

**Assessing the Accuracy of OpenStreetMap Data in South Africa for the
Purpose of Integrating it with Authoritative Data**

By Lindy-Anne Siebritz under the supervision of Dr George Sithole

A dissertation presented for the degree
Master of Science in Engineering



Department of Architecture, Planning and Geomatics
University of Cape Town
February 2014

The copyright of this thesis vests in the author. No quotation from it or information derived from it is to be published without full acknowledgement of the source. The thesis is to be used for private study or non-commercial research purposes only.

Published by the University of Cape Town (UCT) in terms of the non-exclusive license granted to UCT by the author.

PLAGIARISM DECLARATION

1. I know that plagiarism is wrong. Plagiarism is to use another's work and pretend that it is one's own.
2. I have used the Harvard convention for citation and referencing. Each contribution to, and quotation in this thesis from the works of other people has been attributed, and has been cited and referenced.
3. This thesis is my own work.
4. I have not allowed, and will not allow, anyone to copy my work with the intention of passing it off as his or her own work.

Signature _____

Date _____

ACKNOWLEDGEMENTS

This dissertation is unto the glory of the Lord. All of the honour belongs to Him.

I am grateful to my supervisor, Dr George Sithole for his patience and guidance. I would like to thank my parents and good friend, Simonne for their prayers and encouragement.

Ahmad Desai and Thabo Ntsoku provided me with technical assistance. The Open-StreetMap database administrators Grant Slater and Frederick Ramm provided me data and answered all my questions. I would like to extend my thanks to all of them.

Finally, I would like to thank my colleagues who provided me with information, Aslam Parker, Mark McLachlan, Steve Jansen, Raoul Duesimi, Sissiel Kay and Heinrich Du Plessis.

Contents

1	INTRODUCTION	12
1.1	Introduction	12
1.2	Background	14
1.2.1	The Chief Directorate: National Geo-Spatial Information Integrated Topographical Information System	14
1.2.2	OpenStreetMap	15
1.2.3	OpenStreetMap Data Model	16
1.3	Related Work	16
1.3.1	Status of the South African Spatial Data Infrastructure	16
1.3.2	VGI in the SDI Context	17
1.3.3	Quality, Quality Assurance and Quality Control	18
1.3.4	Quality Assessment	18
1.4	Problem Identification	19
1.5	Research Objectives	19
1.5.1	Research Questions	19
1.6	Methodology	20
1.6.1	Quantitative Assessments	20
1.6.2	Qualitative Assessments	20
1.7	Scope	22
1.8	Outcomes	22
2	LITERATURE REVIEW	24
2.1	Introduction	24
2.2	Volunteered Geographic Information	24
2.2.1	Advantages of VGI	24
2.2.2	Disadvantages of VGI	27
2.2.3	Examples of VGI Initiatives Available in South Africa	28
2.2.4	Factors Leading to the Success of VGI	30
2.2.5	Factors Opposing VGI Success	31
2.3	Spatial Data Infrastructure and VGI	33
2.3.1	The Purpose of SDI	33
2.3.2	Integrating VGI and Authoritative Data	33
2.3.3	Global Examples of VGI and SDI/NMA Integration	34
2.4	Methods to Assess Quality	35
2.4.1	Positional Accuracy of Linear Features	35
2.4.2	Geometric Accuracy of Polygon Features	39
2.4.3	Previous Investigations into the Positional Accuracy of OSM Roads Using the Goodchild and Hunter (1997) Method	40
2.4.4	Semantic Accuracy	41

2.4.5	Completeness Assessments	41
2.5	Analysis and Discussion	43
3	CHIEF DIRECTORATE: NATIONAL GEO-SPATIAL INFORMATION AND OPENSTREETMAP SPATIAL DATA MODELS	45
3.1	Introduction	45
3.2	Spatial Data Standards	45
3.2.1	Spatial Data Standards in South Africa	45
3.2.2	Duties of the Committee for Spatial Information	46
3.2.3	CD: NGI Spatial Data Standards - Contributions from External Standards	47
3.2.4	The CD: NGI Internal Spatial Data Standards	47
3.3	CD: NGI Topographical Data Structure	47
3.3.1	Topographic Feature Compilation	47
3.3.2	Standards Governing Topographic Feature Compilation	48
3.3.3	CD: NGI Quality Control	49
3.3.4	CD: NGI Quality Topology	49
3.3.5	Distribution of the CD: NGI Data	49
3.4	OpenStreetMap Data Structure	51
3.4.1	OpenStreetMap Data Elements	51
3.4.2	OpenStreetMap Data Model Components	52
3.4.3	OpenStreetMap Quality Control	53
3.4.4	OpenStreetMap Topology	54
3.4.5	Distribution of Data	57
3.5	Analysis and Discussion	57
4	METHODOLOGY	60
4.1	Introduction	60
4.2	Data	61
4.2.1	Data Sources	61
4.2.2	Co-ordinate System and Projection	61
4.2.3	Selection of Test Areas	61
4.3	Data Cleaning	64
4.3.1	Filtering	64
4.4	Method for Quantitative Assessment	66
4.4.1	Positional Accuracy of Roads	66
4.4.2	Geometric Accuracy Of Polygon Buildings	70
4.4.3	Semantic Accuracy of Roads	73
4.4.4	Completeness of Roads	76
4.5	Method for Qualitative Assessment	76
4.5.1	OSM Currency	76
4.5.2	OSM Uniformity in Acquisition	79
4.6	Analysis and Discussion	79
5	RESULTS AND ANALYSIS	83
5.1	Introduction	83
5.2	Results for Quantitative Assessments	83
5.2.1	Positional Accuracy of Roads	83
5.2.2	Geometric Accuracy of Amenity Buildings	88
5.2.3	Semantic Accuracy of Roads	94

5.2.4	Completeness	96
5.3	Results for Quantitative Assessments	98
5.3.1	OSM Currency	98
5.3.2	OSM Uniformity of Point Acquisition	101
5.3.3	Analysis and Discussion	103
6	INTEGRATION	106
6.1	Introduction	106
6.2	Previous Investigations into Integrating Authoritative Data and OSM Data	106
6.3	Technical Considerations	107
6.3.1	Different Reference Systems	107
6.3.2	Different Representations of Topographical Features	108
6.3.3	Duplication of Features	109
6.3.4	Omission of Attribute and Metadata	109
6.3.5	Different File Formats	112
6.3.6	The Physical and Structural Differences of the Databases	112
6.4	Policies, Licensing and Spatial Data Standards	112
6.4.1	The CD: NGI Policy and OSM Licensing Concerning Data Distribution	112
6.4.2	Adherence to the CD: NGI Spatial Data Standards	113
6.5	Differences in Quality Assurance and Quality Control Processes	113
6.6	The Process for Acquiring and Processing Ancillary Data	114
6.6.1	Using OSM Data for Change Detection	115
6.6.2	Proposed Integration of the CD: NGI and OSM Data	115
6.6.3	Institutional Reorganisation Needs	117
6.7	Analysis and Discussion	117
7	CONCLUSIONS AND RECOMMENDATIONS	119
7.1	Introduction	119
7.2	OSM Positional Accuracy	119
7.3	Qualitative Aspects of the OSM Data	120
7.3.1	Heterogeneity in OSM Data Volumes	120
7.3.2	Heterogeneity in OSM Data Acquisition	121
7.4	Integration Opportunities	122
7.5	Analysis and Discussion	122
7.6	Recommendations and Future Work	123
	APPENDICES	124
A1.	Appendix A: Secondary Activity Level	125
B1.	Appendix B: CD: NGI Process for Compilation of Topographical Features	126
C1.	Appendix C: Data flow for after approved compilation task	127
D1.	Appendix D: OSM Data Model Components	128
E1.	Appendix E: Data flow for acquisition and macro processing of ancillary data: ad hoc	129
F1.	Appendix F: Sample Python scripts	130

List of Figures

1.1	Extract from Wikimapia	13
1.2	Extract from OpenStreetMap	14
1.3	Integration of the CD: NGI data through the iTIS	15
1.4	Map showing twenty-seven test areas	19
1.5	Flowchart outlining the methodology	21
2.1	Example of update via MapShare	29
2.2	Example of the on-line Waze application	30
2.3	Depiction of the BOS and Goodchild and Hunter method	37
2.4	Depiction of the Hausdorff and average distance	38
2.5	Example of the minimum bounding rectangle around a polygon	40
3.1	Extract of CD: NGI feature classification	50
3.2	OSM feature classification	54
3.3	Example of OSM lines overlap errors	55
3.4	Example of OSM polygons overlap errors	56
3.5	Example of OSM dangles	57
4.1	Example of Residential land use - high urban density area	62
4.2	Example of Residential land use - low urban density area	63
4.3	Example of commercial and industrial storage area	63
4.4	Example of features with multiple negative OSM IDs	65
4.5	Removing holes from OSM polygons	66
4.6	Method for removing unwanted OSM road sections	68
4.7	Example of discontinuities in OSM multi-lane road	69
4.8	Identifying the corresponding roads between the CD: NGI and OSM . . .	70
4.9	Identifying corresponding polygons between the CD: NGI and OSM . . .	71
4.10	Example of multiple polygon matches	71
4.11	Example of incorrect polygon matching	73
4.12	Point deletions, additions, modifications and no change	77
4.13	Line deletions, additions, modifications and no change	78
4.14	Identifying corresponding polygons between consecutive data sets	79
4.15	Example of the line matching technique failing	80
4.16	CD: NGI and OSM point data for Western Cape Commercial test area . .	81
4.17	Comparing CD: NGI and OSM point data	82
5.1	Percentage overlap for OSM roads	84
5.2	Depiction of the Mpumalanga data set with north-westerly shift	85
5.3	Scatter plots comparing the CD: NGI and OSM road lengths	87
5.4	Example of generalisation applied to polygons at the CD: NGI	89

5.5	Comparing the CD: NGI and OSM compactness values for commercial test areas	92
5.6	Comparing the CD: NGI and OSM compactness values for residential test areas	92
5.7	Comparing compactness for the Free State residential data set	93
5.8	Distribution of the compactness differences for the Free State residential test area	93
5.9	Graph comparing the standard errors for the CD: NGI and OSM road class matches	96
5.10	Comparing completeness of OSM line data for commercial areas	97
5.11	Comparing completeness of OSM line data for residential areas	97
5.12	Comparing completeness of OSM line data for low urban density areas . .	98
5.13	Comparing OSM additions 2006-2012	99
5.14	Comparing OSM unchanged data 2006-2012 for commercial areas	100
5.15	Comparing OSM unchanged data 2006-2012 for residential areas	100
5.16	Comparing OSM unchanged data 2006-2012 for low urban density areas .	101
5.17	Distribution of point contributions for commercial areas	102
5.18	Distribution of point contributions for residential areas	102
5.19	Distribution of point contributions for low urban density areas	103
6.1	Example of different feature types representing the same feature	109
6.2	Example of JOSM file format containing the source information	111
6.3	Process for ingesting OSM data into the CD: NGI iTIS	116

List of Tables

3.1	Description of OSM data elements	52
3.2	Comparison of CD: NGI and OSM data models	59
4.1	Quality measures undertaken for each OSM data element	61
4.2	Sample matrix comparing the number of road matches	74
4.3	Comparing some of the CD: NGI and OSM road classes	75
4.4	Positive road class matches between the CD: NGI and OSM	76
5.1	Percentage overlap for OSM roads per province	86
5.2	Percentage overlap for OSM roads per settlement category	86
5.3	Comparing the Hausdorff distances for polygons	88
5.4	Comparing the area ratios of matching polygons	90
5.5	Comparing compactness differences for commercial test areas	91
5.6	Comparing compactness differences for residential test areas	91
5.7	Comparing the elongation differences for commercial test areas	94
5.8	Comparing the elongation differences for residential test areas	94
5.9	Percentages for the CD: NGI and OSM road class matches	95
7.1	Table of OSM compatibility	123

List of Acronyms

API	Application Programming Interface
BOS	Buffer-Overlay-Statistics
CD: NGI	Chief Directorate: National Geo-Spatial Information
CSI	Committee for Spatial Information
GPS	Global Positioning System
ICT	Information and Communication Technology
iTIS	integrated Topographic Information System
ISO/TC211	International Organisation for Standardisation's Technical Committee for Geographical Information
M.D.C.U.	Minimum Data Capture Unit
NMA	National Mapping Agency
ODbL	Open Database License
OSM	OpenStreetMap
OS	Ordnance Survey
POI	Points of Interest
PAIA	Promotion of Access to Information Act, No 2 of 2000

QA	Quality Assurance
SA	South Africa
SABA	South African Bureau of Standards
SAGDaD	South African Geospatial Data Dictionary
SANS 1880	South African National Standards 1880
SASDI	South African Spatial Data Infrastructure
SDI	Spatial Data Infrastructure
UGC	User-Generated Content
USGS	United States Geological Survey
VGI	Volunteered Geographic Information

ABSTRACT

The introduction and success of Volunteered Geographic Information (VGI) has gained the interest of National Mapping Agencies (NMAs) worldwide. VGI is geographic information that is freely generated by non-experts and shared using VGI initiatives available on the Internet. The NMA of South Africa i.e. the Chief Directorate: National Geo-Spatial Information (CD: NGI) is looking to this volunteer information to maintain their topographical database; however, the main concern is the quality of the data.

The purpose of this work is to assess whether it is feasible to use VGI to update the CD: NGI topographical database. The data from OpenStreetMap (OSM), which is one the most successful VGI initiatives, was compared to a reference data set provided by the CD: NGI. Corresponding features between the two data sets were compared in order to assess the various quality aspects. The investigation was split into quantitative and qualitative assessments. The aim of the quantitative assessments was to determine the internal quality of the OSM data. The internal quality elements included the positional accuracy, geometric accuracy, semantic accuracy and the completeness. The first part of the qualitative assessment was concerned with the currency of OSM data between 2006 and 2012. The second part of the assessment was focused on the uniformity of OSM data acquisition across South Africa.

The quantitative results showed that both road and building features do not meet the CD: NGI positional accuracy standards. In some areas the positional accuracy of roads are close to the required accuracy. The buildings generally compare well in shape to the CD: NGI buildings. However, there were very few OSM polygon features to assess, thus the results are limited to a small sample. The semantic accuracy of roads was low. Volunteers do not generally classify roads correctly. Instead, many volunteers prefer to class roads generically. The last part of the quantitative results, the completeness, revealed that commercial areas reach high completeness percentages and sometimes exceed the total length of the CD: NGI roads. In residential areas, the percentages are lower and in low urban density areas, the lowest. Nonetheless, the OSM repository has seen significant growth since 2006.

