UHASSELT

KNOWLEDGE IN ACTION

School for Transportation Sciences

Master of Transportation Sciences

Masterthesis

Imputation of missing values in OSM networks

Ahmad Adeel

Thesis presented in fulfillment of the requirements for the degree of Master of Transportation Sciences, specialization Mobility Management

SUPERVISOR :
Prof.dr.ir Tom BELLEMANS

CO-SUPERVISOR :
Prof. dr. ir. Bruno KOCHAN

MENTOR :
De heer Glenn CICH

2016
2017

UHASSELT
KNOWLEDGE IN ACTION

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

School for Transportation Sciences

Master of Transportation Sciences

Masterthesis

Imputation of missing values in OSM networks

Ahmad Adeel

Thesis presented in fulfillment of the requirements for the degree of Master of Transportation Sciences, specialization Mobility Management

SUPERVISOR :

Prof.dr.ir Tom BELLEMANS

CO-SUPERVISOR :

Prof. dr. ir. Bruno KOCHAN

MENTOR :

De heer Glenn CICH

PREFACE

"If you want to predict the future then you must have ability to shape the future" said by Mr. Eric Hofer

This master thesis is an essential requirement for the completion of Masters in Transportation Sciences at Hasselt University. I was offered different topics in the domain of Transportation Sciences from my university but I have chosen this topic because of its prime importance in transportation modelling. Traffic assignment is one of the most important aspects of transportation modelling and my work is an input for traffic assignment. The focus of my work is on OpenStreetMap which is a rich crowdsourced database providing information related to transport networks. Missing values in the OpenStreetMap road network dataset affects its potential utility in the traffic assignment studies. My goal is to fill missing values in the OSM road network dataset in order to use it in traffic assignment effectively.

Primarily, I would like to thank GOD for granting me strength to complete this master thesis. I would like to thank my supervisor Prof dr. ir. Tom Bellemans, co-supervisor Prof. dr. ir. Bruno Kochan.

I would like to pay special gratitude to my mentor Mr. Glenn Cich for motivating me and helping me at every stage of this work. He provided me sufficient assistance and time and his valuable comments and feedback proved constructive in this research work.

I would also like to thank Higher Education Commission of Pakistan (HEC), who awarded me the scholarship and helped me to avail this opportunity to study at University of Hasselt.

AHMAD ADEEL

Dated 11-Aug-17

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

List of Code Example

SUMMARY

The focus of this research is the completion of the road network attribute dataset of OpenStreetMap (OSM). OSM is based on the concept of crowdsourcing in which geographical data is collected by the different volunteers using local knowledge. Geographical information related to a road network provided by OSM can be used in traffic and transportation studies. Utility of this geographical information in traffic assignment models depends upon its completeness. The OSM dataset contains information related to a road network such as road type, number of lanes, speed limit, public transport route, presence of bicycle lane, road capacity, elevation, one way or two way etc. But this information is not available for all the roads, most of the times these attributes are missing.

This study focuses on filling the gaps in the OSM dataset of roads in order to be used in traffic assignment studies. This study paves the way towards effective utilization of the OSM dataset in route assignment studies. An extensive literature has been reviewed to assess the quality and completeness of the OSM data. Beside the literature review, different areas of Flanders have been explored through OSM and availability of information in the OSM dataset of Flanders have been assessed. Specifically, in case of Flanders there are lot of missing values in the road network attribute dataset. Only two attributes of roads are available for some roads that is number of lanes and maximum speed.

In general, there are several techniques to impute missing values in a dataset. Out of these techniques, multiple imputation proves to be the best technique to handle the problem of missing data. With the help of regression imputation models, missing values will be predicted in the OSM dataset so that information extracted from OSM can be used for traffic assignment in an effective manner. There are different software packages available that are used for development of imputation models such as Amelia and Mice. In future, these softwares will be used for the multiple imputation of missing data values in OSM dataset.

1. INTRODUCTION

1.1. Background

OpenStreetMap (OSM) provides global, up-to-date, vector based geographical data which is freely accessible and editable. OSM is based on the concept of crowdsourcing in which geographical data is collected by the different volunteers using arterial imagery, local knowledge and GPS devices (SinghSehra, Singh, & Singh Rai, 2013). OSM maintains a database of different geographic elements comprising ways, nodes, buildings, landmarks and relations (Eugster & Schlesinger, 2013). OSM contains information related to many points of interest however there is no certainty about the completeness and quality of the data. OSM has many contributors therefore changes are reflected very fast compared to the alternative bureaucratic sources (Maier, 2014).

One of the most important applications of the OSM database is the rendering of the geographic features and data to raster images such as the OSM map on the website. However, the project of OSM also provides an Application Programming Interface (API) for fetching raw data and saving it to the database of OSM. The OSM project provides geographic data in XML format which comprises three elements that are nodes, ways and relations (Ciepluch, Jacob, Mooney, & Winstanley, 2010).

"Node" is a very basic element and it contains the attribute longitudes and latitudes. "Ways" refers to the ordered interconnection of the nodes. Linear features and areas are described by this element for example streets and buildings. "Relation" refers to the grouping of elements (nodes and ways) which are geographically related to provide valuable information such as cycle and bus routes. Further each element has an element ID and an arbitrary number of tags (key-value pairs) which explain the element (Eugster & Schlesinger, 2013).

Geographical information related to a road network provided by OSM can be used in traffic and transportation studies. Utility of this geographical information in traffic assignment models depends upon its completeness and positional accuracy. Completeness refers to the degree to which geographical features, their relationships, their attributes are included in the spatial dataset. Completeness of the information matters a lot while using it in the route choice studies. Incompleteness of the information provided by OSM paves the way of this research.

1.2. Context and Problem Statement

Interest in OSM is increasing day by day and there is a lot of discussion for its geometric accuracy and completeness of the information. (Goodchild, 2008) claims that completeness is one of the most significant aspects of OSM data quality. (Girres & Touya, 2010) assessed the quality of the OSM data in terms of its completeness on the whole area of France and results showed only 6% completeness ratio, when OSM dataset was compared with another dataset named BD TOPO (BD TOPO is a dataset with 1 meter precision and derived from digital photogrammetry). Completeness of the OSM dataset becomes more problematic within rural areas. (Haklay, 2010) investigated the completeness of the OSM dataset in London by length comparison of roads and their attributes with the Ordnance Survey Meridian database and results showed that the OSM dataset is providing coverage for 24.5% of the total area that is covered by Meridian.

OSM contains information related to a road network such as road type, number of lanes, speed limit, public transport route, presence of bicycle lane, road capacity, elevation, one way or two way etc. But this information is not available for all the roads, most of the times these attributes are missing.

Everybody can enter information in OSM and it works with key-value pairs in an .XML format but there is no fixed format for values and keys. (Mooney & Corcoran, 2012) also explored that there is a problem of misspellings in key values that are associated with the features (e.g. highway and land use). These types of errors occur when a user enters a value as a free text instead of using OSM core values provided in drop-down lists. The problem is that there are a lot of keys and values and a lot of missing data in the OSM dataset. These missing values affects the potential usability of OSM by professionals.

1.3. Research Questions

Following research questions will be addressed in this research:

- Is there enough information available in OSM?
- What are the existing techniques to fulfill the missing data values in OSM?
- How can missing data values be completed in OSM?
- How can different statistical models be used to fulfill the missing values based upon the available data values?

1.4. Research Objectives

Objectives of the research are described as follow:

- To have a deep understanding of the OSM data management
- To investigate the completeness of the information provided by OSM
- To document existing techniques to fill the missing data values in the OSM dataset
- To test the usage of regression imputation models for multiple imputation of missing attribute values in the OSM dataset

1.5. Importance of the Study

Completeness of the OSM spatial dataset is one of the significant aspects for using it in traffic assignment. Currently, missing attribute values in the geographical dataset of roads are affecting its potential usability in traffic and transportation studies. This study focuses on filling the gaps in the OSM dataset of roads in order to be used in traffic assignment studies. This study will bridge these gaps and will pave the way towards effective utilization of OSM dataset in route assignment studies.

1.6. Research Methodology

In order to carry out the research in an effective manner different steps have been identified and a methodological framework has been defined to meet the objectives of the research. In the first step, research questions and research objectives have been defined after reviewing the relevant literature. Preliminary literature has been reviewed regarding the OSM data management to investigate the completeness of the OSM geospatial dataset. Literature shows that there is a lot of missing geographical data in the OSM dataset which affects its utility.

After defining the research objectives, comprehensive literature has been reviewed regarding various existing techniques to fill the missing data values in the OSM dataset. This literature review will help to understand the existing techniques and validity of these techniques to fill the gaps in the OSM dataset.

In parallel to the literature review, a case study area in United Kingdom has been selected and information from the OSM is extracted regarding transportation networks. Arc GIS, Quantum GIS and Global Mapper are used to extract information from the OSM dataset. Missing values in link attributes are identified after the extraction of the information. Relevant road network attributes which are important for traffic assignment are selected. Missing values in these link attributes will be imputed in this part of thesis.

Missing data values will be filled based on available information by using different statistical models of R packages such as MICE and Amelia for multiple imputation of missing values in the OSM dataset. These models will be tested by the conducting experiments on the OSM data. Software packages like MICE and Amelia are specifically designed for the imputation of missing values.

In Part 1 of the Master Thesis, literature regarding the "OSM data management" and "missing value imputation" has been reviewed. Deep knowledge has been gained regarding the technique of multiple imputation and regression imputation models.

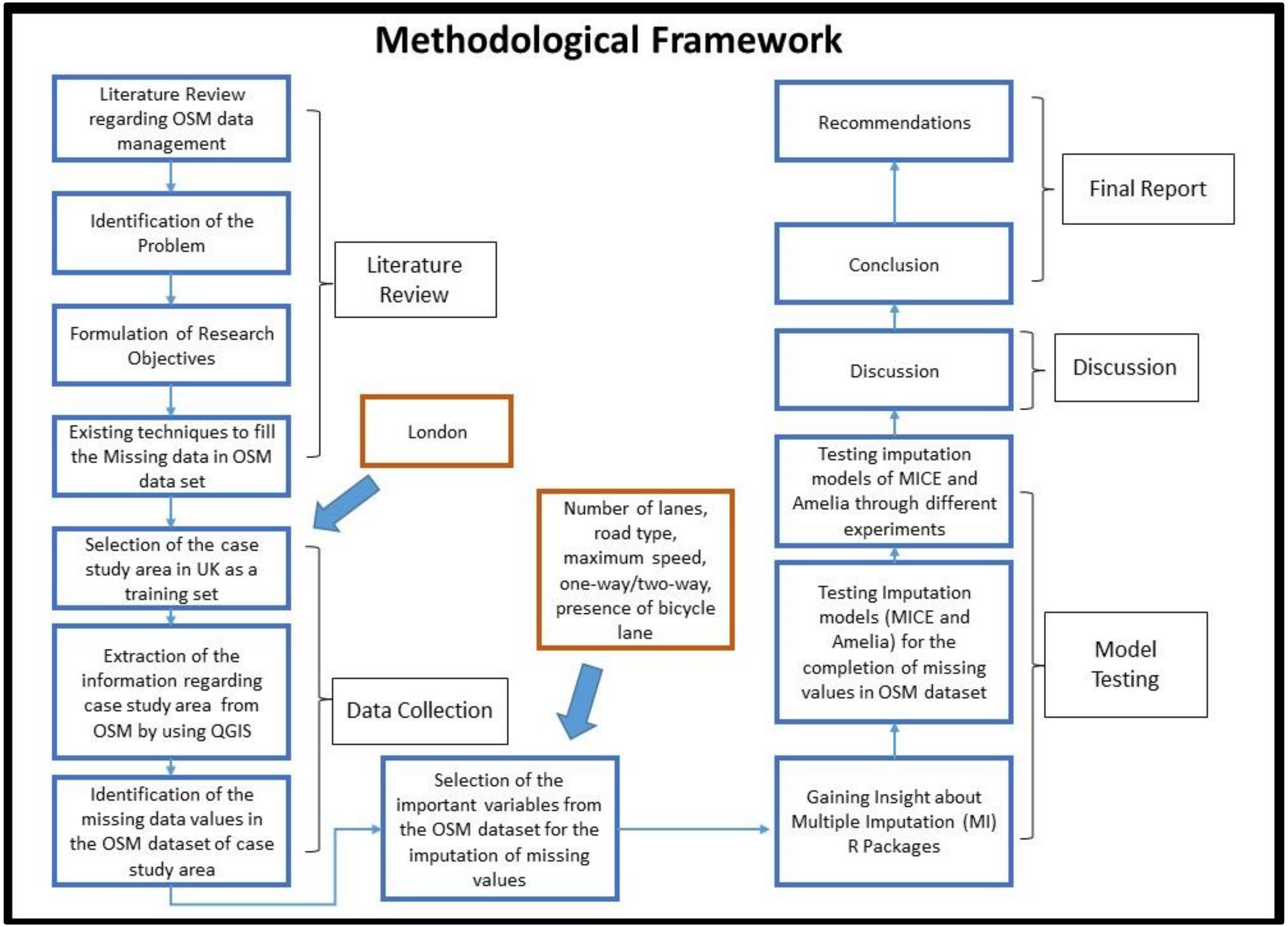
In Part 2 of the master thesis, imputation models of MICE and Amelia are tested by using R which is a programming language for statistical computing. Imputation Models MICE and Amelia predict the missing values based on available transport network information of the selected case study area. Based upon the validity of the predicted information it could be decided that OSM can be used for traffic assignment.

1.7. Limitations of Study

Limitations of study are those influences that a researcher cannot control. They include conditions and shortcomings that cannot be controlled by the researcher and these limitations place restrictions on methodology and conclusions.

In this research study, missing values will be imputed in the OSM datasets in order that OSM can be used for traffic assignment in an effective manner. Missing values are imputed based upon the available information in the OSM dataset. There is no guarantee that already available information in the OSM dataset is 100% correct as this information is provided by the volunteers rather than professionals. If there are errors in the available information, then it is obvious that imputed values will have errors as well.

There are some variables related to the road network which are most important for traffic assignment such as road capacity. This variable is not directly available in the OSM dataset. Some other important variables such as road type, number of lanes, maximum speed are available in the OSM dataset. Hence, the missing values will be imputed for these variables (road type, number of lanes, maximum speed, one-way/two-way and bicycle lane). These variables could help in the determination of road network capacities at later stages.


Methodological Framework
Literature Review regarding OSM data management
Identification of the Problem
Formulation of Research Objectives
Existing techniques to fill the Missing data in OSM data set
Literature Review
London
Selection of the case study area in UK as a training set
Extraction of the information regarding case study area from OSM by using QGIS
Identification of the missing data values in the OSM dataset of case study area
Data Collection
Number of lanes, road type, maximum speed, one-way/two-way, presence of bicycle lane
Selection of the important variables from the OSM dataset for the imputation of missing values
Gaining Insight about Multiple Imputation (MI) R Packages
Testing Imputation models (MICE and Amelia) for the completion of missing values in OSM dataset
Testing imputation models of MICE and Amelia through different experiments
Model Testing
Discussion
Discussion
Conclusion
Recommendations
Final Report

2. LITERATURE LANDSCAPE OF OSM

2.1. Introduction

Last few years have seen many new ways of collecting the geographical information via public initiatives rather than depending on organizations. OSM is a prime example of this crowdsourcing approach which is a collaborative project, just like Wikipedia, initiated in United Kingdom by Steve Coast in 2004. It provides free access to the prime geographical information for many regions of the world. Due to the strong growth in the OSM dataset, many people consider that it can be a potential alternative to the authoritative or commercial data (Girres & Touya, 2010).

OSM was initiated with a focus on the plotting of roads and streets but now it has moved far beyond these elements and it now contains rich variety of geographical information and objects such as land use, buildings and points of interest as the information is uploaded by the thousands of volunteer contributors from all over the world. Development of the cartographic products and collection of the geospatial data is no longer limited to the geographic surveyors, specialists and cartographers that is why OSM has revolutionized the field of cartography (Jokar Arsanjani, Zipf, Mooney, & Helbich, 2015).

(Mooney & Corcoran, 2012) refers to OSM as the Wikipedia map of the world as it is built on same ICT (Information Communication Technology) structures and it offers the project volunteers with the possibility of (a) importing the geographical dataset from Global Positioning System, digital maps and smartphones; (b) immediate updating of the geographic dataset and updating of the associated tools and editing software; (c) permission to gain access to the history of mapping activities in OSM; (d) collaboration and coordination with the other OSM users through different channels such as discussion forums, mailing lists and physical meetings.

Progression of the OSM ecosystem has been very efficacious; an ever- increasing trend of people participating in the OSM project has been observed. OSM had 1.3 million registered contributors in 2013 and 1.85 million registered users/contributors in November, 2014 (SinghSehra, Singh, & Singh Rai, 2014). In 2016 OSM had 3.1 million registered contributors (OSM official website). The OSM community has been involved in many other activities other than collection of geographic data and maintenance of this global database. These activities include open source software development, humanitarian work and building a network to support the OSM users.

Recently, various scientific disciplines such as spatial planning, geoscience, transportation planning, computer science, cartography and ecology apprehended that there is a huge potential in the OSM dataset and it has become the subject of various research activities. Availability of this ever-richer free geographical dataset has gained much interest regarding its usability. There is a range of research questions regarding its potential and limitations. Both the heterogeneity and richness of this geographic dataset has led to many questions on how we can enrich, assess, mine or just use this data in various fields of research (Jokar Arsanjani et al., 2015).

(Hagenauer & Helbich, 2012) posit that geographic information available in the OSM dataset is more complete, semantically and geographically more exact than the corresponding exclusive databases. But it is true for only some areas of the world, there are issues of the completeness and correctness of the OSM data in many regions and these issues may be resolved or mitigated with the help of the specialized approaches (Goodchild, 2013).

2.2. Applications of OSM in Transportation Planning

A lot of special navigation and routing systems use the OSM dataset as an input and these systems are operating on a large scale such as routing for pedestrians, bikes and cars in OpenRouteService (Schmitz, Neis, & Zipf, 2008), wheel chair routing and emergency routing (Neis, Singler, & Zipf, 2010) and three dimensional (3D) indoor routing (Goetz, 2012). Another use of the OSM data set includes the development of Location Based Services (LBS) (Krek, Rumor, Zlatanova, & Fendel, 2009) and these LBS are used for asset tracking, fleet management, and local advertisement. High quality 3D models of cities are generated by using the rich OSM geographical data (Goetz, 2013). There are a lot of mobile applications that provides the routes of public transport within a city such as “OSMtransport”. It means that the OSM dataset is of much importance in the field of transportation planning and mobility management and it is highly useful in route choice studies.

2.3. OSM Data Management

The data model used in OSM is different from other data models used in Geographic Information System (GIS). The OSM data model uses three primitive types coupled with a free-form tagging scheme that allows to label any geographical feature accurately. Topology of the geographical features is also described in this data model that explains the connection between different geographical features. Connections between the different geographical features are significant while using OSM for routing purposes and this feature is not available in traditional GIS. Simplicity in the OSM data model always surprises people who are used to the traditional GIS because models used in typical GIS have many data layers and each geographical feature is described independently of others.

The API (Application Programming Interface) and the data model of OSM are as simple as possible and it is very easy to edit or create new data in the OSM dataset. Some processing is required to OSM data while using this information for any application and many tools are already developed by the OSM community for this purpose.

2.4. Data Primitives

The OSM data model has three basic primitives which includes ways (polylines), nodes (points) and relations (linking nodes and ways with tags). Geographical features that are represented by these primitives can be described with the help of tags and key value pairs. For example, a road can be represented with the tags such as Highway= “Primary”, name= “M10” and one-way= “yes”. There are some attributes which are common to each primitive type such as each primitive has a numerical ID and these are only unique within one type of primitives. Hence, it is possible that there is a way, node and relation with the same numerical ID (Bennett, 2010).

In simple terms, the OSM data model is a mixed graph and it comprises of various edges and vertices. Depending upon the modelling of the real-world features, different parts of the graph are isolated or connected (Girres & Touya, 2010).

The default format to retrieve the data from the OSM server is XML format and it is the mostly used format for the representation of the OSM data model. Other formats can also be used to represent the data model. For instance, Potlatch which is a flash based online editor of OSM, uses the Flash’s action message format to communicate with the server.

2.4.1. Nodes

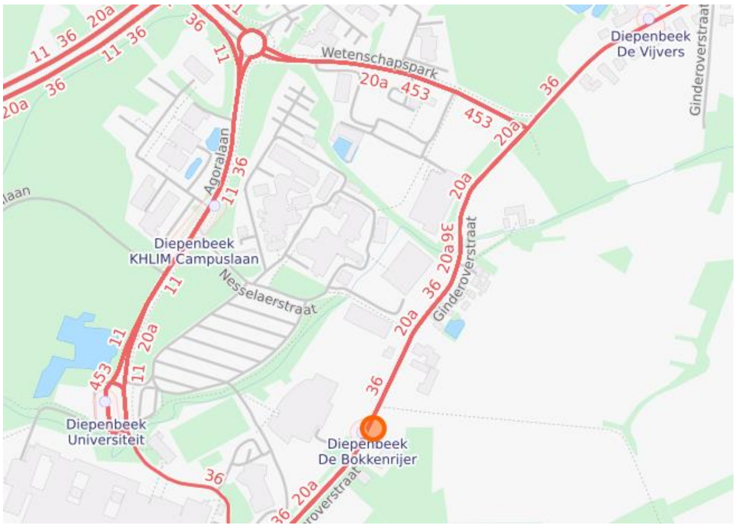
Nodes represent points in space and out of three data primitives, only nodes contain positional information and all other primitives rely on these nodes for the locational information. Nodes can be used to represent a junction, a point of interest or to mark a change in the direction of a road (Barron, Neis, & Zipf, 2014). The XML file of a bus stop (node) near Hasselt University is presented in Code Example 1.

