for the low number of points found is that it is not used as a source to consult geographic information, unlike Google, which has services such as Maps, Earth, and Street View.

The values obtained were discretized into the same intervals used to present IQVU official results: [0,0.5), [0.5, 0.6), [0.6, 0.7), [0.7, 0.8), [0.8, 1], labeled "1", "2", "3", "4" and "5", respectively. After classifying the interval results, accuracy, precision, and recall were calculated for IOL indicators using a multi-class approach [215], testing all datasets against official 2016 IOL results.

Table 5.4 shows the results of the calculations for the Local Availability index. The lowest accuracy values obtained for the Facebook dataset were for the indicator *drugstore*, with a value of 0.18; the highest was for the indicator *Post Office*, with a value of 0.75. For the Foursquare dataset, the lowest value found was for the indicator *Grocery store and*

Table 5.4: Accuracy (acc.), Precision (prec.) and Recall (rec.) of indicators calculated from Foursquare, Google Places, Yelp, and the integrated dataset using the MUTN data model. The Precision and Recall values are the mean of the five IQVU classes

Indicators	Foursquare			GPlaces			Yelp			MUTN		
	acc.	prec.	rec.	acc.	prec.	rec.	acc.	prec.	rec.	acc.	prec.	rec.
Hyper and supermarket	0.46	0.41	0.45	0.41	0.32	0.23				0,63	0.61	0.64
Grocery store and similar	0.12	0.11	0.15	0.34	0.14	0.17	0.30	0.19	0.20	0.41	0.37	0.27
Cultural equipment	0.74		0.58				0.89			0.92	0.32	
Bookstore and stationery	0.53		0.20	0.37	0.10	0.13	0.54	0.19	0.22	0.44	0.32	0.39
Movie rental store				0.48	0.34	0.32	0.54	0.38	0.37	0.54	0.44	0.47
Magazine stand	0.56	0.26	0.28				0.58	0.28	0.32	0.64	0.29	0.33
Sport court, field	0.48	0.43	0.42	0.39	0.31	0.30				0.51	0.45	0.48
Green area										0.92		
Health centers	0.22	0.25	0.20	0.27	0.21	0.21	0.22	0.24	0.20	0.39	0.38	0.32
Other health care services				0.74	0.35	0.42	0.84	0.48	0.40	0.89	0.56	0.45
Dental services				0.64	0.30	0.31	0.88	0.49	0.46	0.90	0.50	0.45
Bank agency	0.72	0.29	0.35	0.65	0.40	0.49	0.50	0.15	0.17	0.83	0.39	0.38
Gas station	0.59	0.40	0.39	0.36	0.25	0.27				0.64	0.42	0.41
Drugstore	0.44	0.23	0.29	0.29		0.14	0.60	0.53	0.53	0.63	0.55	0.59
Post office	0.73	0.38	0.54	0.65	0.42	0.57	0.73	0.28	0.26	0.70	0.34	0.30

similar, with a value of 0.12, while the highest was for the indicator *Cultural Equipment*, with a value of 0.74. For the Google Places (GPlaces) dataset, the value with the lowest accuracy was *Drugstore* (0.29) while the highest was for the indicator *Other health care services*, with a value of 0.74. In the case of the Yelp dataset, the indicator with the lowest accuracy value was *Grocery store and similar* with a value of 0.30 and the highest was for the indicator *Cultural Equipment* with a value of 0.89 accompanied by the indicator *Dental Services*, with a value of 0.88. For the MUTN integrated dataset, the lowest accuracy value found was for the indicator *Grocery store and similar* with a value of 0.41 while the highest values were for the indicators *Cultural equipment* and *Green Area* with a value of 0.92. In the MUTN dataset, the indicators *Other Health care services*, *Dental services*, and *Bank Agency* were with accuracy above 0.8 with values of 0.89, 0.90, and 0.83, respectively.

Except for OpenStreetMap, all other datasets use only points to record locations. As the indicator *Green Area* is calculated using the area of the geometry representing parks and forests, it was only possible to calculate it for the MUTN dataset that has this data integrated. As this type of feature usually does not appear in large numbers in urban areas, and tgiven that they can cover large areas, they are usually well mapped in services like OpenStreetMap. This fact may explain its accuracy value of 0.92 obtained in the *Green Area* indicator for the integrated dataset.

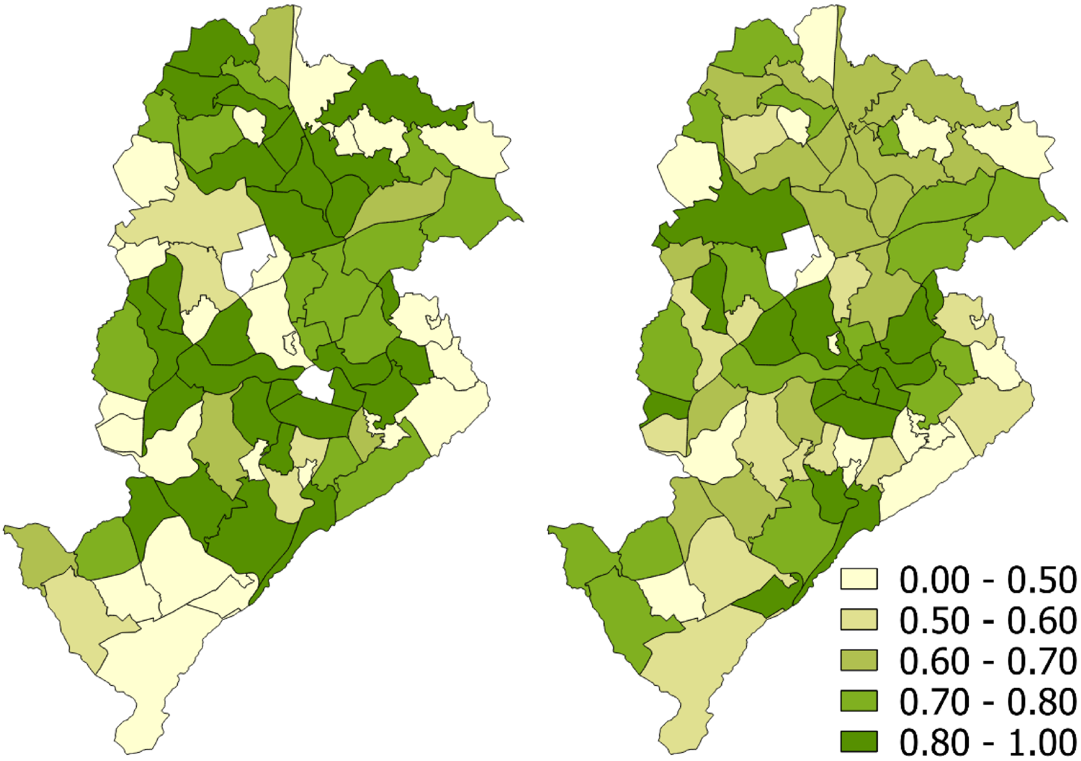
In general, good accuracy was achieved, but results for precision and recall were poor. One possible cause is that intervals “1” and “5” have a wider range of values, taking into account 70% of the possible values, while intervals “2”, “3”, and “4” only answer for 30%. This unbalanced distribution and the five intervals considered could decrease the mean value of recall and precision. This suggests that if a binary classification is used, for example, to identify deprived and well-off regions, better results could be achieved.

In order to allow a visual inspection of the spatial patterns of the results, maps were generated to compare the IOL values for the indicators that were calculated using the integrated dataset on MUTN. Figure 5.3 shows IOL results of variable *Hyper and supermarket*.

5.3.1 Limitations

This work has at least three limitations that need to be observed. First, LBSN data can be biased since it is crowdsourced. LBSNs and other online social network data are tied to the profile of their users, generally young and technology-friendly [179]. This aspect can cause the coverage of those locations to be better around neighborhoods

Figure 5.3: Comparison of results for IOL Indicator *Hyper and supermarket*. Official data is on the left and MUTN data is on the right.



Source: Made by the author.

where the young population lives, works, or has fun [184]. Not having many locations on an LBSN does not necessarily indicate that the ground truth is likewise deprived. However, as smartphone usage increases, it should be increasingly important for businesses to participate in LBSN catalogs such as the ones used in this work.

The second limitation regards the lack of some types of information from LBSN. It was not possible to obtain all indicator values from the LBSNs used in this work, but this will not necessarily always be true. Currently, collecting the missing data required by this work depends on open data policies in place for government information-producing organizations. The authors believe that this can be fixed with upcoming open government data initiatives and legislation that supports information dissemination in machine-readable formats and services. PDF files containing images of tables and maps do not qualify as such. Active crowdsourcing, in volunteered geographic information initiatives directed at those categories, might fill this gap [149], although new biases may be introduced, as in the first limitation.

The third limitation is related to the lack of more frequently updated results of IQVU for comparison. All results for this work are based on data collected in the third quarter of 2015, but the latest IQVU results are from 2016, so we are comparing current LBSN data with official data that is three years old. Improving the results of this work, and possibly calculating a quality of urban life index with online and easily available social networks and official government data may enable us to produce frequent results in the future.

5.3.2 Final Remarks

This work shows the potential use of LBSNs as data sources to calculate the quality of urban life index. Besides the limitations of LBSN data, results encourage the expansion of research work toward improving the data quality and methods for calculating the index.