The qualitative results showed that because the OSM repository has continued to grow since 2006, the level of currency has increased. In South Africa, the most contributions were made between 2010 and 2012. The OSM data set is thus current after 2012. The amount and type of contributions are however not uniform across the country for various reasons. The number of point contributions was low. Thus, the relationship between the type of contribution and the settlement type could not be made with certainty.

Because the OSM data does not meet the CD: NGI spatial accuracy requirements, the two data sets cannot be integrated at the database level. Instead, two options are proposed. The CD: NGI could use the OSM data for detecting changes to the landscape only. The other recommendation is to transform and verify the OSM data. Only those features with a high positional accuracy would then be ingested. The CD: NGI currently has a shortage of staff that is qualified to process ancillary data. Both of the options proposed thus require automated techniques because it is time consuming to perform these tasks manually.

Chapter 1

INTRODUCTION

1.1 Introduction

The role of National Mapping Agencies (NMAs) is to provide current, reliable geographic information. NMAs have a mandate to produce data according to spatial mapping standards (Goodchild, 2009). As technology has evolved, the methods of spatial data creation, manipulation and management have changed drastically and the standards and specifications involved have become stricter in the accuracies they allow. NMAs either implement the International Organisation for Standardisation's Technical Committee for Geographical Information (ISO/TC211) standards and specifications, absolutely or they use them as a base to formulate standards, which are more specific to the organisation.

With the obligation of NMAs worldwide to adhere to mapping standards, they have been very strict regarding data sources. External data must be at a level of accuracy that complies with NMA standards and this has put a limit on the acquisition of spatial data. Until the recent success of Volunteered Geographic Information (VGI), NMAs were not concerned with this limitation, as their products were still up to date and of acceptable accuracy.

The act of volunteering geographic information has been around for a number of years. However, the term VGI was first introduced in 2007 by Goodchild (Goodchild, 2007) and has since grown exponentially, not just in the geomatics sphere. VGI is geographic information that is freely generated by the public and shared over VGI initiatives via the Internet. Users are not required to have any specialised skills in order to contribute data (Haklay and Ellul, 2010). Two examples of VGI initiatives are Wikimapia and OpenStreetMap (see figures 1.1 and 1.2 respectively). Two primary technologies propelled this growth in VGI, the first being Web 2.0, which differed from Web 1.0, in that it allows data sharing via the Internet by anyone who has the necessary resources available (Cooper, Rapant, Hjelmager, Laurent, Iwaniak, Coetzee, Moellering and Duren, 2011). Secondly, commercially available Global Positioning Systems (GPSs) built into mobile devices allow for easy spatial data collection (Cooper, Rapant, Hjelmager, Laurent, Iwaniak, Coetzee, Moellering and Duren, 2011). The combined technology allow for easy upload and sharing of geospatial information via the Internet.

The success of VGI initiatives leaves NMAs with the concern of providing data that is less current than that of VGI repositories, which receive daily spatial data updates; something that NMAs cannot achieve. The counter argument is that NMAs remain the

custodians of reliable information. One of the most researched aspects regarding VGI is the quality or credibility of the information, because no standards are applied and measures to control data submission have not been implemented (Flanagin and Metzger, 2008; Coleman, Nkhwanana and Sabone, 2010; Ostermann and Spinsanti, 2010). The quality perception of the data is very dependent on the purpose for which it is intended to be used, thus a study of VGI quality will have to be from a NMA perspective. This includes assessment of: 1) position and height accuracy, 2) attribute accuracy, 3) currency, 4) completeness, 5) logical consistency and 6) lineage (Cooper, Rapant, Hjelmer, Laurent, Iwaniak, Coetzee, Moellering and Duren, 2011). Even though these cannot be answered with certainty as yet, more and more people are moving away from traditional government produced maps to those produced by the VGI community (Flanagin and Metzger, 2008). This has led to many investigations into the uncertainties of VGI so as to determine the feasibility of using it as a data source for updating authoritative topographical databases (Kounadi, 2009; Girres and Touya, 2010; Al-Bakri and Fairbairn, 2011).



Figure 1.1: Extract from the Wikimapia Application Programming Interface (API) showing an area in the Cape Town region. The roads and erven layers are overlaid onto the image backdrop²

²<http://Wikimapia.org>

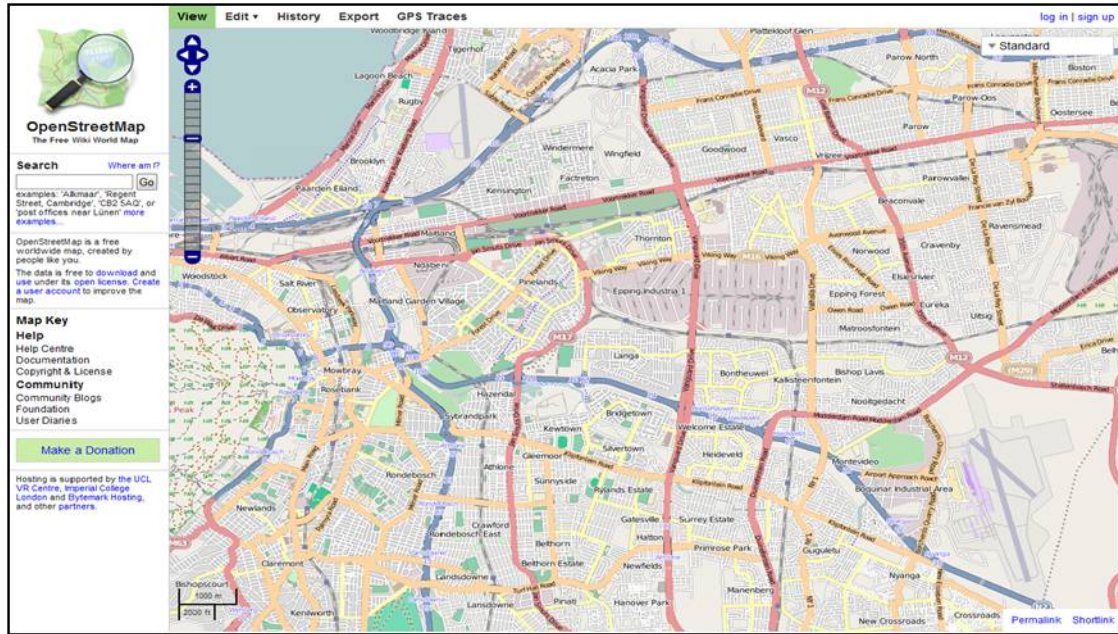


Figure 1.2: Extract from the OpenStreetMap API showing another area in the Cape Town region. The “standard” layer is the active layer showing roads and properties⁴

1.2 Background

The Chief Directorate: National Geo-Spatial Information (CD: NGI) is the NMA of South Africa and is thus mandated to provide topographical information to South Africa (SA). CD: NGI prides itself in supplying topographical information that is accurate and current.

1.2.1 The Chief Directorate: National Geo-Spatial Information Integrated Topographical Information System

The CD: NGI is currently in the process of migrating its topographical data to a new database structure, the integrated Topographic Information System (iTIS). The iTIS is compliant with the minimum requirements of the South African Bureau of Standards (SABS) Geographic Information Standards. SABS in turn is compliant with the minimum requirements of the relevant International Organization for Standardization: Technical Committee 211 (ISO/TC211) family of standards.

The iTIS is a multi-user server system that is able to integrate all the CD: NGI data (e.g. topographic vector data, raster information, DEMS etc., see figure 1.3) (Vorster and Duesimi, 2010). Topographical vector data is stored in an Oracle 10g database. The topographical data is accessed and managed via the Geomedia Transaction Management (GTM) system (Vorster and Duesimi, 2010). Features within the database are classed and described according the South African National Standards 1880 South African Geospatial Data Dictionary (SAGDaD) (Vorster and Duesimi, 2010).

⁴www.openstreetmap.org

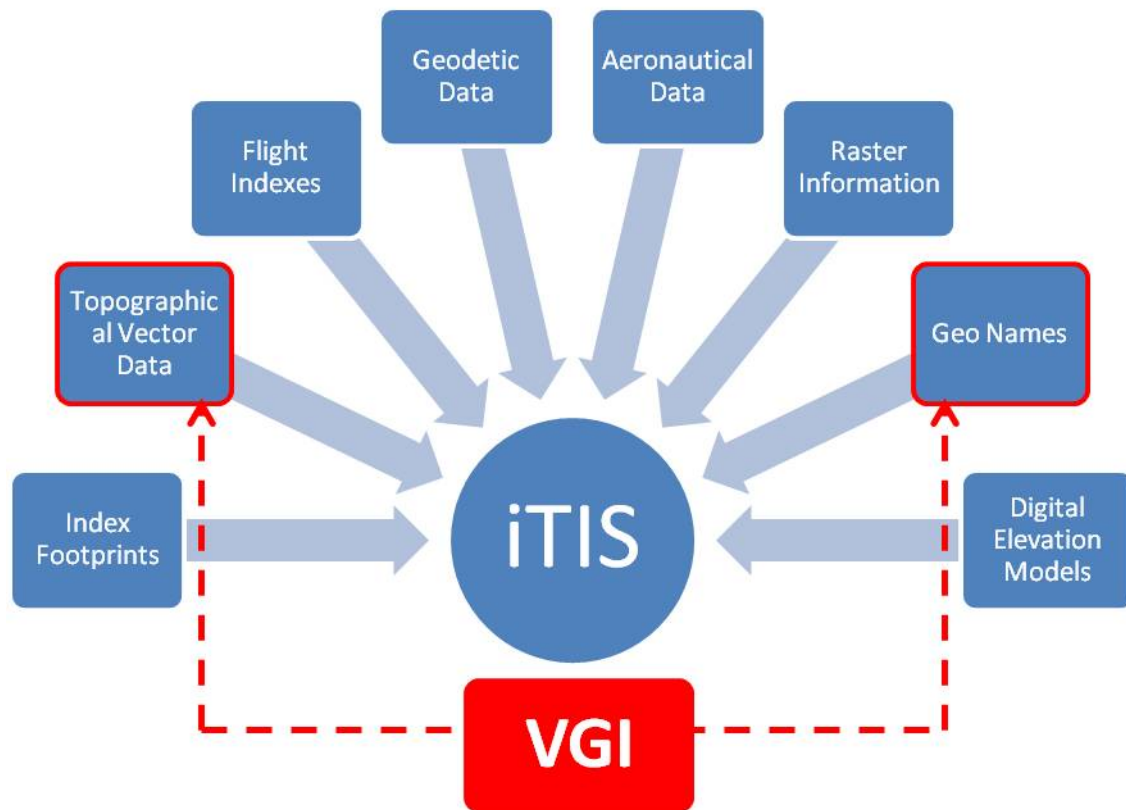


Figure 1.3: Integration of the CD: NGI data through the iTIS and possible VGI integration opportunities (Vorster and Duesimi, 2010)

The move toward an integrated system will facilitate the implementation of the South African Spatial Data Infrastructure (SASDI). A discussion on the SASDI will be provided later. Despite this move to a new integrated system, recent developments in the democratisation of geographical information (i.e. VGI), and the availability of the tools and resources to access that information, has placed the CD: NGI, and NMAs in general, at a disadvantage.

Over the last two years, CD: NGI has expressed a need for establishing innovative ways to detect changes on the earth's surface in order to update its topographical database more regularly. The CD: NGI identified VGI, specifically OpenStreetMap data, as a potential source of current geographic data. As can be seen from figure 1.3, VGI has the potential to be a valuable source of topographical vector data and geographical names (i.e. place names) for the CD: NGI's topographical database.

1.2.2 OpenStreetMap

The OpenStreetMap (OSM) VGI initiative has been tested by numerous researchers (Ather, 2009; Anand, Morley, Jiang, Heshan and Hart, 2010; Haklay, 2010). The OSM repository has seen a rapid increase in volunteer contributions over the years. The types of contributions constitute GPS data as collected by the public, vector data digitised off aerial and satellite imagery and mass geospatial information uploads made available by authoritative sources (Mooney, Corcoran and Winstanley, 2010). The following sections

provide a brief description of the OSM data model.

Web-mapping repositories have a bottom-up approach unlike authoritative database structures, which have a hierarchical structure (Genovese and Roche, 2010). What this means for web-mapping repositories, is that data is decentralised and the public is free to contribute and use the data as they like (Haklay and Ellul, 2010). One of the main reasons why users are motivated to contribute spatial data, leading to the success of a project like OSM, is because users possess this freedom (Budhathoki, Nedovic-Budic and Bruce, 2010), (Ramm, Topf and Chilton, 2011, : 6).

1.2.3 OpenStreetMap Data Model

The OSM data model is complex and is comprised of various Structured Query Language (SQL) databases and the Extensible Markup Language (XML) schema of the API (Ramm *et al.*, 2011, :51). The OSM API where users contribute and view the data, is separate from the actual database (Cooper, 2010). The API allows the user access to the data within the main database (Ramm *et al.*, 2011). Third party websites like Geofabrik⁵ and Cloudmade⁶ make extracts of the full OSM repository, (i.e. either divided by country or update period) available for downloading (Behrens, 2011).

The basic elements or data primitives of OSM are nodes, ways, tags and relations. Formats of the elements include binary, human-readable and database schemes (Behrens, 2011). Nodes are point features and ways are a series of nodes forming a linear feature (Ramm *et al.*, 2011, :52). The OSM model does not include the polygon data type (Ramm *et al.*, 2011, : 83). Instead, a closed way can be tagged as a polygon feature type. However, when OSM data is exported to the Esri shapefile format, closed ways are transformed into polygons.

Tags are used to store information about a feature in the form of keys and values (Ramm *et al.*, 2011, : 54), where a key can be seen as the class that a feature belongs to (e.g. Tourism) and the value can be seen as the attribute (e.g. Hotel) (Behrens, 2011). Relations are the latest addition to the OSM data model and are used to join multipart features (Ramm *et al.*, 2011, : 54).

1.3 Related Work

1.3.1 Status of the South African Spatial Data Infrastructure

The purpose of any Spatial Data Infrastructure (SDI) is to establish “a framework of technology, policies, standards and human resources necessary to acquire, process, store, distribute and improve the utilisation of geographic information” (Craglia, 2007). The need for SDIs arose when governmental departments recognised their inability to produce and maintain high quality spatial data (Mcdougall, 2010). There was a need to integrate data from different organisations so as to improve resource management (Mcdougall, 2010).