CODE EXAMPLE 1: XML file of a node in OpenStreetMap (OSM.org).

```
<node id="1511877409" visible="true" version="3" changeset="26765793" times
tamp="2014-11-
13T21:37:31Z" user="Polyglot" uid="15188" lat="50.9264282" lon="5.3966376">

  <tag k="bus" v="yes"/>

  <tag k="highway" v="bus_stop"/>

  <tag k="name" v="Diepenbeek De Bokkenrijer"/>

  <tag k="network" v="DLLi"/>

  <tag k="operator" v="De Lijn"/>

  <tag k="public_transport" v="platform"/>

  <tag k="ref:De_Lijn" v="401432"/>

  <tag k="route_ref:De_Lijn" v="36;741"/>

  <tag k="zone:De_Lijn" v="04"/>

</node>
```

Above mentioned XML of a node (public transport stop) near Hasselt University contains the latitudes, longitudes, tags and key value pairs. Longitudes and Latitudes are stored in decimal format, up to seven decimal places and this positional information is accurate enough to use for any purpose. XML of a feature is created by the OSM API or by an editing application hence, there is no need to create this information by hand. Above mentioned node is rendered on the map (shown in orange color) in the Figure 1.

FIGURE 1: Map representation of XML file (OSM.org).

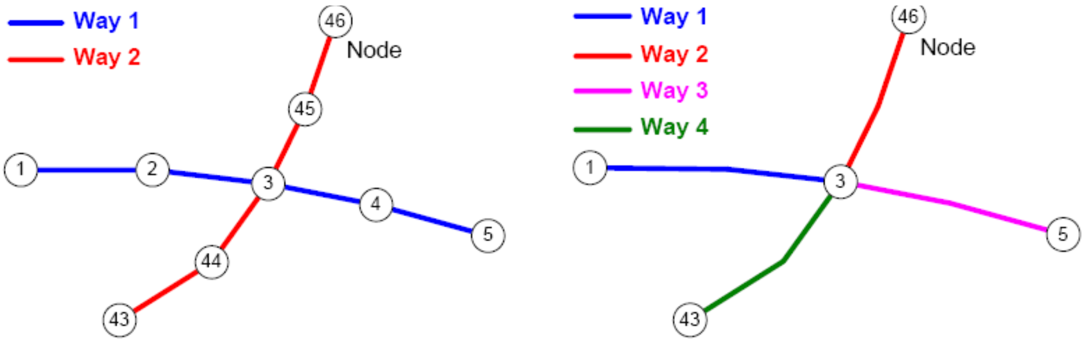


2.4.2. Ways

(Bennett, 2010) explains that ways are an ordered list of nodes and these are used to describe linear features such as cycle paths, roads, waterways. Ways are also used to describe areas in OSM. Ways are closed to form areas. Ways are placed down the center line of the geographical feature when these are used to describe a linear feature. Ways are closed at the perimeter of a polygon to describe any area.

Ways can have 2000 nodes at maximum and at least two nodes for their existence. Upper limit is imposed so that too long ways cannot be created as it effects the performance of the server. A node can belong to more than one way. For example, at junctions nodes are shared by two or more ways. Ordering of nodes is very vital in ways as this ordering specifies the direction of the way from first node to the last node as shown in Figure 2 (Schmitz et al., 2008). The ordering of the nodes from 1 to 5 shows the direction of the way starting from 1 and ending at 5.

FIGURE 2: Routing graph from OSM data (Schmitz et al., 2008).



To represent any "area" by the primitive "way", first and last node of the way must be the same. Tags and key value pairs can be used to give the description of any area. More complex shapes can be created by using different ways. For example, a building with a patio in the middle, can be shown by different ways and a "relation" between them. XML of a parking site near Hasselt university is shown in Code Example 2.

CODE EXAMPLE 2: XML of a Way in OpenStreetMap (OSM.org).

```
<way id="101371600" visible="true" version="3" changeset="39721582" timesta
mp="2016-06-01T22:16:45Z" user="Ralpha" uid="3128251">
  <nd ref="2387628082"/>
  <nd ref="2387628080"/>
  <nd ref="2387628007"/>
  <nd ref="2387628000"/>
  <nd ref="2387627980"/>
  <nd ref="2387627974"/>
  <nd ref="1170556960"/>
  <nd ref="2387627929"/>
  <nd ref="1170556957"/>
  <nd ref="1170556970"/>
  <tag k="amenity" v="parking"/>
  <tag k="fee" v="no"/>
</way>
```

2.4.3. Relation

There are some features that cannot be described by using a single way or node or the situations where same type of features are overlapping (Bennett, 2010). These types of features and relationships are modeled with the help of “Relation” in OSM. Relations define geographical and logical relationships between features (Neis & Zipf, 2012). For example, turn restrictions at intersections, complex branching roads and long distance routes are modelled with the help of Relation (Perkins, 2014) . The XML of a “relation” is given in Code Example 3.

CODE EXAMPLE 3: XML file of a Relation in OSM (OSM.org).

```
<osm version="0.6" generator="OSM server">

  <relation id="113421" visible="true" timestamp=

    "2016-11-03T10:08:27Z" version="2" change set="3851469"

    user="Ahmad Adeel" uid="5854">

    <member type="node" ref="270186" role="via"/>

    <member type="way" ref="4568757" role="from"/>

    <member type="way" ref="4758963" role="to"/>

    <tag k="restriction" v="no_right_turn"/>

    <tag k="type" v="restriction"/>

  </relation>
```

These relations are not shown on the rendered map of OSM. Relations are used by route planning applications and routing algorithms, for instance to prevent a user from making an illegal turn. (Bennett, 2010) calls it the tiniest established part of the OSM data model. Relation contains all the common attributes of the other primitives. Above mentioned code is showing the turning restriction between two roads.

2.5. Change Sets in OSM

The OSM API contains another data structure which keeps the record of the changes that are made to the data primitives. In order to identify the changes in the OSM dataset, change sets are introduced. Before the introduction of the change sets, it was very difficult to identify the changes in the OSM data set. Editor applications of OSM automatically opens the change sets when the map is being edited and these change sets are then closed when changes are being done. Change Sets keep the record of the changes made in the OSM data set by minutes, hours and days (Neis & Zipf, 2012).

2.6. Tagging

The real power of the OSM data model lies in tags which explain the characteristics of the geographical features. Three data primitives represent the real-world features and tags provide information about the of real-world features. Tags are simply key-value pairs and attributes of the geographical features are described by using these structured key-value pairs. “Key” is analogous to the attribute (for example “amenity”) and “value” describes the value of the attribute for given geographical object (for example “cafe”) (Vandecasteele & Devillers, 2015).

2.6.1. Tags for Roads

Most mapped geographical features in OSM are roads and in some regions of the map, roads are the only feature available. All footpaths, cycle paths, roads and land-based routes use the highway=* key. The word "highway" can be confusing for some areas where it is used for a special class of road. It is mostly used in OSM and it means any public road. Even the word highway is used for the unsurfaced roads in OSM. (Bennett, 2010) refers to main values for `highway=*`

`highway=motorway` for high-speed, long-distance roads with access restrictions.
`highway=primary` for the major roads
`highway=secondary` for the minor roads
`highway=residential` for the access roads in residential areas.
`highway=footway` for the footpath whether it is unsurfaced or surfaced

Similarly, additional information related to a road network can be tagged by using tags such as `lanes=*` and `Maximum speed=*`. The best practice is to give the source of speed limit in tags (" Tagging Guidelines - OSM Wiki," 2016).

2.6.2. Tags for Amenities

The next most mapped geographical feature in OSM are amenities and it could include anything from small bins to shops. Mostly amenities are shown using nodes. Some common values for `amenity=*` are:

`amenity=parking` for car parking sites. With the help of the further tags it could be specified that whether this parking is indoor or outdoor and how many parking spaces are available
`amenity=fuel` for the fueling station
`amenity=place_of_worship` for representing the religious building of any religion
`amenity=hospital` for any hospital and other information related to the availability of emergency services may be provided in an extra tag.

2.6.3. Tags for Settlements

Settlements are mapped using the `place=*` tag, the main values for which are:

```
place=country
place=village
place=hamlet
place=town
place=suburb
```

There are more than forty thousand tags that are recorded in the OSM data set and it is difficult to describe all the tags. Rich collaborative approach and current free state of tagging causes many errors as well when the OSM dataset is used by some other application (Vandecasteele & Devillers, 2015).

2.6.4. Tagging Problems

The way contributors of OSM annotate or tag geographical features affects the quality of the OSM data set. (Mooney & Corcoran, 2012) investigated the issues arising due to improper tagging of the objects. The major reason of the errors and missing values in the OSM data set is caused by the contributors who choose the values from OSM ontology and spell these values incorrectly. Tagging problems have a detrimental effect on the quality of the OSM data and it damages the perception of the OSM data in the professional community. Flexibility in the tagging process and lack of mechanism to check the adherence to the OSM ontology is the core reason of this problem.

Although collaborative approach and flexibility in tagging allows a rich description of the geographic features according to local knowledge. But It also creates semantic and geometric heterogeneity. (Vandecasteele & Devillers, 2015) define semantic heterogeneity as the amount and diversity of tags used to describe the same object in the OSM data set. As given in Table 1 various tags are used to describe the same road. Similarly, a university can be represented by a point or a polygon. It is a case of geometric heterogeneity. Similarly, various tags are used to describe the maximum speed of a link such as `Maximum speed=50`, `Maximum speed=50 km/h`, `Maximum speed=50 miles/hour`. The Table 1 shows that how same road is tagged by various contributors in Germany.

TABLE 1: Assigned different values to a road in Germany (Mooney & Corcoran, 2012).

Version	**Highway**	**Contributor ID**
1	Primary	16634
2	Tertiary	65545
3	Primary	21456
4	Residential	22556

The use of various tags for the same object causes problems when the OSM dataset is used for other applications. The OSM API is not sensitive to these different tags but other application such as route planning applications are not able to read the OSM data correctly (Bennett, 2010). That is why diversity of tagging has a detrimental effect on the quality of the OSM data.

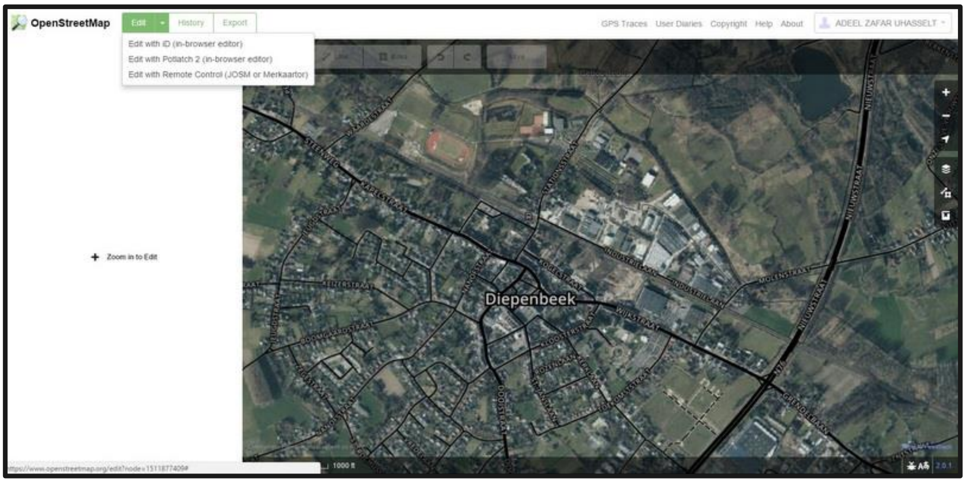
2.7. OSM Editing Techniques

OSM contributors gather geographical information in the form of GPS traces, spoken or written notes, photographs and various types of recorded surveys and then they turn this information into data and contribute to the OSM database. There are various types of editing applications available to contribute data to the OSM database. Most famous editing applications of OSM are given as follow:

JOSM- Java-based desktop editor
Potlatch- web-based editor
Merkaartor- desktop editing application

OSM editing applications can be distinguished between two categories. The first category includes ID editors or Potlatch, which is accessible via a web browser and it is based on web architecture. The second category includes Java OSM (JOSM) or Merkaartor which are desktop applications to edit and contribute to the OSM dataset (Vandecasteele & Devillers, 2015). OSM map editors available on OSM official website are shown in Figure 3.

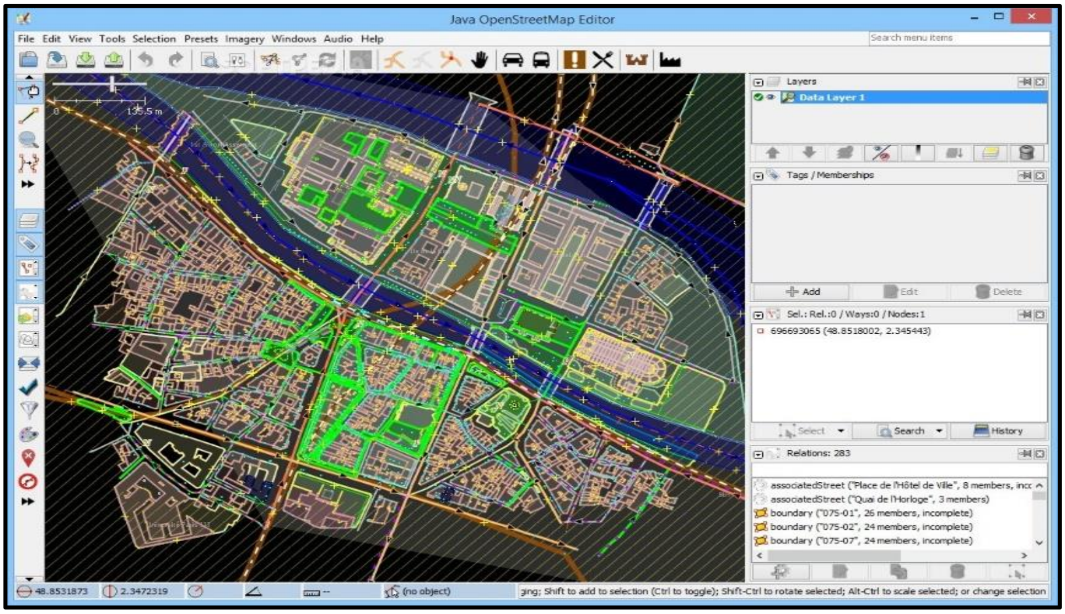
FIGURE 3: OpenStreetMap Editors (OSM.org).



2.7.1. JOSM

JOSM is an open-source, free standalone application that permits to edit, create or delete data from OSM. Statistics shows that JOSM has been used by more than 40% of the contributors. JOSM API is flexible enough to create various plugins to extend JOSM functionalities. To help the editing of the map JOSM has a menu, which is called “presets”. This menu contains list of common tags. This tagging list is organized by different categories such as sports, facilities and highways. After the creation of the geographical feature, OSM contributors can choose the suitable tags from the list. Limited number of tags are available within the “presets” menu. This tagging list can be freely extended by adding tags manually. This manual tagging creates some serious issues as described in the previous section. “Presets” also reminds the contributor about the additional tags to describe the geographical feature. For example, if a contributor is tagging a road then additional tags would be related to the maximum speed and number of lanes. Plugins of JOSM are also available that analyzes the current tag and recommends related tags such as “OSM Semantic Plugin”. Figure 4. shows the interface of Java Open Street Map editor.

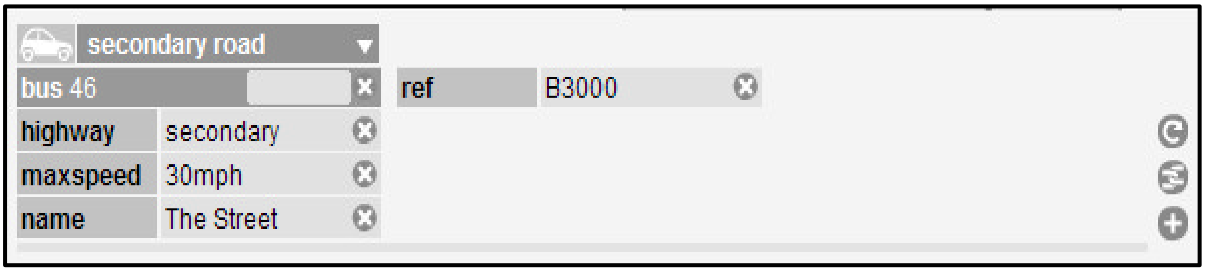
FIGURE 4: Java OpenStreetMap Editor (Vandecasteele & Devillers, 2015).


Java OpenStreetMap Editor
File
Edit
View
Tools
Selection
Presets
Imagery
Windows
Audio
Help
Search menu items
135.5 m
Layers
Data Layer 1
Tags / Memberships
Add
Edit
Delete
Sel.: Rel.:0 / Ways:0 / Nodes:1
696693065 (48.8518002, 2.345443)
Select
Search
History
Relations: 283
associatedStreet ("Place de l'Hôtel de Ville", 8 members, inc
associatedStreet ("Quai de l'Horloge", 3 members)
boundary ("075-01", 26 members, incomplete)
boundary ("075-02", 24 members, incomplete)
boundary ("075-07", 24 members, incomplete)
48.8531873
2.3472319
(no object)
ging; Shift to add to selection (Ctrl to toggle); Shift-Ctrl to rotate selected; Alt-Ctrl to scale selected; or change selection

2.7.2. Potlatch

It is an online flash-based editor for OSM and its advantage is that it does not require any installation for its usage. This editing application also has a system of “presets” that makes the tagging process faster and easier. However, it also allows to enter the free form values and keys for tags. In Potlatch, tags attached to a geographical feature are presented in Figure 5.

FIGURE 5: Tags attached to a road in Potlatch Editor (Bennett, 2010).



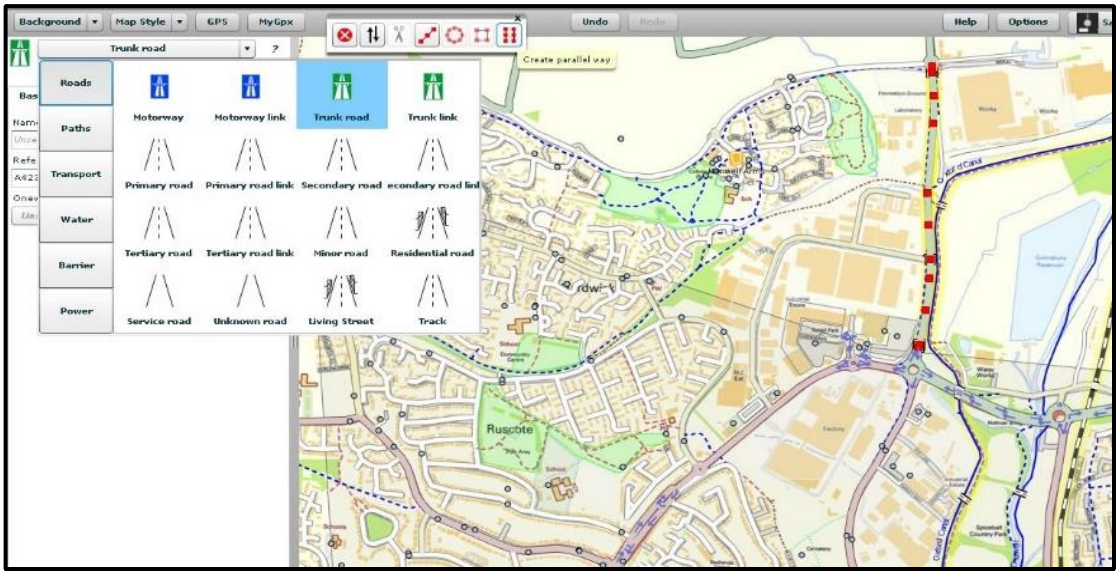
secondary road

bus 46

highway	secondary
maxspeed	30mph
name	The Street
ref	B3000

Tags attached to the way, show that a secondary road named The Street, with a 30mph speed limit, and the road number B3000. The dark gray box above the tags with the text bus 46 shows that this way is a member of a relation. By clicking anywhere on the map a “node” is placed and a “way” can be drawn by placing a start node and then continuing node by node. Figures 6 shows the interface of the Potlatch editor.

FIGURE 6: Potlatch editor of OSM (OSM.org).



3. QUALITY ASSESSMENT OF OSM

Quality plays a vital role while working with the geographical data, especially in data assessment and production (Veregin, 1999) or exchange (Goodchild, 2013). Quality of the data becomes an important aspect in case of OSM as the geographical data is collected by the volunteers who have no restrictions during the annotation and data collection process. Various research studies have been conducted on the quality assessment of OSM data and different researches have different point of views.

(Welser et al., 2011) remark that there is a perception in the domain of scientific research that crowdsourced projects such as “Wikipedia” will not be sufficient because such kind of projects rely on non-professional volunteers with unknown identity and same issue prevails in the GIS community regarding the quality of OSM dataset. (Qian et al., 2009) claim that one of the major negative impacts of the crowdsourced data is that it is collected by non-experts with non-professional equipment which means that there is no guarantee of the quality of the OSM data until and unless it is compared with some other source of geodata. (Ballatore & Bertolotto, 2011) concluded that OSM data is semantically poor but spatially rich. This issue has been described in the previous sections as well because there are issues of semantic heterogeneity with the OSM data set.

3.1. Quality Assessment Parameters

According to ISO (19113) quality of the spatial data (OSM dataset) can be assessed on following elements (Haklay, 2010).

- *Completeness*

Completeness of the geographical data refers to the degree to which geographical elements, their relationships and their attributes are included in the spatial dataset. It also includes information related to other relevant rules of mapping, selection criteria’s and used definitions.

- *Positional Accuracy*

Positional accuracy refers to the absolute and relative accuracy of the coordinate values in the data set. In simple words, it defines the closeness of the locational information with the exact position. Relative accuracy is the accuracy of relative positions of geographical features with respect to the relative positions being true.

- *Logical Consistency*

Logical consistency defines the accurateness of the relations established within a geographical dataset.

- *Temporal Accuracy*

Temporal accuracy refers to the historical evolution of the geographical dataset.

- *Thematic Accuracy*

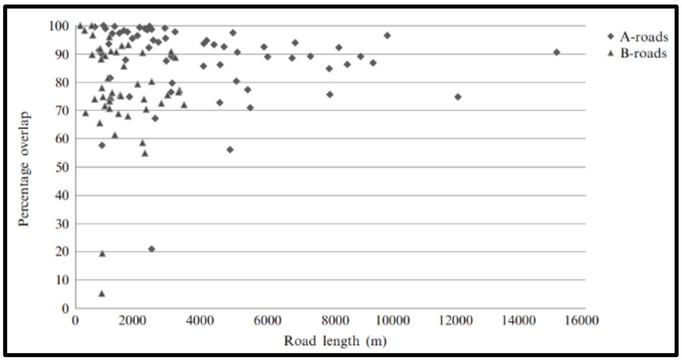
Certain attributes are assigned to the geographical features in a spatial data set. Thematic Accuracy defines the correctness of the attributes assigned to a specific geometry.