The data integration methods and the consolidation into a dataset within the defined data model facilitate the development of techniques for more frequently calculating the urban quality of life indices. As expected, the integrated MUTN dataset obtained better accuracy results than all other datasets used in isolation, indicating a path that can be followed and improved for more accurate results in the future.

Chapter 6

Conclusion

This thesis has presented a multimodal transport network data model representing urban transport infrastructure, considering individual and collective transportation modes. In constructing the MUTN, we used data from official and alternative sources. Among the alternative data sources, several services in which users are responsible for mapping (OpenStreetMap) or registering points of interest (Google Places, Yelp, Foursquare, Facebook) were used. As highlighted throughout this work, these datasets generally have a tradeoff for quality, production cost, and updating criteria. While official data is usually considered better on the first criterion, it is often at a disadvantage on the last two. Hence the opportunity was created to use techniques for integrating these available data to balance these criteria to maximize quality and timeliness and minimize cost.

Methods for integrating the geographic data to compose the urban multimodal transportation network have been proposed, both for the transportation network and for points of interest that can come from different sources and be consolidated into the proposed data model. While the official data was created focusing on the displacement of motor vehicles, the multimodal transportation network built by data integration allows routing using different modes of transportation. The multimodal network created allows better control over the parameters of how the routing is done and the updated status of the data used. This information is not always available when third-party routing services such as Google Maps are used. In addition, using MUTN does not incur additional costs for using the service, which can happen with a large volume of requests, as is common in situations such as building an accessibility matrix with a more extensive set of details for each route.

A case study was conducted to implement a multimodal transportation network for Belo Horizonte from official data and services like OpenStreetMap, Facebook, Foursquare, Google, and Yelp to demonstrate the use of the proposed data model and methods. The data integration allowed the creation of a road network with more data than the separate datasets. The official data was focused on motor vehicle traffic, and to create the multimodal network, semantic information from the OpenStreetMap data was used to develop the road network for pedestrians and cyclists. In addition, official data in the form of GTFS files containing information on bus lines in the metropolitan area were integrated into the MUTN network. Ultimately, data of points of interest from different datasets were used and integrated into the MUTN to create routes where users can drive a part of the stretch and then walk to the destination or integrate with public transport. Testing and validating data integration results is complex due to the lack of a comparable alternative that allows knowing the exact parameters of the routing algorithms and even getting the dataset in a format that will enable its direct comparison with our results. One way found for comparison was to create routes between 80 points representing the closest point to each centroid of a city planning unit in the road network. All MUTN routes for cars, pedestrians, and public transport were compared with their equivalents on Google Maps, considering time and distance. We found differences in route times ranging from 9.5%, 9.9%, and 15.5% for the pedestrian, vehicle, and public transport modes, respectively. Differences of 4.7%, 7.3%, and 18.4% for distances were found. As established in Section 4.11, these differences can be explained, in part, by the availability of sections of the respective road networks. Observing the segments with the most significant disparities confirmed this hypothesis.

Creating a MUTN makes it possible to monitor and analyze many aspects of the dynamics of life in urban areas. To show one of these possibilities, an investigation was conducted to determine the quality of life indicators using data integration methods, and the MUTN was created for Belo Horizonte. One of the challenges of traditionally assessing the quality of life index is gathering data from large organizations with different spatial and temporal granularities, making it difficult to calculate these indices more often. As seen in Section 5.1, the IQVU had the last results made available in 2016, even though it was proposed as an index that could be calculated more frequently. Our proposal to use crowdsourced data integrated into MUTN to estimate quality indicators proved promising and opened possibilities for creating indices that can be determined dynamically and with different spatial and temporal granularities.

Several limitations were identified in the development of this work. The first is that at different points in the data integration process, parameters must be established for the algorithms that often depend on the characteristics of the datasets participating in the integration to obtain good results or to be processed with good performance. For example, integrating datasets with good positional accuracy may use a lower value for searching nearby elements. Other parameters that can be adjusted are those used for semantic matching. Differences in how the datasets are produced (crowdsourced, official) or the language itself may impact the optimal values for, for example, string similarity matching. Another limitation is that it is impossible to update the MUTN from changes in the original datasets automatically. One way to mitigate this would be to create a separate dataset with only the new data and use it to integrate with the MUTN network. Another limitation is the lack of an appropriate interface to apply the integration techniques developed. During the development of the thesis work, the source codes were produced in languages such as Python and R, and it is necessary to use database servers with support for spatial data, such as PostGIS. Unfortunately, there is not yet a simplified and unified interface for using all the methods, requiring specific knowledge of the technologies used.

The limitations identified serve to point out possibilities for future work. Developing techniques or methods to characterize and analyze the datasets that will be integrated concerning their level of detail may facilitate finding optimal parameters for the integration methods and make them more efficient. Another possibility for future work is to create mechanisms for incremental updates of the data integration from updates of the original datasets without re-running the entire integration process. Also needed is the implementation of a user-friendly interface for the execution of the entire integration process that allows the adjustment of parameters and visualization of intermediate and final results for better control of the entire process. The interface implementation should be accompanied by a refactoring of the source code that can serve to find possible points that can be better implemented to improve the process's overall performance.

The use of the MUTN allows various applications to use as a basis for analyzing problems pertinent to the population and public administration in cities. One work that can be developed is the creation of a quality-of-life index that can be determined exclusively with data available in open platforms. The MUTN could be used as a base data model to integrate data from the selected sources. The capabilities of the multimodal network and the other integrated data would allow quality indexes to be determined with lower spatial granularity and higher frequency. Instead of indexes for an entire planning unit, as in the IQVU in Belo Horizonte, indexes could be determined for neighborhoods or even blocks, for example. Since "quality of life" can be subjective, the MUTN could facilitate the calculation of dynamic indexes in which the user could establish the weights for the indicators of interest. For example, a family looking for a place to live will give more weight to safety and education indicators. In comparison, a young single professional may give more weight to culture and sports.

References

[1] Rifaat Abdalla. Geospatial Data Integration. In *Introduction to Geospatial Information and Communication Technology (GeoICT)*, pages 105–124. Springer Nature, 2016.

[2] Ehsan Abdolmajidi, Ali Mansourian, Julian Will, and Lars Harrie. Matching Authority and VGI Road Networks Using an Extended Node-Based Matching Algorithm. *Geo-spatial Information Science*, 18(2-3):65–80, July 2015.

[3] T. M. Adams, N. A. Koncz, and A. P. Vonderohe. *Guidelines for the Implementation of Multimodal Transportation Location Referencing Systems.* National Academy Press, 2001.

[4] Maythm Al-Bakri and David Fairbairn. Assessing similarity matching for possible integration of feature classifications of geospatial data from official and informal sources. *International Journal of Geographical Information Science*, 26(8):1437–1456, August 2012.

[5] Helmut Alt, Christian Knauer, and Carola Wenk. Comparison of distance measures for planar curves. *Algorithmica*, 38(1):45–58, October 2003.

[6] Javad Jomehpour Chahar Aman and Janille Smith-Colin. Transit deserts: Equity analysis of public transit accessibility. *Journal of Transport Geography*, 89:102869, December 2020.

[7] Suchith Anand, Jeremy Morley, Wenchao Jiang, Heshan Du, and Glen Hart. When worlds collide : combining Ordnance Survey and Open Street Map data. In *Proceedings of Association for Geographic Information GeoCommunity Conference*, Stratford-Upon-Avon, UK, October 2010.

[8] V. Antoniou. *User Generated Spatial Content: An Analysis of the Phenomenon and its Challenges for Mapping Agencies.* Phd, University College London, 2011.

[9] Aaron Antrim, Sean J. Barbeau, and Others. The Many Uses of GTFS Data - Opening the Door Transit and Multimodal Application. *Location-Aware Information Systems Laboratory at the University of South Florida*, pages 1–24, 2013.

[10] Philippe Apparicio, Marie-Soleil Cloutier, and Richard Shearmur. The case of Montréal's missing food deserts: Evaluation of accessibility to food supermarkets. *International Journal of Health Geographics*, 6(1):4, 2007.

[11] E.M. Arkin, L.P. Chew, D.P. Huttenlocher, K. Kedem, and J.S.B. Mitchell. An efficiently computable metric for comparing polygonal shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(3):209–216, March 1991.

[12] Jamal Jokar Arsanjani, Christopher Barron, Mohammed Bakillah, and Marco Helbich. Assessing the Quality of OpenStreetMap Contributors together with their Contributions. In *16th AGILE International Conference on Geographic Information Science.*, pages 14–17, 2013.

[13] Jamal Jokar Arsanjani, Peter Mooney, Marco Helbich, and Alexander Zipf. An Exploration of Future Patterns of the Contributions to OpenStreetMap and Development of a Contribution Index. *Transactions in GIS*, 19(6):896–914, March 2015.

[14] Aamer Ather. A quality analysis of openstreetmap data. mathesis, University College London, May 2009.

[15] Hannah M. Badland, Jerome N. Rachele, Rebecca Roberts, and Billie Giles-Corti. Creating and applying public transport indicators to test pathways of behaviours and health through an urban transport framework. *Journal of Transport & Health*, February 2017.

[16] Yoonsik Bang, Jaebin Lee, Jiyoung Kim, and Kiyun Yu. An iterative algorithm for matching two road network data sets. In *MAKE S. ASPRS/MAPPS Fall Conference, The International Archives of the Photogrammetry: Remote Sensing and Spatial Information Sciences. San Antonio:[sn]*, pages 19–25, 2009.