⁵www.geofabrik.de

⁶download.cloudmade.com

In the mid 1980's, South Africa attempted to implement a spatial data infrastructure i.e. the South African Spatial Data Infrastructure (SASDI), but due to a lack of human resources, sufficient digital information and human expertise, these attempts failed (Clarke, 2011). It was only in February 2004 that the SDI Act (Act 54 of 2003) was passed in South Africa (Clarke, 2011). The act provides regulations regarding the creation, management, standardisation and interdepartmental synchronisation of spatial data in South Africa (Government Gazette, 2004). Before the passing of the Act, between 1998 and 2004 there were many attempts to implement the ideals of an SDI (Clarke, 2011). There were many positive outcomes as a result and these probably led to the establishment of the SDI Act. However, from 2005 to 2009 there was very little further development to the SASDI and thus a new approach was sought (Clarke, 2011).

In 2010, the responsibility of implementing the SASDI was again entrusted to the CD: NGI (Clarke, 2011). One of the main outcomes of this change was the establishment of the Committee for Spatial Information (CSI) as specified by the act. The responsibilities of the CSI include prescribing standards for managing, integrating and distributing spatial data across organisations (Government Gazette, 2004; Cooper and Eloff, 2011). Also, providing support to ensure the implementation and functioning of the SASDI (Government Gazette, 2004). It appears that the shift in responsibility to CD: NGI has brought about a noticeable acceleration in the implementation of the SASDI. Makanga and Smit (2010) do however state that the implementation of National SDIs across the African continent has been very slow compared to the rest of the world.

1.3.2 VGI in the SDI Context

The consideration of VGI as an official data source cannot be done outside the context of SDI. A SDI is focused on standards so that when data from various organisations are integrated, a global spatial data standard exists. VGI enforces no spatial data standards. Cooper, Rapant, Hjelmager, Laurent, Iwaniak, Coetzee, Moellering and Duren (2011) state, “an effective SDI should generate participatory VGI because it provides value to end users and hence stimulates them to contribute to the SDI.”

Cooper, Rapant, Hjelmager, Laurent, Iwaniak, Coetzee, Moellering and Duren (2011) proposed a SDI model where six different stakeholders involved in the SDI were identified. However, after considering the success of VGI, the authors reconsidered the responsibilities of each stakeholder to include VGI into the SDI model. There are existing SDIs, which already integrate VGI in various degrees (Guelat, 2009) as cited in (Cooper, Rapant, Hjelmager, Laurent, Iwaniak, Coetzee, Moellering and Duren, 2011). In most cases though, SDI and VGI have remained separate entities because integrating spatial data with varying quality continues to be a great challenge (Al-Bakri and Fairbairn., 2010).

South Africa may be a long way from implementing the SASDI, but the country is working toward it. Thus, considering VGI separate to SDI would be impractical. Superseding this possible integration of VGI into SDI, is gaining knowledge on the quality of VGI.

1.3.3 Quality, Quality Assurance and Quality Control

The internal quality refers to what the data is really like and is not dependent on the perception of the user (Mostafavi et al. 2004). The external quality may be described as how well the product meets the user needs (Mostafavi et al. 2004). Another way of describing the external quality is its fitness for use (Girres and Touya, 2010).

According to Goodchild and Li (2012), NMAs maintain their quality assurance by implementing quality control procedures at the various stages of data production. In addition, there are procedures to perform quality assessments of the compiled data (Goodchild and Li, 2012). The CD: NGI implements various quality assurance procedures. These procedures are there to ensure that the elements of quality are within the prescribed tolerances. The CD: NGI aspires to a high positional accuracy, attribute accuracy, consistency, completeness, semantic accuracy and temporal accuracy.

1.3.4 Quality Assessment

Amongst the numerous VGI initiatives, OSM has gained the most attention. Previous researchers have attempted to assess the quality of OSM data by comparing it to a reference data set (Haklay and Ellul, 2010; Girres and Touya, 2010; Zielstra and Zipf, 2010a). In most cases, the reference data set is from an authoritative source. This type of assessment yields results about the internal quality. However, the fitness for use is also being determined, where the NMA is the user and the use is updating of their topographical database. Researchers will investigate those quality elements that are of most importance to them. However, the positional accuracy, semantic accuracy and completeness are considered to be of greater importance.

There are two popular buffer techniques that may be used to assess the positional accuracy of linear features. The technique used most often to assess VGI was introduced by Goodchild and Hunter (1997). The second technique is called the buffer-overlay-statistics (BOS) method by Tveite and Langaas (1999). The positional accuracy of polygons may be determined by distance computations, for example, the Hausdorff distance or the two-dimensional surface distance.

The semantic accuracy determines the correctness of feature classification. Previous researchers have determined the semantic accuracy of VGI automatically (Al-Bakri and Fairbairn, 2011; Patwardhan, Banerjee and Pedersen, 2003). Although it may be more time consuming, the semantic accuracy may also be determined manually.

Completeness was one of the first quality measures performed on VGI data. It looks at how much data is missing and how much extra data exists in the test data when compared to a reference data set (Haklay, 2010), (Haklay and Ellul, 2010) (Girres and Touya, 2010).

The varied results obtained in the above-mentioned studies showed that the location and extent; the reference data set; the method and execution of testing and the date of the VGI extraction influence the results considerably.

1.4 Problem Identification

The literature has highlighted two important facts: i) the type of quality assessments undertaken is dependent on the end use of the VGI and ii) the results are relative to the reference data set. The results from this spatial data quality study are intended to provide insight into VGI and authoritative data integration. At the time of writing, no results were applicable to South Africa.

1.5 Research Objectives

The objectives of this dissertation are to investigate:

- the quality of VGI within a South African context
- the integration of OSM data into the CD: NGI topographical database

The working hypothesis is that VGI meets the CD: NGI national mapping standards and can thus be used for updating the CD: NGI topographical database. The assumption is made that the reference data has a higher quality than the OSM data.

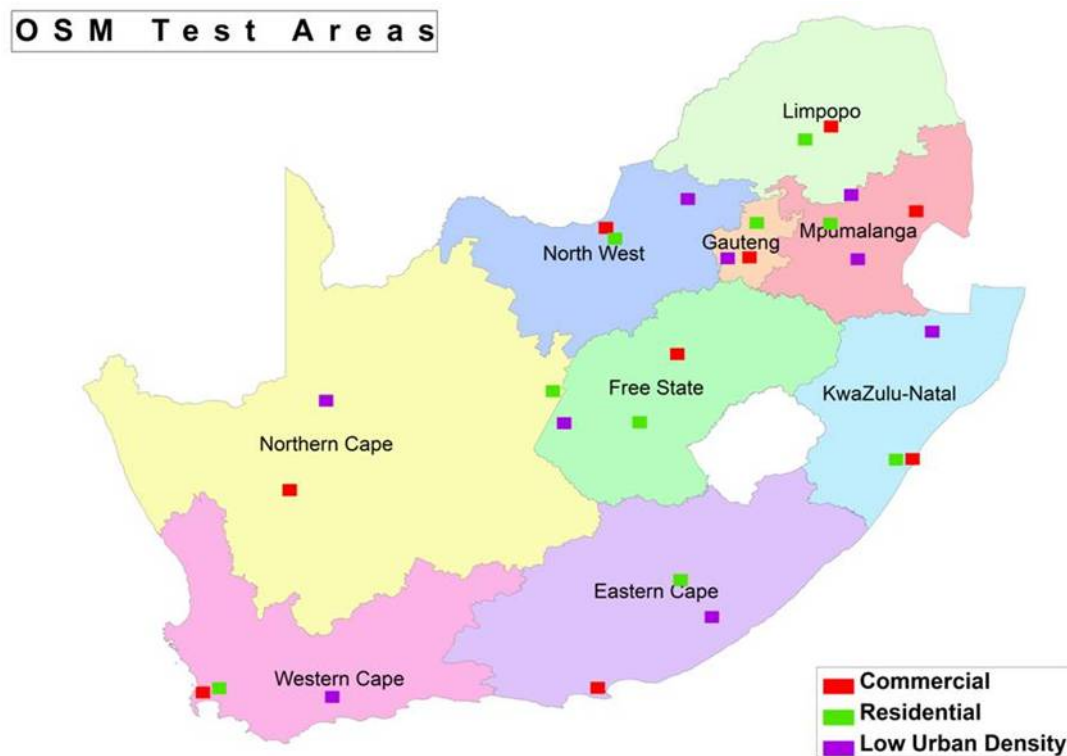


Figure 1.4: Map showing the twenty-seven test areas across South Africa

1.5.1 Research Questions

The first objective, investigating the quality of VGI within a South African context, will be addressed by answering the following research questions:

- Which accuracy assessments are most important within a NMA context?
- Does the OSM data meet the CD: NGI spatial data standards?

The second objective, investigating the integration of OSM data into the CD: NGI topographical database, will be addressed by answering the next set of questions:

- Is it necessary to meet all the accuracy requirements in order to use the OSM data?
- Is the data acquired evenly across the country?
- What is the rate of data generation and is it sufficient from a NMA perspective?
- What is the most frequently generated data type and is this useful to the CD: NGI?
- How can VGI be used within the CD: NGI mapping structure?

Both quantitative measures and qualitative assessments are necessary to answer the research questions.

1.6 Methodology

A flowchart outlining the methodology is provided in figure 1.5. The techniques used are grouped in order to clearly answer the questions associated with the quantitative and qualitative assessments. The combination of quantitative and qualitative assessments was chosen based on the CD: NGI accuracy requirements. Due to technical and time constraints, the assessments do not respond to all the requirements.

1.6.1 Quantitative Assessments

The quantitative measures will be used to answer the first set of research questions that address the first objective. It includes the:

- Geometric Accuracy —includes both the positional and shape accuracy (for polygons) (Al-Bakri and Fairbairn, 2011).
- Semantic Accuracy —how accurately a feature is represented in the database when compared to how it is interpreted (Haklay, 2010).
- Completeness —the omission of data and the presence of excess data (that is, commission) (Girres and Touya, 2010).

1.6.2 Qualitative Assessments

The qualitative measures will be used to answer the second set of research questions that address the second objective. It includes the:

- OSM currency —investigates the the evolution of the OSM point, line and polygon data from 2006 to 2012.
- Uniformity in data acquisition —investigates whether OSM data is contributed evenly across the country.

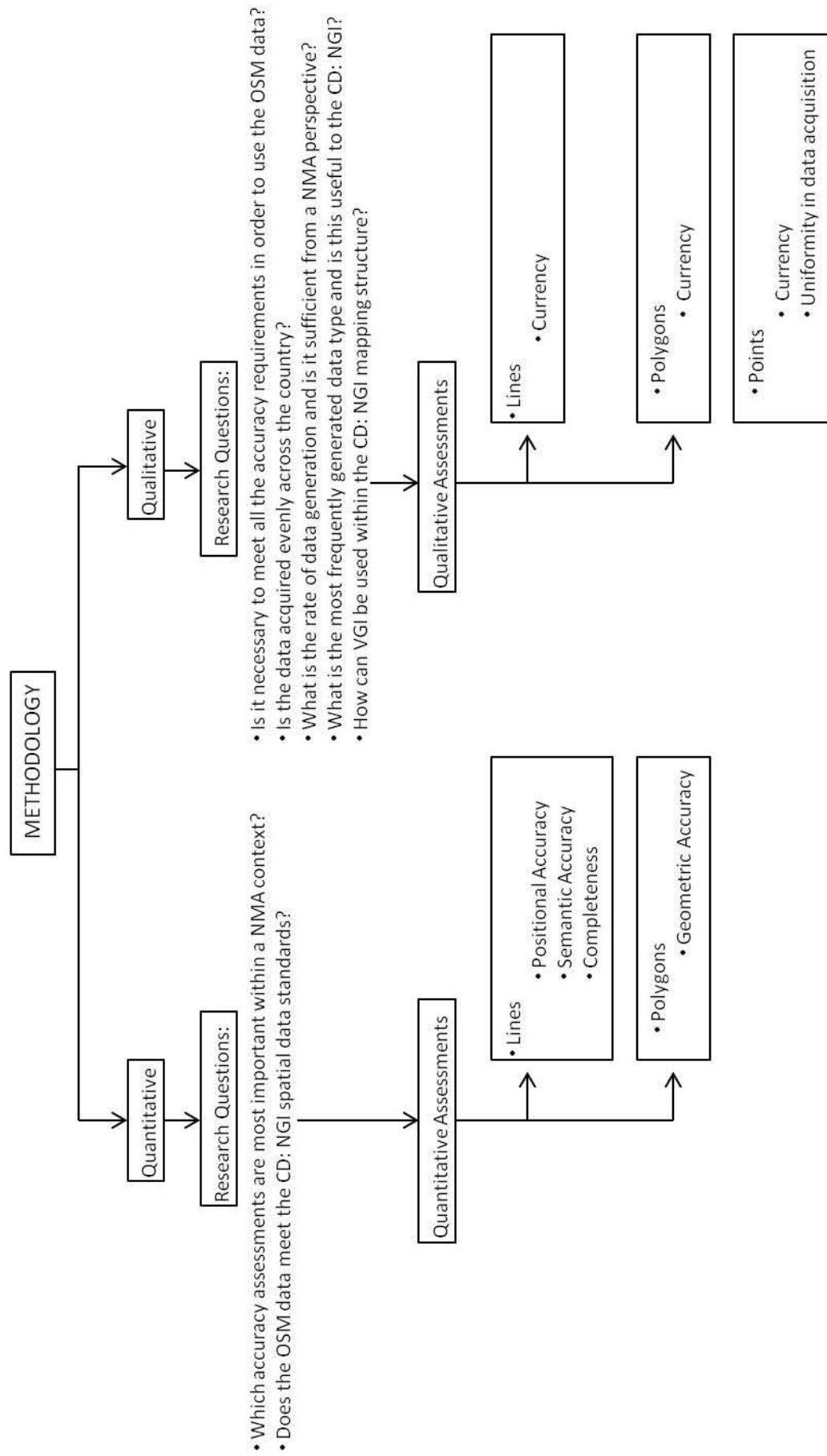


Figure 1.5: Flowchart outlining the methodology

1.7 Scope

The extent of the testing was limited to South Africa. Twenty-seven test areas were sampled, three from each of the nine provinces. The test areas included high-and-low urban density areas and residential areas. The OSM VGI initiative was chosen as the test data set for this investigation, because: i) it is easily and freely available for testing, ii) of the rapid growth rate since it started in 2004 and iii) the results obtained in this study may be compared to the literature. The reference vector data set was obtained from the CD: NGI. Both the test and reference data sets were exported to and processed in shapefile format. The Esri ArcMap GIS software will be the main processing software. All OSM data elements was analysed, this includes points (or nodes), lines (or open ways) and polygons (or closed ways).