3.2. Scientific Studies regarding Quality Assessment of OSM

The increasing availability of the collaboratively and voluntarily collected geographical data leads towards numerous research studies with a focus on the OSM spatial data. In the beginning, scientific investigations were only focused on the road network of OSM. A most commonly used method to assess the quality of the OSM data is the comparison with some referenced data (Haklay, 2010; Zielstra & Zipf, 2010; Girres & Touya, 2010; Mooney et al. 2010a; Neis et al. 2012). However, accessibility to commercial and high quality rich data sets for the purpose of comparison is often limited due to high costs and copyright issues.

Accuracy of the OSM dataset were checked with the help of the referenced spatial dataset. For instance, (Haklay, 2010) compared the road network of OSM with the Ordnance survey (OS) meridian 2 dataset for England. Results of the study explained that the information provided by OSM is accurate within 6 meters of position recorded by the OS dataset and approximately 80% of the overlap of motorways between two datasets. This study further concluded that OSM covers almost 29% area of England and almost 24% roads have complete attributes. The graph in Figure 7 shows the accuracy of OSM to capture B-type and A-type roads, which are small roads in England. For A-type roads there is 88% overlapping between the OSM dataset and the OS meridian 2 datasets. For B-type roads there is 77% overlapping between the two datasets.

FIGURE 7: Comparison of A-type and B-type roads between OSM and Ordanance Survey Map (Haklay, 2010).



Zielstra & Zipf (2010) compared the OSM dataset with TeleAtlas-MultiNet dataset in Germany and the results showed that 80% of the road network data is overlapping with the referenced dataset in the bigger cities of Germany. Researchers further concluded that information provided by OSM is accurate enough to be used for routing applications. Overlapping percentage fluctuates between 50% to 80% within medium towns. There exists correlation between the number of contributors and quality of the OSM data. Similar scientific studies were conducted by Ludwig, Voss, & Krause-Traudes in 2011 and Neis & Zipf in 2012 and the OSM dataset was compared with Navteq and TomTom MultiNet dataset respectively.

Girres & Touya (2010) assessed the quality of the French OSM data set by comparing it with BD TOPO dataset (BD TOPO is a dataset with 1-meter precision and derived from digital photogrammetry) and they concluded that there are issues of heterogeneity in the OSM dataset because of the contributor’s freedom. When two datasets were compared, only 49% of the secondary roads were semantically correct in OSM data set as compared to the BD TOPO dataset. The errors were made by contributors who classified some roads as residential or tertiary roads although those roads were secondary. OSM Roads of zone-1 and zone-2 showed the completeness ratio of 45% and 37% with respect to the BD TOPO dataset of roads.

Other features of OSM data have also been the subject of interest. (Mooney, Corcoran, & Winstanley (2010) compared the land cover features of OSM with the Ordnance Survey Ireland data by using the shape similarity test.

All the scientific studies regarding the quality assessment of OSM shows that OSM data has a high locational accuracy and most of the details are found for urban areas with higher number of contributors. Rural areas show lower quality level of OSM data as the number of contributors are also limited.

3.2.1. Intrinsic OSM Quality Analysis

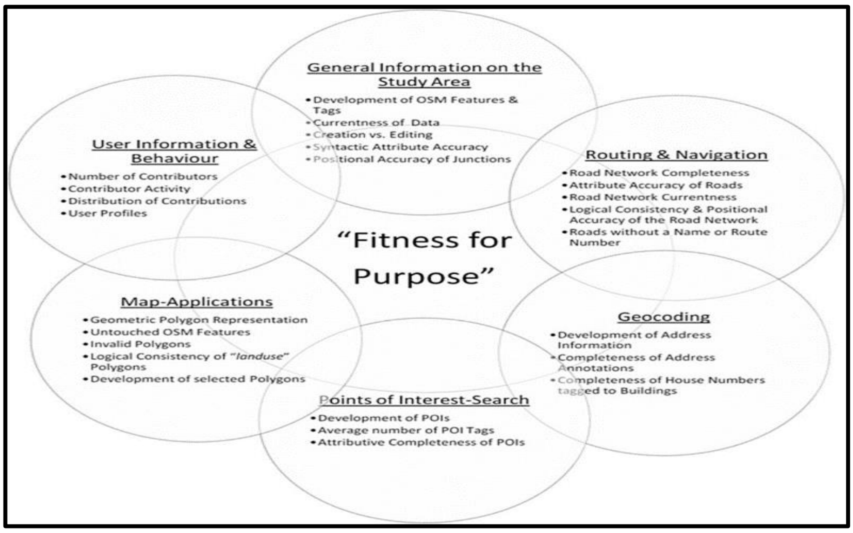
A lot of studies have evaluated the quality of the OSM data but in most of the cases OSM data is compared with some authoritative dataset. (Batini, Cappiello, Francalanci, & Maurino, 2009) state that analysis of the intrinsic data quality captures the inherent quality of data and it can assess the objectivity, accuracy and believability of the data. (Mooney & Corcoran, 2012) assessed the quality of the OSM data by analyzing the heavily edited objects. Researchers used the full-history-dump file of OSM to examine the heavily edited objects.

Full editing history of the OSM objects can be accessed via OSM-Full-History-Dump. Whenever the geometry of a feature is changed then a new version of the OSM object is created. Modifying, adding or deleting a tag also leads towards the creation of a new version number. OSM-Full-History-Dump can be used as a sole input for the intrinsic quality analysis of the OSM data. Rather than comparing the OSM data with some referenced dataset, temporal dimension of the OSM data can be considered. Absolute statements regarding the quality of data cannot be made while analyzing intrinsic quality however approximate statements can be made based on the relative indicators.

(Barron et al., 2014) developed a framework for the intrinsic quality analysis of the OSM data called iOSM Analyzer. iOSM Analyzer works on the history file of OSM data set and it is implemented as a command line-based tool running on the Linux operating system. It is written in the Python programming language and based solely on open source components. Researcher divided the results of iOSM Analyzer into different categories based upon the "Fitness of Purpose" approach. These categories include General information of the area, navigation and routing, Points of Interests, geocoding, user information and behavior and map applications.

According to the following framework of intrinsic quality assessment, evolution of the OSM features provides a first insight about the quality of the OSM data in an area. For instance, number of lines, points and polygons added per month gives a more diverse and general impression of the area. Similarly, if there are no further additions for particular road network category during some previous years then it can be taken as an indication of the completed road network. One of the intrinsic quality indicator is "user information and behavior". For instance, more number of contributor in a particular area indicates the quality of the OSM data. Polygon attributes can be checked in terms of the completeness of the information such as house numbers and house coordinates which gives an insight on the quality of the OSM data. Number of POIs in an area can be used as a first quantitate indicator to check the quality of the OSM data.

FIGURE 8: OSM Intrinsic Quality Indicators (Barron et al., 2014).


General Information on the Study Area
• Development of OSM Features & Tags
• Currentness of Data
• Creation vs. Editing
• Syntactic Attribute Accuracy
• Positional Accuracy of Junctions
User Information & Behaviour
• Number of Contributors
• Contributor Activity
• Distribution of Contributions
• User Profiles
Routing & Navigation
• Road Network Completeness
• Attribute Accuracy of Roads
• Road Network Currentness
• Logical Consistency & Positional Accuracy of the Road Network
• Roads without a Name or Route Number
"Fitness for Purpose"
Map-Applications
• Geometric Polygon Representation
• Untouched OSM Features
• Invalid Polygons
• Logical Consistency of "*landuse*" Polygons
• Development of selected Polygons
Geocoding
• Development of Address Information
• Completeness of Address Annotations
• Completeness of House Numbers tagged to Buildings
Points of Interest-Search
• Development of POIs
• Average number of POI Tags
• Attributive Completeness of POIs

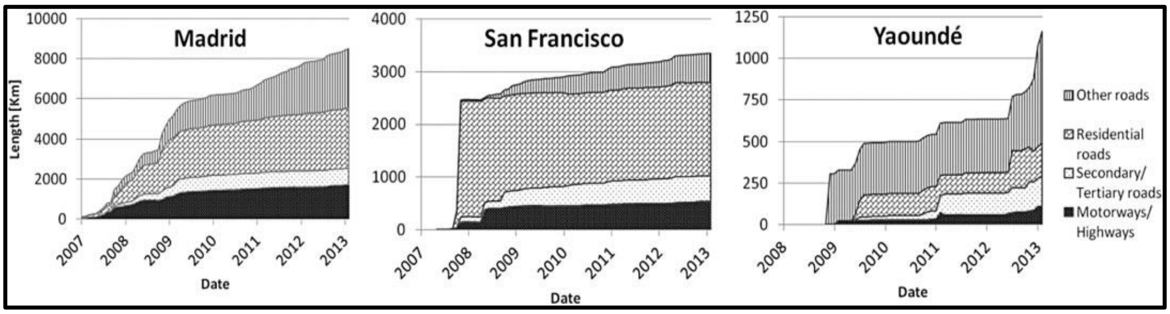
Out of the above-mentioned indicators in Figure 8 for the intrinsic quality assessment of OSM data, routing and navigation is the most important indicator in the context of this research.

The completeness of the road network is one of the most important aspects in route planning and navigation. As the intrinsic approach does not allow the usage of any referenced data therefore by analyzing the creation and editing history of the road network a particular category of road can be stated as “Close to Completion”. The logical reason behind this statement is that monthly increase in the length of the particular road category is very small or very close to zero even when there are a lot of contributors in the area.

Logical consistency is also an important element for using the OSM data for routing purposes. By means of the internal tests three topological errors can be checked: (1) roads are not connected at junctions. (2) duplicate network geometry (3) passing roads without a shared node. Results of the intrinsic study of (Barron et al., 2014) regarding completeness of the OSM road network are shown in the Figure 9.

Results for San Francisco show that there are stable lengths for highways and motorways since 2011 and length of the secondary, tertiary roads and residential roads did not increase significantly since 2012. These two categories of roads can be referred as “**close to completion**”. Although the category of other roads is still incomplete in the OSM dataset. For Madrid, except secondary roads all the other roads are not completed yet and the graph for Yaounde shows that all the road categories are not completely mapped yet.

FIGURE 9: Development of OSM road network by road category (Barron et al., 2014).

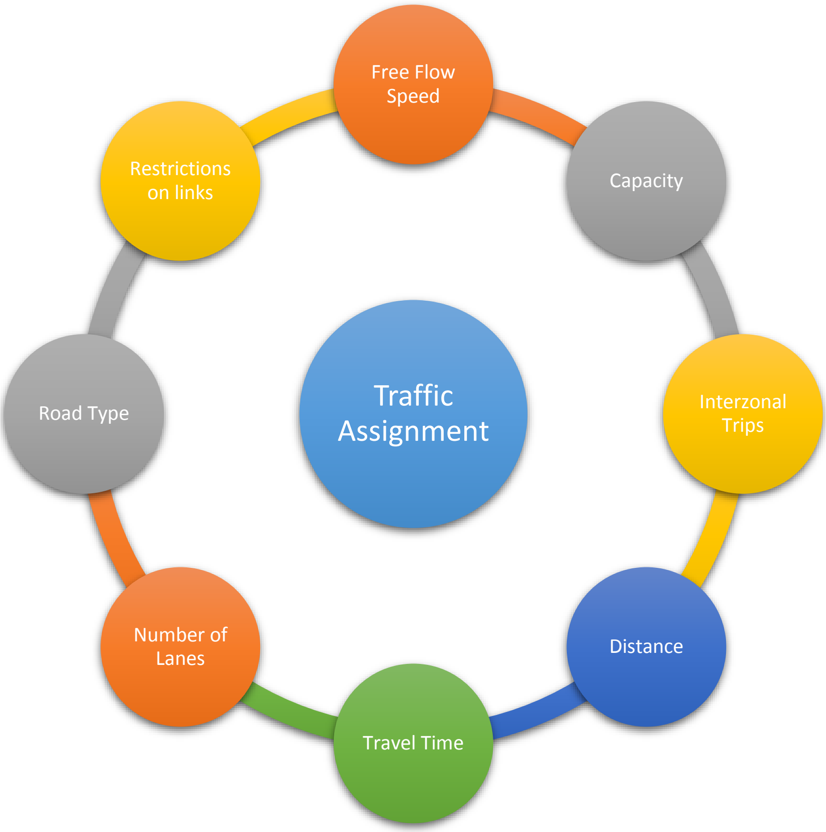


4. ROAD NETWORK ATTRIBUTES IN OSM

4.1. Important Variables for Traffic Assignment

Traffic Assignment is the part of the process which allocates a given set of trip interchanges to a specific transport system or network. Traffic Assignment requires the complete information of the existing transportation system and a matrix of interzonal trips (Patrickson, 1994)**.** Road network attributes which are very important for traffic assignment are shown in Figure 10.

FIGURE 10. Variables important for traffic assignment.



A set of rules are defined in traffic assignment for the identification of desirable routes to connect the origin and destination and then systematically allocating Origin-Destination Trips on these routes so that certain features of reality could be achieved (Ortúzar S. & Willumsen, 2011). These set of principles and rules are used to load a fixed trip matrix onto the road network which generates link flows. There are several important objectives of traffic assignment but all of them do not receive the same emphasis in all situations. Main objectives include:

- To obtain reasonable link flows and identification of severely congested links
- To estimate zone to zone, travel times and travel cost for a given level of demand
- To obtain turning movements which helps in junction design and remodeling
- To identify that which O-D pairs use a particular link

The OSM dataset contains a lot of important variables such as free flow speed, number of lanes, road type, one-way/two-way road, distance between two points, travel time, public transport routes and information regarding bicycle lanes. However, information related to road capacity and interzonal trips is not directly available in the OSM dataset. OSM datasets which are extracted from the shapefiles of Flanders include limited information regarding some important variables: free flow speed, number of lanes, road type, one-way/two-way road and bicycle lanes. All of these variables are considered as significant for traffic assignment and these variables are also helpful for the calculations of some other important variables such as number of lanes and free flow speed help in the calculations of road network capacity. Missing values in all above-mentioned variables (free flow speed, number of lanes, road type, one-way/two-way road and bicycle lanes) will be imputed by using some statistical techniques in order that OSM can be used in route planning studies in a more effective manner.

4.2. Exporting Road Network Attributes

OSM data is freely available in ESRI shape file format and XML format which can be downloaded from GeoFabrik web service (http://download.geofabrik.de). This data is updated frequently. An advantage of using the OSM dataset is that its most up-to-date version can be downloaded freely from the GeoFabrik web service for any region of the world. Continent-wise and region-wise data is available.

Figure 11 shows the procedure of exporting road network attributes from OSM. After downloading the OSM data in XML format from the above-mentioned web service. This data is processed in Quantum GIS software to convert it into a spatial lite DB file. Then this DB file is processed by the function “export OSM topology to spatial lite” in QGIS to convert it into an output layer. Attribute table of this output layer shows the attributes of the road network of a specific area for which the XML file is downloaded. The attribute table contains various fields such as lanes, maximum speed, cycle way, highway, road type etc. which provide information regarding a road. Table 2 shows the attributes of a road network in Antwerp. All the attributes of a road are not available in OSM dataset.

FIGURE 11. Exporting road network data from OSM.

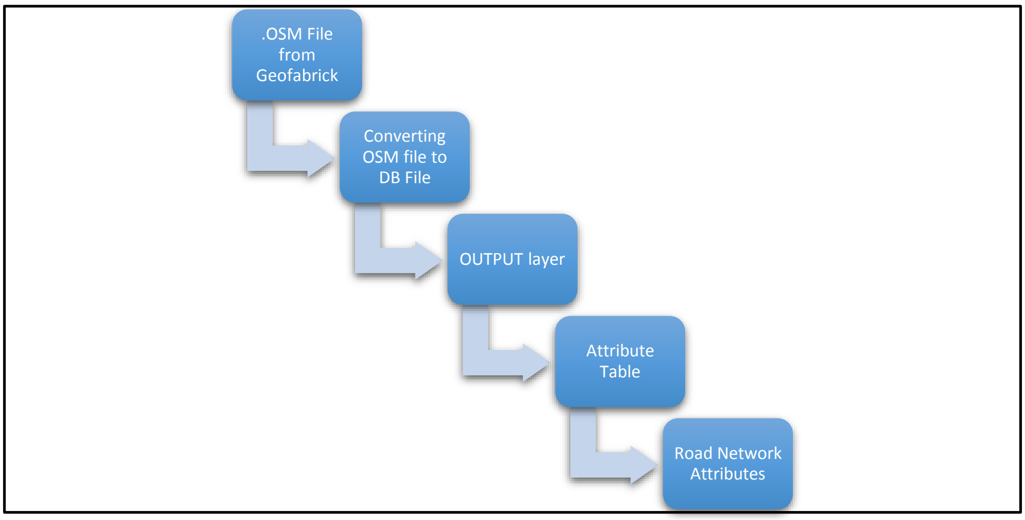


Figure 12 & 13 show “link type” statistics for Belgium available in the OSM dataset. It explains that out of total 762057 digitized links, 12431 are tagged as “motorways”, similarly 30261 links are tagged as “primary”.

FIGURE 12. "Link type" statistics in the OSM dataset of Belgium.

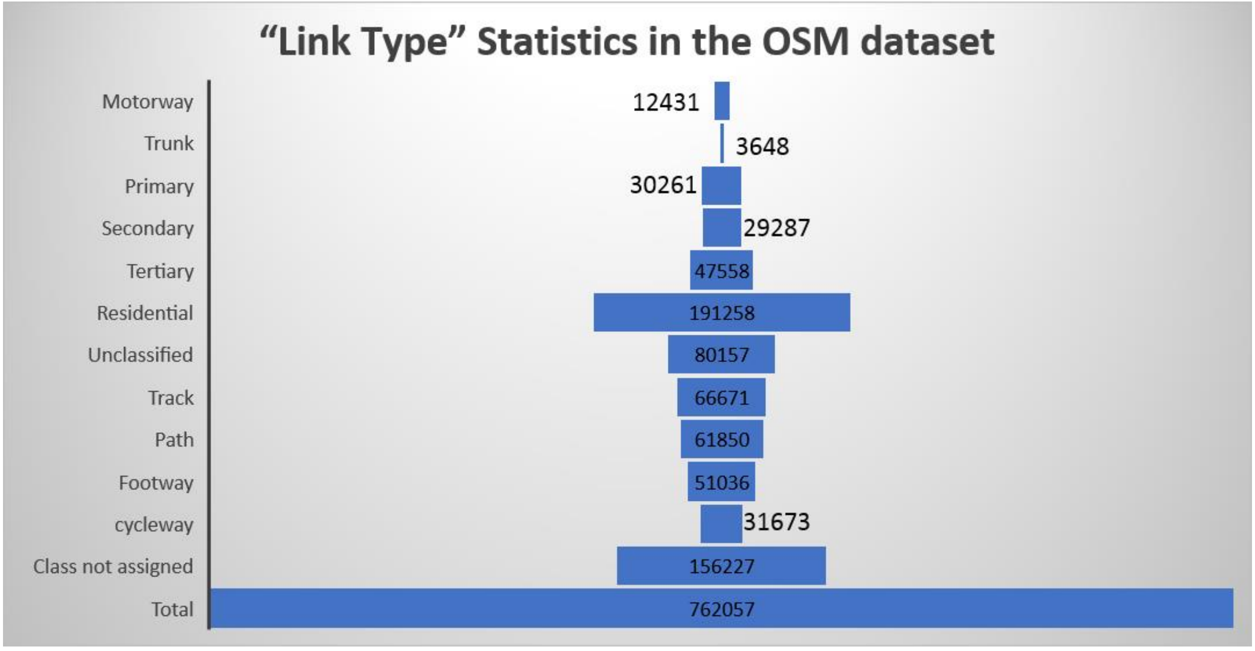


FIGURE 13. Variable count in the OSM dataset of Belgium.

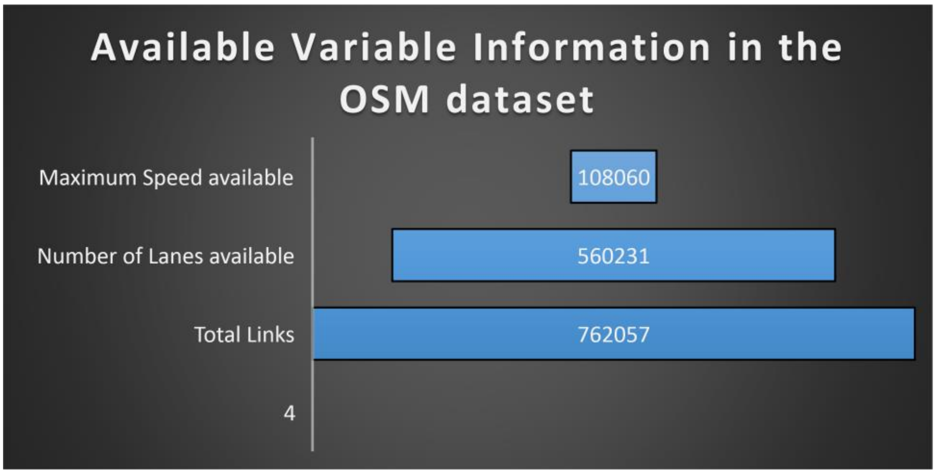


Figure 13 shows that information regarding "maximum speed" is available only for 108060 roads in the OSM dataset of Belgium and information regarding "number of lanes" is only available for 560231 roads. These figures explain that there are a lot of missing values in the road network attributes of the OSM dataset as out of 762057 links only 108060 roads have information regarding maximum speed and 56231 roads have information related to number of lanes.

Missing values in the OSM data set affects the quality of OSM data and limits its usage for the route planning applications. Table 2 shows road network attributes of OSM dataset of some part of Antwerp. There are a lot of roads in Antwerp for which "maximum speed" and "number of lanes" are not available.

4.3. GIS Attribute Table of OSM Road Network

*TABLE 2. Attribute table of road network of some part of Antwerp (*Open Street Map).