[17] Christopher Barron, Pascal Neis, and Alexander Zipf. A comprehensive framework for intrinsic OpenStreetMap quality analysis. *Transactions in GIS*, 18(6):877–895, December 2013.

[18] M. Batty, K. W. Axhausen, F. Giannotti, A. Pozdnoukhov, A. Bazzani, M. Wachowicz, G. Ouzounis, and Y. Portugali. Smart cities of the future. *The European Physical Journal Special Topics*, 214(1):481–518, November 2012.

[19] Richard A Becker, Ramon Caceres, Karrie Hanson, Ji Meng Loh, Simon Urbanek, Alexander Varshavsky, and Chris Volinsky. A tale of one city: Using cellular network data for urban planning. *IEEE Pervasive Computing*, 10(4):18–26, April 2011.

[20] Catriel Beeri, Yaron Kanza, Eliyahu Safra, and Yehoshua Sagiv. Object fusion in geographic information systems. *Proceedings of the 30th VLDB Conference*, pages 816–827, 2004.

[21] Alberto Belussi, Barbara Catania, Eliseo Clementini, and Elena Ferrari. Spatial data on the web: Issues and challenges. In *Spatial Data on the Web*, pages 1–12. Springer Berlin Heidelberg, 2007.

[22] Joschka Bischoff and Kai Nagel. Integrating explicit parking search into a transport simulation. *Procedia Computer Science*, 109:881–886, 2017.

[23] Samuel D. Blanchard and Paul Waddell. UrbanAccess. *Transportation Research Record: Journal of the Transportation Research Board*, 2653(1):35–44, January 2017.

[24] Jens Bleiholder and Felix Naumann. Data fusion. *ACM Computing Surveys (CSUR)*, 41(1):1, 2009.

[25] Geneviève Boisjoly and Ahmed El-Geneidy. Daily fluctuations in transit and job availability: A comparative assessment of time-sensitive accessibility measures. *Journal of Transport Geography*, 52:73–81, April 2016.

[26] Omar Boucelma, Mehdi Essid, and Yassine Lassoued. A quality-enabled spatial integration system. In *Spatial Data on the Web*, pages 133–157. Springer Berlin Heidelberg, 2007.

[27] Burak Boyac and Nikolas Geroliminis. Estimation of the Network Capacity for Multimodal Urban Systems. *Procedia - Social and Behavioral Sciences*, 16:803–813, 2011.

[28] Patrick Brosi. Real-Time Movement Visualization of Public Transit Data. mathesis, University of Freiburg, 2014.

[29] Maria Antonia Brovelli, Marco Minghini, Monia Molinari, and Peter Mooney. Towards an automated comparison of OpenStreetMap with authoritative road datasets. *Transactions in GIS*, March 2016.

[30] Agustina Buccella, Alejandra Cechich, and Pablo Fillottrani. Ontology-driven geographic information integration: A survey of current approaches. *Computers & Geosciences*, 35(4):710–723, 2009.

[31] Nama Raj Budhathoki, Bertram (Chip) Bruce, and Zorica Nedovic-Budic. Reconceptualizing the role of the user of spatial data infrastructure. *GeoJournal*, 72(3-4):149–160, July 2008.

[32] Matthias Butenuth, Guido v. Gösseln, Michael Tiedge, Christian Heipke, Udo Lipeck, and Monika Sester. Integration of heterogeneous geospatial data in a federated database. *ISPRS Journal of Photogrammetry and Remote Sensing*, 62(5):328–346, October 2007.

[33] J. Allison Butler. *Designing Geodatabases for Transportation.* ESRI PR, August 2008.

[34] Francesco Calabrese, Mi Diao, Giusy Di Lorenzo, Joseph Ferreira, and Carlo Ratti. Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. *Transportation Research Part C: Emerging Technologies*, 26:301–313, January 2013.

[35] Francesco Calabrese, Giusy Di Lorenzo, Liang Liu, and Carlo Ratti. Estimating origin-destination flows using mobile phone location data. *IEEE Pervasive Computing*, 10(4):36–44, April 2011.

[36] Silvana Camboim, João Bravo, and Claudia Sluter. An Investigation into the Completeness of, and the Updates to, OpenStreetMap Data in a Heterogeneous Area in Brazil. *ISPRS International Journal of Geo-Information*, 4(3):1366–1388, August 2015.

[37] Ching-Chien Chen and Craig A. Knoblock. Conflation of geospatial data. In *Encyclopedia of GIS*, pages 133–140. Springer US, 2008.

[38] Daniel Chen, Anne Driemel, Leonidas J. Guibas, Andy Nguyen, and Carola Wenk. Approximate map matching with respect to the fréchet distance. In *2011 Proceedings of the Thirteenth Workshop on Algorithm Engineering and Experiments (ALENEX)*, pages 75–83. Society for Industrial and Applied Mathematics, January 2011.

[39] Hainan Chen and Volker Walter. Hierarchical quality inspection of spatial data by data integration. In *ASPRS 2010 Annual Conference*, pages 94–105, San Diego, USA, 2010.

[40] Shaopei Chen, Jianjun Tan, Christophe Claramunt, and Cyril Ray. Multi-scale and multi-modal GIS-t data model. *Journal of Transport Geography*, 19(1):147–161, January 2011.

[41] Vincent Chin, Mariam Jaafar, Jason Moy, Maria Phong, Shenya Wang, Matthew McDonnell, and Irfan Prawiradinata. Unlocking cities: The impact of redesharing in southeast asia and beyond. Technical report, The Boston Consulting Group, 2017.

[42] Peter Christen. *Data Matching - Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection.* Data-Centric Systems and Applications. Springer Berlin Heidelberg, 2012.

[43] Błażej Ciepłuch, Peter Mooney, Ricky Jacob, Jianghua Zheng, and Adam C Winstanely. Assessing the quality of open spatial data for mobile location-based services research and applications. *Archiwum Fotogrametrii, Kartografii i Teledetekcji*, 22:105–116, 2011.

[44] David J. Coleman. Potential contributions and challenges of VGI for conventional topographic base-mapping programs. In *Crowdsourcing Geographic Knowledge*, pages 245–263. Springer Netherlands, June 2012.

[45] David J Coleman, Yola Georgiadou, Jeff Labonte, Earth Observation, and Natural Resources Canada. Volunteered Geographic Information : The Nature and Motivation of Produsers. *International Journal of Spatial Data Infrastructures Research*, 4(4):332–358, 2009.

[46] Antony Kyle Cooper. *An exposition of the nature of volunteered geographical information and its suitability for integration into spatial data infrastructures.* phdthesis, University of Pretoria, 2016.

[47] Paolo Crucitti, Vito Latora, and Sergio Porta. Centrality in networks of urban streets. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 16(1):015113, March 2006.

[48] Jan De Leeuw, Mohammed Said, Lapezoh Ortegah, Sonal Nagda, Yola Georgiadou, and Mark DeBlois. An Assessment of the Accuracy of Volunteered Road Map Production in Western Kenya. *Remote Sensing*, 3(12):247–256, February 2011.

[49] Merkebe Getachew Demissie, Gonçalo Homem de Almeida Correia, and Carlos Bento. Exploring cellular network handover information for urban mobility analysis. *Journal of Transport Geography*, 31:164–170, July 2013.

[50] R. Devillers, Y. Bédard, R. Jeansoulin, and B. Moulin. Towards spatial data quality information analysis tools for experts assessing the fitness for use of spatial data. *International Journal of Geographical Information Science*, 21(3):261–282, March 2007.

[51] Thomas Devogele, Christine Parent, and Stefano Spaccapietra. On spatial database integration. *International Journal of Geographical Information Science*, 12(4):335–352, June 1998.

[52] Thomas Devogele, Jenny Trevisan, and Laurent Raynal. Building a multi-scale database with scale-transition relationships. In *International symposium on spatial data handling*, pages 337–351, 1996.

[53] Hong-Hai Do and Erhard Rahm. Coma: A system for flexible combination of schema matching approaches. In *Proceedings of the 28th international conference on Very Large Data Bases*, pages 610–621. VLDB Endowment, 2002.

[54] Michael W. Dobson. VGI as a compilation tool for navigation map databases. In *Crowdsourcing Geographic Knowledge*, pages 307–327. Springer Netherlands, June 2012.

[55] H.E. Dongcai. A study on theory and method of spatial vector data conflation. *Research Journal of Applied Sciences, Engineering and Technology*, 5(2):563–567, January 2013.

[56] Helen Dorn, Tobias Törnros, and Alexander Zipf. Quality Evaluation of VGI Using Authoritative Data—A Comparison with Land Use Data in Southern Germany. *ISPRS International Journal of Geo-Information*, 4(3):1657–1671, September 2015.

[57] Heshan Du. *Matching disparate geospatial datasets and validating matches using spatial logic.* Phd, The University of Nottingham, 2015.

[58] Heshan Du, Natasha Alechina, Michael Jackson, and Glen Hart. A method for matching crowd-sourced and authoritative geospatial data. *Transactions in GIS*, 21(2):406–427, May 2016.

[59] Heshan Du, Suchith Anand, Natasha Alechina, Jeremy Morley, Glen Hart, Didier Leibovici, Mike Jackson, and Mark Ware. Geospatial information integration for authoritative and crowd sourced road vector data. *Transactions in GIS*, 16(4):455–476, May 2012.