1.8 Outcomes

The results of this dissertation will be specific to the test areas; however, the method can be repeated for as many areas as necessary, in order to provide a better understanding of the quality of OSM in South Africa. If the OSM data meet the CD: NGI national mapping standards, then the OSM repository will provide a rich geographic data source. Whether it is necessary to meet all the mapping standards or only a subset will be discussed at a later stage.

The qualitative assessments provide an overview of how users contribute data across South Africa. The literature has already established that contributions will vary across a country. What this implies is that NMAs cannot adopt one standard integration process, but rather it should be specific to the settlement type or province.

The purpose of assessing the quality of VGI in a NMA context is ultimately for updating authoritative databases. In theory, the VGI would feed directly into the databases but in reality, this may not be feasible. Perhaps the better option is to propose a new data flow model, where VGI is incorporated and a quality assessment is performed on all new data before it feeds into the main database. This may prove to be a difficult task in the immediate future, but it is certainly worth investigating. What is plausible at this stage is using VGI as a method of detecting changes to the landscape, but this should not be its final purpose. Reporting on the differences between authoritative geographic information and VGI, which is what this thesis will provide, aids in identifying of how far off the CD: NGI is from integrating VGI into its current data production flow.

The main outcome of this investigation is a summary of OSM quality in reference to the CD: NGI geo-spatial data. This will be the deciding factor whether it is feasible to integrate OSM and the CD: NGI data.

A review of the literature related to the investigation is presented in chapter two. The third chapter provides a comparison of the CD: NGI and OSM data models. The methodology undertaken is presented in chapter four and the results in chapter five. Chapter six gives a summary of the processes needed to transform the OSM data in order to resemble the CD: NGI data. In addition, a proposal is made for the CD: NGI and OSM data integration. The final chapter provides a discussion on the conclusions and the

recommendations.

Chapter 2

LITERATURE REVIEW

2.1 Introduction

The terms “crowdsourcing” and “user-generated content” (UGC) are closely related to VGI and are often used interchangeably. Literature shows that there are clear distinctions between these three terms. Schenk and Guittard (2009) define crowdsourcing as, “a form of outsourcing not directed to other companies but to the crowd by means of an open tender (open call) via an Internet platform. It is important to emphasize that the call should not be limited to experts or preselected candidates.” The definition of UGC varies, but Cooper, Coetzee, Kaczmarek, Kourie, Iwaniak and Kubik (2011) describe it as the broader context, which includes any information that is being relayed from one person to the next, whether documented or given verbally. In terms of the Internet, the definition becomes specific, where the information is shared over Web sites (Goodchild, 2008*b*). Another important part to the definition, is that the information is created by non-experts using non-expert methods (Wunsch-Vincent and Vickery, 2007) as cited in (Cooper, Coetzee, Kaczmarek, Kourie, Iwaniak and Kubik, 2011). VGI may be seen as a branch of UGC where the content being generated is comprised of geographic information (Goodchild, 2007). In contrast to this, authoritative data is geographical information created by experts within a NMA environment. The NMA may or may not operate under a SDI.

This chapter provides a summary firstly, of the various aspects associated with VGI, including the need for VGI, the advantages and disadvantages of VGI and the factors that influence the success of VGI. Secondly, examples of VGI and authoritative data integration are discussed and thirdly, a more specific discussion on the methods of quality assessment of OSM data is provided.

2.2 Volunteered Geographic Information

2.2.1 Advantages of VGI

Increasing and Frequent Demand for Spatial Data

In many developing countries, spatial data sets are either incomplete, outdated or non-existent (Bishop, Escobar, Karuppannan, Suwarnarat, Williamson, Yates and Yaqub, 2000). SDIs were supposed to address this problem, but in reality there are many obstacles prohibiting the implementation of a successful SDI especially in developing countries

(Bishop *et al.*, 2000). There are cases where SDIs have been highly successful in developed countries, because the country has what is required to implement and maintain a SDI (Musinguzi, Bax and Tickodri-Togboa, 2004). Examples of countries who have had success with SDI implementation are Australia with its Australian Spatial Data Infrastructure (ASDI); The United States of America's United States Spatial Data Infrastructure (US SDI) and the European Union's Infrastructure for Spatial Information in the European Community (INSPIRE) initiative (Antoniou, 2011; Musinguzi *et al.*, 2004).

Amongst the many reasons for the lack of spatial data in developing countries, is the unavailability of spatial data in a digital format (Bishop *et al.*, 2000). They simply do not have the resources, whether technical or financial, to shift to a digital system (Bishop *et al.*, 2000). VGI thus provides access to rich spatial data to those countries that are in desperate need. Genovese and Roche (2010) investigated the potential of using VGI for struggling developing countries and found several factors that hinder the integration of VGI into authoritative data. They suggest that it may not be practical to use VGI at least until the crucial questions about VGI have been confidently answered, for example: will the public continue to show interest and thus participate in VGI (Genovese and Roche, 2010)? The authors do however conclude that VGI in developing countries have the potential to increase the social utility value of SDIs; where social utility can be defined as the usefulness of an object to its users (Roche, Sureau and Caron, 2003; Genovese and Roche, 2010).

NMAs who used to be the primary providers of spatial data, are now finding it an even greater challenge to maintain the currency of spatial information since the emergence open initiatives (McLaren, 2012). The challenge will however be greater in developing countries. The CD: NGI, for example, works on a pre-determined change detection program. Aerial imagery is generated for a third of the country annually, instead of detecting changes as they occur on the ground. Thus, if a change has occurred just after the completion of the first year of the cycle, in an area that is set for the first year of flying, that change will only be captured in the following cycle (i.e. three years later). Note that this is only the capturing of the aerial image and not the detection of that change. Detecting the change will occur at the compilation stage and this in itself is a lengthy procedure due to a lack in human resources. Other NMAs may have the delay at a different stage of the production process, but the result is the same - an outdated map (Goodchild and Glennon, 2010).

While official mapping is declining worldwide, (Mcdougall, 2009), (Gould, 2007) as cited in (Genovese and Roche, 2010) the demand for spatial data is rising (Mcdougall, 2009). Instead of competing with open initiatives, NMAs are realising that open initiatives present a great opportunity for collaboration (McLaren, 2012). Goodchild (2007) as cited Genovese and Roche (2010) states that VGI could even replace the traditional methods of spatial data capture in operation at NMAs, whereas Elwood, Goodchild and Sui (2012) suggest that although non-experts are providing vast amounts of spatial data, it could perhaps serve as a means to detect changes to the landscape and not necessarily replace official data and mapping processes.

Reducing Map Production Costs

It is a well-known fact that authoritative map production is costly (Budhathoki, Bruce and Nedovic-Budic, 2008; Goodchild, 2008*b*; Goodchild, 2009) and in some parts of the world the information is expensive to obtain (Haklay, Singleton and Parker, 2008; Budhathoki *et al.*, 2010). According to Elwood *et al.* (2012), the high cost of map production incurred by NMAs is one of the main reasons for the emergence and rapid growth of VGI. High production costs are experienced by NMAs worldwide. As a result, NMAs have tendered out compilation work to the private sector for many years incurring additional external production costs (McLaren, 2012; Mcdougall, 2009).

Flanagin and Metzger (2008) state that generally, digital network technologies have caused a decrease in the cost of spatial data production and dissemination (Goodchild, 2007; Zielstra and Zipf, 2010*b*). The next step is to transfer this reduced cost to NMAs because currently, it is benefiting the users, who use to obtain spatial data from mapping organisations. NMA operational policies vary from one country to the next, thus the extent to which an NMA is able to save on production cost will depend on how they are able to make use of this freely available spatial data. The highest instance of reduced costs would be when VGI is directly ingested or ingested with minimal editing and automated verification techniques. Whether this is possible in reality has yet to be demonstrated.

There is a general view amongst researchers on the topic that any level of VGI ingestion will result in reduced map production costs (Antoniou, 2011; Goodchild, 2008*b*). Johnson and Sieber (2013, :71) give a different perspective, they state that the previous researchers have made it seem as if the costs involved in integrating VGI are minimal, but this may not be the case. There are costs involved in re-aligning the NMA workflow to incorporate VGI, Internet costs for accessing the latest updates and funds for training staff in VGI usage (Johnson and Sieber, 2013, :71). It is thus necessary that NMAs perform an extensive cost analysis to weigh up whether there will in fact be a reduction in costs.

Community Involvement

In addition to the possible reduction in production costs for governmental mapping organisations, VGI also presents an opportunity for the community to become involved in decision-making processes. Gouveia and Fonseca (2008) believe that community involvement is very necessary. Johnson and Sieber (2011) and Ryttersgaard (2001) state that community involvement promotes governmental transparency.

Location of VGI contributors will determine the correctness and richness of spatial data (Carrera and Ferreira, 2007), (Johnson and Sieber, 2013, :66). Those living closest to the spatial feature will be able to contribute the most valuable spatial description of that feature. However, in most cases the contribution will not come from the locals, especially in poorer communities. A previous study on the OSM user activity has shown that the number of OSM members is continuously increasing with about 150 new members daily since the start of 2011 (Neis, Zielstra and Zipf, 2012). Even though there are thousands of registered members, only a small amount actually contributes. Less than 10% contribute data and only 1% actively contribute on a continuous bases (Neis *et al.*, 2012).

Haklay, Basiouka, Antoniou and Ather (2010) investigated whether the accuracy of OSM data increases as the number of volunteers increase. The results showed that for OSM the accuracy increases with more contributors, but not linearly. They obtained an accuracy of better than 6 m when the number of contributors was more than fifteen. What they also found was that the greatest improvement on positional accuracy in an area of 1 km² is obtained by the contributions of the first five volunteers.

These are the statistics of but one initiative, but it is one the most successful VGI initiatives, so it does give an indication of community involvement for future initiatives. Although only a few of the registered members contribute, thousands of people are aware that they can be part of the great VGI movement. Perhaps in years to come, more people will be willing to contribute spatial data or have access to the resources needed to contribute to open initiatives.

2.2.2 Disadvantages of VGI

There are two main disadvantages of VGI: i) malicious contributions and ii) the uncertainty in quality. The latter is one of the most popular topics of the investigation amongst researchers. At this phase of VGI investigations, it appears as though malicious contributions are not prevalent. It is however an important issue especially when considering integration of VGI and authoritative data. Authoritative sources like NMAs may be held responsible for disseminating incorrect information, but with VGI there are no consequences for users contributing malicious content (Cooper, Coetzee, Kaczmarek, Kourie, Iwaniak and Kubik, 2011; Goodchild, 2008b). The most that can be done is to prohibit the user from contributing any data in future.

Malicious Content

There have been no significant developments for methods to detect malicious content. Quality assessments could provide some indication about malicious content, because malicious information is in effect incorrect information. Whether or not the data is incorrect because of human error or because of malicious intent would require further investigation. As Cooper, Coetzee, Kaczmarek, Kourie, Iwaniak and Kubik (2011) state, malicious content is likely to have sufficient metadata to appear credible. Thus, the task of detecting malicious content may prove to be complicated.

In 2010, the Department of Computer Science at the University of Pretoria reported on a project whereby they used formal concept analysis (FCA) to assess taxonomies (or classifications) for various user-generated-content (Cooper, Kourie and Coetzee, 2010). Cooper *et al.* (2010) state that the FCA methodology could be extended to assess the quality of VGI and therefore to detect malicious content.

Detection of malicious content within VGI would be easier to identify than most user-generated content, because of spatial context (Elwood *et al.*, 2012), (Sui, 2004) as cited in (Goodchild and Glennon, 2010). Goodchild and Li (2012) explain that there is a geographic approach when observing the quality of VGI. The approach is based on the law that all spatial features are related; where the measure of relatedness decreases with

distance (Goodchild and Li, 2012). Logical consistency which refers to the how a spatial data set complies with logical rules could be used to detect inconsistent spatial contributions. Coleman, Georgiadou and Labonte (2009) as cited in Antoniou (2011) state that the quality of VGI may be indicative of the motivation of contributors. This implies that malicious motivations degrade the quality of VGI. Quantitatively it cannot be said how much malicious content exists within any given VGI data set, which makes it difficult to say exactly what the effect of malicious content has been on the quality of VGI thus far.

VGI Quality Assurance

Quality includes a certain level of trust, which is missing from VGI as opposed to the level of trust gained by NMAs (Goodchild and Glennon, 2010). A person, who uses spatial data to generate an income, will typically choose to use official data even if it is not free or cheap. There must be that confidence that what has been produced is correct. VGI is a long way from gaining the level of trust that NMAs have. Simply put, VGI currently does not carry the assurance of quality (Goodchild and Li, 2012).

2.2.3 Examples of VGI Initiatives Available in South Africa

Examples of VGI initiatives available in South Africa are mentioned in the following section. The definition of VGI encompasses a whole range of initiatives. For example, some initiatives acquire contributions without the user being aware of their participation, i.e. involuntarily (Elwood *et al.*, 2012). These are passive contributors, while other initiatives are based solely on the active participation from users like the OSM initiative. Some initiatives that make use of passive contributions, obtain the permission of users first. Below are three examples of VGI initiatives available in SA.

TomTom

The TomTom in-car navigation initiative is a good example of contributions that are made involuntarily, but with the consent of their clients. TomTom generates 35 million kilometres of road, covering 104 countries (Grobbe, 2012). Users may report on possible mapping errors via the on-line map editing application, MapInsight (Coleman *et al.*, 2010). In addition to this, the MapShareTM application was also launched in 2007, where users can add updates either on a desktop computer with Internet access; with a commercial GPS device or with any android cell phone or iPhone (Coleman *et al.*, 2010). Users of the applications must be registered, but the application is not free. As registered users drive, TomTom automatically generates routing data. All updates are verified, before they are accepted. Figure 2.1 shows a user making an update using the MapShareTM application. This specific update was not accepted after being verified.