Id	**Road Type**	**One-Way**	**Maximum speed Km/h**	**Lanes**	**Bicycle-Lane**
4292149	residential	yes	50	missing	missing
4295454	residential	missing	30	missing	missing
4295463	residential	yes	30	missing	missing
4295468	residential	missing	30	missing	missing
4295469	residential	missing	missing	missing	missing
4295470	residential	missing	missing	missing	missing
4295471	residential	missing	missing	missing	missing
4295472	residential	yes	50	missing	missing
4295836	residential	missing	30	missing	missing
4295848	secondary	yes	50	1	no
4295899	secondary	yes	50	2	no
4310130	motorway_link	yes	70	1	missing
4310131	motorway_link	yes	100	1	missing
4310132	motorway_link	yes	100	1	missing
4310134	motorway_link	yes	100	1	missing
4310135	motorway_link	yes	missing	2	missing
4310136	motorway_link	yes	missing	3	missing
4310141	motorway_link	yes	missing	2	missing
4310143	motorway_link	yes	missing	3	missing

Table 2 explains that there are a lot of missing values in the attributes of roads in the OSM dataset of Antwerp. Secondly there are some mistakes in already filled values such as roads with ID 4310141, 4310132 and 4310131 showing the maximum speed of 100km/h with only one lane. 100 km/h is not present in the speed limit categories in Belgium. For, school areas, built up areas, national roads and motorways, speed limits are 30km/h, 50km/h, 90km/h and 120km/h respectively.

FIGURE 14. Road Network Map of some part of Antwerp

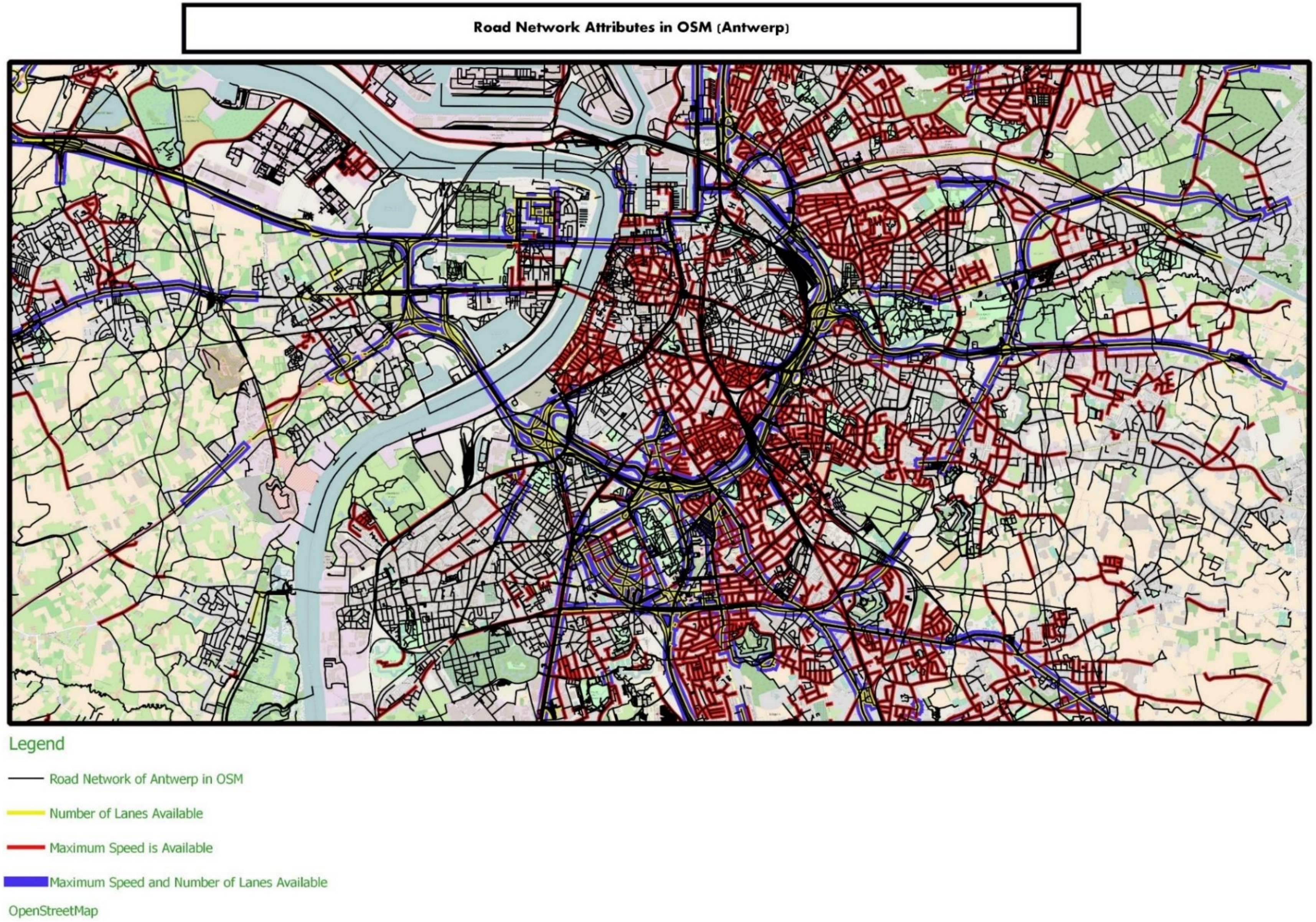
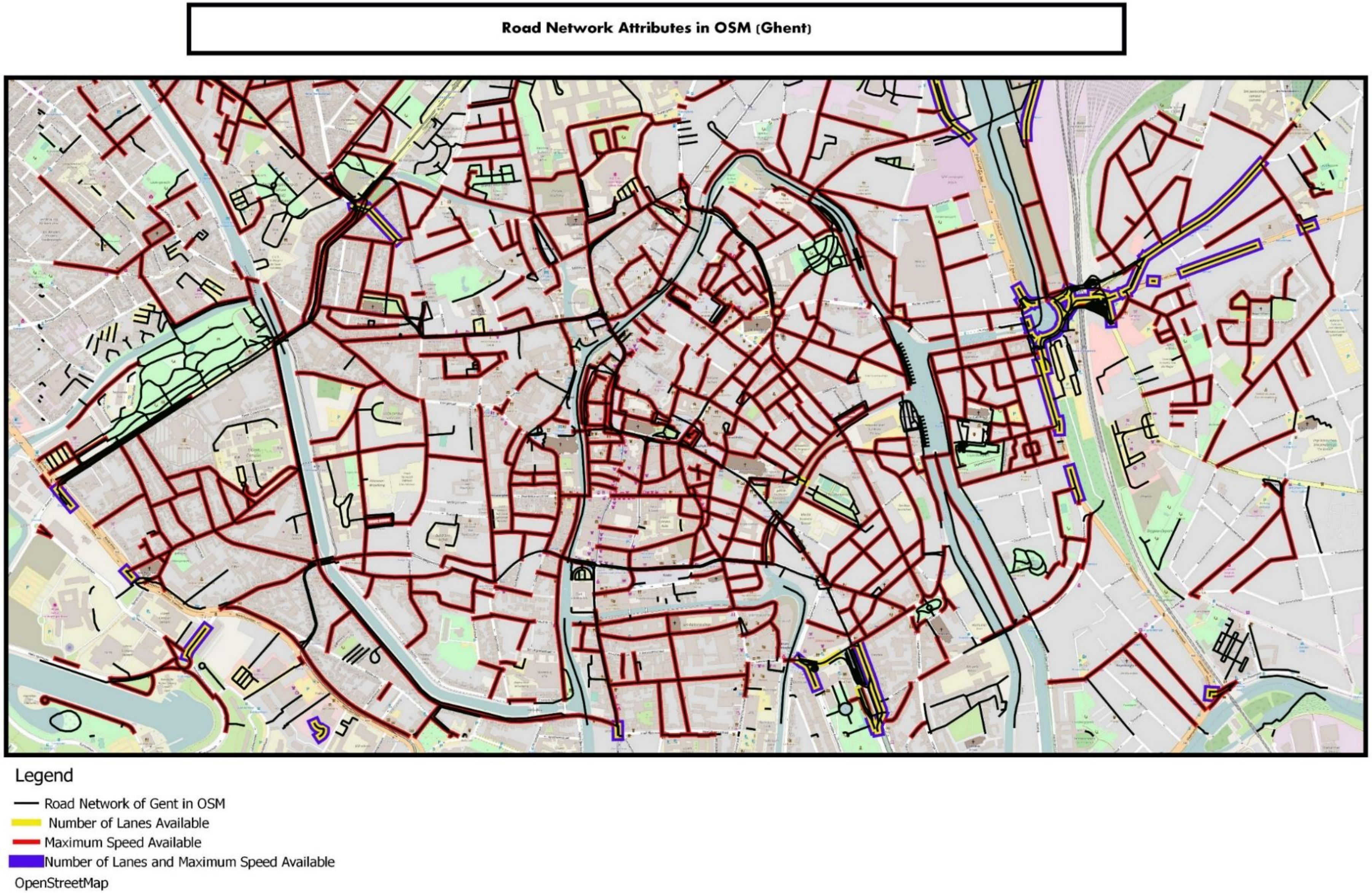

Road Network Attributes in OSM (Antwerp)
Legend
Road Network of Antwerp in OSM
Number of Lanes Available
Maximum Speed is Available
Maximum Speed and Number of Lanes Available
OpenStreetMap

FIGURE 15. Road Network map of Some part of Ghent


Road Network Attributes in OSM (Ghent)
Legend
Road Network of Gent in OSM
Number of Lanes Available
Maximum Speed Available
Number of Lanes and Maximum Speed Available
OpenStreetMap

5. MISSING VALUES IMPUTATION IN THE OSM DATASET

Missing data values is a persistent problem faced by most of the researchers. A lot of methods have been developed to draw inferences from a dataset of missing values (A. Little & Rubin, 1987). In order to deal with the missing data values, multiple imputation is recommended as the best practice by statisticians (Little & Rubin, 2002). Multiple imputation is known as the gold standard among statisticians for the treatment of missing values (Baraldi & Enders, 2010). There are several imputation approaches to deal with the missing data and these are quite useful practically because once the dataset is free from missing values then the complete data set may be used for several useful purposes (Raghunathan, Lepkowski, Van Hoewyk, & Solenberger, 2001). The OSM road network dataset have a lot of missing values which affects its utility for professional purposes. Several route planning applications operate on the basis of the OSM dataset but missing attribute values affects the utility of OSM in such applications.

Advancement in the statistical methods and development of softwares to complete the missing datasets allow the researchers to generate their own code according to specific missing data problem. It enables the researchers to estimate substantive models to predict the missing values in the dataset (Van Buuren, 2012). Some techniques for the imputation of missing values are discussed in the following sections. Before the introduction of imputation techniques, missing data mechanisms are needed to be explained.

5.1. Missing Data Mechanisms

5.1.1. Missing Completely at Random (MCAR)

In case of Missing Completely at Random, probability of being missing a data value is same for all the cases. It means that reasons of missing data are unrelated to the data. An example of MCAR is that when a random sample is taken from a population then each person of that population has the same chance of being selected in the sample. Those members of the population who are not selected in the population are Missing Completely at Random (Van Buuren, 2012).

Let Y be a nXp matrix that contains the data values on p variables and there are n units in the sample. There are some observed values (Y_{obs}) and some missing values (Y_{mis}) and matrix R stores the locations of missing values in Y. Elements of R and Y are denoted as r_{ij} and y_{ij}. If y_{ij} is observed then r_{ij} =1 and if y_{ij} is missing then r_{ij} =0 So the distribution of R depends upon Y= ($Y_{obs,}$ Y_{mis}). Let Ψ consists of the parameters of missing data model then the expression for missing data model can be written as $\Pr(R\,|\, Y_{obs,} Y_{mis,} \Psi)$ and data is said to be MCAR if

$$\Pr(R=0\,|\, Y_{obs,} Y_{mis,} \Psi) = \Pr(R=0\,|\Psi)$$

It means that probability of being missing a data value depends only on model parameters.

5.1.2. Missing at Random (MAR)

In case of Missing at Random, probability of being missing a data value is same within groups that are defined by the observed data. Missing at Random (MAR) is a broader class than Missing Completely at Random (MCAR). An example of MAR is that when a sample is selected from a population and probability of being selected depends on some known characteristics. MAR is more realistic and general as compared to the MCAR and most of the software packages for the imputation of missing values start from the assumption that data is missing at random (Van Buuren, 2012). The data is said to be MAR if the probability of being missing is dependent on observed information

$$\Pr(R=0 \mid Y_{\mathrm{obs}}, Y_{\mathrm{mis}}, \Psi) = \Pr(R=0 \mid Y_{\mathrm{obs}}, \Psi)$$

5.1.3. Missing not at Random (MNAR)

In case of Missing not at random, probability of being missing a data value depends on the reasons that are unknown to us. An example of MNAR is that public opinion research is carried out and some persons are not responding due to unknown reasons. MNAR is considered as complex case and strategy to handle MNAR is to find some data about the causes of missingness. In case of MNAR probability of being missing also depends on un-observed information and it includes Y_{mis} itself (Van Buuren, 2012).

$$\Pr(R=0 \mid Y_{\mathrm{obs}}, Y_{\mathrm{mis}}, \Psi)$$

Dataset extracted from OSM (OSM) has a lot of missing values. Missing data could have two major reasons. One is that there are less number of contributors in an area that is why OSM data set is not complete. Secondly there are some tagging problems that causes missingness in the data set of OSM so that is why OSM data is MAR. Techniques to handle missing data values are discussed in the following section.

5.2. Mean Imputation

It is a very simple method to deal with the missing data values. In this method missing data values are replaced by mean. This method has a serious disadvantage as it distorts the distribution in many ways. Mean Imputation disturbs the relations between variables, underestimates the variance, creates biasness in estimates. It is not a recommended method to deal with the missing values by statisticians due to its disadvantages (Donders, van der Heijden, Stijnen, & Moons, 2006).

5.3. Regression Imputation

A regression imputation produces smarter imputations as it integrates the knowledge of other variables. It involves the estimation of fitted model based on the observed data values. Fitted model helps to estimate the missing data values which then serve as replacement for the missing data.

Regression model produces unbiased estimates of the means and regression weights if the explanatory variables are complete. In case values are missing at random (MAR) then unbiased estimates are produced subject to the condition that factors that influence missing values are the part of the regression model. In regression imputation, variability of the imputed data is underestimated in a systematic way. The degree of underestimation depends upon the proportion of missing cases and explained variance (A. Little & Rubin, 2002). The Normal linear regression model is given as follow:

$$\text{Predict. } \dot{y} = \hat{\beta}_0 + X_{\mathrm{mis}}\, \hat{\beta}_1$$

- X_{mis} is the complementing subset of rows of X for which y is missing.
- The imputed values in Y are shown by ẏ.
- $\hat{\beta}_0$ and $\hat{\beta}1$ are estimates calculated from the observed data.

5.4. Stochastic Regression Imputation

Stochastic Regression Imputation (SRI) is basically the improvement of the regression imputation. SRI method estimates the slope, intercept and residual variance through a linear model and then imputes the missing values according to these provisions. Hence the improvement is made by adding a suitable amount of random noise in the predicted value. Let’s consider that observed data values are normally distributed around the regression line and estimated standard deviation is equal to σ^2 with a mean zero. A random value is drawn from the normal distribution with standard deviation σ and a mean of zero. This random value is added to the predicted value. In this way amount of uncertainty in the predicted value is reduced (Van Buuren, 2012).

$$\text{Predict + noise. } \dot{y} = \hat{\beta}_0 + X_{mis}\,\hat{\beta}_1 + \varepsilon$$

- Where ε is randomly drawn from normal distribution as $\varepsilon \sim N(0, \sigma^2)$

The main advantage of the Stochastic Regression Imputation is that it produces not only the unbiased estimates of regression weights but also it preserves the correlation between the variables. Stochastic Regression Imputation does not handle all the problems and there are a lot of refinements that are required. However, the main idea to draw from the residuals is very powerful that forms the basis for further research.

5.5. Indicator Method

A popular method to deal with the missing data values is the missing-indicator method. A new indicator or dummy (0/1) variable is created for each independent variable that has missing values, “1” describes that there is a missing value on the original variable and “0” indicates the observed value. In simple words, each missing value is replaced by “0”. To estimate the association between the outcome variable and independent variable, indicator variable is always included along with the original variable although it is coded. Benefit of using the indicator method is that all the subjects are included and there is no need to exclude any subject. But this is not a recommended method by the statisticians to deal with the missing values even when the values are missing completely at random as it produces biased regression estimates (Donders et al., 2006). Indicator method produces biased regression estimates even when there is low amount of missing values in the dataset (Van Buuren, 2012).

Although single imputation of missing values, using regression imputation and stochastic regression imputation produces good results but still there is some room for improvement. Missing values can be imputed more accurately by using multiple imputation. Also single imputation underestimates the standard error of estimates (Zhang, 2003b).

5.6. Multiple Imputation

Multiple imputation (MI) is considered as the best method to deal with missing data values as it offers a mechanism to deal with the inherited uncertainty of imputations (Van Buuren, 2012). It was proposed by Rubin back in 1970 to deal with the incomplete data. Rubin stressed that incomplete data should be dealt based on some principal methods. A principal method to deal with the missing data is multiple imputation which consists of three steps (Rubin, 1977).

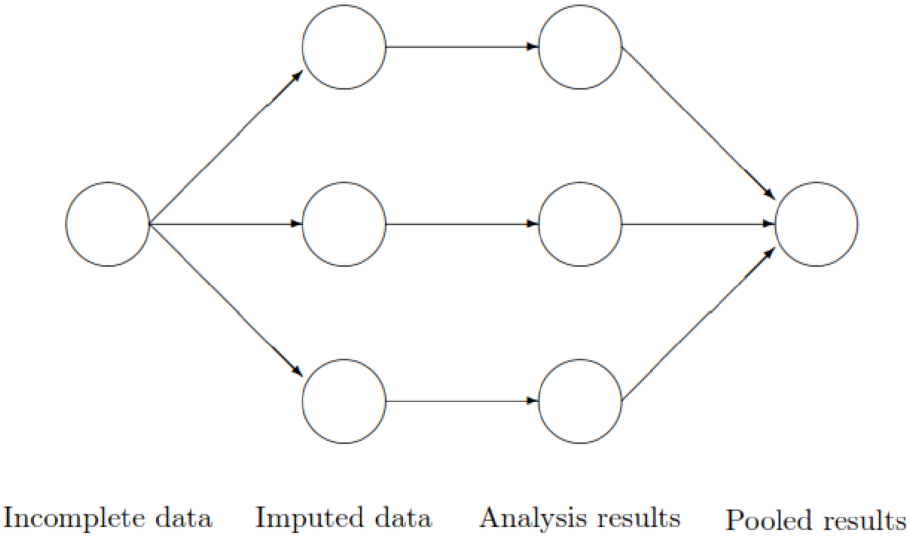
Rubin argued that it is not the correct way to impute one value for a missing value in general. Whenever a model is estimated from the observed data then even for that model Imputed values could not be calculated with certainty. His solution was really simple to create multiple imputations for a missing data value so that uncertainty of imputations may be reflected.

Analysis starts with observed values of incomplete data. Several versions of completed data are produced by the multiple imputation through replacement of missing data with the plausible data values. A distribution is being modelled for each missing entry and plausible data values are drawn from that distribution (Van Buuren, 2012).

Basically, it is a complex process to impute the missing values in a data set as normally datasets have complex patterns. At the start, a researcher develops a statistical model and relate all the variables in a dataset. After that researcher uses the observed data values and that model to find out the distributions of model parameters such as slope, intercept, and regression variance. Based upon the estimated values of these distributions, plausible values of model parameters are generated. Using the observed data values and plausible values of model parameters, values for missing data are generated. In this way one complete data set is created and this process is repeated m times.

In order to generate imputations from scratch it is necessary to have the knowledge of Markov Chain Monte Carlo method and Bayesian statistics. Fortunately, there are some software packages that make it needless to generate imputations from scratch (Salkind & Rasmussen, 2007)**.** Main steps involved in Multiple imputation are explained in the Figure 16.

FIGURE 16: Scheme of main steps in MI source flexible imputation of missing data (Honaker et al., 2011).



5.6.1. Step 1

It involves the imputation of missing values to create “m” complete data sets by using a proper imputation model based on observed values.

The first step is to draw the random samples from the imputation model. This is the most important part of the MI. This task involves developing an imputation model and drawing random samples from it. An imputation model should preserve the distributional relationships between the observed data values and missing values. In this way inferences obtained from imputed data set are not biased. For example, if imputation model does not include those variables that are to be used in the inferences from the imputed complete data, then it will create biasness.

5.6.2. Step 2

It involves the analysis (by fitting a regression model) of the complete imputed datasets by treating every complete imputed dataset as a real complete dataset. Softwares and complete-data procedures may be utilized directly.

5.6.3. Step 3

In this step analysis results from "m" imputed datasets are combined in a suitable way to obtain repeated imputation inference. Variances of this combined result comprises of "between imputation variance" and "within imputation variance". In this way uncertainties in the "m" imputed datasets are correctly merged in the final inference.

The theory for making a repeated imputation inference is derived from a Bayesian model (Rubin,1987). Let Q be a generic scalar quantity to be estimated, such as treatment effect, odds ratio or regression coefficient. Then the observed-data posterior distribution of Q is given by

$$P(Q|Y_{\text{obs}}) = \int P(Q|Y_{\text{obs}\,.}\;Y_{\text{mis}})P(\,Y_{\text{mis}}|Y_{\text{obs}})dY_{mis}$$

$P(Q|Y_{\text{obs}})$ is posterior distribution of Q given the observed data (Y_{obs}). $P(Q|Y_{\text{obs}\,.}\;Y_{\text{mis}})$ is the posterior distribution of Q in the (hypothetically) completed data set. $P(\,Y_{\text{mis}}|Y_{\text{obs}})$ is the posterior distribution of missing data given the observed data. In order to interpret the above-mentioned equation it is more convenient to read it from right to left. $P(\,Y_{\text{mis}}|Y_{\text{obs}})$ is used to make imputations for the Y_{mis} . $P(Q|Y_{\text{obs}\,.}Y_{\text{mis}})$ means that quantity of interest is being calculated from the dataset that is being completed hypothetically, which is $P(\,Y_{\text{mis}}|Y_{\text{obs}})$ (Van Buuren, 2012). These steps are then repeated for the new draws of $Y_{\text{mis.}}$ Multiple imputation under normal linear imputation model can be presented as:

$$\text{Bayesian multiple imputation. } \dot{y} = \hat{\beta}_0 + X_{\text{mis}}\,\hat{\beta}_1 + \varepsilon$$

- Where Ɛ is randomly drawn from normal distribution as Ɛ ~ N (0, σ^2)
- $\hat{\beta}_0$, $\hat{\beta}_1$ and σ are drawn from the posterior distribution

5.6.4. Simple Example

Assume there is a variable Y that is a function of X and Z such as

$$Y \sim X\,,Z$$

$$Y \sim \beta_o + \beta_1 X + \beta_2\,Z$$

There are some missing values in X and there is a need to impute the missing values in X, general imputation model for imputing the values of X using regression is presented as follow

$$X \sim \beta_o + \beta_1 Z + \beta_2\,Y$$

Using the above-mentioned imputation model and observed values of variables, values of X are imputed. Values of the coefficients $\beta_{o,}$ β_1, β_2 are estimated from the observed data and then these values are used to impute the missing values of X. When the values X are imputed first time then it is possible to estimate the values of coefficients $\beta_{o,}$ β_1, β_2 again from the imputed dataset. Using the estimates of variance and estimated coefficients of imputed X it is possible to pick m values of β^ out of its distribution. A matrix called VCV matrix contains all the estimated coefficients $\beta_{o,}$ β_1, β_2 of imputation model. VCV matrix represents the uncertainty of β coefficients which are used to predict the missing values. When there are m values of β^ then it possible to impute m values for the missing values and creating m many datasets.

$$\begin{tikzcd} Y \sim \beta_o + \beta_1 X + \beta_2 Z \\ X \sim \beta_o + \beta_1 Z + \beta_2 Y \arrow[u] \end{tikzcd}$$

Using each of the imputed data sets of X, coefficients of model $Y \sim \beta_o + \beta_1 X + \beta_2 Z$ are calculated and value of Y is calculated with more certainty.