[60] Yubin Duan and Jie Wu. Spatial-Temporal Inventory Rebalancing for Bike Sharing Systems With Worker Recruitment. *IEEE Transactions on Mobile Computing*, 21(3):1081–1095, March 2022.

[61] M. Duckham and M. Worboys. An algebraic approach to automated geospatial information fusion. *International Journal of Geographical Information Science*, 19(5):537–557, May 2005.

[62] Kenneth J. Dueker and J. Allison Butler. Gis-t enterprise data model with suggested implementation choices. *Journal of the Urban and Regional Information Systems Association*, 10:12–36, 1997.

[63] Khalid L A El-Ashmawy. Testing the positional accuracy of OpenStreetMap data for mapping applications. *Geodesy and Cartography*, 42(1):25–30, January 2016.

[64] ESRI. Transportation data model. Online, 2016.

[65] Hongchao Fan, Bisheng Yang, Alexander Zipf, and Adam Rousell. A polygon-based approach for matching OpenStreetMap road networks with regional transit authority data. *International Journal of Geographical Information Science*, 30(4):748–764, November 2015.

[66] Hongchao Fan, Alexander Zipf, Qing Fu, and Pascal Neis. Quality assessment for building footprints data on OpenStreetMap. *International Journal of Geographical Information Science*, 28(4):700–719, January 2014.

[67] Steven Farber, Melinda Z. Morang, and Michael J. Widener. Temporal variability in transit-based accessibility to supermarkets. *Applied Geography*, 53:149–159, September 2014.

[68] Steven Farber, Benjamin Ritter, and Liwei Fu. Space-time mismatch between transit service and observed travel patterns in the wasatch front, utah: A social equity perspective. *Travel Behaviour and Society*, 4:40–48, May 2016.

[69] V. O. Fernandes, E. N. Elias, and A. Zipf. Integration of authoritative and volunteered geographic information for updating urban mapping: Challenges and potentials. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLIII-B4-2020:261–268, August 2020.

[70] FIRJAN. Os custos da mobilidade nas regiões metropolitanas do Rio de Janeiro e São Paulo. techreport, FIRJAN, July 2014. Accessed on 2022-03-23.

[71] Stefan Foell, Santi Phithakkitnukoon, and Gerd Kortuem. Catch me if you can: predicting mobility patterns of public transport users. In *17th International IEEE Conference on Intelligent Transportation Systems (ITSC 2014)*, Qingdao, China, 2014.

[72] Mohammad Forghani and Mahmoud Delavar. A quality study of the OpenStreetMap dataset for tehran. *ISPRS International Journal of Geo-Information*, 3(2):750–763, May 2014.

[73] Foursquare. About Us, 2016. https://pt.foursquare.com/about. Accessed on 2016-07-14.

[74] Ayelet Gal-Tzur, Susan M. Grant-Muller, Einat Minkov, and Silvio Nocera. The Impact of Social Media Usage on Transport Policy: Issues, Challenges and Recommendations. *Procedia - Social and Behavioral Sciences*, 111:937–946, February 2014.

[75] Irene Garcia-Martí, Raul Zurita-Milla, Arno Swart, Kees C. van den Wijngaard, Arnold J.H. van Vliet, Sita Bennema, and Margriet Harms. Identifying environmental and human factors associated with tick bites using volunteered reports and frequent pattern mining. *Transactions in GIS*, 21(2):277–299, May 2016.

[76] Karst T. Geurs and Bert van Wee. Accessibility Evaluation of Land-Use and Transport Strategies: Review and Research Directions. *Journal of Transport Geography*, 12(2):127–140, June 2004.

[77] Jorge Gil. Building a multimodal urban network model using OpenStreetMap data for the analysis of sustainable accessibility. In Jamal Jokar Arsanjani, Alexander Zipf, Peter Mooney, and Marco Helbich, editors, *OpenStreetMap in GIScience: Experiences, Research, Applications*, Lecture Notes in Geoinformation and Cartography, pages 229–251. Springer International Publishing, 2015.

[78] Jorge Gil. Urban modality. *A+BE — Architecture and the Built Environment*, 6(1):1–434, 2016.

[79] Jean François Girres and Guillaume Touya. Quality Assessment of the French OpenStreetMap Dataset. *Transactions in GIS*, 14(4):435–459, August 2010.

[80] Fausto Giunchiglia, Pavel Shvaiko, and Mikalai Yatskevich. Semantic schema matching. In *Lecture Notes in Computer Science*, pages 347–365. Springer Berlin Heidelberg, 2005.

[81] Michael F. Goodchild. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4):211–221, November 2007.

[82] Michael F. Goodchild and Gary J. Hunter. A simple positional accuracy measure for linear features. *International Journal of Geographical Information Science*, 11(3):299–306, April 1997.

[83] Gv Gösseln and M. Sester. Integration of geoscientific data sets and the german digital map using a matching approach. *International Archives of Photogrammetry and Remote Sensing*, 35:1249–1254, 2004.

[84] Anita Graser. Integrating open spaces into OpenStreetMap routing graphs for realistic crossing behaviour in pedestrian navigation. *GI_Forum*, 1:217–230, 2016.

[85] Thomas R Gruber. Toward principles for the design of ontologies used for knowledge sharing? *International journal of human-computer studies*, 43(5-6):907–928, 1995.

[86] Ji-Hong Guan, Shui-Geng Zhou, Jun-Peng Chen, Xiao-Long Chen, Yang An, Wei Yu, Rong Wang, and Xu-Jun Liu. Ontology-based GML schema matching for spatial

information integration. In *Proceedings of the 2003 International Conference on Machine Learning and Cybernetics (IEEE Cat. No.03EX693)*. IEEE, 2003.

[87] Bin Guo, Zhiwen Yu, Daqing Zhang, and Xingshe Zhou. Cross-community sensing and mining. *IEEE Communications Magazine*, 52(8):144–152, August 2014.

[88] Yuval Hadas. Assessing public transport systems connectivity based on google transit data. *Journal of Transport Geography*, 33:105–116, December 2013.

[89] Mordechai Haklay. How good is volunteered geographical information? a comparative study of OpenStreetMap and ordnance survey datasets. *Environment and Planning B: Planning and Design*, 37(4):682–703, August 2010.

[90] Mordechai (Muki) Haklay, Sofia Basiouka, Vyron Antoniou, and Aamer Ather. How Many Volunteers Does it Take to Map an Area Well? The Validity of Linus' Law to Volunteered Geographic Information. *The Cartographic Journal*, 47(4):315–322, November 2010.

[91] Muki Haklay. Citizen science and volunteered geographic information: Overview and typology of participation. In *Crowdsourcing Geographic Knowledge*, pages 105–122. Springer Netherlands, June 2012.

[92] Patrick A. V. Hall and Geoff R. Dowling. Approximate string matching. *ACM Computing Surveys*, 12(4):381–402, December 1980.

[93] Francis Harvey, Adam Iwaniak, Serena Coetzee, and Anthony Cooper. SDI Past, Present and Future: A Review and Status Assessment. In Abbas Rajabifard and David Coleman, editors, *Spatially enabling government, industry and citizens. Research and development perspectives*, chapter 2, pages 23–38. GSDI Association Press, Needham, MA, USA, 2012.

[94] Peyman Hashemi and Rahim Ali Abbaspour. Assessment of logical consistency in OpenStreetMap based on the spatial similarity concept. In *Lecture Notes in Geoinformation and Cartography*, pages 19–36. Springer International Publishing, 2015.

[95] J T Hastings. Automated conflation of digital gazetteer data. *International Journal of Geographical Information Science*, 22(10):1109–1127, 2008.

[96] Suining He and Kang G. Shin. Information Fusion for (Re)Configuring Bike Station Networks With Crowdsourcing. *IEEE Transactions on Knowledge and Data Engineering*, 34(2):736–752, February 2022.

[97] Marco Helbich, Björn Schadenberg, Julian Hagenauer, and Maartje Poelman. Food deserts? healthy food access in amsterdam. *Applied Geography*, 83:1–12, June 2017.

[98] Sascha Hoogendoorn-Lanser, Rob van Nes, and Serge Hoogendoorn. Modeling Transfers in Multimodal Trips: Explaining Correlations. *Transportation Research Record: Journal of the Transportation Research Board*, 1985(1):144–153, January 2006.

[99] Yong Huh, Kiyun Yu, and Joon Heo. Detecting conjugate-point pairs for map alignment between two polygon datasets. *Computers, Environment and Urban Systems*, 35(3):250–262, 2011.

[100] Ángel Ibeas, Luigi dell'Olio, Borja Alonso, and Olivia Sainz. Optimizing bus stop spacing in urban areas. *Transportation Research Part E: Logistics and Transportation Review*, 46(3):446–458, May 2010.

[101] IBGE. Instituto Brasileiro de Geografia e Estatística - Censo 2010, 2022.

[102] M a Ismail and M N Said. Integration of geospatial multi-mode transportation Systems in Kuala Lumpur. *IOP Conference Series: Earth and Environmental Science*, 20:012027, June 2014.

[103] ISO. ISO14825:1996 - geographic data files (GDF). techreport, International Organization for Standardization, December 1996.

[104] ISO. ISO19113:2002 - geographic information–quality principles. Standard, International Organization for Standardization, 2002.