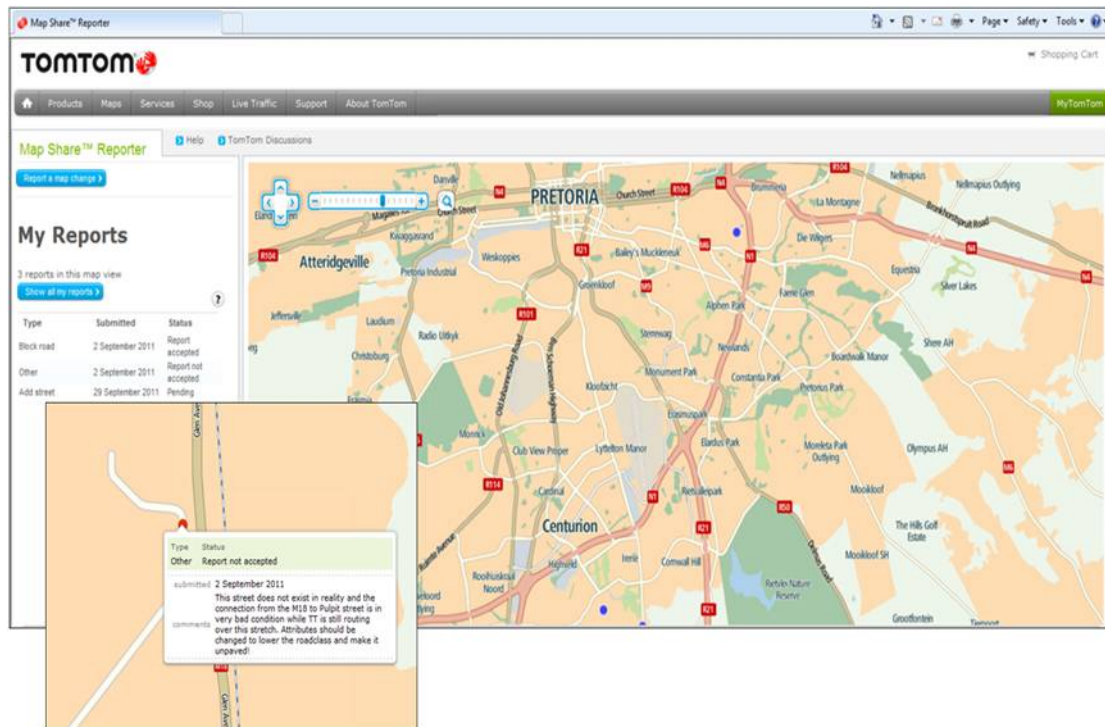


Figure 2.1: Example of a user making an update using the MapShareTM application (TeleAtlas, 2013)

NAVTEQ

NAVTEQ is similar to TomTom in that it provides navigation data to paying users only (Mcdougall, 2009; Coleman *et al.*, 2009). Apart from their other means of obtaining road data, the registered users are used to generate spatial data via cell phone and in-car GPS devices. All errors and updates submitted by users are individually verified before any changes are made to the base map (Kounadi, 2009; Coleman *et al.*, 2009).

Waze

The Waze navigation application collects routing information via their on-line and mobile application (see figure 2.2) (Coleman, 2010). But unlike the afore-mentioned initiatives, Waze spatial data is freely available to the public. The user community contributes all spatial data and traffic updates and there is no verification process for new updates.

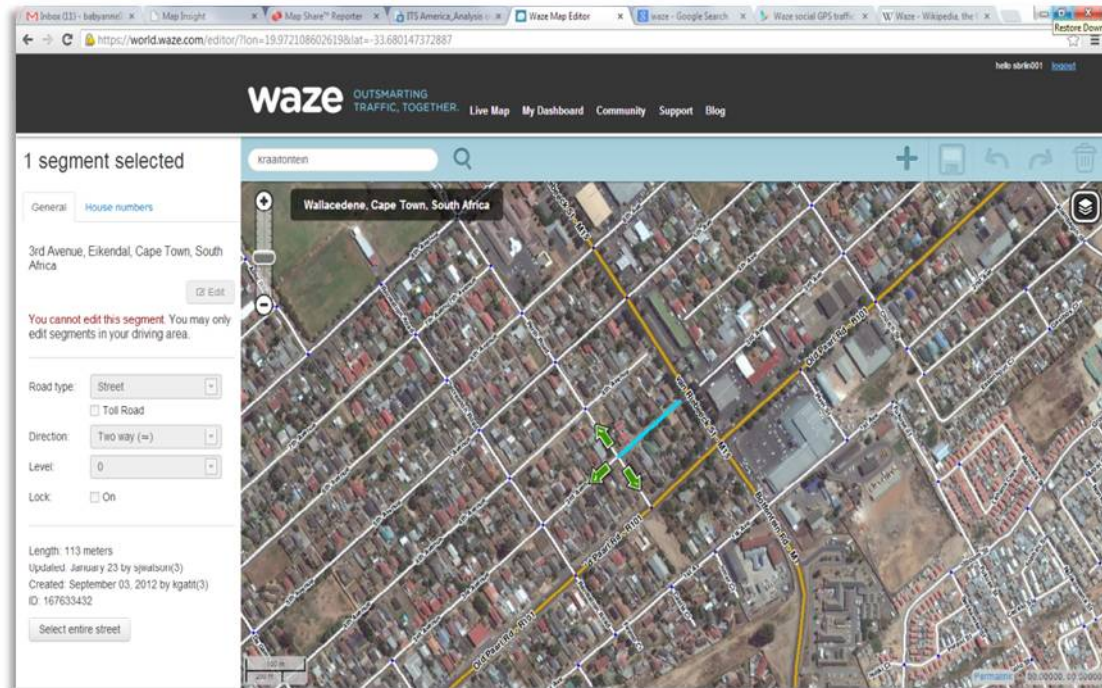


Figure 2.2: Example of the on-line Waze application

2.2.4 Factors Leading to the Success of VGI

Only three VGI initiatives have been discussed, but there are many other initiatives worldwide that have been successful. The factors that have led to the success of VGI initiatives include amongst others, user motivation and the availability of resources.

User Motivation

Previous researchers have presented various motivations behind user contributions (Coleman *et al.*, 2010; Shekhar, 2010; Budhathoki *et al.*, 2010). Coleman *et al.* (2009) grouped volunteers into five main categories based on their motivations. The categories ranged from the passionate non-expert that has no background knowledge, to the expert authority whose livelihood may depend on the credibility of him or her providing information within this sphere of knowledge (Coleman *et al.*, 2009).

It is useful to consider user motivations, as it may lead to a better understanding of the mechanisms of and elements involved in a VGI initiative and how these may be improved to better suit the volunteers. Coleman *et al.* (2009) state that amongst the positive motivations for contributing data, are also negative motivations. These include agenda, malice and mischief, where contributors will contribute incorrect or false information. Goodchild (2008a) on the topic of spatial accuracy states that, there will always be people who are disruptive users when it comes to Web activity. Although, malicious contributions may be more likely to occur where spatial data is linked to sensitive data (e.g. political content) or where the data could be used to paint a picture that may influence the decisions or viewpoint of the users of the information (Flanagin and Metzger, 2008; Antoniou, 2011).

Thus, understanding user motivations is also useful when determining the credibility of VGI (Flanagin and Metzger, 2008). The definition of credibility as referred to here is the believability of a source or message, which is composed of two primary dimensions: trustworthiness and expertise (Hovland, Janis and Kelley, 1953) as cited in (Flanagin and Metzger, 2008). In an early look at user motivations by Goodchild (2007), self-promotion without any incentive, is presented as the main reason why people would contribute information.

Technological Developments

Three major technological developments have allowed for VGI. The first of these was the introduction of Web 2.0, which has provided the opportunity for the user to create content and make it available on the Internet (Goodchild, 2008*a*; Haklay *et al.*, 2008). In other words, there is a two-way sharing of information as opposed to Web 1.0 where the user is being served information. The advantage of Web 2.0 and the fact that computer hardware and broadband connections to the Internet has become cheaper over the years, has greatly encouraged user interaction (Flanagin and Metzger, 2008).

The second advancement in technology is increased access to GPS services, which have been made available through commercial sources. These include car-navigation devices and GPS-enabled cameras allowing any person to add co-ordinates to their image (Goodchild, 2007). Anyone with a handheld GPS device can generate a track either actively or passively. Examples of VGI initiatives that have been created for GPS track uploads are Waze and Google Map Maker.

Finally, non-experts are able to download free GIS software (i.e. freeware) allowing them to view and create vector data for contributing to a VGI initiative. The OSM initiative for example, has an off-line Java Editor called JOSM where users can access and make edits to the existing base map (Ramm *et al.*, 2011, :109).

2.2.5 Factors Opposing VGI Success

The obstacles that hinder VGI and SDI integration are numerous and although there are common factors across the world, there are those factors which are specific to a country. One of the greatest limiting factors that is common across the globe is the digital divide because VGI cannot exist where technological infrastructure is unavailable. This in itself is the result of a number of factors, the most obvious being the lack of finances.

The Digital Divide

Williams (2001) as cited in Genovese and Roche (2010) describes the digital divide as the “gap between people with adequate access to digital information and technology versus those with very limited or no access at all”. It has been said that the main digital divide at a global level is between the developed countries in the northern hemisphere and the undeveloped countries in the southern hemisphere (Keniston, 2003). Governments have drawn on various information and communication technology (ICT) solutions to address

the digital divide (Baskaran and Muchie, 2006). Only a small percentage of the world has managed to minimise the gap significantly, while other parts like sub-Saharan Africa still have not seen much improvement (Genovese and Roche, 2010; Norris, 2000; Keniston, 2003).

Wilson (2006) as cited in Fuchs and Horak (2008) describes the digital divide as a multifaceted concept with eight aspects linked to various social demographic dimensions. The identified aspects include: “access to ICT services”; “availability of relevant applications and information on-line” and “capacity to produce one’s own content” (Wilson, 2006) as cited in (Fuchs and Horak, 2008). When there is an imbalance in or a lack of any of these aspects, it positions a community or country on the less advantageous side of the digital divide. In developing countries, there is an imbalance between providing telecommunication infrastructure and the access citizens have to it. Many developing countries have been able to provide telecommunication services due to the liberalisation of the telecommunications market. Fuchs and Horak (2008) completed a study on Internet access in Ghana and South Africa. They argue that liberalising the telecommunications market does not solve or contribute greatly to the digital divide problem because those living in poorer communities are not able to afford access to these technologies.

South Africa is definitely on the disadvantaged side of the digital divide, but it has seen more progress in access to ICT services than other developing countries like India. In 2005/6 South Africa was ranked at the 23rd place in terms of telecommunication networks and ranked number one across the African continent (Baskaran and Muchie, 2006, :193).

The status of a country’s ICT services will determine how successful a VGI initiative could be and what growth may be expected in the years to come. The results from previous investigations regarding the growth pattern of OSM data for different settlement areas, are in agreement with this statement (Haklay and Ellul, 2010; Zielstra and Zipf, 2010a; Neis *et al.*, 2012). They found that although OSM has had great global success there is still a clear difference in the volume of contributed data between affluent and poorer communities.

Difference in Culture and Interests

Users within a community will map those features, which they deem to be more important (Siebritz, Sithole and Zlatanova, 2011). Their viewpoint of what is important, is influenced by experiences (Elwood, 2008) as cited in (Basiouka and Potsiou, 2012), cultural traits and education (Siebritz *et al.*, 2011). This in itself does not oppose VGI success as such, but rather it may contribute to inconsistent mapping (or under-mapping) of certain topographical features. More than this, in communities where volunteer mapping is a foreign concept because of a lack of knowledge or where it is not promoted, gaps may be left in the volunteer base data.

2.3 Spatial Data Infrastructure and VGI

2.3.1 The Purpose of SDI

As stated in section 1.3, a SDI can be defined as “a framework of technology, policies, standards and human resources necessary to acquire, process, store, distribute and improve the utilisation of geographic information” (Craglia, 2007). The implementation of a well-defined, functioning SDI leads to a spatially-enabled government (Masser, Rajabifard and Williamson, 2008). The SDI must function in such a way that it is open to its users. This has proven to be a difficult task as NMAs have always been operating as information silos (Masser *et al.*, 2008). McDougall (2010) states that developing countries are more notorious for information silos and these silos are the cause of unsuccessful SDI implementation. SDI policies are aimed at providing uniform spatial data and this is opposed by the open nature of VGI.

2.3.2 Integrating VGI and Authoritative Data

Although in most cases, it is the NMA that is in charge of the development of a country’s SDI, the distinction should be made here that VGI for integrating with SDI is quite different from VGI for integrating with NMA workflows (Budhathoki *et al.*, 2008). As discussed in section 2.1, an SDI encompasses much more than just producing and providing spatial data. This section will first discuss VGI in the context of SDI and then governmental mapping agencies.

Integrating VGI into a SDI

What makes integration of VGI into a NMA workflow simpler than with a SDI is the fact that a SDI is far more complex than the operation of a NMA. Also, a NMA may operate without compliance to national or international prescribed standards. This makes it easier to incorporate VGI because there is no obligation to conflate the VGI with their data.

Goodchild (2007) and Elwood *et al.* (2012) believe that VGI has the potential to aid in successful SDI development. This is however dependent on how a country’s SDI is defined. For some SDI definitions, VGI may be ideal, but the fact remains that in many cases SDI and VGI initiatives are being developed in isolation. As with most things developed in isolation, integration becomes a difficult task.

Other researchers are of the opinion that the definitions and role players of SDI need to be reworked to include VGI so that the users of spatial data become active by contributing information (Cooper, Rapant, Hjelmager, Laurent, Iwaniak, Coetzee, Moellering and Duren, 2011; Budhathoki *et al.*, 2008). Rajabifard, Binns, Masser and Williamson (2006) believe that a new generation of SDI is necessary; one that includes input from the local community. The term used for this type of SDI is “user-centric” (Wytzisk and Sliwinski, 2004). As community input increases, so the need for discerning between the roles of producers and users become less important (Elwood *et al.*, 2012). Elwood *et al.* (2012) state that instead of having citizens as sensors, they should be promoted to partners so that they are involved in decision-making processes instead of simply contributing information.

VGI has gained much more interest and thus interaction from the users than SDIs have, even though the purpose of both VGI and SDI is spatial data sharing (Budhathoki *et al.*, 2008). The reasons for this phenomenon will vary for different countries, one reason may perhaps be due to users not being informed about the operations of the SDI or perhaps the adherence to policies and procedures may seem too cumbersome.

Integrating VGI into a NMA Workflow

Johnson and Sieber (2013, :70-75) describe three problems faced by governmental mapping organisations concerning VGI integration. The first challenge (also mentioned in section 2.2) is the cost of integrating VGI into existing mapping workflows. Beside re-organisation of hardware and software, there is also the human resource aspect to be considered. Employees may not have the necessary spatial analysis skills for VGI, because the structure and functioning is so different to authoritative data (Johnson and Sieber, 2013, :71).

The second challenge is for government mapping agencies to accept spatial data produced by non-experts (Johnson and Sieber, 2013, :72-73). It is a known fact that the public deem authoritative spatial information to be without error, even if this is not necessarily the case. The opposite holds true for most, not all, NMAs—they do not deem volunteer information to be of high quality. Johnson and Sieber (2013, :72) says the reason for government mapping agencies' reluctance to accept volunteer information is the legal implications of incorrect information being disseminated to the public.