5.6.5. Multivariate Imputation through Chained Equations

Most of the times there are many variables that have missing values in a dataset. In case of OSM almost all the variable which are important for traffic assignment have missing values. Multivariate Imputation through chained equations is the name of the software that is being used for the imputation of missing values in a multivariate data by Fully Conditional Specification (FCS). FCS is an algorithm of multiple imputation. In this method, a separate imputation model is specified for each variable that has missing values. In this way, missing values are imputed sequentially starting from the first variable that has missing values.

5.6.5.1. Procedure with an Example

- **STEP 1**: First of all, for all the missing values in the dataset, fill out the missing values through random draws from the observed values. Assume there is a dataset with five variables X, Y, Z, W, Q and there are missing values in these variables as shown in the Table 3.

TABLE 3: MICE Explanation.

X	Y	Z	W	Q
2	3	-	8	4
-	1	8	-	6
4	2	3	-	1

According to the first step, missing values are filled with some random numbers drawn from the observed values. So, the table 4 showing the filled values which were missing in Table 3.

TABLE 4: Imputation of missing value in variable Y.

Y	X	Z	W	Q
2	3	2	8	4
4	1	8	8	6
4	2	3	1	1

- **STEP 2**: Next step is to move through columns and impute each single variable by using some method. In this case missing value of variable Y can be imputed by the imputation model

Y ~ X , Z , W, Q

An important point to note is that missing value of Y is being imputed by the values of other variables X, Z, W, Q. Although values 2, 8, 1 were randomly filled in the missing values of variables Z and W but these are also contributing in the imputation of missing value of variable Y as shown in Table 4. Missing value in the variable Y was randomly filled as 4. But after imputation it is possible that value is changed to 3.

TABLE 5: Imputation of missing value in variable Z.

Y	**X**	**Z**	**W**	**Q**
2	3	**2**	8	4
3 (imputed value)	1	8	**8**	6
4	2	3	**1**	1

Similarly, next missing value exists in the variable Z which was randomly filled as 2 in the start. The imputation model for the missing values of Z could be written as

Z ~ X , Y , W, Q

An important point to note is that value of variable Z is being imputed by the values of other variables X, Y, W, Q. For the imputation of missing value of Z, values of all other variables are contributing particularly imputed value of variable Y is also contributing for the imputation of missing value in Z. Similarly randomly filled values of variable W which are 8 and 1 are also contributing for the missing value imputation of Z as shown in table 5.

TABLE 6: Imputation of missing values in all columns.

Y	**X**	**Z**	**W**	**Q**
2	3	**1.8** (imputed value)	8	4
3 (imputed value)	1	8	**8**	6
4	2	3	**1**	1

Similarly missing values of variable W can be imputed by the imputation model W ~ x , Y , Z, Q and imputed values of X and Y which are 3 and 1.8 respectively, will contribute for the imputation of missing values in variable W as shown in table 6 . All the missing values of variable will be imputed in this way.

- **STEP 3**: Next step is to replace all the randomly filled missing values with the fitted values. Step 3 is repeated again and again until certain number of cycles are completed

TABLE 7: Imputed Values.

Y	X	Z	W	Q
2	3	1.8 (imputed value)	8	4
3 (imputed value)	1	8	7 (imputed value)	6
4	2	3	2 (imputed value)	1

- Do this process m times to create m imputed data sets and in this way missing values are imputed multiple times.

5.6.6. Reasons to use Multiple Imputation

- Multiple imputation is exclusive in the sense that it offers a mechanism to deal with the inherent uncertainty of the imputations.
- It separates the solution of the missing data problem from the solution of the complete data problem. The missing data problem is solved first, the complete data problem next.
- It considers randomness of data and that is why its results are unbiased

5.7. Software Packages in R to deal with the missing data

Following software packages are available in R to deal with the missing values in a dataset

- Package Amelia
- Package Mice
- Knn Imputation
- MIMCA
- Miss MDA
- MIssForest

Explanation of all these packages is present in Table 8. Each imputation package has its own specification with rest to the type of variable to be imputed. After comparison of all these packages Amelia and MICE are selected for the imputation of missing values in OSM dataset. Amelia and MICE are best to use in case of multiple imputation of missing values in the OSM dataset as missing values exist in all types of variables (Catagorical, Binary, Continuous) in the OSM data set.

Both these packages use the technique of Multiple Imputation to fill the missing data values. Package MICE works based upon the principle of chained equations as explained previously. In multiple imputation, random draws of parameters are taken from the posterior distribution. In Amelia, EM algorithm is combined with the bootstrap approach to take draws from the posterior distribution. Advantage of Amelia is that it combines the comparative speed and ease of use of algorithm with the power of multiple imputation.

TABLE 8: Comparison of Missing Values Imputation Packages Available in R

kNN Imputation	MIMCA	Miss MDA	MissForest	MICE	Amelia
• Knn imputation uses k-nearest approach in order to impute missing values in a dataset. • For every missing value in a dataset, kNN imputation identifies k closest observations based on the Euclidean distance and calculates the weighted average. • Library (DMwr) is available in R package to use kNN imputation for the completion of missing values. • Advantage is that all the missing values are imputed in all variables in one call only • Restricted to only one type of variable, it cannot deal with different types of variables (Stekhoven & Buhlmann, 2012).	• Its a multiple imputation package to deal with the missing values in the categorical data • It uses the Multiple Correspondence Analysis (MCA) to complete the missing values in categorical data • MIMCA is less time consuming on the datasets of higher dimensions as compared to the other datasets • Major disadvantage is that, it works only with categorical data. All types of variables cannot be dealt using MIMCA (Audigear, Agrocampus, & Josse, 2015).	• This package imputes the missing values in a multivariate dataset • Imputation methods include principal component analysis for continuous variables, multiple correspondence analysis for the categorical data and multiple factor analysis for the multi table data. • Primarily it is designed for the single imputation of missing values in a multivariate data set (Josse & Husson, 2016).	• MissForest is an R package that imputes the missing values in different types of variables simultaneously • Imputation method of this package averages many regression trees and a random forest constitutes a multiple imputation scheme • MissForest has high computational efficiency and attractive package • Its working is identical to the MICE package • Less efficient as compared to the results of MICE package (Stekhoven & Buhlmann, 2012).	• Primarily designed to use the technique of Multiple Imputation (MI) • Algorithm explained in section 5.5.5 • It allows the column wise specification of the imputation method • Allows the selection of predictor matrix. In this way irrelevant variables can be excluded from the imputation model • It multiply imputes the missing values in all types of variables simultaneously and preserves the relationship among different variables (Van Buuren & Groothuis-Oudshoorn, 2011).	• Primarily designed to use the technique of Multiple Imputation (MI) to deal with missing values • Algorithm explained in section 5.7 • It is very simple to use and less time consuming. A person who does not know R programming language can use this package for Multiple Imputation of missing values • Presents very efficient results and works with all types of variables. Even it is used for the imputation of time series variables (Honaker et al., 2011). •

5.7.1. Package Amelia

The Amelia package developed by Harvard university, allows the users to complete the missing values in the incomplete datasets so that complete dataset can be used for the fruitful purposes. In this way, it helps to avoid inefficiencies, biases and incorrect estimates that can result due to the presence of the missing values in the dataset. As in case of OSM dataset, missing values affect its potential usability in the route planning applications.

The main advantage of the Amelia package is that it performs multiple imputations. As discussed in the previous section is that this imputation technique is much better as compared to other imputation methods and it increases the efficiency to a great extent. Furthermore, ad-hoc imputation methods such as mean imputation can lead towards inefficiency and biasness in covariances and variances.

Creation of multiple imputations is a burdensome process as the algorithms involved in this process are very technical in nature. Amelia package helps the users in the application of these algorithms to apply the imputation models and generated multiply imputed datasets. Package Amelia goes numerous steps beyond the capabilities of other packages to deal with the missing values. It uses the bootstrap based EMB algorithm which can impute all types of variable and with many more observations in a limited amount of time (Honaker, Joseph, King, Scheve and Singh., 1998-2002). EMB Algorithm is the combination of Expectation-Maximization (EM) Algorithm and Bayesian model. Amelia's EMB algorithm makes it unique and simple and faster than other alternative packages. Features of Amelia package also allows the users to make accurate and much more valid imputations for time series, cross sectional data sets (Honaker and Singh., 2010).

Imputation model in the package Amelia assumes that completed dataset which contain both observed and unobserved values, are multivariate normal. If there is dataset (n × k) donated as D with observed values D^{obs} and missing values D^{mis} then is

$$D \sim Nk(\mu, \Sigma)$$

which states that dataset D has a multivariate normal distribution with mean vector μ and covariance matrix Σ. Multivariate normal distribution is a crude approximation of the true distribution of the data. Multivariate normal distribution is used to describe any set of correlated real valued random variables each of which clusters around a mean value. Imputation models of Amelia works well even if there are missing values in the categorical variables and mixed datasets.

5.7.1.1. Algorithm

In the generation of multiple imputations for an incomplete dataset, the most important concern is the complete data parameters (θ). In simple words, it means that when the dataset is completed using any imputation model then further imputations are generated based upon the parameters of recently completed dataset. In order to write down the model of the data, it is assumed that observed data is D^{obs} and M is the missingness matrix. Thus, the likelihood of the missing data and observed data is

$$p(D^{obs}, M \mid \theta)$$

This likelihood can be broken into two components

$$p(D^{obs}, M \mid \theta) = p(M \mid D^{obs})p(D^{obs} \mid \theta)$$

In order to impute the missing values multiple times, the most important thing is the complete data parameters that is why likelihood can be written as follow

$$L(\theta|\ D^{obs}) \propto p(D^{obs}|\theta)$$

By using the law of iterated expectations, it can be rewrite as

$$p(D^{obs}|\theta) = \int p(D|\theta)\ dD^{mis}.$$

With this likelihood and a flat prior on θ, the posterior can be written as follow

$$p(\theta|D^{obs}) \propto p(D^{obs}|\theta) = \int p(D|\theta)dD^{mis}$$

One of the main computational difficulty in the multiple imputation of the missing data is taking the draws from the posterior. EM algorithm is one of the simple computational approach that helps to take draws from the posterior (Zhang, 2003b). EMB algorithm that is being used in the package Amelia combines the boostrap approach and EM algorithm to take draws from the posterior. For each draw, data is bootstrapped to simulate estimation uncertainty and then EM algorithm is applied to find out the mode of the posterior. Posterior model parameters are estimated and data is multiply imputed to minimize the uncertainty of the imputations (Honaker, King, Blackwell, & others, 2011).

Amelia-II exists as a package for the R statistical programming language. Using the knowledge of the R language, Amelia-II can be run on command line. Alternatively, Amelia-II is also available as AmeliaView, where Interactive Graphical User Interface (GUI) allows the users to set options and run Amelia without the knowledge of R programming language.

5.7.1.2. Advantages

- Package Amelia goes numerous steps beyond the capabilities of other packages to deal with the missing values. It uses the bootstrap based EMB algorithm which can impute all types of variable and with many more observations in a limited amount of time.
- Features of Amelia package also allows the users to make accurate and much more valid imputations for time series, cross sectional data sets.
- It is very simple to use and less time consuming. A person who does not know R programming language can use this package for Multiple Imputation of missing values

5.7.2. Package MICE

Multivariate imputation by chained equations is the name of the software package in R for the imputation of incomplete values in a dataset. The main advantage of the MICE package is that it efficiently imputes the missing values in multivariate dataset where there are missing values in more than one variables. MICE package uses the Fully conditional specification (FCS) algorithm and it has the functions to impute the multilevel data, data handling, automatic predictor selection, specialized pooling and model selection. In order to impute the missing values in the categorical data, MICE package works really well (Van Buuren & Groothuis-Oudshoorn, 2011).

Missing data values that occur in more than one variables presents a special challenge. An ideal approach to deal with the missing data in more than one variables is fully conditional specification. FCS specifies the multivariate imputation model on variable by variable basis by a set of conditional densities. It means that one imputation model is being specified for each individual missing variable. The basic idea of FCS algorithm is quite old and it has been proposed by using different names such as Variable by variable imputation (Zhang, 2003b), Stochastic Relaxation (Zhang, 2003b), Sequential Regressions (Raghunathan et al. 2001) chained equations and fully conditional specification (Zhang, 2003b). Package MICE has three major functions

- Generating Multiple Imputations
- Analyzing the Imputed datasets
- Pooling the analysis Results

5.7.2.1. Advantages

Specific advantages of the MICE package are

- Column wise specification of the imputation model
- Subset selection of predictors
- Efficiently imputes missing values in more than one variables
- Effectively deals with the missing values in categorical variables
- Preserves the relation in the data
- Preserves the uncertainty about these relations.

5.7.2.2. Imputation Methods in MICE

An elementary imputation model is being specified for each of the incomplete variables in MICE. Specified Imputation method takes a set of predictors which are complete at that time and returns a single imputation for the missing entry which is present in the target column. There are many imputation models that are supplied by MICE package. All of these imputation models have different names which are used for the identification of imputation model. Table 9 describes the built-in imputation functions available in MICE package (Van Buuren & Groothuis-Oudshoorn, 2011) .

TABLE 9: Imputation Methods available in MICE.

Method	Description	Scale Type
pmm	Predictive Mean Matching	Numeric
norm	Bayesian Linear Regression	Numeric
Norm.nob	Linear Regression, non-Bayesian	Numeric
mean	Unconditional Mean Imputation	Numeric
2l.norm	Two level linear model	Numeric
logreg	Logistic Regression	Factor 2 level
polyreg	Polytomous Regression	Factor > 2 level
lda	Linear Discriminant Analysis	Factor
sample	Random sample from observed data	any

All the above-mentioned methods have utility according to the type of variable that is being imputed. For example, in order to impute a categorical variable, "*polyreg*" method could be used. Similarly, for the imputation of binary variable, "logreg" method could be used. "*Norm*"and "pmm" could be used for the imputation of numeric type of variable.

In the dataset extracted from OSM, there are different types of variables (Categorical, Continuous, Binary) which have missing values such as number of lanes, maximum speed, one way, road Type etc. Hence, missing values in each variable could be filled according to the type of variable using MICE and Amelia packages.

6. DATA PREPARATION AND ANALYSIS

6.1. Data Collection

This chapter explains the process followed for the collection and preparation of the data. Later the methodology adopted for performing different experiments on OSM datasheet has been described. Finally, the utilization of the OSM data sheet in the imputation models of Amelia and MICE has been discussed. Moreover, this chapter, dives deep into the complicated and subtle changes needed to be made in the dataset for accurate outcomes.

6.2. OSM Dataset of London City

Initially it was decided to work on the OSM dataset of Flanders region in Belgium but an analysis related to missing values was performed on the OSM dataset of Belgium and it was concluded that percentage of missing values was very high in case of Belgium. This analysis has been described in the previous chapter. With that high percentage of missing values in the OSM dataset it was unfair to test the imputation models of Amelia and MICE Packages. The percentage of missing values in the OSM dataset is highly linked with the number of contributors/volunteers. As in case of Belgium percentage of OSM contributors is not so high that is why there are a lot of missing values that exist in the OSM dataset of Belgium. Literature review showed that percentage of OSM contributors is very high in case of United Kingdom as compared to Belgium that is why there is less percentage of missing values in the OSM datasets of UK.

London city was selected as a case study area and the OSM dataset of London city was downloaded using QGIS software. A special plugin 'Open layer Plugin' in QGIS was used to download the extensive road network OSM dataset of London city. There are various variables which are available in the attribute table of the shapefiles of London City. Out of these variables, five variables (Road Network Attributes) were selected for further processing which are important with respect to traffic assignment and these are listed below

- Number of Lanes
- Maximum Speed
- Road Type
- One way
- Bicycle Lane

6.3. Selection of Road Segments

To analyze the performance of imputation models of MICE and Amelia, it was necessary to collect the OSM road network dataset of London City in which attribute values of above mentioned variables were complete. In order to select those roads which, have full attribute values, whole geographical area of London city was selected in QGIS and the following filter was applied.

*'Number of Lanes is not null **AND** Maximum Speed is not null **AND** Road Type is not null **AND** One way is not null **AND** Bicycle lane is not null'*

6.4. Experimentation

Missing values were generated intentionally in the OSM dataset of London so that performance of various imputation models could be assessed. In this regard, it was decided to perform various experiments in which missing values were generated in different variables and performance of different imputation models were analyzed.

TABLE 10: Experiments on the OSM Dataset to Complete Missing Values

Experiments	**Maximum Speed**	**Number of Lanes**	**Road Type**	**One Way**	**Bicycle Lane**
Experiment 1	🟥	🟩	🟩	🟩	🟩
Experiment 2	🟩	🟥	🟩	🟩	🟩
Experiment 3	🟩	🟩	🟥	🟩	🟩
Experiment 4	🟩	🟩	🟩	🟥	🟩
Experiment 5	🟩	🟩	🟩	🟩	🟥
Combination Experiments					
Experiment 6	🟥	🟥	🟩	🟩	🟩
Experiment 7	🟩	🟥	🟥	🟩	🟩
Experiment 8	🟥	🟩	🟥	🟩	🟩

Table 10 shows the difference between difference experiments that were performed on the OSM dataset of London. Red color shows those variables in which missing values were created intentionally and green color shows variables in which there are no missing values in a particular experiment. Later, these missing values were filled using different imputation packages.

OSM volunteers follow a free form tagging scheme while adding information to the OSM database that is why there are a lot of different road categories are in the OSM datasets. For example, some person is tagging a road as 'primary road' and some other person is tagging the road as 'primary link'. Similarly, a lot of different names can be seen in the OSM road network dataset for the same category of road. It was difficult to manage the variable 'road type' due to the presence of many number of road categories in the OSM dataset. These different road categories were merged into five basic road categories so that variable 'road type' can be used in the further process in an efficient manner. These five basic road categories are presented as follow: -

- Motorway
- Primary Road
- Secondary Road
- Tertiary Road
- Residential Road/Access Road

6.5. Coding

To process the data sheet in R statistical programming software, it was necessary to code different nominal variables and these variables should be specified to the R packages. In our OSM road network dataset there were three nominal variables and it was necessary to code these variables so that these variables can be processed in R software. Coding is presented in Table 11.

Road Type

TABLE 11: Variable Coding (Road Type)

Road Type	Codes
Motorway	1
Primary Road	2
Secondary Road	3
Tertiary Road	4
Access Road	5

One Way

Similarly, for the variables ‘One way’ coding is presented as follow: -

TABLE 12: Variable Coding (One Way)

Variable (One Way)	Code
Road is One way (Yes)	1
Road is One way (No)	0

Bicycle Lane

Similarly, for the variables ‘Bicycle lane’ coding is presented as follow: -

TABLE 13: Variable Coding (Bicycle Lane)

Variable (Bicycle Lane)	Code
Separate Bicycle lane is present (Yes)	1
Separate Bicycle lane is present (No)	0

6.5.1. Example

The table 14 presents the example of a particularly coded road. It means that it is a primary one-way road with a maximum allowed speed of 50 Km/h, and number of lanes are 3 and there is no separate bicycle lane.

Table 14: Coding Example

Road ID	Road Type	Maximum Speed (Km/h)	Number of Lanes	One Way	Bicycle Lane
12432	2	50	3	1	0

6.6. General Procedure

Experiment 1-5

In the experiments 1-5, missing values were generated in one variable in the OSM dataset through R software. A sample of 1000 roads with full road attribute values were taken and 30% values were missed in one variable in a particular experiment. All other variables were having complete attribute values.

Experiment 6-8 (Combinations)

In the experiments 6-8, missing values were generated in two variables through R software. A sample of 1000 roads with full road attribute values were taken and 30% values were missed in two variables in a particular experiment. All other variables were having complete attribute values.

6.7. Experiment 1: Maximum Speed Missing

In experiment 1, missing values were generated in the variable 'maximum speed' and same procedure is adopted as explained in section 6.6. R code for generating missing values is given in Code Example 4.

6.7.1. R Code

CODE EXAMPLE 4: Generation of Missing Values

```
insert_nas <- function(x) {
 len <- length(x)
 n <- sample(1:floor(0.3*len), 1)
 i <- sample(1:len, n)
  x[i] <- NA }
 Maximum Speed<- sapply(Maximum Speed, insert_nas)
 Maximum Speed
  write.csv(Maximum Speed, "D:/Masterthesis/Extra/sheet.csv")
```

After generation of the missing values in the data set of 1000 roads, these missing values were imputed using the imputation models of MICE and Amelia. The Table 15 presents an example of the datasheet with the missing values.