[105] ISO. ISO14825:2004 - intelligent transport systems – geographic data files (GDF) – overall data specification. Standard, International Organization for Standardization, February 2004.

[106] ISO. ISO14825:2011 intelligent transport systems – geographic data files (GDF) – GDF5.0. Standard, International Organization for Standardization, July 2011.

[107] Ben Jestico, Trisalyn Nelson, and Meghan Winters. Mapping ridership using crowdsourced cycling data. *Journal of Transport Geography*, 52:90–97, April 2016.

[108] Junfeng Jiao. Identifying transit deserts in major Texas cities where the supplies missed the demands. *The Journal of Transport and Land Use*, pages 1–12, 2017.

[109] Junfeng Jiao and Maxwell Dillivan. Transit Deserts: The Gap between Demand and Supply. *Journal of Public Transportation*, 16(3):23–39, 2013.

[110] Junfeng Jiao, Anne V. Moudon, Jared Ulmer, Philip M. Hurvitz, and Adam Drewnowski. How to Identify Food Deserts: Measuring Physical and Economic Access to Supermarkets in King County, Washington. *American Journal of Public Health*, 102(10):e32–e39, October 2012.

[111] Steffen John, Stefan Hahmann, Adam Rousell, Marc-O Löwner, and Alexander Zipf. Deriving incline values for street networks from voluntarily collected GPS traces. *Cartography and Geographic Information Science*, 44(2):152–169, June 2016.

[112] Mohsen Kalantari and Veha La. Assessing OpenStreetMap as an open property map. In *Lecture Notes in Geoinformation and Cartography*, pages 255–272. Springer International Publishing, 2015.

[113] Hassan a. Karimi and Piyawan Kasemsuppakorn. Pedestrian network map generation approaches and recommendation. *International Journal of Geographical Information Science*, 27(5):947–962, May 2013.

[114] Piyawan Kasemsuppakorn and Hassan A Karimi. Pedestrian network data collection through location-based social networks. In *Proceedings of the 5th International ICST Conference on Collaborative Computing: Networking, Applications, Worksharing*, pages 1–9. IEEE, 2009.

[115] Leyla Kazemi and Cyrus Shahabi. GeoCrowd: Enabling Query Answering with Spatial Crowdsourcing. In *Proceedings of the 20th International Conference on Advances in Geographic Information Systems - SIGSPATIAL '12*, page 189, New York, New York, USA, 2012. ACM Press.

[116] Khoa Tran, Edward Hillsman, Sean Barbeau, and Miguel A. Labrador. Go_Sync - A framework to synchronize crownd-sourced mapping contributions from online communities and transit agency bus stop inventories. In *ITS World Congress*, pages 1–17, Orlando, Florida, USA, 2011.

[117] Giti Khoshamooz and Mohammad Taleai. Multi-Domain User-Generated Content Based Model to Enrich Road Network Data for Multi-Criteria Route Planning. *Geographical Analysis*, pages 1–29, April 2017.

[118] Birgit Kieler, Wei Huang, Jan-Henrik Haunert, and Jie Jiang. Matching river datasets of different scales. In *Advances in GIScience*, pages 135–154. Springer Berlin Heidelberg, 2009.

[119] Jung Ok Kim, Kiyun Yu, Joon Heo, and Won Hee Lee. A new method for matching objects in two different geospatial datasets based on the geographic context. *Computers and Geosciences*, 36(9):1115–1122, 2010.

[120] Rob Kitchin. The real-time city? Big Data and Smart Urbanism. *GeoJournal*, 79(1):1–14, February 2014.

[121] Kittelson & Assoc, Inc., Parsons Brinckerhoff, Inc., KFH Group, Inc., Texas A&M Transportation Institute, and Arup. Transit Capacity and Quality of Service Manual. Technical report, Transportation Research Board, Washington, DC, USA, 2013.

[122] Brian Klinkenberg. The true cost of spatial data in canada. *The Canadian Geographer/Le G?ographe canadien*, 47(1):37–49, March 2003.

[123] Mohammad Kolahdouzan, Ching-Chien Chen, Cyrus Shahabi, and Craig a. Knoblock. GeoMatchMaker: automatic and efficient matching of vector data with spatial attributes in unknown geometry systems. In *Proceedings of UCGIS 2005 Summer Assembly*, 2005.

[124] Nicholas A. Koncz and Teresa M. Adams. A Data Model for Multi-dimensional Transportation Location Referencing Systems. *Urban and Regional Information Systems Association Journal*, 14(2):27–41, 2002.

[125] Nicholas A. Koncz and Teresa M. Adams. A data model for multi-dimensional transportation applications. *International Journal of Geographical Information Science*, 16(6):551–569, September 2002.

[126] Qing-Jie Kong, Qiankun Zhao, Chao Wei, and Yuncai Liu. Efficient traffic state estimation for large-scale urban road networks. *IEEE Transactions on Intelligent Transportation Systems*, 14(1):398–407, March 2013.

[127] Thomas Koukoletsos, Mordechai Haklay, and Claire Ellul. Assessing Data Completeness of VGI through an Automated Matching Procedure for Linear Data. *Transactions in GIS*, 16(4):477–498, 2012.

[128] Billy Pik Lik Lau, Sumudu Hasala Marakkalage, Yuren Zhou, Naveed Ul Hassan, Chau Yuen, Meng Zhang, and U. Xuan Tan. A survey of data fusion in smart city applications. *Information Fusion*, 52:357–374, December 2019.

[129] Hye Kyung Lee, Junfeng Jiao, and Seung Jun Choi. Identifying spatiotemporal transit deserts in Seoul, South Korea. *Journal of Transport Geography*, 95:103145, July 2021.

[130] Maurício Borges Lemos, Otávio de Avelar Esteves, Rodrigo Ferreira Simões, et al. A methodology for making an urban quality of life index (in Portuguese). *Nova Economia*, 5(2):157–176, 1995.

[131] Daniela Antunes Lessa, Carlos Lobo, and Leandro Cardoso. Accessibility and urban mobility by bus in Belo Horizonte/Minas Gerais – Brazil. *Journal of Transport Geography*, 77:1–10, May 2019.

[132] Jun Li, Huayang Dai, Zhang o Yuan, Qiming Qin, and Hongbo Jiang. Extracting map information from trajectory and social media data. In *2015 2^{nd} IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services (ICSDM)*. IEEE, July 2015.

[133] L. Li and Michael F. Goodchild. Automatically and accurately matching objects in geospatial datasets. In *Proceedings of Joint International Conference on Theory, Data Handling and Modelling in GeoSpatial Information Science*, pages 98–103, Hong Kong, China, 2010.

[134] Linna Li and Michael F. Goodchild. An optimisation model for linear feature matching in geographical data conflation. *International Journal of Image and Data Fusion*, 2(April 2012):309–328, 2011.

[135] James J.H. Liou, Chao-Che Hsu, and Yun-Shen Chen. Improving transportation service quality based on information fusion. *Transportation Research Part A: Policy and Practice*, 67:225–239, September 2014.

[136] Lu Liu. *Data model and algorithms for multimodal route planning with transportation networks*. phdthesis, TECHNISCHE UNIVERSITÄT MÜNCHEN, 2011.

[137] Pengyuan Liu and Filip Biljecki. A review of spatially-explicit GeoAI applications in Urban Geography. *International Journal of Applied Earth Observation and Geoinformation*, 112:102936, August 2022.

[138] Tong Liu, Yu Zheng, Lubin Liu, Yanchi Liu, and Yanmin Zhu. Methods for sensing urban noises. Technical report, Microsoft, May 2014.

[139] Paul A. Longley, Michael F. Goodchild, David J. Maguire, and David W. Rhind. *Geographic Information Systems and Science*. Wiley, 2005.

[140] Xuechen Luan. A structure-based approach for matching road junctions with different coordinate systems. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, I-4:41–46, July 2012.

[141] Ina Ludwig, Angi Voss, and Maike Krause-Traudes. A comparison of the street networks of navteq and OSM in germany. In *Lecture Notes in Geoinformation and Cartography*, pages 65–84. Springer Berlin Heidelberg, 2011.

[142] Deelesh Mandloi and Jean-Claude Thill. Object-oriented data modeling of an indoor/outdoor urban transportation network and route planning analysis. In *GeoJournal Library*, pages 197–220. Springer Science Business Media, 2010.

[143] Daniela Mantel and Udo Lipeck. Matching cartographic objects in spatial databases. *Proceedings of XXth ISPRS Congress*, pages 172–176, 2004.

[144] Robert W Marans. Quality of urban life & environmental sustainability studies: Future linkage opportunities. *Habitat International*, 45:47–52, 2015.

[145] Karel Martens. Promoting bike-and-ride: The Dutch experience. *Transportation Research Part A: Policy and Practice*, 41(4):326–338, May 2007.

[146] Irene Garcia Martí, Luis E. Rodríguez, Mauricia Benedito, Sergi Trilles, Arturo Beltrán, Laura Díaz, and Joaquín Huerta. Mobile application for noise pollution monitoring through gamification techniques. In *Lecture Notes in Computer Science*, pages 562–571. Springer Nature, 2012.

[147] Ariane Mascret, Thomas Devogele, Iwan Le Berre, and Alain Hénaff. Coastline matching process based on the discrete fréchet distance. In *Progress in Spatial Data Handling*, pages 383–400. Springer Berlin Heidelberg, 2006.