Lastly, NMAs are faced by jurisdictional issues. VGI crosses local and national jurisdictional boundaries in order for higher levels of government to address problems (Johnson and Sieber, 2013, :74-75). The belief is that this jumping of scale may cause government to lose control resulting in the government being overruled in decision-making processes (Johnson and Sieber, 2013, :75).

2.3.3 Global Examples of VGI and SDI/NMA Integration

National Map Corps

The United States Geological Survey (USGS) started a program, the National Map Corps, in 1994 where the public could submit annotations and corrections to hardcopy topographic maps (Coleman *et al.*, 2010; United States Geological Survey, 2011). The program was highly successful in terms of people continuously contributing spatial data. In later years, an on-line image viewer was developed, so that users could contribute updates digitally (Coleman, 2013). The program was however stopped in 2008 due to insufficient funding (United States Geological Survey, 2011). Other reasons include the incapacity to assess all the incoming contributions in the allocated time (Coleman *et al.*, 2010).

Victoria Department of Sustainability and Environment - Notification for Edit Service (NES)

The State of Victoria in Australia initiated a program, allowing government users of the state's Corporate Spatial Data Library (CSDL) to annotate updates or corrections to the base spatial data set via an on-line editor (Coleman *et al.*, 2010; Thomas, Hedberg, Thompson and Rajabifard, 2009). These changes and updates are sent to the data custodians for verification (Coleman *et al.*, 2010). The program is still running and the results have shown a considerable decrease in map update time (Coleman *et al.*, 2010).

Regional Spatial Data Infrastructure of North-Rhine Westphalia, Germany

The German ministry of the interior and land surveying service, North-Rhine Westphalia employs Public Private Partnerships (Bernard, 2002; Riecken, Bernard, Portele and Remke, 2003). Through these partnerships, the ministry allows amongst others, users of geographic information to contribute spatial data (Coleman, 2010; Riecken *et al.*, 2003).

2.4 Methods to Assess Quality

NMAs are mandated to ensure that their data has a high internal and external quality. The internal quality refers to the integrity of the data, in other words, the absence of errors (Boin and Hunter, 2006; Devillers, Jeansoulin and Moulin, 2007). The level of internal quality is instituted by organisational spatial data standards. A further discussion on the topic of spatial data standards is provided in section 3.2. The external quality is concerned with the fitness for use (Girres and Touya, 2010). The fitness for use is determined by the satisfaction of the client needs. Clients are satisfied if the; features being mapped are relevant to their needs; products are easily available in various formats; technical support is readily available and if the product prices are reasonable. Determining the external quality is a difficult task because it can only be assessed by user-needs surveys. Users are not always willing to participate in surveys. The internal quality however, may be measured by comparing data sets from different sources. For this investigation, the OSM internal quality is determined by comparing it to the CD: NGI data.

From a NMA perspective, the positional accuracy of spatial data is typically the most sought after element of accuracy (Coleman *et al.*, 2010). Second and third to that would be the semantic accuracy and the completeness (Coleman *et al.*, 2010). The definitions for these quality elements were provided in section 1.6. But before that, it is necessary to discuss the methods for assessing the quality of linear and polygon features. The following sections provide a discussion on the techniques used to determine the positional accuracy of lines and the geometric accuracy of polygons in previous studies.

2.4.1 Positional Accuracy of Linear Features

This section will discuss how results vary due to the different methods employed in assessing the positional accuracy of linear features. Although there are various methods available, the focus will be on buffer techniques and distance between linear features

techniques, because these methods were also used for this investigation

Buffer Techniques

The buffer technique originated from the epsilon error band as introduced by Perkal (1966). The purpose of the epsilon error band is to account for spatial uncertainties in linear features (Perkal, 1966). These spatial uncertainties exist as a result of the cartographic processes used to represent line features using point data (Perkal, 1966). The epsilon error band can be understood as a buffer generated around a linear feature with the width of the buffer being equal to epsilon (Goodchild and Hunter, 1997). By computing the average of the variances of all the error sources, the average standard deviation can be found (Perkal, 1966). Epsilon will equal approximately two times this standard deviation (Perkal, 1966). Two main buffer techniques are discussed below.

The buffer-overlay-statistics (BOS) method was introduced by Tveite and Langaas (1999) where two line data sets, the test and reference data sets, are compared by generating buffers iteratively around both data sets. The use of two buffers introduces a weighting of the errors, so that the parts of a line feature, which are further from the reference line will have a larger weight (Tveite and Langaas, 1999). In this way, this method builds on from the original definition of the epsilon error band (Tveite and Langaas, 1999). The buffer iteration is run until the percentage of the reference data set that lies within the buffer meets the requirements as chosen by the operator (Tveite and Langaas, 1999). The intersections of the buffer areas and the total line lengths are used to compute the average displacement and the oscillation of the test data set (see figure 2.3 (a)) (Tveite and Langaas, 1999). Computation of the oscillation is useful because it identifies the bias (Tveite and Langaas, 1999).

The second buffer technique was introduced by Goodchild and Hunter (1997) and is the most widely used method to assess the accuracy of linear features. The authors state that the use of the epsilon band in previous studies was sensitive to outliers (Goodchild and Hunter, 1997). In this method, the test data is also compared to the reference data sets, but buffers are iteratively generated only around the reference data set (Goodchild and Hunter, 1997). In this way, no assumption is made about the accuracy of either test data set (Goodchild and Hunter, 1997). The initial buffer width chosen may be based on knowledge of the reference data positional accuracy standards (Goodchild and Hunter, 1997). Or the buffer width could be generated iteratively until the desired result is achieved. The accuracy of the test data can be known by computing what proportion of the test linear feature lies within the final buffer (see figure 2.3 (b)).

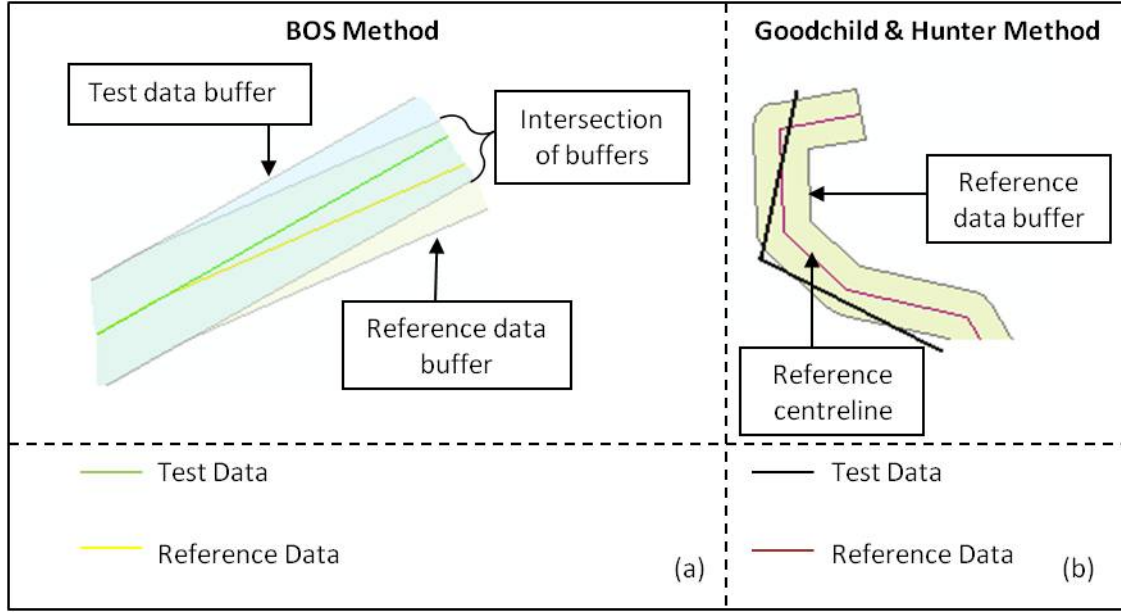


Figure 2.3: (a) depicts the BOS method and (b) the Goodchild and Hunter (1997) method

Advantages and Disadvantages of Buffer Techniques

The latter method assumes that corresponding line features are present in both data sets (Tveite and Langaas, 1999). The BOS method does not make this assumption and therefore allows for the computation of the completeness and removal of miscodings to be done before the displacement is calculated (Tveite and Langaas, 1999).

The BOS method on the other hand, assumes that the line features in the test and reference data sets have a homogeneous accuracy across the test area. Tveite and Langaas (1999) did a practical assessment of their method and found that the bias behaved systematically. They state that for this reason, the bias can be identified. The method by Goodchild and Hunter (1997) makes no assumption about the positional accuracy of either data set. Another reason why the Goodchild and Hunter (1997) method is so widely used, is because it is insensitive to outliers.

Distance Techniques

Other, less popular ways of assessing the accuracy of linear features is the use the Hausdorff distance and the average distance. The Hausdorff distance gives the “maximum distance of a set to the nearest point in the other set” (Gregoire and Bouillot, 1998). The average distance provides the mean distance between two line features (Girres and Touya, 2010). The Hausdorff distance can also be used to compute the distance between two polygon features. The equations for the Hausdorff and average distances are given by:

1. Hausdorff Distance: $DH(A, B) = \text{Max}(\sup_{x \in A} d(x, B), \sup_{x \in B} d(x, A))$

- where the distance from A to B: $D_{A \rightarrow B} = (\sup_{x \in A} (\inf_{y \in B}) d(x, y))$
- where the distance from B to A: $D_{B \rightarrow A} = (\sup_{y \in B} (\inf_{x \in A}) d(x, y))$
- and: $DH = \text{Max}(D_{A \rightarrow B}, D_{B \rightarrow A})$, (see figure 2.4 (a))
(Hangouët, 1995)

2. Average Distance:

$$dM = \frac{S}{\frac{L1 + L2}{2}}$$

(McMaster, 1986) as cited in (Girres and Touya, 2010)

- where S represents the area between the two line features
- and L1 and L2 are the lengths of the features being compared (see figure 2.4 (b))

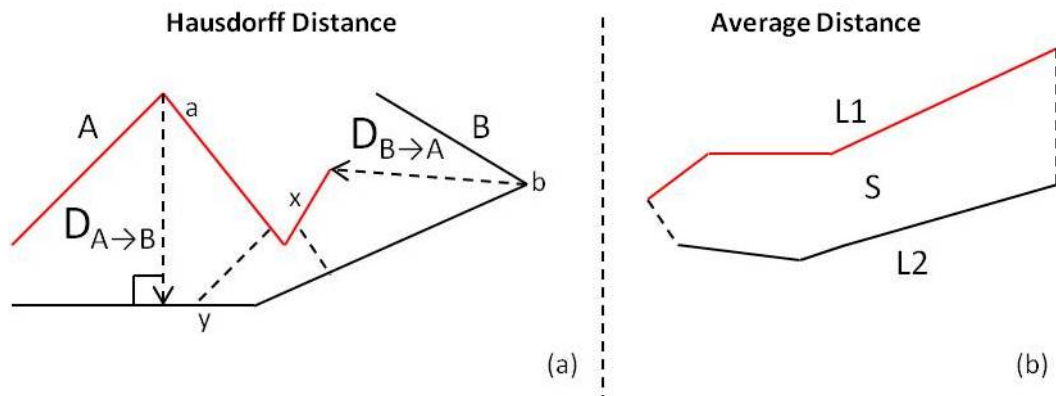


Figure 2.4: (a) shows the Hausdorff distance between line features A and B (Hangouët, 1995); (b) shows the average distance between two line features with lengths L1 and L2 (Girres and Touya, 2010)

Complications with the computing the Hausdorff distance arise when the distance is taken between any two points on the line features and not necessarily between the vertices (Hangouët, 1995). In general, the Hausdorff distance is computationally intensive (Rotter, Skulimowski and Kotropoulos, 2005). Computations with end of line vertices also make for an unstable result (Hangouët, 1995).

Other Techniques

Goodchild and Hunter (1997) mention another technique to measure the accuracy of linear features where the point topologies of two data sets are compared. Antoniou (2011) demonstrated this method of node matching. A node topology was built in the ArcMap environment and the nodes from two data sets were compared based on their positions and street names (Antoniou, 2011). The results showed that 60% of the nodes matched (Antoniou, 2011).

This method may work well for rectilinear roads with the well-defined nodes, but presents a problem for curved roads (Goodchild and Hunter, 1997). The number of nodes and the position of them will differ more for curved roads, making point matching a difficult task (Goodchild and Hunter, 1997).

2.4.2 Geometric Accuracy of Polygon Features

Less complicated methods have been used to assess the positional accuracy of polygon features. Referring specifically to OSM, not many researchers have focused on polygon data because OSM was primarily designed for road networks. Polygons from different sources may be compared in position and shape. The Hausdorff distance (see section 2.4.1) and surface distance may be used to compute the positional accuracy (Vauglin, 1997) as cited in (Girres and Touya, 2010). The two-dimensional surface distance between polygons is computed as a ratio of the intersection and union of them. The distance is expressed as a value between zero and one (Girres and Touya, 2010). Polygons that are closer to each other will have a value close to zero, while the polygons further away will result in a surface distance close to one (Girres and Touya, 2010). The equation for the surface distance is given by:

1. Surface Distance: $dS = 1 - \frac{S(A \cap B)}{S(A \cup B)}$

–where A and B are two different polygons
(Vauglin, 1997) as cited in (Girres and Touya, 2010)

Shape descriptors use the geometric dimensions of the shapes being compared. Examples of shape descriptors include: elongation, centre of gravity, solidity, compactness etc. (Mingqiang, Kidiyo and Joseph, 2008; Al-Bakri and Fairbairn, 2011).

Compactness

Compactness is a region-based shape descriptor that measures how much the shape of a polygon varies from a predefined shape; for example a circle, which is the most compact shape (Lee and Wan, 2004; Maceachren, 1985). Many methods for measuring compactness have been investigated and introduced (Maceachren, 1985; Lee and Wan, 2004; Li, Goodchild and Church, 2013). The compactness computed in terms of the ratio of the polygon area and perimeter will be discussed here. This method is the most common method and is suitable for regular shapes (Maceachren, 1985). The equation for compactness is given by:

1. Compactness: $C = \frac{\text{Area}}{(0.282 \times \text{perimeter})^2}$

Because the square of the perimeter is taken, the compactness result will not vary as the size of the polygon varies (Maceachren, 1985).

Elongation

Elongation is the ratio of the width and length of the smallest rectangle containing every point making up the polygon (Mingqiang *et al.*, 2008). This is referred to as the minimum bounding rectangle (see figure 2.5).