6.7.2. Datasheet Example with Missing Values

TABLE 15: Datasheet Example

Road Type	Maximum Speed	One Way	Bicycle Lane	Number of Lanes
2	40	1	0	3
3	20	1	1	1
2	30	1	1	3
3	30	1	0	2
3	20	1	1	1
2	50	1	0	3
2	30	1	1	3
2	30	1	1	3
3	20	1	1	1
3	**NA**	1	1	2
3	**NA**	1	1	2
2	30	1	1	3
3	30	1	1	2
2	**NA**	1	1	3
3	**NA**	1	1	1
3	30	1	0	2
2	50	1	0	2
2	30	1	1	3
2	30	1	1	3
2	50	1	0	2
3	30	1	1	2
1	100	1	1	6
2	**NA**	1	1	3
1	100	1	1	5
2	**NA**	1	0	2

6.7.3. Missing Values Imputation Using Amelia Package

At first Package Amelia was used for the imputation of missing values in the OSM datasheet. In package Amelia, it is necessary to specify nominal or ordinal variables if they exist in the datasheet. As three variables were nominal in the datasheet that is why these variables were specified to Amelia as nominal variables. In the OSM dataset, three variable nominal variables are Road Type, One way and Bicycle Lane. Amelia package imputes missing values multiple times so it is necessary to mention the number of imputation in the R code. Each missing value is imputed 5 times in the OSM dataset and then the results are combined. R code for using Amelia is presented in the Code Example 5.

CODE EXAMPLE 5: Multiple Imputation using Amelia Package.

```
 > require(Amelia)

 > data (OSM_datasheet)

 > summary(OSM_datasheet)

 >a.out2 <- amelia (OSM_datasheet, m = 5, noms= “Road Type”, “One Way”, “Bicycle Lane”.
p2s=0)

>a.out2

>save(a.out, file = "imputations.RData")

>write.amelia(obj=a.out, file.stem = "outdata")
```

6.7.4. Missing Values Imputation Using MICE Package

Later, Imputation model of MICE package was used for the completion of missing values in the OSM datasheet. MICE package allows to use different imputation models for different variables. While writing the R code, argument mice () specifies the imputation method to be used for the completion of missing values. If the method is specified in one string then all the data columns will be imputed by the elementary function being indicated in that one string. Argument “ “ specifies those columns in the datasheet where there are no missing values. R codes for using MICE is presented in Code Example 6.

CODE EXAMPLE 6: Multiple Imputation Through MICE package.

```
>library(mice)

>OSM_Datasheet

>md.pattern(OSM_Datasheet)

> imp <- mice(OSM_Datasheet)

>imp <- mice(OSM_Datasheet, meth= c(“norm”, “ “, “ “, “ “))

>complete(imp)
```

6.7.5. Root Mean Square Error

Root mean square error is calculated to assess the difference between the actual values and the predicted values. It is basically the measure of the difference between the values predicted by a model or an estimator and the values that are actually observed. Individual differences between the actual values and the predicted values are called residuals and RMSE assists to aggregate them into a single measure of the predictive power (Chai & Draxler, 2014). The formula for calculating the RMSE is given in the following equation

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(X_{obs,i} - X_{model,i})^2}{n}}$$

RMSE has the same units as the units of the quantity being estimated, and it cannot be calculated unless the true value is known and. Missing values imputed by the imputation models of MICE and Amelia were compared with the actual observed values through the calculation of RMSE and It was analyzed that which imputation model gives more valid results.

6.7.6. Actual Values and Imputed Values (Maximum Speed)

TABLE 16: Actual and Imputed Values (Experiment 1)

Actual Value (Maximum Speed)	Predicted Value MICE	Predicted Value Amelia
20	25	34
30	30	33
30	30	40
30	30	20
30	30	39
30	40	43
30	30	39
30	30	39
30	30	28
30	30	37
30	30	33
30	30	41
30	30	16
30	40	40
…	…	…

Table 16 shows the actual values of maximum speed and predicted values of maximum speed. Some values are shown in this table, as total sample was consisting on the data of one thousand roads that cannot be shown here. Using the perditions of MICE and Amelia, RMSE is calculated and performance of the imputation models of MICE and Amelia is being analyzed and results are presented in Table 17. It is pertinent to mention that lower value of the RMSE shows the better performance of the model.

TABLE 17: Experiment 1 Results

Experiment	Missing Variable	Missing Values Percentage	Value Range	RMSE (Amelia)	RMSE (MICE)
Experiment 1	Maximum Speed	30%	20-100	9.19	6.16

There is no perfect value for root mean square error it is always dependent on the maximum and minimum values in the dataset. Root mean square error has the same units as the quantity being estimated. As In case of variable “maximum speed” the highest value is 100 km/h and the lowest value is 20 Km/h. The value of RMSE 6.16 km/h, in case of MICE, is acceptable. Overall, RMSE calculated in case of Amelia shows a higher value which means that imputation model of MICE performs in a better way as compared to Amelia.

6.8. Experiment 2: Missing Number of Lanes

In the experiment 2, missing values were generated in the variable number of lanes and same procedure was adopted as explained for Experiment 1

TABLE 18: Experiment 2 Datasheet

Road Type	Maximum Speed	One Way	Bicycle Lane	Number of Lanes
2	40	1	0	3
3	20	1	1	**NA**
2	30	1	1	**NA**
3	30	1	0	**NA**
3	20	0	1	1
2	50	1	0	3
2	30	1	1	**NA**
2	30	1	1	3
3	20	1	1	**NA**
3	30	1	1	2
…	…	…	…	…

Similarly, Amelia and MICE packages were used for the imputation of missing number of lanes in the OSM dataset. Table 19 shows the actual number of lanes in the OSM dataset and these are compared with the imputed number of lanes through Amelia and MICE.

TABLE 19: Actual and Predicted Values in Experiment 2

Number of Lanes (Observed Values)	Imputed Values with MICE	Imputed Values with Amelia
3	3	2
1	1	2
3	2	2
2	1	2
1	1	2
3	3	2
3	3	2
3	3	5
1	1	2
2	2	3
2	2	2
3	3	1
2	2	1
3	3	2
1	1	2
2	2	2
2	2	2
3	3	2
…	…	…

6.8.1. RMSE Calculation

RMSE was calculated to assess the difference between the actual value of number of lanes and predicted through multiple imputation packages, MICE and Amelia. Results are shown in Table 20.

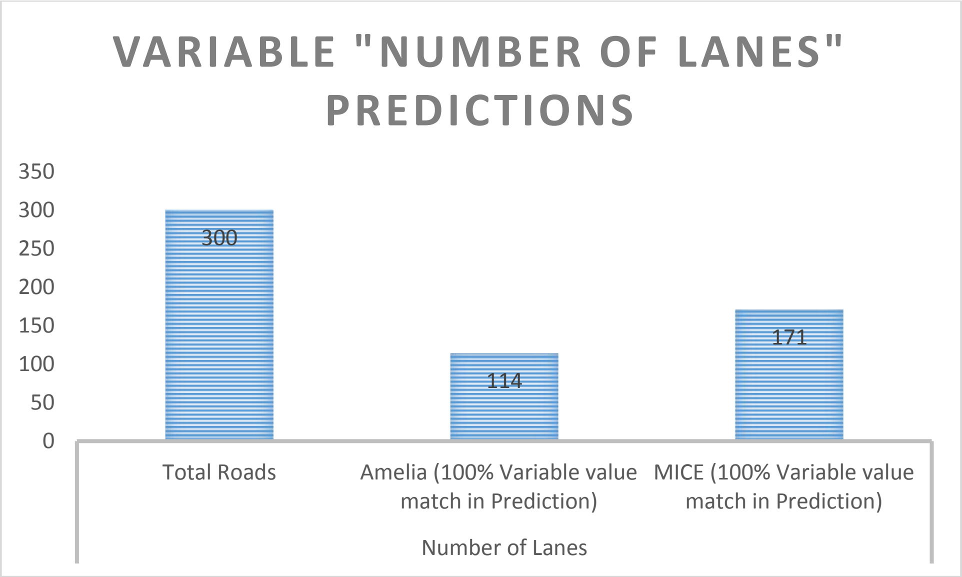
TABLE 20: Experiment 2 Results

Experiment	Missing Variable	Missing Values Percentage	Value Range	RMSE (MICE)	RMSE (Amelia)
Experiment 2	Number of Lanes	30%	1-7	0.68	1.03

The value of the RMSE explains that, imputed values of MICE are better as compared to the imputed values of Amelia. Generally, a trend has been observed in the imputations, whenever a difference exists between actual value and predicted value, the predictions of MICE show (generally) a difference of 1 lane from the actual number of lanes. For example, if a road has 4 lanes in actual, and when these lanes are imputed through MICE then MICE can predict 3 number of lanes for the same road but in case of Amelia, sometimes there exist a difference of two lanes in the predicted value from the actual value. MICE and Amelia also imputed same number of lanes as actually existing in original dataset. This trend has been shown in the Figure 17.

In case of Amelia the value of the RMSE is bigger that is 1.03 which means that predictions of the imputation model of Amelia is not accurate enough as compared to the predictions of MICE. The difference of two lanes from the actual number of lanes asks questions about the accuracy of the predictions. Overall, from the experiment 2, it was concluded that lane predictions of MICE are better than Amelia.

Figure 17: Perfect Value Match in Predictions



6.9. Experiment 3: Missing Road Type

In the experiment 3, missing values were generated in the variable ‘Road Type’. MICE differentiate between three types of variables: numeric, binary variable (factor with 2 levels) and categorical variable (factor with more than 2 levels). Each type has a default imputation method, which are indicated in the Table 9. R code for the imputation of missing values in the nominal variable with more than 2 categories was different as compared to the previous experiments. In MICE, an imputation method "Polyreg" was used for the imputation of missing values in the categorical variable (Road Type). In case of Amelia there is a need to specify nominal variables. Datasheet with missing values is presented in the Table 21.

TABLE 21: Datasheet for Experiment 3.

Road Type	Maximum Speed	One Way	Bicycle Lane	Number of Lanes
2	40	1	0	3
3	20	1	1	1
2	30	1	1	3
3	30	1	0	2
NA	20	1	1	1
2	50	1	0	3
2	30	1	1	3
2	30	1	1	3
3	20	1	1	1
3	20	1	1	2
3	30	1	1	2
NA	30	1	1	3
NA	30	1	1	2
…	…	…	…	…

Similarly, Amelia and MICE packages were used for the imputation categorical variable "Road Type" in the OSM dataset. Table 22 shows the actual road type in the OSM dataset and the imputed "Road Type" through Amelia and MICE.

TABLE 22: Actual and Imputed Values in Experiment 3

Actual Value (Road Type)	Predicted Value MICE	Predicted with Amelia
3	3	4
2	2	3
3	4	3

3	3	3
2	2	3
2	2	2
3	2	3
1	1	3
2	2	3
3	4	4
2	2	3
2	2	3
3	3	3
3	3	3
3	2	3
3	3	2
5	5	3
…	…	…

6.9.1. RMSE Calculation

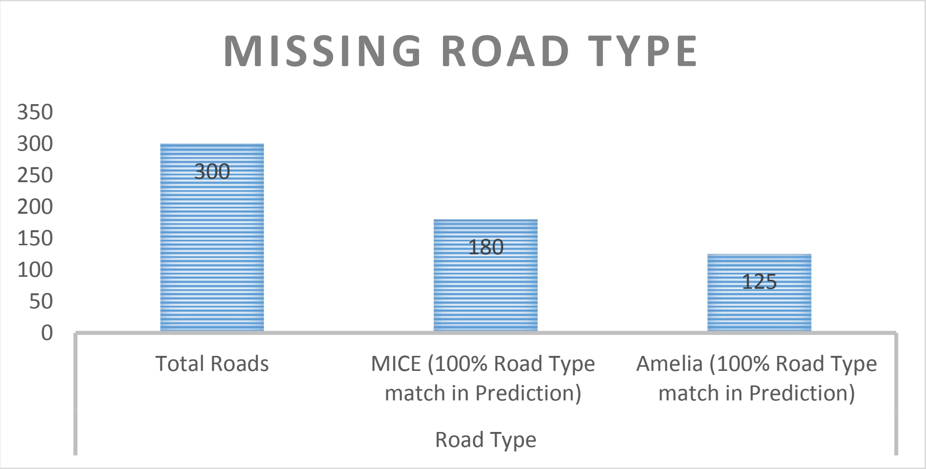
RMSE was calculated to assess the difference between the actual value of variable "Road Type" and predicted through multiple imputation packages, MICE and Amelia. Results are shown in Table 23.

Table 23: Results of Experiment 3

Experiment	Missing Variable	Missing Values Percentage	Value Range	RMSE (Amelia)	RMSE (MICE)
Experiment 3	Road Type	30%	1-5	0.87	0.67

Value of RMSE shows the better Predictions of the MICE as compared to Amelia. Generally, Amelia and MICE both show a difference of 1 in the predicted value (coded value) of the variable "Road Type" from the actual value. For example, if a road has been coded as 2 in actual dataset then it is a secondary road but the predictions of MICE and Amelia can code it as primary or tertiary road. Another interesting fact is that, out of the dataset of 300 roads with missing "Road Type", 180 roads showed a perfect match with the actual road type in case of MICE and 125 roads showed a perfect match in case of Amelia. This fact is shown in the Figure 18.

FIGURE 18: Perfect Value Match in Prediction



6.10. Experiment 4: Missing One-Way

In the experiment 4, missing values were generated in the variable ‘One Way. One way is a binary variable and “1” means that road in one way and “0” means that road is not one way. Binary variables have a default imputation method in MICE, which are indicated in the Table 9. R code for the imputation of missing values in the nominal variable with 2 categories was different as compared to the previous experiments. In MICE, an imputation method “logreg” was used for the imputation of missing values in the binary variable (one way). In case of Amelia there is a need to specify nominal variables through the expression “noms” in the R code. Datasheet with missing values is presented in the Table 24.

TABLE 24: Experiment 4 Datasheet

Road Type	Maximum Speed	One Way	Bicycle Lane	Number of Lanes
2	40	1	0	3
3	20	1	1	1
2	30	1	1	3
3	30	NA	0	2
3	20	1	1	1
2	50	NA	0	3
3	20	1	1	1
3	20	0	1	2
3	30	1	1	2
…	…	…	…	…

Amelia and MICE packages were used for the imputation of binary variable “One Way” in the OSM dataset. Table 25 shows the actual coding of variable “One Way” in the OSM dataset and the imputed code of variable “One way” through Amelia and MICE.

TABLE 25: Actual and Predicted Values in Experiment 4

Actual Value	Predicted Value MICE	Predicted value Amelia
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	0
1	1	1
0	0	1
1	1	1
1	1	1
1	1	1
	…	…

6.10.1. RMSE Calculation

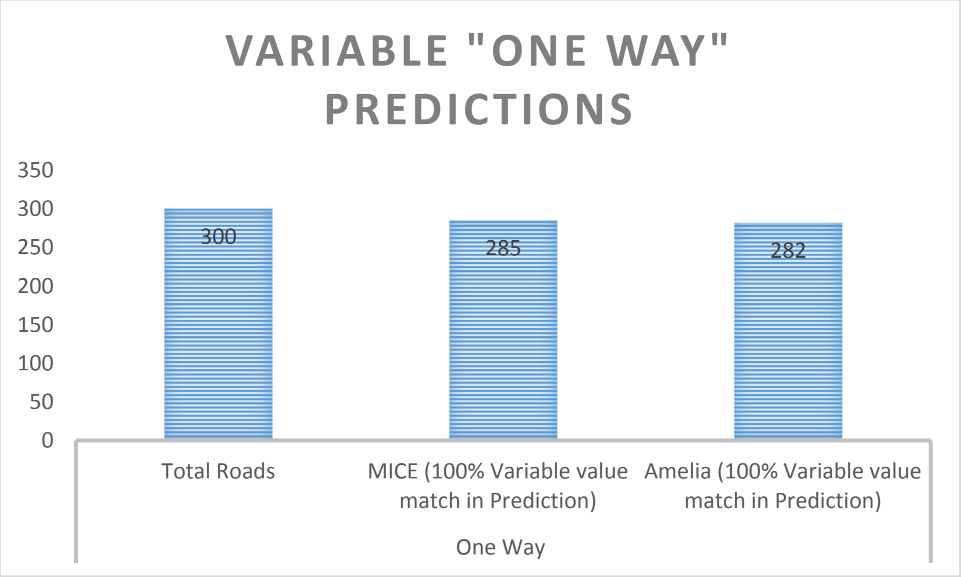
RMSE was calculated to assess the difference between the actual value of variable “one way” and predicted through multiple imputation packages, MICE and Amelia. Results are shown in Table 26.

TABLE 26: Results Experiment 4

Experiment	Missing Variable	Missing Values Percentage	Value Range	RMSE (Amelia)	RMSE (MICE)
Experiment 4	One Way	30%	0, 1	0.25	0.21

In case of binary variables (0, 1), the value of the RMSE should be much lower for good predictions. RMSE value for MICE is 0.21 and for Amelia it is 0.25, these values are lower that is why both packages are generating good predictions for the variable "One way". Out of the dataset of 300 roads with missing variable "One Way", 282 roads showed a perfect match with the actual value of variable in the predictions of Amelia. In case of MICE 285 roads showed a perfect with the original value of the variable. This fact is shown in the Figure 19.

FIGURE 19: Perfect Value Match in Prediction (One Way)



6.11. Experiment 5: Missing Bicycle Lane

In the experiment 5, missing values were generated in the variable 'Bicycle Lane. "Bicycle Lane" is a binary variable and "1" means that separate bicycle lane is present on the road and "0" means that there is no separate bicycle lane. In MICE, an imputation method "logreg" was used for the imputation of missing values in the binary variable (Bicycle Lane). In case of Amelia there is a need to specify nominal variables through the expression "noms" in the R code. Datasheet with missing values is presented in the Table 27.

TABLE 27: Datasheet for Experiment 5

Road Type	Maximum Speed	One Way	Bicycle Lane	Number of Lanes
2	40	1	0	3
3	20	1	1	1
2	30	1	**NA**	3
3	30	1	**NA**	2
3	20	1	1	1
2	50	1	0	3
2	30	1	1	3

2	30	1	1	3
3	20	1	1	1
3	20	1	**NA**	2
3	30	1	1	2
…	…	…	…	…

Amelia and MICE packages were used for the imputation of binary variable “Bicycle Lane” in the OSM dataset. Table 28 shows the actual coding of variable “Bicycle Lane” in the OSM dataset and the imputed code of variable “Bicycle Lane” through Amelia and MICE.

TABLE 28: Actual and Predicted Values in Experiment 5

Actual Value	Predicted Value MICE	Predicted Value Amelia
1	1	1
0	0	0
1	1	1
1	1	1
1	0	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
0	1	1
1	1	1
1	1	1
0	0	1
…	…	…

6.11.1. RMSE Calculation

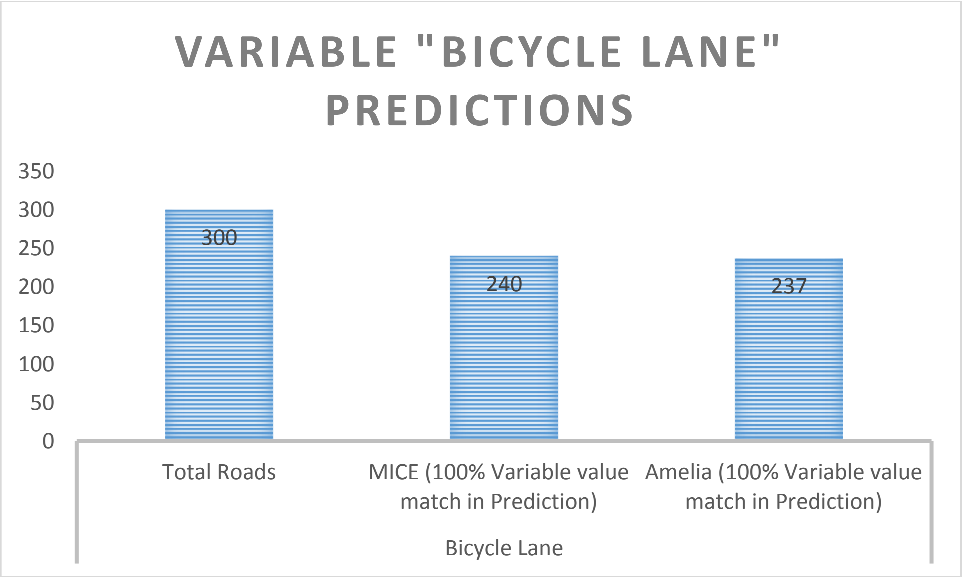
RMSE was calculated to assess the difference between the actual value of variable "bicycle lane" and predicted through multiple imputation packages, MICE and Amelia. Results are shown in Table 29.

TABLE 29: Results Experiment 5

Experiment	Missing Variable	Missing Values Percentage	Value Range	RMSE (Amelia)	RMSE (MICE)
Experiment 5	Bicycle Lane	30%	0, 1	0.43	0.42

In this experiment value of RMSE is greater for both imputation packages as compared to the previous experiment. Low value of RMSE shows better predictions the imputation models. These values of the RMSE show that discrepancies exist in the predictions of the MICE and Amelia in case of this variable "Bicycle Lane". Out of the dataset of 300 roads with missing variable "Bicycle lane", 237 roads showed a perfect match with the actual value of variable in the predictions of Amelia. In case of MICE 240 roads showed a perfect match with the original value of the variable. This fact is shown in the Figure 20.

FIGURE 20: Perfect Value Match in Prediction (Bicycle Lane)



6.12. Experiment 6: Missing Maximum Speed and Number of Lanes

In the experiment 6, missing values were generated in two variables 'Maximum Speed" and "Number of Lanes". Same procedure was adopted for missing values generation as discussed in section 6.6. Datasheet for experiment 6 is shown in Table 30.