[148] A. Paolo Masucci and Carlos Molinero. Robustness and closeness centrality for self-organized and planned cities. *The European Physical Journal B*, 89(2), February 2016.

[149] Guilherme Vezula Mateveli, Natália Gonçalvez Machado, Mirella M. Moro, and Clodoveu Augusto Davis Jr. Taxonomia e Desafios de Recomendação para Coleta de Dados Geográficos por Cidadãos. In *Proceedings of the 30th Brazilian Symposium on Databases*, pages 105–110, Petrópolis, RJ, 2015.

[150] Bibiana McHugh. Pioneering open data standards: The gtfs story. In Brett Goldstein and Lauren Dyson, editors, *Beyond Transparency - Open Data and Future of Civic Innovation*, chapter 10, pages 125–136. Code for America Press, San Francisco, CA, USA, 2013.

[151] G McKenzie, K Janowicz, and B Adams. A weighted multi-attribute method for matching user-generated Points of Interest. *Cartography and Geographic Information Science*, 41(2):125–137, 2014.

[152] Tong Meng, Xuyang Jing, Zheng Yan, and Witold Pedrycz. A survey on machine learning for data fusion. *Information Fusion*, 57:115–129, May 2020.

[153] Harvey J. Miller and Shih-Lung Shaw. *GIS-T Data Models.* Oxford Univ Pr, November 2001.

[154] Harvey J Miller and Shih-Lung Shaw. Geographic Information Systems for Transportation in the 21st Century. *Geography Compass*, 9(4):180–189, April 2015.

[155] Hossein Mohammadi. *The Integration of Multi-source Spatial Datasets in the Context of SDI Initiatives.* Phd, University of Melbourne, 2008.

[156] Hossein Mohammadi, Abbas Rajabifard, and Ian P. Williamson. Development of an interoperable tool to facilitate spatial data integration in the context of SDI.

International Journal of Geographical Information Science, 24(4):487–505, March 2010.

[157] Sébastien Mustière and Thomas Devogele. Matching networks with different levels of detail. *GeoInformatica*, 12(4):435–453, 2008.

[158] Maria Inês Pedrosa Nahas. *Theoretical Basis, Calculation Methodology and Applicability of Intra-Urban Indicators in the Municipal Management of the Quality of Urban Life in Large Cities: The Case of Belo Horizonte (in Portuguese).* Phd thesis, Universidade Federal de São Carlos, July 2002.

[159] Pascal Neis, Dennis Zielstra, and Alexander Zipf. The Street Network Evolution of Crowdsourced Maps: OpenStreetMap in Germany 2007-2011. *Future Internet*, 4(4):1–21, December 2012.

[160] Pascal Neis, Dennis Zielstra, and Alexander Zipf. Comparison of Volunteered Geographic Information Data Contributions and Community Development for Selected World Regions. *Future Internet*, 5(2):282–300, June 2013.

[161] Rob Van Nes. *Design of multimodal transport networks: A hierarchical approach.* IOS Press, Delft, Netherlands, 2002.

[162] Kenji Nozaki, Teruhisa Hochin, and Hiroki Nomiya. Semantic schema matching for string attribute with word vectors. In *2019 6th International Conference on Computational Science/Intelligence and Applied Informatics (CSII).* IEEE, May 2019.

[163] OGC. OGC OWS-9 Cross Community Interoperability (CCI) Conflation with Provenance Engineering Report. Technical report, OGC, 2013.

[164] A Olteanu, Sébastien Mustière, and Anne Ruas. Matching imperfect spatial data. In *Caetano, M., Painho, M.(Es.), Proceedings of 7th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences. Lisbon*, pages 7–9. Lisboa:Instituto Geográfico Português, 2006.

[165] Ana-Maria Olteanu. Matching geographical data using the Theory of Evidence. *Proceedings of 20th*, 201(Gesbert 2005):5–9, 2007.

[166] L. M. Ostresh Jr. SPA: A shortest path algorithm. In G. Rushton, M. F. Goodchild, and M. Ostresh Jr., editors, *Computer Programs for location-allocation problems*, pages 141–162. Department of Geograpy. The University of Iowa., Iowa City, 1973.

[167] Bei Pan, Yu Zheng, David Wilkie, and Cyrus Shahabi. Crowd sensing of traffic anomalies based on human mobility and social media. In *Proceedings of the 21st*

ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - SIGSPATIAL'13. Association for Computing Machinery (ACM), 2013.

[168] Jeffrey Partyka, Pallabi Parveen, Latifur Khan, B. Thuraisingham, and Shashi Shekhar. Enhanced geographically typed semantic schema matching. *Journal of Web Semantics*, 9(1):52–70, March 2011.

[169] Guo Peng and Sun Yanling. Study on urban traffic incident GIS-T data model. In *2008 International Workshop on Education Technology and Training & 2008 International Workshop on Geoscience and Remote Sensing*. IEEE, December 2008.

[170] Rafael Moraes Henrique Pereira and Tim Schwanen. Tempo de deslocamento casa-trabalho no Brasil (1992-2009): diferenças entre regiões metropolitanas, níveis de renda e sexo. techreport, IPEA, 2013.

[171] Kenneth Perrine, Alireza Khani, and Natalia Ruiz-Juri. Map-Matching Algorithm for Applications in Multimodal Transportation Network Modeling. *Transportation Research Record: Journal of the Transportation Research Board*, 2537:62–70, January 2015.

[172] Lawrence Philips. Hanging on the Metaphone. *Computer Language*, 7(12 (December)):39, 1990.

[173] Gustavo da Rocha Barreto Pinto, Sergio Palma J. Medeiros, Jano Moreira de Souza, Julia Celia Mercedes Strauch, and Carlete Rosana Ferreira Marques. Spatial data integration in a collaborative design framework. *Communications of the ACM*, 46(3):86–90, March 2003.

[174] Tatiana Pontes, Marisa Vasconcelos, Jussara Almeida, Ponnurangam Kumaraguru, and Virgilio Almeida. We know where you live. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing - UbiComp'12*. ACM Press, 2012.

[175] Claire Prudhomme, Timo Homburg, Jean-Jacques Ponciano, Frank Boochs, Christophe Cruz, and Ana-Maria Roxin. Interpretation and automatic integration of geospatial data into the semantic web. *Computing*, 102(2):365–391, February 2020.

[176] Srinivas S. Pulugurtha and Mahesh Agurla. Assessment of Models to Estimate Bus-Stop Level Transit Ridership using Spatial Modeling Methods. *Journal of Public Transportation*, 15(1):33–52, March 2012.

[177] Zhao Qiankun, Kong Qingjie, Xia Yingjie, and Liu Yuncai. An improved method for estimating urban traffic state via probe vehicle tracking. In *Proceedings of the 2011 30th Chinese In Control*, pages 5586–5590. IEEE, 2011.

[178] Mohammed A. Quddus, Washington Y. Ochieng, and Robert B. Noland. Current map-matching algorithms for transport applications: State-of-the art and future research directions. *Transportation Research Part C: Emerging Technologies*, 15(5):312–328, October 2007.

[179] Daniele Quercia and Diego Saez. Mining urban deprivation from Foursquare: Implicit crowdsourcing of city land use. *IEEE Pervasive Computing*, 13(2):30–36, 2014.

[180] José Alberto Quintanilha. Qualidade em sistemas de informação geográficas. In *IV Simpósio Brasileiro de Geoprocessamento*, 1997.

[181] Jean-Paul Rodrigue, Claude Comtois, and Brian Slack. *The Geography of Transport Systems*. Routledge, New York, NY, USA, 3 edition, 2013.

[182] Robert J. Rogerson. Quality of Life and City Competitiveness. *Urban Studies*, 36(5-6):969–985, May 1999.

[183] Zev Ross, Iyad Kheirbek, Jane E. Clougherty, Kazuhiko Ito, Thomas Matte, Steven Markowitz, and Holger Eisl. Noise, air pollutants and traffic: Continuous measurement and correlation at a high-traffic location in new york city. *Environmental Research*, 111(8):1054–1063, November 2011.

[184] Mattias Rost, Louise Barkhuus, Henriette Cramer, and Barry Brown. Representation and communication: Challenges in interpreting large social media datasets. In *Proc. of the 2013 Conference on Computer Supported Cooperative Work*, pages 357–362. ACM, 2013.

[185] Wade Roush. Welcome to Google Transit: How (and Why) the Search Giant is Remapping Public Transportation. *Community Transportation*, pages 20–29, 2012.

[186] William Rucklidge. *Efficient Visual Recognition Using the Hausdorff Distance*. Springer, November 1996.

[187] J J Ruiz-Lendínez, F J Ariza-López, and M A Ureña-Cámara. Automatic positional accuracy assessment of geospatial databases using line-based methods. *Survey Review*, 45(332):332–342, September 2013.

[188] Juan J. Ruiz-Lendínez, Francisco J. Ariza-López, and Manuel A. Ureña-Cámara. A point-based methodology for the automatic positional accuracy assessment of geospatial databases. *Survey Review*, 48(349):269–277, 2016.

[189] Jes Ryttersgaard. Spatial Data Infrastructure: Developing Trends and Challenges. In *Committee on Development Information II*, September 2001.

[190] Alan Saalfeld. Conflation automated map compilation. *International journal of geographical information systems*, 2(3):217–228, January 1988.