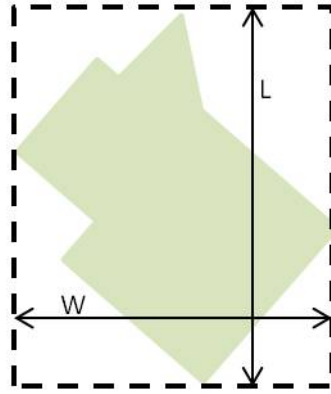


Figure 2.5: Example of the minimum bounding rectangle around a polygon

2.4.3 Previous Investigations into the Positional Accuracy of OSM Roads Using the Goodchild and Hunter (1997) Method

Three dissertation studies were done in 2009 by Ather (2009), Kounadi (2009) and Sabone (2009), where the authors investigated the positional accuracy of the OSM road data by comparing it to authoritative data. The assumption in all three studies is that the authoritative data sets are of better quality. In all three investigations, the method by Goodchild and Hunter (1997) was used to compute the percentage of OSM line features that are within the reference data buffer. Sabone (2009) covered a much smaller area compared to Ather (2009) and Kounadi (2009). The exact size is not mentioned, but rather a schematic is presented outlining the test area, which is a subset of Fredericton City in Canada. Kounadi (2009) covered an area of 25 km² in Athens, Greece and Ather (2009) four times this area, located across London. These differences in the geographical extent and location contribute to varied results. Furthermore, the method used and the execution of the method will also contribute to the results.

Although these studies employed the same method, there were differences in the preparation of the data and execution of the method. Ather (2009) generated a 1 m buffer for the OSM data set allowing for comparison to the Haklay (2010) 2009 study. A buffer was not generated for the OSM data sets in the other two investigations. Different buffer sizes were generated for the reference data sets, depending on the documented positional accuracy and the road width. Kounadi (2009) states that because roads of greater width have wider buffers, the margin for error is increased. Other preparations of the data involved matching streets between the test and reference data sets by their names to ensure correct matching of line features (Ather, 2009; Kounadi, 2009). In addition to name matching, Kounadi (2009) performed a splitting process for OSM roads that were digitised and named incorrectly.

The method was executed, either using separate tiles and the results summed (Ather, 2009; Zielstra and Zipf, 2010a) or as a single data set (Sabone, 2009). Different GIS packages will use different clipping algorithms and can contribute to the final result (Haklay and Ellul, 2010).

The results from the Sabone (2009) study yielded a 94.04% average overlap for a 10 m buffer width. Kounadi (2009) used 7.5 m, 5 m and 4 m buffer widths for three different

road classes and obtained a combined average overlap of 89.54%. Ather (2009) used two buffer widths for the two road classes examined, 3.75 m and 5.6 m. The results for the two road classes were combined and an average computed for each of the four test areas. The average overlap percentages were 80.80%, 81.03%, 85.19 and 85.80%, respectively. Compared to Haklay (2010) who obtained an average overlap of 80% for with a 20 m buffer width.

2.4.4 Semantic Accuracy

The semantic accuracy, which is the measure of correct feature classification, may be found by measuring the semantic relatedness or semantic similarity between concepts from different data sets. Semantic similarity describes how much one word resembles another, while semantic relatedness includes a wider range of relationships between words (Patwardhan *et al.*, 2003). Previous researchers have used the WordNet::Similarity software to automatically assess the semantic accuracy of a data set (Al-Bakri and Fairbairn, 2011; Patwardhan *et al.*, 2003; Pedersen, Patwardhan and Michelizzi, 2004). WordNet is an on-line lexical database containing four parts of speech (noun, verb, adjective and adverbs) where related words are organised into synsets (or synonym sets) (Patwardhan *et al.*, 2003). The investigation by Al-Bakri and Fairbairn (2011) compared Ordnance Survey (OS) semantics to OSM semantics to determine whether there was any similarity. The authors realised that even before the similarity function could be run in Word::Similarity, some manipulation of the OSM data was necessary. For example, where a term consisted of more than one word, the root was used to replace the original term. Even so, the results were negative; the semantics from the two data sets were found to be dissimilar.

Baglatzi, Kokla and Kavouras (2012) argue that the freedom allowed in OSM tagging results in semantic interoperability problems. They proposed aligning OSM tags with a structured ontology, namely, the Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE). They introduce a questionnaire that helps contributors to distinguish between spatial features and thereby ensuring the correct tag value is added according to the DOLCE.

Girres and Touya (2010) followed a simple method of determining the semantic accuracy by computing the percentage of correctly matched road classes between OSM and BD TOPO. The results of their investigation showed that in most cases contributors will class roads incorrectly due to underestimation.

2.4.5 Completeness Assessments

Girres and Touya (2010) define completeness as, “ the absence of data (omission) and the presence of excess data”. A completeness measure in terms of VGI demonstrates its usefulness as contributed by non-professionals (Haklay, 2010). Haklay (2010) was the first to produce statistics on the completeness of OSM data against an official data set, the Ordnance Survey (OS) Meridian 2 data set. The OSM data set was extracted in August of 2008. Completeness was measured by dividing the total length of OSM data by the total length of the Meridian 2 data within a given extent (Haklay, 2010). The author measured completeness for the whole of England. The OSM data set was found

to be 69% complete compared to the Meridian 2 data set. In other words, the Meridian 2 data set had 31% more roads in 2008 than the OSM data set. Haklay (2010) found that the amount of contributions made to OSM is influenced by the affluence of an area and the availability of imagery.

Haklay and Ellul (2010) extended the Haklay (2010) completeness study by using three OSM data sets instead of one, for the period 2008 to 2009, covering England. The aim of the study was to determine how population density, geographic location and socio-economics affect OSM contributions, respectively. This was achieved by incorporating other geographical data sets, such as the Index of Deprivation data into the completeness computations for different areas. The first completeness measure was done for all OSM roads within the selected area. The second completeness measure only included those OSM road classes that could be matched to the Meridian 2 road classes. What this means is that only OSM road features with attribute information about the road type (i.e. tagged roads) were used.

The results of Haklay and Ellul’s (2010) study show that in areas where the population densities are higher, the attribute information for features is more complete. OS provide higher detail compared to OSM, but overall the OSM repository is growing much faster (Haklay and Ellul, 2010). For the first completeness measure, the first OSM data set was 65% complete; while the second and third data sets exceeded the total length of the reference data set, resulting in values of 102% and 124%. When the completeness values exceed 100%, perhaps it is more correct to say that (in this instance) the total OSM road lengths are 2% and 24% longer than the reference data sets, respectively. The results from these two completeness measures within this study were vastly different, with the completeness values for tagged roads being much lower. What was further revealed in this study regarding the relationship between the affluence of an area and the number of contributions, was that there is an even bigger gap between areas of varying affluence when the attribute information of features were taken into account.

Zielstra and Zipf (2010a) performed a comparative study between the OSM and TeleAtlas MultiNet data sets for a number of cities in Germany. The TeleAtlas data set is not an authoritative data set, but it is professionally produced with an accuracy of up to 1 m (Zielstra and Zipf, 2010a). The investigations were completed on three data sets for 2009. The overall completeness measure showed that the total road length for the OSM data set was less than the TeleAtlas data set for all three test periods. The OSM data set went from being 71% to 82% and finally to 93% complete in April, July and December of 2009, respectively. Although the OSM data is less than the TeleAtlas data in all three instances, the OSM growth rate is higher.

A completeness study was also undertaken in France by Girres and Touya (2010) by comparing the French OSM data to BD TOPO data from Institut Géographique National (IGN). The study area covered the whole of France. The overall average completeness measure was 37%.

The results from the four different studies show that because the OSM repository is growing at a fast pace, the completeness measure is specific to the date of data extraction. Also, the term “completeness” becomes futile when the test data exceeds the reference data, especially since the explicit assumption is made that the reference data

set is of better quality.

As stated before, the location and extent of the data sets, also influence the results. OpenStreetMap started in London, England (Haklay, 2010), thus the assumption is that the number of contributions would be higher for England in earlier years when OSM was first started, compared to the rest of the world. The assumption appears to be true when comparing completeness values across studies for data sets extracted around the same date. Haklay and Ellul (2010) provides a completeness measure of 102% and 124% for two different time periods, compared to (Zielstra and Zipf, 2010b) who obtained 71% and 93% for similar time periods.

The number of contributions differ between countries, but also within a country. Haklay and Ellul (2010) and Zielstra and Zipf (2010a) highlighted the heterogeneity of contributions between rural and urban areas. Girres and Touya (2010) noted that there is an overall heterogeneity in the OSM data due to heterogeneous contributors. The same study also found that the number of contributions made to OSM is dependent on the density of contributors within a given area.

2.5 Analysis and Discussion

Globally, there is an increasing demand for current spatial data. In the developing parts of the world, there is already a shortage of digital spatial data. Therefore, the increasing demand for spatial data is placing pressure on NMAs. VGI presents itself as a possible solution to the problem. Possible decreased mapping costs and involvement from the community in decision-making processes are but two of the advantages presented by VGI. VGI also has disadvantages associated with it, the first of these being contributions made with malicious intent. Malicious content does however not seem to be a great obstacle at this point in time. Secondly, the aspect that has gained the most interest is the quality of VGI. In terms of VGI quality investigations, OSM has been tested the most.

The previous studies on the quality of OSM data provide a good basis to work from. The OSM repository is growing at such a fast rate that even one year difference between studies yield significantly different results. The four factors which influence the final results are: i) location ii) extent iii) the study period and iv) the methodology used.

The success of VGI initiatives, like OSM is as a result of various factors. The most obvious is the advancement in technology like Web 2.0, commercially available GPS devices, GPS-enabled cellphones and GIS freeware. Positive user motivations have also proved to be advantageous, as most users contribute data altruistically. On the other hand, technological advancement has not benefited everyone. The digital divide separates those who have access to technology from those who do not. Moreover, because VGI is technology-driven, those who are disadvantaged cannot be part of the phenomenon. Associated with this is cultural differences and interests amongst volunteers. For example, communities who are not exposed to VGI may not find the act of contributing data to be important.

The South African SDI Act does not make provision for VGI and full implementation of the SASDI is still a long way off, which means that the opportunity for integrating VGI

is still a long way off. The CD: NGI on the other hand, is free to make use of spatial data produced by volunteers in the mean time. Overall, integrating VGI into authoritative data is a complicated process, but as was discussed in this chapter, others have managed to do so. A good understanding of the VGI data structure and a well-defined integration workflow could make for a successful collaboration.

Chapter 3

CHIEF DIRECTORATE: NATIONAL GEO-SPATIAL INFORMATION AND OPENSTREETMAP SPATIAL DATA MODELS

3.1 Introduction

In order to integrate data from different sources, it is necessary to have an understanding of their data models. This includes how the data is acquired, stored and the standards and policies governing the data.

In a NMA context, spatial data models are generally governed by spatial data standards and policies. The purpose of standards is to minimise the type of errors and uncertainties inherent in spatial data (Elwood *et al.*, 2012). The CD: NGI is working toward spatial data standards that better suit the requirements of the SASDI. For the CD: NGI to find the balance between establishing spatial data standards and addressing the VGI data model, which functions without spatial data standards is a difficult task. It is necessary to understand the data models of each of these in order to determine the degree to which the data structures differ and hence the feasibility of integrating the two data sets.

3.2 Spatial Data Standards

3.2.1 Spatial Data Standards in South Africa

The International Organization for Standardization: Technical Committee 211 (ISO/TC 211), founded in 1947, is responsible for developing and publishing international standards pertaining to geographic information or geomatics (Coetzee, Cooper and Strydom, 2007). South Africa, represented by the South African Bureau of Standards (SABS) is one of the 35 participating members of ISO/TC 211 (*ISO/TC 211 Presentations.*, 2012). The SABS is a government mandated statutory body established by the Standards Act, 2008 (Act No. 8 of 2008) (Government Gazette, 2008). When a new

standard has been developed and sent out for review, elected members of the SABS may participate in the reviewing and voting processes.

ISO/TC 211 relies on technical and sub-committees for the creation, management and approval of all geographical information standards. Sub Committee 71 E (SC 71 E) Geographic Information represents the SABS as the body responsible for establishing geographic information standards in SA (Coetzee *et al.*, 2007). The committee either adopts the standard from ISO/TC 211 as is, or uses the existing international standard to develop a standard that is applicable to South Africa (Coetzee *et al.*, 2007).

South African governmental departments, including the CD: NGI, who provide and maintain spatial data are not obligated to adhere to international standards or standards with a South African profile. At the time of writing, no data custodians had been officially appointed. The result is incoherent, duplicated spatial data sets between the various departments.

3.2.2 Duties of the Committee for Spatial Information

The purpose of the South African Spatial Data Infrastructure (SASDI) is to establish integrated and uniform spatial data sets across various data custodians in South Africa. The CSI as the representing body for the SASDI has amongst others, the following duties in terms of SDI Act (Government Gazette, 2004):

- Defining the base data set (Committee for Spatial Information, 2012)
- Identifying and appointing data custodians, where data custodians are organs of State (Committee for Spatial Information, 2012)
- Ensuring that all data custodians provide base data sets that adhere to the specified spatial data standards and regulations concerned with the “functions of data capture, validation, maintenance, management, archiving and documenting” (Committee for Spatial Information, 2012)

An organ of state may only be identified as a data custodian upon receiving mandated responsibility, i.e. a request by the CSI. They must have the necessary infrastructure and resources to provide the base data set. Spatial data standards are used to set out the requirements of the base data set.

The SASDI will force organisations to comply with CSI specified standards. Should the CSI successfully implement their duties, the SASDI will result in great benefits. Probably, the most important benefit is reduced production costs due to data sharing amongst government organisations. Governmental agencies are currently looking to the map users as a source of funds (Goodchild, 2007). Compliance to data standards as defined by the CSI includes the generation of metadata, which leads to another important benefit and that is reduced data custodian liability.

3.2.3 CD: NGI Spatial Data Standards - Contributions from External Standards

The South African National Standards 1880 (SANS 1880): South African Geospatial Data Dictionary (SAGDaD) and its application is one of the standards that was adapted to create a profile for South Africa, based on the ISO 19110: 2005 standard (Coetzee *et al.*, 2007). The SAGDaD “provides a list of feature types, and definitions, to which a unique code has been allocated” (Vorster, 2009).

The CD: NGI does not directly comply with the SANS 1880, but it was used during the design phase of the iTIS database model because of the comprehensive feature classification it presents. The CD: NGI did however implement its own feature classification, but the description of each feature was matched to the corresponding feature description contained in the SAGDaD. A feature reference code was used to relate the descriptions from the two sources.

There are many reasons why the CD: NGI chose not to comply with international standards and even those standards created for the South African profile. These reasons include: i) in-house standards are specific to the products provided by the CD: NGI; ii) limited existing standards necessary for other spatial data production processes (e.g. the orthorectification process) employed by the CD: NGI; iii) the high costs of complying with a standard and iv) the lengthy approval processes of standards.