Table 30: Experiment 6 Datasheet

Road Type	Maximum Speed	One Way	Bicycle Lane	Number of Lanes
2	40	1	0	3
3	20	1	1	2
2	**NA**	1	1	**NA**
3	**NA**	1	0	**NA**
3	20	1	1	1
2	50	1	0	3
2	**NA**	1	1	**NA**
2	30	1	1	3
3	**NA**	1	1	**NA**
3	20	1	1	2
3	20	1	1	2
2	30	1	1	3
…	…	…	…	…

6.12.1. RMSE Calculation

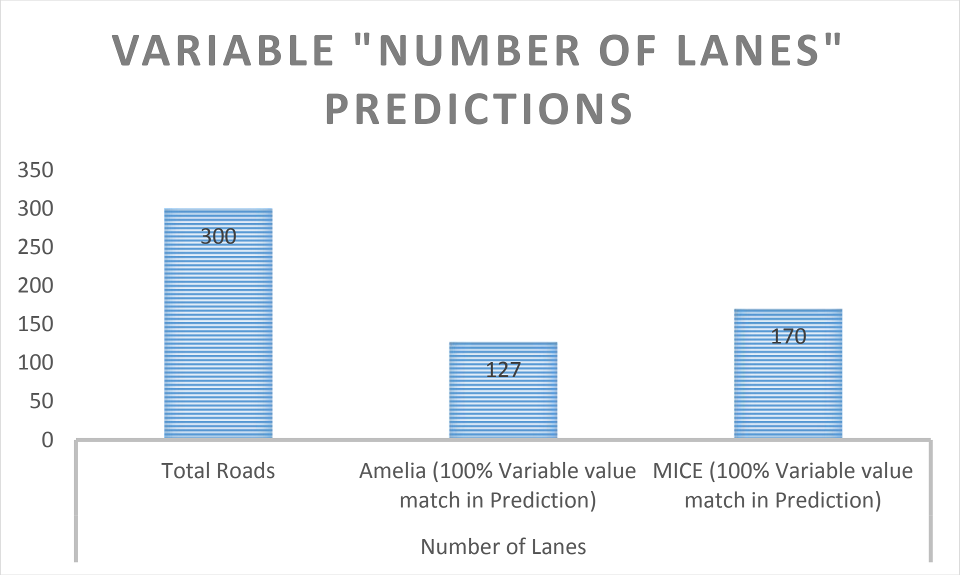
Amelia and MICE packages were used for the imputation of both variables in the OSM dataset. Results of the experiment 6 are presented in Table 31.

TABLE 31: Experiment 6 Results

Experiment	Missing Variable	Missing Values Percentage	Value Range	RMSE (Amelia)	RMSE (MICE)
Experiment 6	Maximum Speed	30%	20-100	10.21	6.42
	Number of Lanes		1-6	0.83	0.76

In case of the variable "maximum speed", in the experiment 1, where there were missing values only in the variable maximum speed, value of the RMSE was 6.16 Km/h. In this experiment, there were missing values in two variables therefore there is a slight increase in the value of the RMSE. Overall the value of 6.37 Km/h is quite acceptable. Same trend is observed in case of Amelia, value of the RMSE has increased slightly. But if the predictions of MICE are compared with the speed predictions of Amelia then MICE perform better. Experiment 1 and Experiment 6 show the same results and same trend is observed in case of the variable number of lanes. Figure 20 shows the number of predicted values that are 100% matched with the actual value in experiment 6.

FIGURE 21: Perfect Value Match in Prediction (Number of Lanes)



6.13. Experiment 7: Missing Road Type and Number of Lanes

In the experiment 7, missing values were generated in two variables 'Number of Lanes" and "Road Type". Same procedure is adopted as discussed in section 6.6. Table 33 shows some part of the datasheet.

TABLE 32: Datasheet Experiment 7

Road Type	Maximum Speed	One Way	Bicycle Lane	Number of Lanes
2	40	1	0	3
NA	20	1	1	**NA**
NA	30	1	1	**NA**
NA	30	1	0	**NA**
4	20	1	1	1
2	50	1	0	3
2	30	1	1	2

2	30	1	1	3
NA	20	1	1	**NA**
3	20	1	1	2
3	30	1	1	2
3	30	1	1	3
3	30	1	1	2
…	…	…	…	…

6.13.1. RMSE Calculation

Amelia and MICE packages were used for the imputation of both variables in the OSM dataset. RMSE was calculated and the results are presented in Table 33.

TABLE 33: Experiment 7 Results

Experiment	Missing Variable	Missing Values Percentage	Value Range	RMSE (Amelia)	RMSE (MICE)
Experiment 7	Road Type	30%	1-5	0.93	0.81
	Number of Lanes		1-6	1.09	0.71

In case of the combination of missing values in both variables, Value of the RMSE has increased for both MICE and Amelia predictions. Most of the times, MICE Predictions of the variable "number of lanes" show that lane difference between actual value and predicted value is 1. But in case of Amelia, predictions of the number of lanes sometimes show a difference of 2 lanes from the original value. Even in case of combination of missing values, MICE impute missing values in an efficient manner as compared to Amelia. Figure 22 and Figure 23 clarify the picture of the imputations of both packages in case of combination of missing values and show the number of predicted values that are 100% matched with the actual value in experiment 7.

FIGURE 22: Perfect Value Match in Predictions (Number of Lanes)

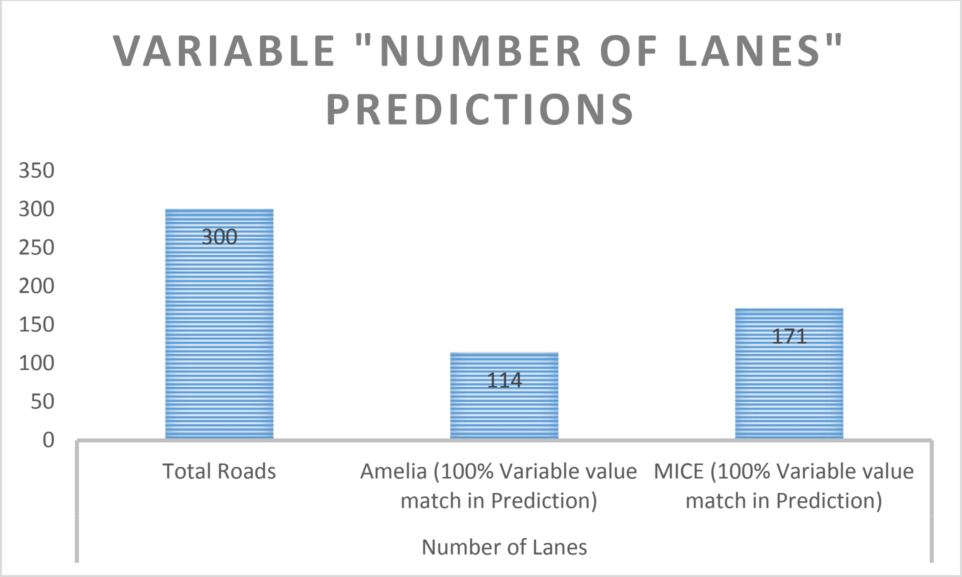
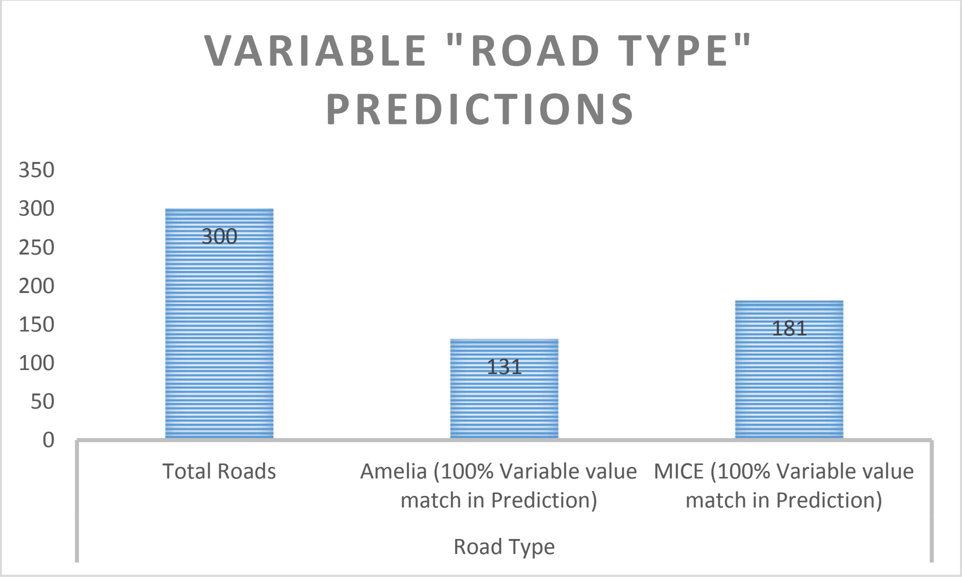


FIGURE 23:Perfect Value Match in Predictions (Road Type)



6.14. Experiment 8: Missing Road Type and Maximum Speed

In the experiment 8, missing values were generated in two variables 'Maximum Speed" and "Road Type". Table 34 shows some part of the datasheet for Experiment 8.

TABLE 34: Experiment 8 Datasheet

Road Type	Maximum Speed	One Way	Bicycle Lane	Number of Lanes
2	40	1	0	3
3	20	1	1	1
2	30	1	1	3

3	30	1	0	2
NA	**NA**	1	1	1
2	50	1	0	3
2	30	1	1	3
2	30	1	1	3
3	20	1	1	1
NA	**NA**	1	1	2
NA	**NA**	1	1	2
NA	**NA**	1	1	3
3	30	1	1	2
…	…	…	…	…

6.14.1. RMSE Calculation

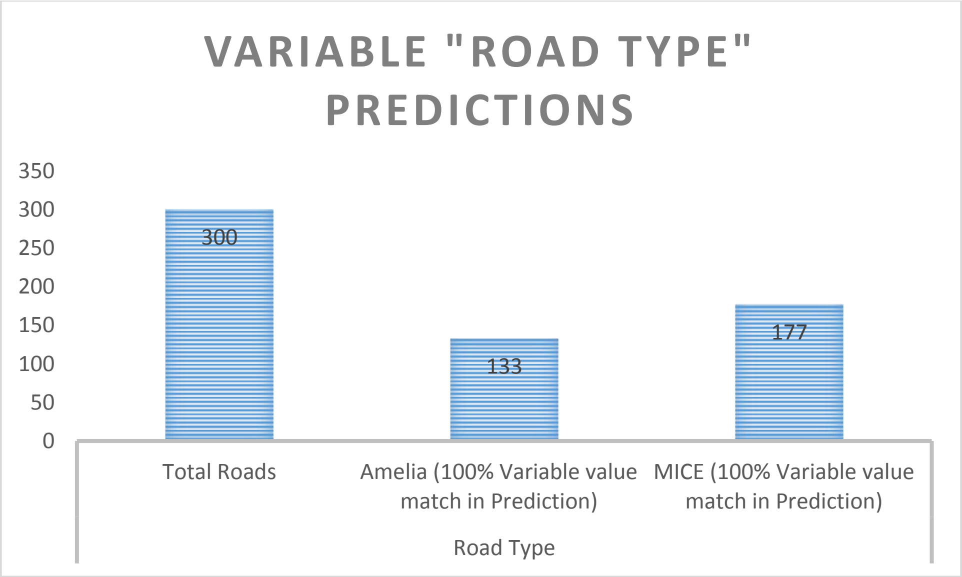
Amelia and MICE packages were used for the imputation of both variables in the OSM dataset. Results of experiment 8 are presented in Table 35.

TABLE 35: Experiment 8 Results

Experiment	Missing Variable	Missing Values Percentage	Value Range	RMSE (Amelia)	RMSE (MICE)
Experiment 8	Road Type	30%	1-5	0.90	0.84
	Maximum Speed		20-100	9.31	8.06

In case of combination experiment 8, similar to the previous experiments value of the RMSE has increased slightly as compared to the individual experiments. In case of Amelia, a general trend has been observed in the imputed datasheet that predicted speeds show a difference a 10 Km/h from the actual value. Most of the times, in case of MICE, the difference is less than 10 km/h from actual value. Overall, predictions of the Imputation model of MICE is better as compared to Amelia. In case of the variable ‘Road Type’, MICE predictions showed a perfect value match for 190 roads and Amelia predictions showed a perfect value match for 140 roads as shown in Figure 24.

FIGURE 24: Perfect Value Match in Predictions (Road Type)



7. Overall Results Table

TABLE 36: Comparison of all Experiments

Experiment	Missing Variable	Missing Values Percentage	Value Range	RMSE (Amelia)	RMSE (MICE)
Experiment 1	Maximum Speed (Km/h)	30%	20-100	9.19	6.16
Experiment 2	Number of Lanes	30%	1-7	1.03	0.68
Experiment 3	Road Type	30%	1-5	0.87	0.67
Experiment 4	One Way	30%	0, 1	0.25	0.21
Experiment 5	Bicycle Lane	30%	0,1	0.43	0.42
Experiment 6	Maximum Speed	30%	20-100	10.21	6.42
	Number of Lanes		1-6	0.83	0.76
Experiment 7	Road Type	30%	1-5	0.93	0.81
	Number of Lanes		1-6	1.09	0.71
Experiment 8	Road Type	30%	1-5	0.90	0.84
	Maximum Speed		20-100	9.31	8.06

Table 36 shows the results of all 8 experiments. In all these experiments 30% were missing in the datasheet. Results of all experiments showed that predictions of MICE are better as compared to Amelia. MICE package can impute missing values in an efficient manner as compared to other packages if there are different types of variables in the datasheet and missing values exist in all types of variables simultaneously. But in case of OSM dataset, more than 30% missing values exist in some countries. Like in case of Belgium there are huge number of missing values in the OSM dataset. That is why it is necessary to check the performance of the imputation models of MICE and Amelia in case of higher percentage of missing values. All 8 experiments were repeated with 50% missing values in the OSM dataset and comparison is shown in the TABLE 37. Overall results are discussed in the Section 8.

7.1. Comparison Table

Table 37: Comparison Table

Experiment	Missing Variable	30% Missing Values		50% Missing Values	
		RMSE (Amelia)	RMSE (MICE)	RMSE (Amelia)	RMSE (MICE
Experiment 1	Maximum Speed (Km/h)	9.19	6.16	10.20	7.82
Experiment 2	Number of Lanes	1.03	0.68	1.06	0.74
Experiment 3	Road Type	0.87	0.67	0.91	0.70
Experiment 4	One Way	0.25	0.21	0.26	0.22
Experiment 5	Bicycle Lane	0.43	0.42	0.44	0.43
Experiment 6	Maximum Speed	10.21	6.42	10.41	7.76
	Number of Lanes	0.83	0.76	0.88	0.77
Experiment 7	Road Type	0.93	0.81	0.95	0.82
	Number of Lanes	1.09	0.71	1.12	0.74
Experiment 8	Road Type	0.90	0.84	0.96	0.88
	Maximum Speed	9.31	8.06	9.47	8.10

8. Discussion and Conclusion

The principle aim of this research was to find out the ways to complete the missing values in the OSM dataset. Moreover, exploring various software packages available in R that are being used for the imputation of missing values in a dataset. To realize this mission different experiments were performed on the OSM dataset so that best imputation package can be selected for imputation of missing values in the OSM dataset.

In order to fill the missing values in the OSM dataset, a simple method is to estimate a regression model based upon the observed data and fill out the missing values using regression equation. But this method is not preferred as compared to the technique of multiple imputation. Multiple Imputation is the best way to complete the missing values in a dataset as it deals with the inherited uncertainty associated with the imputations and its results are unbiased (Zhang, 2003). A series of regression models are being estimated in the multiple imputation for the completion of each missing value in the dataset (Raghunathan et al., 2001). Amelia and MICE are two best suitable multiple imputation packages for the completion of missing values in the OSM dataset as these packages are useful for all types of variables (Categorical, Binary, Continuous).

Results have shown that MICE package outperforms Amelia package for the completion of missing values in the OSM dataset. In each experiment, the value of the root mean square error is lower for the predictions of MICE as compared to Amelia. Even in case of 50% missing values in the OSM dataset (See Table 37), results of both packages are not calamitous. When missing values are increased from 30% to 50%, the value of the RMSE has increased slightly (for MICE) because of the increased variability in the data. This slight increase in the RMSE values show that MICE package can be used for the completion of missing values in the OSM dataset even for higher percentages of missing data.

Although predictions of MICE are not 100% perfect but overall its results are quite acceptable. Generally, most of predicted values (not all) for the variable “maximum speed” shows a difference of less than 10 Km/h from the actual values of the variable both in case of 30% and 50% missing data. Similarly, for the variable “number of lanes”, generally predictions of MICE do not show a difference of more than one lane from the actual value. Sometimes differences of two lanes exist between the actual value and predicted value but these events are quite rare. But in case of Amelia, two lane differences are not a rare event. In case of binary variables (One way and Bicycle lane) and categorical variables (Road Type), more than 60% of MICE predictions showed a perfect value match with original value in all experiments.

As a cross check, a regression model has been estimated for the variable ‘maximum Speed’ based upon the complete data values (see Annexure-I). Estimated regression equation has been used for the prediction of missing values in the dataset. The values of the RMSE was higher (14.1 km/h) than both MICE and Amelia and this approach is not preferred in literature that is why further regression models are not estimated and multiple imputation was preferred.

It can be concluded that MICE package suites best for the completion of missing values in the OSM dataset due to its highly effective performance when missing values exist in more than one variables. This package has separate imputation methods for continuous, binary and categorical variables. It allows the column wise specification of the imputation methods according to the type of variable. Experiments have shown that predictions of MICE do not show terrible results when missing values are increased that is why it suites best to OSM dataset.

Completeness of the OSM dataset highly depends upon the number of contributors within an area that is why missing values in the OSM dataset varies from country to country depending upon the number of volunteers contributing to the OSM dataset. Number of volunteers are less in case of Belgium that is why missing values are huge in the OSM dataset of Belgium.

OSM dataset is the crowd sourced information which is collected by the non-experts so there are problems of data quality. There are always chances of error in the crowd sourced information. Second major reason of missing values is the tagging problems and issues of semantic heterogeneity. Different persons tag the same link with different keys and values which creates problems. Hence, missing values in the OSM dataset affect its potential usability in traffic and transportation studies. Hence, these missing values can be filled using the technique of multiple imputation.

9. Recommendations

Recommendations clarify that either the research is practically relevant and it takes into account the interests of stake holders. This research study agrees with the approach suggested by Van Buuren, 2012 that multiple imputation is the best way to deal with problems of missing data. It is strongly recommended to use MICE package for the completion of missing values in the OSM dataset. This package is useful for any dataset where there are missing values in more than one type of variables.

OSM is a source of free and rich geographical information related to the road network, therefore local authorities should launch awareness and promotional campaigns so that number of volunteers can be increased and this rich information can be used for fruitful purposes.

There should be simple OSM editors for the addition of useful information in the OSM road network dataset. Tag recommender systems can help the volunteers in correct tagging of various objects.

In this research, various experiments have been performed considering 30% and 50% missing values in the dataset. These experiments could be performed even for higher percentages of missing values (60% or 70%) to analyze the performance of imputation models.

Another R package MIssForest which imputes missing values in the multivariate datasets can be used for the completion of missing values in the OSM data and the results can be compared with this research study.

BIBLIOGRAPHY

Audigear, V., Agrocampus, Q., & Josse, J. (2015). Multiple Imputation for catagorical variables with multiple correspondence analysis, 30.

Ballatore, A., & Bertolotto, M. (2011). Semantically Enriching VGI in Support of Implicit Feedback Analysis. In K. Tanaka, P. Fröhlich, & K.-S. Kim (Eds.), *Web and Wireless Geographical Information Systems* (Vol. 6574, pp. 78–93). Berlin, Heidelberg: Springer Berlin Heidelberg

Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. Journal of statistical software, 45(3).

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Computing Surveys*, *41*(3), 1–52.

Bennett, J. (2010). *OpenStreetMap: be your own cartographer*. Birmingham: Packt Publishing Bermingham ISBN 978-1-847197-50-4

Barron, C., Neis, P., & Zipf, A. (2014). A Comprehensive Framework for Intrinsic OpenStreetMap Quality Analysis. *Transactions in GIS*, *18*(6), 877–895.

Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology*, *48*(1), 5–37.

Brand JPL (1999). Development, implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets. Erasmus University, Rotterdam.

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, *7*(3), 1247–1250.

Ciepluch, B., Jacob, R., Mooney, P., & Winstanley, A. C. (2010). Comparison of the accuracy of OSM for Ireland with Google Maps and Bing Maps. In Proceedings of the Ninth International Symposium on Spatial Accuracy Assessment in Natural Resuorces and Enviromental Sciences 20-23rd July 2010 (p.337). University of Leicester

Donders, A. R. T., van der Heijden, G. J. M. G., Stijnen, T., & Moons, K. G. M. (2006). Review: A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology*, *59*(10), 1087–1091.

Dempster, Arthur P., N.M. Laird and D.B. Rubin. 1977. “Maximum likelihood estimation from incomplete data via the em algorithm.” Journal of the Royal Statistical Society B 39:1–38

Eugster, J. A., & Schlesinger, (Thomas. (2013). Osmar:OpenStreetMap and R, The R Journal Vol 5/1 June 2013

Goetz, M. (2013). Towards generating highly detailed 3D CityGML models from OpenStreetMap. *International Journal of Geographical Information Science*, *27*(5), 845–865.

Girres, J.-F., & Touya, G. (2010). Quality Assessment of the French OpenStreetMap Dataset: Quality Assessment of the French OpenStreetMap Dataset. *Transactions in GIS*, *14*(4), 435–459.

Goodchild, M. F. (2013). The quality of big (geo)data. Journal of *Dialogues in Human Geography*, *3*(3), 280–284. 2

Haklay, M. (2010). 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.

Hagenauer, J., & Helbich, M. (2012). Mining urban land-use patterns from volunteered geographic information by means of genetic algorithms and artificial neural networks. *International Journal of Geographical Information Science*, *26*(6), 963–982.

Honaker, J., King, G., Blackwell, M., & others. (2011). Amelia II: A program for missing data. *Journal of Statistical Software*, *45*(7), 1–47.

Honaker, James, Anne Joseph, Gary King, Kenneth Scheve and Nau- nihal Singh. 1998-2002. "AMELIA: A program for missing data."

Honaker, James and Gary King. 2010. "What to do about missing values in time series cross-section data." American Journal of Political Science 54(2):561–581.

Jokar Arsanjani, J., Zipf, A., Mooney, P., & Helbich, M. (Eds.). (2015). *OpenStreetMap in GIScience*. Cham: Springer International Publishing. Retrieved from http://link.springer.com/10.1007/978-3-319-14280-7

Josse, J., & Husson, F. (2016). missMDA: a package for handling missing values in multivariate data analysis. *Journal of Statistical Software*, *70*(1), 1–31.

Jokar Arsanjani, J., Zipf, A., Mooney, P., & Helbich, M. (Eds.). (2015). *OpenStreetMap in GIScience*. Cham: Springer International Publishing.