[191] E. Safra, Y. Kanza, Y. Sagiv, and Y. Doytsher. Ad hoc matching of vectorial road networks. *International Journal of Geographical Information Science*, 27(1):114–153, 2013.

[192] Eliyahu Safra and Yerach Doytsher. Using matching algorithms for improving locations in cadastral maps. In *XXIII FIG Congress*, pages 1–16, Munich, Germany, 2006.

[193] Arnaud Sahuguet, John Krauss, Luis Palacios, and David Sangokoya. Open Civic Data : Of the People, By the People, For the People. *Bulletin of the Technical Committee on Data Engineering*, 37(4):15–26, 2014.

[194] Fatiha Saïs, Nathalie Pernelle, and Marie-Christine Rousset. Combining a logical and a numerical method for data reconciliation. In *Lecture Notes in Computer Science*, pages 66–94. Springer Berlin Heidelberg, 2009.

[195] Maria Salonen and Tuuli Toivonen. Modelling travel time in urban networks: Comparable measures for private car and public transport. *Journal of Transport Geography*, 31:143–153, July 2013.

[196] F. Samadzadegan. Data Integration related to Sensors, Data and Models. In *XXth ISPRS Congress*, Istambul, Turkey, 2004.

[197] Ashok Samal, Sharad Seth, and Kevin Cueto. A feature-based approach to conflation of geospatial sources. *International Journal of Geographical Information Science*, 18(5):459–489, July 2004.

[198] Reinaldo Onofre dos Santos, Diego Ferreira Fonseca, Júlia de Carvalho Nascimento, Nina Ferraz Tolentino, Rodrigo Nunes Ferreira, Gustavo Libério de Paulo, and Jussara da Silva Rocha. Relatório geral sobre o cálculo do Índice de qualidade de vida urbana de belo horizonte (iqvu-bh) para 2016. Technical report, Prefeitura de Belo Horizonte, 2018.

[199] Salatiel Ribeiro dos Santos, Clodoveu Augusto Davis Jr., and Rodrigo Smarzaro. Integration of data sources on traffic accidents. In Claudio Campelo and Laércio Namikawa, editors, *Brazilian Symposium on Geoinformatics - GeoInfo 2016*, pages 192–203, Campos do Jordão, SP, 2016.

[200] Salatiel Ribeiro dos Santos, Clodoveu Augusto Davis Jr., and Rodrigo Smarzaro. Analyzing Traffic Accidents based on the Integration of Official and Crowdsourced Data. *Journal of Information and Data Management*, 8(1):67–82, 2017.

[201] Mohamed Sarwat, Justin J. Levandoski, Ahmed Eldawy, and Mohamed F. Mokbel. LARS*: An efficient and scalable location-aware recommender system. *IEEE Transactions on Knowledge and Data Engineering*, 26(6):1384–1399, June 2014.

[202] Paul Scarponcini. Standardization and modeling of transportation infrastructure semantics. In Peter van Oosterom and Sisi Zlatanova, editors, *Creating Spatial Information Infrastructures: Towards the Spatial Semantic Web*, chapter 4. CRC Press, 2008.

[203] Theo J. H. Schoemaker, Kaspar Koolstra, and Piet H. L. Bovy. Traffic in the 21st century - a scenario analysis for the traffic market in 2030. In M.P.C. Weijnen and E.F. Ten Heuvelhof, editors, *The infrastructure playing field in 2030*, pages 175–194. Deft University Press, 1999.

[204] Sukhjit Singh Sehra, Hardeep Singh Rai, and Jaiteg Singh. Quality assessment of crowdsourced data against custom recorded map data. *Indian Journal of Science and Technology*, 8(33), December 2015.

[205] Monika Sester, Jamal Jokar Arsanjani, Ralf Klammer, Dirk Burghardt, and Jan-Henrik Haunert. Integrating and Generalising Volunteered Geographic Information. In *Abstracting Geographic Information in a Data Rich World*, pages 119–155. Springer International Publishing, 2014.

[206] Shih-Lung Shaw. Geographic information systems for transportation: from a static past to a dynamic future. *Annals of GIS*, 16(3):129–140, November 2010.

[207] Amit Sheth. Citizen sensing, social signals, and enriching human experience. *IEEE Internet Computing*, 13(4):87–92, July 2009.

[208] Amit P. Sheth and James A. Larson. Federated database systems for managing distributed, heterogeneous, and autonomous databases. *ACM Computing Surveys*, 22(3):183–236, 1990.

[209] Wenzhong Shi, Peter Fisher, and Michael F. Goodchild, editors. *Spatial Data Quality*. CRC, September 2002.

[210] P. Shvaiko and J. Euzenat. Ontology matching: State of the art and future challenges. *IEEE Transactions on Knowledge and Data Engineering*, 25(1):158–176, January 2013.

[211] Anne Dorothée Slovic, Diego Bogado Tomasiello, Mariana Giannotti, Maria de Fatima Andrade, and Adelaide C. Nardocci. The long road to achieving equity: Job accessibility restrictions and overlapping inequalities in the city of São Paulo. *Journal of Transport Geography*, 78:181–193, June 2019.

[212] Rodrigo Smarzaro, Clodoveu A. Davis, and José Alberto Quintanilha. Creation of a multimodal urban transportation network through spatial data integration from authoritative and crowdsourced data. *ISPRS International Journal of Geo-Information*, 10(7):470, July 2021.

[213] Rodrigo Smarzaro, Tiago França Melo de Lima, and Clodoveu Augusto Davis Jr. Could data from location-based social networks be used to support urban planning? In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, pages 1463–1468, Republic and Canton of Geneva, Switzerland, 2017. International World Wide Web Conferences Steering Committee.

[214] Rodrigo Smarzaro, Tiago França Melo de Lima, and Clodoveu Augusto Davis Jr. Quality of Urban Life Index from Location-Based Social Networks Data: A case study in Belo Horizonte, Brazil. In *Volunteered Geographic Information and the Future of Geospatial Data*, pages 185–207. IGI Global, 2017.

[215] Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, 45(4):427–437, 2009.

[216] Wenbo Song, Timothy L. Haithcoat, and James M. Keller. A snake-based approach for TIGER road data conflation. *Cartography and Geographic Information Science*, 33(4):287–298, 2006.

[217] Statista. Number of average monthly unique visitors to Yelp.com from 2019 to 2022, by device, 2023. https://www.statista.com/statistics/1326159/number-of-monthly-visitors-to-yelp-by-device/. Accessed on 2023-02-23.

[218] Enrico Steiger, Timothy Ellersiek, and Alexander Zipf. Explorative public transport flow analysis from uncertain social media data. *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information - GeoCrowd '14*, pages 1–7, 2014.

[219] Daniel Steiner, Hartwig Hochmair, and Gernot Paulus. Quality Assessment of Open Realtime Data for Public Transportation in the Netherlands. *GI_Forum*, 1:579–588, 2015.

[220] Peter R. Stopher and Arnim H. Mayburg. *Urban Transportation Modeling and Planning*. Lexington Books, 1975.

[221] H. Sung, S. Lee, and S. Cheon. Operationalizing Jane Jacobs's Urban Design Theory: Empirical Verification from the Great City of Seoul, Korea. *Journal of Planning Education and Research*, 35(2):117–130, June 2015.

[222] E Talen and L Anselin. Assessing spatial equity: An evaluation of measures of accessibility to public playgrounds. *Environment and Planning A: Economy and Space*, 30(4):595–613, April 1998.

[223] Sui Tao, Jonathan Corcoran, Iderlina Mateo-Babiano, and David Rohde. Exploring Bus Rapid Transit passenger travel behaviour using big data. *Applied Geography*, 53:90–104, September 2014.

[224] Sui Tao, David Rohde, and Jonathan Corcoran. Examining the spatial-temporal dynamics of bus passenger travel behaviour using smart card data and the flowcomap. *Journal of Transport Geography*, 41:21–36, December 2014.

[225] Henrikki Tenkanen, Perttu Saarsalmi, Olle Järv, Maria Salonen, and Tuuli Toivonen. Health research needs more comprehensive accessibility measures: integrating time and transport modes from open data. *International Journal of Health Geographics*, 15(1):23, 2016.

[226] Kaitlin Toms and Wei Song. Spatial analysis of the relationship between levels of service provided by public transit and areas of high demand in jefferson county, kentucky. *Papers in Applied Geography*, 2(2):147–159, April 2016.

[227] Xiaohua Tong, Dan Liang, and Yanmin Jin. A linear road object matching method for conflation based on optimization and logistic regression. *International Journal of Geographical Information Science*, 28(4):824–846, 2014.

[228] Xiaohua Tong, Wenzhong Shi, and Susu Deng. A probability-based multi-measure feature matching method in map conflation. *International Journal of Remote Sensing*, 30(20):5453–5472, 2009.

[229] Miguel J. Torres-Ruiz and Miltiadis Lytras. Urban Computing and Smart Cities Applications for the Knowledge Society. *International Journal of Knowledge Society Research*, 7(1):113–119, January 2016.

[230] Harry T Uitermark, Peter J M Van Oosterom, Nicolaas J I Mars, and Martin Molenaar. Ontology-Based Geographic Data Set Integration Introduction : Context , Related Work and Overview A Conceptual Framework for Ontology-Based Geographic Data Set Integration. In M. H. Böhlen, C. S. Jensen, and M. O. Scholl, editors, *International Workshop on Spatio-Temporal Database Management STDBM'99*, volume 1678, pages 60–78, Edinburgh, 1999. Springer.

[231] Harry T. Uitermark, Peter J.M. van Oosterom, Nicolaas J.I. Mars, and Martien Molenaar. Ontology-based integration of topographic data sets. *International Journal of Applied Earth Observation and Geoinformation*, 7(2):97–106, August 2005.

[232] A Ulubay and MO Altan. A different approach to the spatial data integration. *International Archives Of Photogrammetry Remote Sensing And Spatial Information Sciences*, 34(4):656–661, 2002.

[233] United Nations. World urbanization prospects, the 2014 revision, 2014.

[234] E. Lynn Usery, Michael P. Finn, and Michael Starbuck. Data layer integration for the national map of the united states. *Cartographic Perspectives*, 0(62):28–41, March 2009.

[235] L Vaccari, P Shvaiko, and M Marchese. A geo-service semantic integration in Spatial Data Infrastructures. *International Journal of Spatial Data Infrastructures Research*, 4:24–51, 2009.

[236] Arjan Van Binsbergen and Johan Visser. *Innovation Steps Towards Efficient Goods Distribution Systems for Urban Areas.* Delft University Press, 2001.

[237] Marianne Vanderschuren, Robert Cameron, Alexandra Newlands, and Herrie Schalekamp. Geographical modelling of transit deserts in cape town. *Sustainability*, 13(2):997, January 2021.

[238] Eduardo Alcântara Vasconcellos. *Urban Transport Environment and Equity: The Case for Developing Countries.* Routledge, 2014.

[239] Hugo De Souza Vellozo, Michele Brito Pinheiro, and Clodoveu Augusto Davis Jr. Strepitus: um aplicativo para coleta colaborativa de dados sobre ruído urbano. In *Workshop de Computação Aplicada à Gestão do Meio Ambiente e Recursos Naturais - WCAMA*, pages 1047–1051, 2013.

[240] Steffen Volz. Data-driven matching of geospatial schemas. In Anthony G. Cohn and David M. Mark, editors, *Spatial Information Theory: International Conference, COSIT 2005, Ellicottville, NY, USA, September 14-18, 2005. Proceedings*, chapter 8, pages 115–132. Springer Berlin Heidelberg, Berlin, Heidelberg, 2005.

[241] Steffen Volz. An Iterative Approach for Matching Multiple Representations of Street Data. In *Proceedings of the JOINT ISPRS Workshop on Multiple Representations and Interoperability of Spatial Data*, pages 101–110, Hannover, Germany, 2006.

[242] Paul Waddell and Firouzeh Nourzad. Incorporating Nonmotorized Mode and Neighborhood Accessibility in an Integrated Land Use and Transportation Model System. *Transportation Research Record: Journal of the Transportation Research Board*, 1805(1):119–127, January 2002.

[243] Yilun Wang, Yu Zheng, and Tong Liu. A noise map of new york city. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing Adjunct Publication - UbiComp'14 Adjunct*. Association for Computing Machinery (ACM), 2014.

[244] Yu-hong Wang and Feng-yuan Wei. A schema-matching-based approach to propagating updates between heterogeneous spatial databases. In Lin Liu, Xia Li, Kai Liu, and Xinchang Zhang, editors, *Geoinformatics 2008 and Joint Conference on GIS and Built Environment: Advanced Spatial Data Models and Analyses*. SPIE, October 2009.

[245] Dorota Węziak-Białowolska. Quality of life in cities – Empirical evidence in comparative European perspective. *Cities*, 58:87–96, October 2016.

[246] Michael J. Widener, Steven Farber, Tijs Neutens, and Mark Horner. Spatiotemporal accessibility to supermarkets using public transit: an interaction potential approach in Cincinnati, Ohio. *Journal of Transport Geography*, 42:72–83, January 2015.

[247] Stefan Wiemann and Lars Bernard. Spatial data fusion in spatial data infrastructures using linked data. *International Journal of Geographical Information Science*, 30(4):613–636, September 2015.

[248] Sarah Williams, Adam White, Peter Waiganjo, Daniel Orwa, and Jacqueline Klopp. The digital matatu project: Using cell phones to create an open source data for Nairobi's semi-formal bus system. *Journal of Transport Geography*, 49:39–51, December 2015.

[249] James Wong. Leveraging the General Transit Feed Specification for Efficient Transit Analysis. *Transportation Research Record: Journal of the Transportation Research Board*, 2338(2338):11–19, December 2013.

[250] Emerson M. A. Xavier, Francisco J. Ariza-López, and Manuel A. Ureña-Cámara. A survey of measures and methods for matching geospatial vector datasets. *ACM Computing Surveys*, 49(2):1–34, August 2016.

[251] Bisheng Yang and Yunfei Zhang. Pattern-mining approach for conflating crowdsourcing road networks with POIs. *International Journal of Geographical Information Science*, 29(5):786–805, March 2015.

[252] Bisheng Yang, Yunfei Zhang, and Xuechen Luan. A probabilistic relaxation approach for matching road networks. *International Journal of Geographical Information Science*, 27(2):319–338, February 2013.

[253] Ling Yin and Shih-Lung Shaw. Exploring space-time paths in physical and social closeness spaces: a space-time GIS approach. *International Journal of Geographical Information Science*, 29(5):742–761, January 2015.

[254] S Ying, L Li, Y R Gao, and Y Min. Probabilistic matching of map objects in multi-scale space. In *Proceedings of the 25th International Cartographic Conference*, 2011.

[255] Shuxin Yuan and Chuang Tao. Development of conflation components. *Proceedings of Geoinformatics*, 2(2):1–13, 1999.

[256] Yihong Yuan and Martin Raubal. Extracting dynamic urban mobility patterns from mobile phone data. *Geographic Information Science*, pages 354–367, 2012.

[257] Yihong Yuan and Martin Raubal. Analyzing the distribution of human activity space from mobile phone usage: An individual and urban-oriented study. *International Journal of Geographical Information Science*, 30(8):1594–1621, 2016.

[258] Wei Zeng, Chi-Wing Fu, Stefan Muller Arisona, Alexander Erath, and Huamin Qu. Visualizing mobility of public transportation system. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1833–1842, December 2014.

[259] Fuzheng Zhang, David Wilkie, Yu Zheng, and Xing Xie. Sensing the pulse of urban refueling behavior. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing - UbiComp'13*. Association for Computing Machinery (ACM), 2013.

[260] Meng Zhang. *Methods and implementations of road-network matching.* Phd, Technical University of Munich, 2009.

[261] Xiang Zhang, Tinghua Ai, Jantien Stoter, and Xi Zhao. Data matching of building polygons at multiple map scales improved by contextual information and relaxation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 92:147–163, 2014.

[262] Xiang Zhang, Xi Zhao, Martin Molenaar, Jantien E Stoter, Menno-Jan M.-J. Kraak, and Tinghua Ai. Pattern classification approaches to matching building polygons at multiple scales. *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, I-2(September):19–24, 2012.

[263] Yingjia Zhang, Xueming Li, Aiming Wang, Tongliga Bao, and Shenzhen Tian. Density and diversity of OpenStreetMap road networks in China. *Journal of Urban Management*, 4(2):135–146, December 2015.

[264] Huimin Zhao. Semantic matching across heterogeneous data sources. *Communications of the ACM*, 50(1):45–50, January 2007.

[265] Y. Zheng, W. Wu, Y. Chen, H. Qu, and L. M. Ni. Visual analytics in urban computing: An overview. *IEEE Transactions on Big Data*, 2(3):276–296, September 2016.

[266] Yu Zheng. Location-based social networks: Users. In Yu Zheng and Xiaofang Zhou, editors, *Computing with Spatial Trajectories*, pages 243–276. Springer New York, New York, NY, 2011.

[267] Yu Zheng, Licia Capra, Ouri Wolfson, and Hai Yang. Urban Computing: Concepts, Methodologies, and Applications. *ACM Transactions on Intelligent Systems and Technology*, 5(3):1–55, September 2014.

[268] Yu Zheng, Furui Liu, and Hsun-Ping Hsieh. U-Air: When urban air quality inference meets big data. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '13*, New York, New York, USA, 2013. ACM Press.

[269] Dennis Zielstra and Hartwig Hochmair. Using free and proprietary data to compare shortest-path lengths for effective pedestrian routing in street networks. *Transportation Research Record: Journal of the Transportation Research Board*, 2299:41–47, December 2012.

[270] Dennis Zielstra and Hartwig H. Hochmair. Digital street data: Free vs Proprietary. *GIM International*, 25(7), 2011.

[271] Dennis Zielstra and Alexander Zipf. A Comparative Study of Proprietary Geodata and Volunteered Geographic Information for Germany. In *13th AGILE International Conference on Geographic Information Science*, 2010.

[272] H. Zimmermann. OSI reference model–the ISO model of architecture for open systems interconnection. *IEEE Transactions on Communications*, 28(4):425–432, April 1980.

[273] M.H.P. Zuidgeest, M.J.G. Brussel, A. Arora, S. Bhamidipati, S. Amer, F.A.M. De Souza, and T. Godefrooij. On bus-bike integration Final Consultants ' Report On bus-bike integration. Technical Report October, ITC, Enschede, 2009.