Jokar Arsanjani, J., Zipf, A., Mooney, P., & Helbich, M. (Eds.). (2015). OpenStreetMap in GIScience Experiences Research and Applications. Cham: Springer International Publishing Switzerland.

Krek, A., Rumor, M., Zlatanova, S., & Fendel, E. M. (2009). *Urban and Regional Data Management: UDMS 2009 Annual*. CRC Press.

Kennickell AB (1991). \Imputation of the 1989 survey of consumer nances: Stochastic relaxation and multiple imputation." ASA 1991 Proceedings of the Section on Survey Re- searchMethods, pp. 1(10).

Ludwig, I., Voss, A., & Krause-Traudes, M. (2011). A Comparison of the Street Networks of Navteq and OSM in Germany. In S. Geertman, W. Reinhardt, & F. Toppen (Eds.), Advancing Geoinformation Science for a Changing World (Vol. 1, pp. 65–84). Berlin, Heidelberg: Springer Berlin Heidelberg.

Mooney, P., & Corcoran, P. (2012). The Annotation Process in OpenStreetMap. Journal of Transactions in GIS, 16(4), 561–579.

Mooney, P., Corcoran, P., & Winstanley, A. C. (2010). 18th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM SIGSPATIAL GIS 2010): November 2-5, 2010, San Jose, California. New York: Association for Computing Machinery.

Maier, G. (2014). OpenStreetMap, the Wikipedia Map. The Journal of ERSA powered by WU, REGION, Volume 1, Number 1, 1(1), 3–10

Neis, P., & Zipf, A. (2012). Analyzing the Contributor Activity of a Volunteered Geographic Information Project — The Case of OpenStreetMap. ISPRS International Journal of Geo-Information, 1(3), 146–165.

Neis, P., Singler, P., & Zipf, A. (2010). Collaborative mapping and emergency routing for disaster logistics–case studies from the haiti earthquake and the UN Portal for Afrika.

Ortúzar S., J. de D., & Willumsen, L. G. (2011). Modelling Transport (Fourth edition). Chichester, West Sussex, United Kingdom: John Wiley & Sons.

Perkins, C. (2014). Plotting practices and politics: (im)mutable narratives in OpenStreetMap. Transactions of the Institute of British Geographers, 39(2), 304–317. https://doi.org/10.1111/tran.12022

Patrickson, M. (1994). Traffic Assignment Problem: Models and Methods. Courier Dover Publications, 2015 ISBN 0486802272, 9780486802275.

Ortúzar S., J. de D., & Willumsen, L. G. (2011). Modelling Transport (Fourth edition). Chichester, West Sussex, United Kingdom: John Wiley & Sons.

Qian, X., Li, D., Li, P., Shi, L., & Cai, L. (2009). 10.1109@GEOINFORMATICS.2009.5293442.pdf. Center for Spatial Information Science and Systems George Mason University, 4. https://doi.org/10.1109/GEOINFORMATICS.2009.5293442

Roderick J. A .Little , Donald B Rubin (1987). Statistical Analysis With Missing Data, First Edition ,Wiley Publishers ISBN 0471802549, 9780471802549

Roderick J. A .Little , Donald B Rubin (2002). Statistical Analysis With Missing Data, Second Edition ,Wiley Publishers ISBN ISBN: 978-0-471-18386-0

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. Survey Methodology, 27(1), 85–96.

Raghunathan TE, Lepkowski JM, van Hoewyk J, Solenberger P (2001). \A multivariate technique for multiply imputing missing values using a sequence of regression models." Survey Methodology, 27, 85{95).

Rubin, D. &. A. Little, R . . (1987). LittleRubin_1987_Statistical Analysis with Missing Data

Stef van Buuren (2012), Flexible Imputation of Missing Data, CRC Press, 2012, ISBN 1439868255, 9781439868256

Schmitz, S., Neis, P., & Zipf, A. (2008). New applications based on collaborative geodata—the case of routing. In Proceedings of XXVIII INCA international congress on collaborative mapping and space technology

SinghSehra, S., Singh, J., & Singh Rai, H. (2013). Assessment of OpenStreetMap Data - A Review. International Journal of Computer Applications, 76(16), 17–20

Salkind, N. J., & Rasmussen, K. (Eds.). (2007). Encyclopedia of measurement and statistics. Thousand Oaks, Calif: SAGE Publications.

Stekhoven, D. J., & Buhlmann, P. (2012). MissForest--non-parametric missing value imputation for mixed-type data. Bioinformatics, 28(1), 112–118. https://doi.org/10.1093/bioinformatics/btr597

Tagging Guidelines - OpenStreetMap Wiki. (n.d.). Retrieved November 28, 2016, from http://wiki.openstreetmap.org/wiki/Australian_Tagging_Guidelines#Regional_Roads

Van Buuren, S., & Groothuis-Oudshoorn, (Karin. (2011). Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, VV(II).

Van Buuren, S. (2012). Flexible imputation of missing data. CRC press.

van Buuren S, Brand JPL, Groothuis-Oudshoorn CGM, Rubin DB (2006b). \Fully conditional specification in multivariate imputation." Journal of Statistical Computation and Simula- tion, 76(12), 1049{1064.

Vandecasteele, A., & Devillers, R. (2015). Improving Volunteered Geographic Information Quality Using a Tag Recommender System: The Case of OpenStreetMap. In J. Jokar Arsanjani, A. Zipf, P. Mooney, & M. Helbich (Eds.), OpenStreetMap in GIScience (pp. 59–80). Cham: Springer International Publishing.

Veregin, H. (1999). Data quality parameters. Geographical Information Systems, 1, 177–189.

Welser, H. T., Cosley, D., Kossinets, G., Lin, A., Dokshin, F., Gay, G., & Smith, M. (2011). Finding social roles in Wikipedia. In Proceedings of the 2011 iConference (pp. 122–129). ACM. Retrieved from http://dl.acm.org/citation.cfm?id=1940778

Zhang, P. (2003). International Statistical Review , 71, 3, 581–592, Printed in The Netherlands,

International Statistical Institute

Zielstra, D., Hochmair, H. H., & Neis, P. (2013). Assessing the Effect of Data Imports on the Completeness of OpenStreetMap - A United States Case Study: Assessing the Effect of Data Imports on the Completeness of OpenStreetMap. Transactions in GIS, 17(3), 315–334.

Annexure-A

Experiment 1 (Maximum Speed)		
Actual Value	**Predicted Value Amelia**	**Predicted with MICE**
20	34	30
30	33	30
30	40	30
30	30	30
30	39	30
30	30	40
30	39	30
30	30	30
30	28	30
30	37	30
30	33	30
30	32	30
30	26	30
30	40	40
30	26	30
40	36	20
50	50	50
20	21	30
20	23	30
20	17	30
20	31	30
30	27	30
30	22	20
20	25	30
30	21	30
30	33	30
20	28	30
30	33	30
20	28	30
30	19	20
50	42	50
50	56	50
30	40	30
30	20	30
30	24	30
30	31	30
40	45	50
30	33	30
30	34	30
30	36	30

50	40	50
50	40	50
40	48	40
50	47	40
50	43	50
50	49	50
50	52	40
50	46	50
20	29	30
30	29	20
30	17	30
30	23	20
30	31	30
50	42	50
50	49	50
50	43	50
20	14	20
30	31	20
30	35	20
20	36	20
30	41	30
20	18	30
20	32	30
30	32	30
100	90	100
50	39	40
50	34	50
30	25	20
70	52	70
30	31	30
100	90	100
20	26	30
30	31	20
30	31	30
40	34	50
20	35	20
30	43	30
50	45	50
100	90	100
70	59	70
30	49	30
30	40	30
20	13	20
70	50	70

Annexure-B

Experiment 2 (Number of Lanes)		
Actual Value	**Predicted Value Amelia**	**Predicted Value MICE**
1	2	1
3	2	2
2	2	2
3	2	3
1	2	1
3	2	3
2	2	2
6	5	5
1	2	2
2	3	2
2	2	3
3	1	2
1	1	1
2	2	2
2	2	2
2	2	2
2	2	2
2	2	2
2	2	2
1	2	2
2	2	2
1	1	2
2	1	2
2	4	2
2	2	2
3	1	2
2	2	2
2	3	1
2	2	2
1	2	1
2	2	3
2	3	2
3	3	2
3	2	2
2	2	2
3	2	3
2	2	3
2	3	1
2	2	2
2	3	2

1	2	1
3	2	3
3	1	2
2	3	2
2	2	1
1	2	2
2	3	3
1	2	1
3	3	3
2	4	3
2	3	2
3	2	3
3	2	3
3	2	2
2	1	2
2	2	2
2	2	1
1	2	1
1	1	1
2	2	2
2	1	2
5	4	5
3	2	2
4	4	4
4	3	4
3	2	2
2	2	1
2	1	1
2	3	1
2	3	2
2	2	1
4	4	4
3	2	3
3	2	2
3	2	3
2	2	2
3	2	3
3	2	3

Annexure-C

Experiment 3 (Road Type)		
Actual Value	**Predicted with Amelia**	**Predicted Value MICE**
3	4	3
2	3	2
3	3	3
3	3	3
2	3	2
2	2	2
3	3	2
1	3	1
2	3	2
3	4	3
2	3	2
2	3	2
3	3	3
3	3	3
3	3	2
3	2	3
5	3	5
5	3	3
2	3	2
2	3	2
4	3	3
4	3	3
5	3	4
5	3	5
3	3	3
2	4	3
2	3	3
2	3	3
5	3	4
4	3	3
2	2	2
3	4	3
3	4	2
2	3	3
2	3	2
2	3	2
2	3	2
2	2	2
2	3	2
2	2	2
2	3	2

3	3	3
2	2	2
4	4	3
2	2	2
5	4	5
3	3	3
2	3	2
3	2	4
2	2	2
2	2	2
2	3	2
2	2	2
2	3	3
2	3	2
2	3	3
2	3	2
2	4	3
2	2	4
4	3	3
4	3	5
2	3	2
2	3	2
2	2	2
2	4	2
3	2	3
2	4	3
3	2	3
3	3	3
3	3	4
3	3	2
3	3	2
2	4	2
2	2	2
2	3	2
2	3	2
2	2	2
2	2	2
4	4	3
4	4	3
2	3	4
4	4	3
2	3	2

Annexure-D

Experiment 4 (One Way)		
Actual Value	**Predicted Value Amelia**	**Predicted Value MICE**
1	1	1
1	1	1
1	1	0
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
0	1	0
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
0	1	0
0	1	0
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1

1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
0	1	1
0	1	1
1	1	1
1	1	1
1	1	1
0	1	0
1	1	1
1	1	0
1	1	0
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1

Annexure-E

Experiment 5 (Bicyclye Lane)		
Actual Value	**Predicted Value Amelia**	**Predicted Value MICE**
1	1	0
0	1	1
1	1	1
1	1	1
1	1	0
1	1	1
1	1	1
1	1	1
1	1	0
1	1	1
0	1	1
1	1	1
1	1	1
0	1	0
0	1	1
1	0	1
0	1	0
0	1	1
1	0	1
0	1	0
1	0	1
1	1	1
1	1	0
1	1	1
1	0	1
1	0	1
0	1	0
0	1	1
0	1	1
0	1	0
0	1	0
0	1	0
0	1	0
0	1	0
0	1	1
1	1	1
0	1	1
1	1	1
1	1	0
1	1	1

0	1	0
1	0	1
1	1	1
1	1	0
1	1	1
0	0	1
1	1	1
0	1	1
1	1	0
1	1	1
1	1	0
0	1	1
0	1	0
1	1	0
0	1	1
0	1	0
0	1	1
0	1	0
0	1	1
0	1	0
1	1	1
0	1	1
0	1	1
1	1	0
0	1	1
0	1	1
0	1	1
0	1	1
1	1	0
1	1	1
1	1	1
1	0	1
1	1	0
0	1	0
1	1	1
1	1	1
1	1	1
1	1	0
0	1	1
1	1	1
0	1	1
1	1	1
1	1	1

Annexure-F

Experiment 6					
Number of Lanes			Maximum Speed		
Actual Value	Predicted Value Amelia	Predicted Value MICE	Actual Value	Predicted Value Amelia	Predicted Value MICE
1	2	1	20	28	30
3	2	3	30	29	30
2	1	2	30	35	30
3	2	2	30	26	20
1	3	1	30	33	30
3	3	3	30	45	40
2	2	1	30	38	30
6	4	6	30	30	50
1	2	2	30	28	30
2	2	1	30	27	30
2	2	2	30	30	30
3	2	3	30	36	30
1	2	2	30	45	30
2	2	2	30	34	30
2	2	2	30	27	30
2	2	1	40	31	30
2	2	2	50	43	50
2	2	2	20	29	30
2	1	2	20	14	20
1	1	2	20	14	20
2	3	2	20	32	20
1	2	1	30	27	30
2	2	2	30	27	30
2	3	1	20	19	20
2	3	1	30	27	20
3	3	3	30	38	30
2	2	1	20	32	30
2	2	2	30	32	30
2	1	2	20	27	30
1	2	1	30	22	20
2	2	2	50	38	50
2	2	2	50	45	40
3	3	3	30	23	30
3	2	3	30	21	30
2	2	2	30	26	30
3	2	2	30	28	30
2	3	3	40	43	50
2	2	1	30	28	30
2	4	3	30	35	30

2	2	2	30	36	30
1	2	1	50	36	40
3	3	3	50	40	50
3	3	3	40	52	50
2	2	3	50	40	30
2	2	2	50	39	50
1	1	1	50	44	50
2	2	1	50	46	40
1	2	2	50	43	50
3	2	2	20	29	30
2	3	3	30	27	30
2	3	2	30	23	20
3	2	3	30	29	30
3	2	3	30	32	30
3	2	1	50	43	40
2	2	1	50	44	50
2	1	1	50	38	50
2	2	3	20	28	30
1	2	2	30	35	30
1	2	1	30	38	30
2	3	3	20	30	30
2	2	2	30	29	20
5	3	5	20	26	30
3	3	2	20	36	30
4	4	4	30	31	20
4	3	4	100	96	100
3	3	3	50	44	50
2	2	2	50	44	40
2	2	1	30	27	30
2	2	1	70	72	70
2	3	3	30	34	30
2	2	2	100	93	100
4	3	4	20	21	20
3	3	2	30	30	20
3	3	3	30	29	30
3	2	2	40	45	50
2	2	1	20	26	20
3	2	3	30	37	30
3	2	3	50	46	50
2	2	2	100	100	100
5	4	5	70	69	70
2	3	2	30	40	30
3	3	2	30	37	30

Annexure-G

Experiment 7					
Number of Lanes			Road Type		
Actual Value	Predicted Value Amelia	Predicted Value MICE	Actual Value	Predicted Value Amelia	Predicted Value MICE
1	2	2	3	3	3
3	3	3	2	2	2
2	2	2	3	3	3
3	2	3	3	3	4
1	1	1	2	3	2
3	2	3	2	3	2
2	2	2	3	3	3
6	5	5	1	3	1
1	2	3	2	3	2
2	2	2	3	3	3
2	3	3	2	2	2
3	3	3	2	4	2
1	2	1	3	3	3
2	1	2	3	2	3
2	2	3	3	3	4
2	2	2	3	3	4
2	2	2	5	4	4
2	2	2	5	2	5
2	1	2	2	4	2
1	1	1	2	3	2
2	2	2	4	3	2
1	2	1	4	3	3
2	2	1	5	3	4
2	2	2	5	3	3
2	2	2	3	3	3
3	2	3	2	4	2
2	1	1	2	3	2
2	2	1	2	4	3
2	2	2	5	2	4
1	2	1	4	4	3
2	2	1	2	2	2
2	1	2	3	3	4
3	3	2	3	2	3
3	3	3	2	3	3
2	2	2	2	3	4
3	2	2	2	2	2
2	3	3	2	2	2
2	2	1	2	2	3
2	2	2	2	3	2

2	3	2	2	3	2
1	2	2	2	2	2
3	3	3	3	3	3
3	2	2	2	4	2
2	2	2	4	3	3
2	2	2	2	4	2
1	2	1	5	3	5
2	2	2	3	3	2
1	2	1	2	3	2
3	2	3	3	2	3
2	2	3	2	3	2
2	3	3	2	3	2
3	3	2	2	3	2
3	3	3	2	2	2
3	2	1	2	3	3
2	1	2	2	3	2
2	2	1	2	3	3
2	2	2	2	2	3
1	2	1	2	2	4
1	2	1	2	3	3
2	2	3	4	3	3
2	2	2	4	3	4
5	5	5	2	3	2
3	3	3	2	4	2
4	4	4	2	2	2
4	4	4	2	2	2
3	3	3	3	3	4
2	2	1	2	4	3
2	2	1	3	2	3
2	2	1	3	3	3
2	2	2	3	3	3
2	2	2	3	3	2
4	4	4	3	3	3
3	3	2	2	3	2
3	2	2	2	2	2
3	2	3	2	2	4
2	1	2	2	2	2
3	2	2	2	2	2
3	2	2	2	3	2
2	2	3	4	3	3
5	5	5	4	3	3
2	1	2	2	3	4
3	2	3	4	3	3

Annexure-H

Experiment 8					
Maximum speed			Road Type		
Actual Value	Predicted Value Amelia	Predicted Value MICE	Actual Value	Predicted Value Amelia	Predicted Value MICE
20	31	30	3	3	3
30	24	20	2	3	2
30	32	30	3	3	3
30	29	30	3	3	2
30	39	30	2	4	4
30	44	30	2	3	2
30	34	30	3	3	4
30	44	30	1	3	1
30	30	20	2	3	2
30	26	30	3	3	4
30	38	30	2	2	3
30	35	30	2	3	2
30	24	30	3	3	2
30	30	30	3	3	2
30	30	20	3	3	3
40	31	30	3	3	2
50	45	50	5	3	5
20	26	20	5	3	5
20	17	30	2	2	2
20	23	20	2	2	2
20	23	30	4	3	2
30	33	30	4	4	5
30	34	30	5	4	4
20	34	20	5	5	5
30	28	20	3	3	3
30	35	30	2	3	2
20	30	30	2	3	2
30	31	30	2	4	5
20	19	30	5	4	4
30	31	20	4	4	5
50	42	40	2	3	2
50	52	30	3	3	4
30	33	30	3	3	2
30	24	30	2	3	3
30	27	30	2	2	2
30	23	30	2	2	2
40	44	50	2	2	2
30	32	20	2	3	2
30	40	30	2	3	2

30	36	30	2	2	2
50	37	40	2	2	2
50	46	30	3	3	2
40	32	50	2	2	2
50	38	40	4	4	2
50	41	40	2	3	2
50	49	50	5	4	5
50	40	40	3	3	2
50	40	50	2	2	2
20	31	30	3	2	2
30	29	30	2	2	2
30	21	20	2	2	2
30	20	30	2	3	2
30	31	20	2	3	2
50	43	50	2	2	4
50	50	40	2	3	2
50	38	40	2	3	2
20	24	30	2	2	2
30	31	50	2	3	2
30	37	30	2	3	2
20	28	20	4	3	4
30	35	30	4	2	4
20	21	30	2	3	2
20	27	20	2	2	2
30	31	20	2	3	2
100	90	100	2	2	4
50	37	40	3	3	3
50	39	40	2	4	3
30	23	30	3	3	3
70	73	70	3	4	3
30	27	20	3	3	4
100	85	100	3	3	2
20	29	20	3	3	2
30	22	30	2	2	2
30	30	20	2	3	2
40	40	50	2	3	2
20	29	30	2	3	2
30	40	30	2	3	2
50	50	40	2	2	2
100	76	100	4	2	3
70	71	70	4	3	3
30	43	30	2	2	3
30	33	40	4	3	2

Annexure-I

- Regression model estimation for maximum speed

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	169558.247	7	24222.607	441.939	.000[b]
	Residual	62812.086	1146	54.810		
	Total	232370.333	1153			

a. Dependent Variable: Maximum Speed

b. Predictors: (Constant), Residential, Highway, Teritary, Bicycle Lane, One way, Secondary, Number of lanes

Coefficients[a]

		Unstandardized Coefficients		Standardized Coefficients		
Model		B	Std. Error	Beta	t	Sig.
1	(Constant)	29.333	1.567		18.722	.000
	One way	2.766	1.107	.042	2.499	.013
	Bicycle Lane	-8.521	.518	-.268	-16.450	.000
	Number of lanes	4.189	.376	.264	11.130	.000
	Highway	31.479	1.198	.533	26.287	.000
	Secondary	-3.569	.578	-.121	-6.180	.000
	Teritary	-5.857	.812	-.133	-7.217	.000
	Residential	-8.070	1.407	-.101	-5.737	.000

a. Dependent Variable: Maximum Speed

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.854[a]	.730	.728	7.403

a. Predictors: (Constant), Residential, Highway, Teritary, Bicycle Lane, One way, Secondary, Number of lanes

Auteursrechtelijke overeenkomst

Ik/wij verlenen het wereldwijde auteursrecht voor de ingediende eindverhandeling:
Imputation of missing values in OSM networks

Richting: **Master of Transportation Sciences-Mobility Management**
Jaar: **2017**

in alle mogelijke mediaformaten, - bestaande en in de toekomst te ontwikkelen - , aan de Universiteit Hasselt.

Niet tegenstaand deze toekenning van het auteursrecht aan de Universiteit Hasselt behoud ik als auteur het recht om de eindverhandeling, - in zijn geheel of gedeeltelijk -, vrij te reproduceren, (her)publiceren of distribueren zonder de toelating te moeten verkrijgen van de Universiteit Hasselt.

Ik bevestig dat de eindverhandeling mijn origineel werk is, en dat ik het recht heb om de rechten te verlenen die in deze overeenkomst worden beschreven. Ik verklaar tevens dat de eindverhandeling, naar mijn weten, het auteursrecht van anderen niet overtreedt.

Ik verklaar tevens dat ik voor het materiaal in de eindverhandeling dat beschermd wordt door het auteursrecht, de nodige toelatingen heb verkregen zodat ik deze ook aan de Universiteit Hasselt kan overdragen en dat dit duidelijk in de tekst en inhoud van de eindverhandeling werd genotificeerd.

Universiteit Hasselt zal mij als auteur(s) van de eindverhandeling identificeren en zal geen wijzigingen aanbrengen aan de eindverhandeling, uitgezonderd deze toegelaten door deze overeenkomst.

Voor akkoord,

Adeel, Ahmad

Datum: **11/08/2017**