3.2.4 The CD: NGI Internal Spatial Data Standards

In earlier years, guidelines and regulations for the creation and maintenance of geographic data were developed in-house at the CD: NGI. In later years, the Quality Assurance (QA) division was established and they are responsible for the creation and maintenance of spatial data standards at the CD: NGI. The existing documentation was used for the production of standards, but many of the principles were already outdated. The creation of standards thus provided an opportunity to review current and future geographical data production within the organisation.

When a new standard is created, the QA division collaborates with the respective divisions involved in that specific spatial data process. Upon completion of the draft document, a copy is sent to the entire organisation for review. A standard is approved when all commentary has been dealt with and the necessary amendments have been made.

3.3 CD: NGI Topographical Data Structure

3.3.1 Topographic Feature Compilation

The overview of the topographic feature compilation process is shown in appendix A1. The process includes four main stages: identifying production requirements, preparing for topographic compilation, performing the compilation and finalising the compilation process. The details of the third and fourth stages are provided in appendices B1 and C1, respectively. Various inputs are required before features may be compiled (see appendix

B1). The purpose is to ensure that the compilation is done methodically, according to a pre-defined schedule, avoiding duplication of work. Upon completion of the compilation stage, the compilation appraisal is performed and the data is committed to the iTIS (see appendix C1).

3.3.2 Standards Governing Topographic Feature Compilation

There are currently eighteen spatial data standards governing the various spatial data production processes at the CD: NGI. Most of the standards have been approved while a few are still in the draft stage. As the CD: NGI is still in the process of migrating to the iTIS, the standard governing the operation of the iTIS is still in the developmental stage. Instead of using the incomplete iTIS standard, the approved standard for Capture of Topographic Data will be used in this investigation. This is the most applicable standard for comparing the quality of authoritative and volunteer data because it provides the rules and accuracies for capturing topographical features, as they will be stored within the iTIS. This standard was approved in March 2012 and has since undergone revision. The final approval was given on 18 March 2013. The standard is comprised of three main documents namely: Topological Integrity Rules, Capture of Topographical Data Standard and Data Structure and Attribute Describes.

1. Topological Integrity Rules —provides the topology guidelines i.e. connectivity and adjacency or co-incidence for compiling line and area features. In a GIS context topology can be defined as the rules, which govern how vector features are related to each other (Theobald, 2001). Examples (from section C3, Annexure B-Topographical Data Capture Guidelines, subsection C 3.5 Feature Class: Trans Roads) of the type of topological rules are given below:
 - (a) Roads are drawn with the minimum number of vertices. The aim is to assign vertices only at bend points.
 - (b) Individual roads are captured with a single line segment, whether multi-or-single lane.
 - (c) Tracks, Footpaths, Hiking Trails and other Roads must have coincident vertices with connecting roads.
 - (d) All roads except Tracks, Footpaths, Hiking Trails and other Roads must be split at connecting vertices.
2. Capture of Topographical Data Standard —the requirements for the capture of topographical data. The specification for each feature is provided in terms of the: (i) feature type, (ii) feature ID, (iii) SANS 1880 ID Number, (iv) definition of feature, (v) data capture specifications, (vi) geometry type and (vii) the Minimum Data Capture Unit (M.D.C.U). The following statements are specified regarding the positional and semantic accuracy:

- (a) Features captured by photogrammetric methods must have a positional accuracy not exceeding 10 metres at the 95% confidence level.
- (b) Features vectorised from topographic maps must have a positional accuracy not exceeding 40 metres (0.8 mm at the map scale) at the 95% confidence level. These refer to rivers and contours.
- (c) Features shall be correctly classified at the 90% confidence interval

The CD: NGI follows the traditional top-down feature classification. The standard document specifies the various super classes, sub-classes and features. Figure 3.1 is a graphical representation of the hierarchical structure whereby topographical features are stored. There are 11 super classes (red border), 42 feature classes (green border) and 402 feature types (orange border). Feature types may also have subtypes, but this is not shown in the diagram. Due to the size of the entire model, the diagram only shows an extract of the classification.

3. Data Structure and Attribute Describes —provides the domain or description for each of the attribute values mentioned in the Topological Integrity Rules previously. It also separates the mandatory features to be captured from the features captured by other data custodians and features, which are no longer compiled. Currently, each topographical feature is stored with a unique geographical ID (GID), which does not change for the life of the feature.

3.3.3 CD: NGI Quality Control

Ensuring high quality spatial mapping is vital for NMAs; therefore, quality control processes are put in place at the critical stages of the map production process. Appendix B1 shows that the topographic compiler performs two checks on their own work and then a senior employee who ensures compliance to standards does the final quality control. All the quality checks are performed by visual inspection, which means that no numeric indicator of quality is generated.

3.3.4 CD: NGI Quality Topology

As the features are being captured, the operator complies with the integrity rules discussed in section 3.3.2. Once the data is committed to the iTIS, final checks are performed to ensure data integrity. An automatic topology check is run. The errors are corrected either automatically or manually depending on the type of error.

3.3.5 Distribution of the CD: NGI Data

The CD: NGI distributes spatial data under the Copyright Act, No. 98 of 1978. Clients may use the data as they desire, but they may not sell it or any derivative of it (Government Gazette, 1978). The clause that is included on all map products states; “this map may not be reproduced by any means, including digital, without prior permission.” Data vendors may however apply to the CD: NGI for copyrights. After an

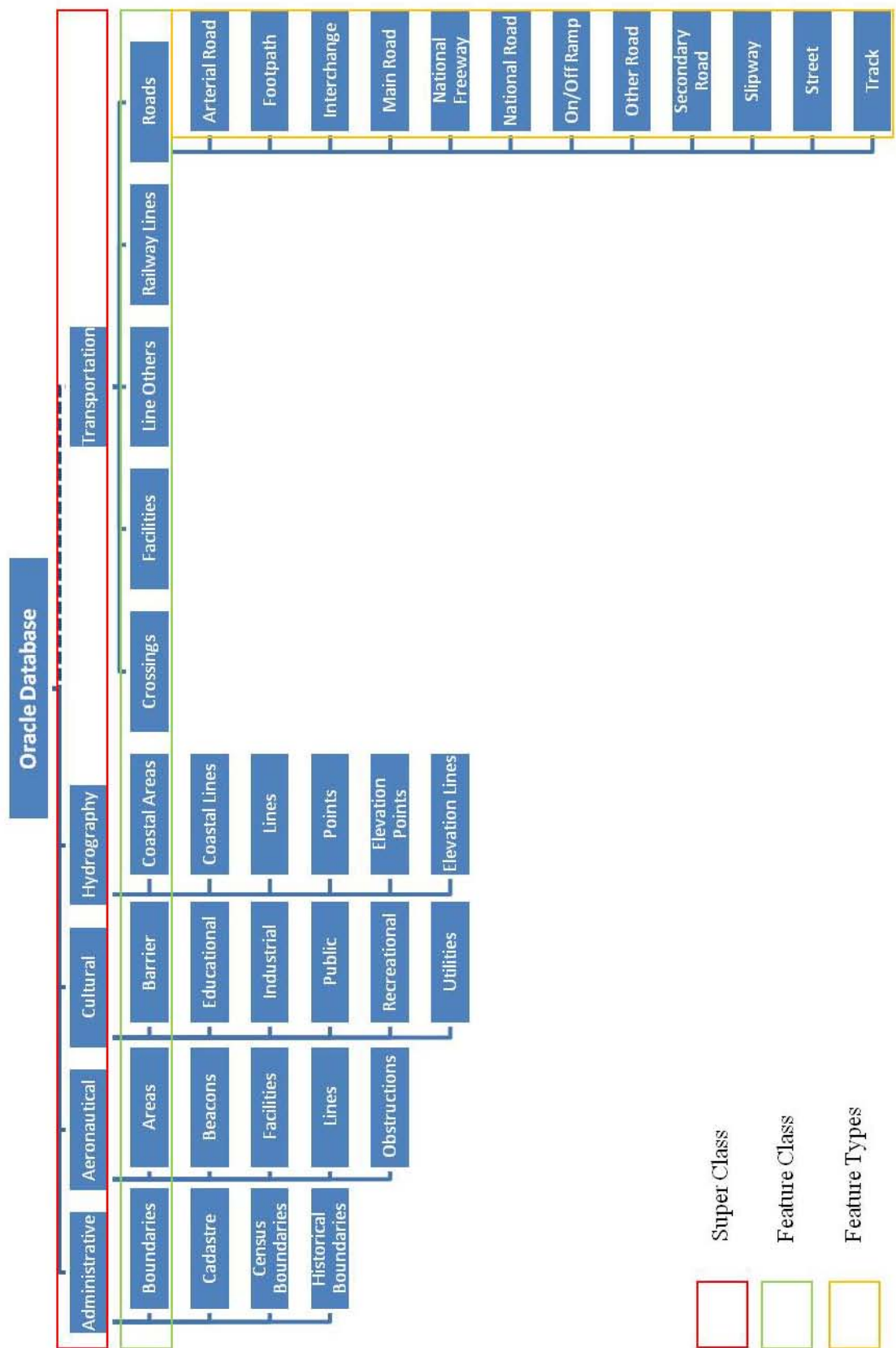


Figure 3.1: Extract of CD: NGI feature classification

agreement between the two parties has been reached, the vendor may sell the CD: NGI products or any derivative thereof.

3.4 OpenStreetMap Data Structure

3.4.1 OpenStreetMap Data Elements

OSM does not follow the traditional point, line and polygon convention instead, OSM map features are represented by: nodes, ways and tags. The most recently added data element is the relation. Table 3.1 shows graphical examples of each data element.

Tags

Tags are string values stored as keys and values (Ramm *et al.*, 2011, :54), where a key can be seen as the feature class (e.g. Tourism) and value as the attribute (e.g. Hotel) (Behrens, 2011). OSM does make available a list of preferred tags, but essentially a tag can be assigned any value (*OpenStreetMap Main Page.*, 2013). Ramm *et al.* (2011, : 61) states that this has been one of the main reasons for the success of OSM. For the purpose of obtaining a more coherent map, users may propose a “standard tag” and users are encouraged to participate in a voting process to render the tag “official” (*OpenStreetMap Main Page.*, 2013). This is not a popular process as most users simply create their own tags.

The purpose of a tag is to add descriptions to topographic features. More than one tag may be assigned to a feature (Ramm *et al.*, 2011, :54), for example, connecting roads classed under different road categories may be drawn as a single feature resulting in more than one tag.

Nodes and Ways

Nodes can be stand-alone point features denoting Points of Interest (POI) like hospitals and restaurants (Ramm *et al.*, 2011, :53). Nodes can also constitute ways in the same way that vertices constitute traditional line segments with the added difference that the node may have a tag (Ramm *et al.*, 2011, :52). Nodes may also be a member of a relation (Ramm *et al.*, 2011, : 54).




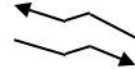
Ways are used to represent line features like roads and boundaries (Ramm *et al.*, 2011, :53). A way can have three states: open, closed or filled. Open ways are used to represent linear features (*OpenStreetMap Main Page.*, 2013). Closed ways, are ways that close on the first node and can be seen as a polygon feature, representing features like buildings (Ramm *et al.*, 2011, :53). Filled closed ways are areas used to represent features like land-use areas (*OpenStreetMap Main Page.*, 2013).

Relations

Relations are the latest addition to the OSM data model and are used to join multipart features (Ramm *et al.*, 2011, : 54). An example is a road comprised of two uncon-

nected ways. The ways are members of the relation. The members may each have a role, where the role is given by text which adds further description about the relation (*OpenStreetMap Main Page.*, 2013). In this example the role of one way may be set to inner, while the other is set to outer describing the directions of traffic flow on a multi-lane road (*OpenStreetMap Main Page.*, 2013). There are different types of relations, depending on the type of features it connects (*OpenStreetMap Main Page.*, 2013). Ramm *et al.* (2011, : 54) state that the OSM database does not read in relations, as these special connections, but rather it ensures that the rules used to connect features maintain their integrity.

Table 3.1: Description of OSM data elements (*OpenStreetMap Main Page.*, 2013; Ramm *et al.*, 2011, :52-55)

OSM Data Elements			
Element	Description	Diagram	Example
Node	represents a single point as a Points of Interest (POI) or constitutes a way; may have a tag; may be a member of a relation	•	Stand-alone Hospital
Way	represents lines or polygons; limited to 2000 nodes per way; can be open or closed way or a closed, filled way; may be a member of a relation		
	open ways are lines that do not close on the first node		Street
	closed ways are lines that end on the first node		Building
	an area is a filled closed way		Land-use area
Tag	represents feature names or descriptions; stored with a key and value ; stores in unicode strings		Tourism (<i>key</i>) = Ho- tel (<i>value</i>)
Relation	describes the relationship between geographical elements; consists of at least one member /tag; nodes and ways may have roles		Joined Roads

3.4.2 OpenStreetMap Data Model Components

The diagram in appendix D1 shows the OSM component overview. All data is stored in the main PostgreSQL database. Users access the database and contribute data via the API. Potlatch2 is the on-line editor used for GPS traces. Merkaator and JOSM are the two off-line editors used for other map editing tasks. The OpenLayers (slippy map) is the map the user sees when they open the OSM homepage. As the user switches

between vector and raster layers, the map layers are displayed using tiles. Tiles divide the slippy map into pre-defined, lower resolution squares that are saved on disk. The tiles are rendered via the “tile server”. The software used to render the tiles is Mapnik.

The OSM features are stored with the classification depicted in figure 3.2. The classification resembles the CD: NGI hierarchical classification in figure 3.1. The top level consists of the key and the next level consists of the value as discussed in section 3.4.1. Although there are primary keys, the total number of keys and values are undefined as users may create new ones, as they require. Unlike the CD: NGI, the OSM classification does not have a third and fourth level.

3.4.3 OpenStreetMap Quality Control

For some time the OSM database administrators performed certain quality checks, but for most parts volunteers were responsible for detecting erroneous or malicious contributions (Keler, Trame and Kauppinen, 2011). Currently, there are various tools available to perform quality control. Volunteers, together with the database administrators have developed various tools for reporting bugs and detecting errors. There are three tools available for reporting bugs: Notes, Openstreetbugs and MapDust (*OpenStreetMap Main Page.*, 2013). The error detecting tools are comprised of a variety of checks. These include Waycheck, which is used to identify open ends and crossings; SomeChecks, which verifies one-ways, rondabouts, double-nodes and areas and NoName Map, which identifies missing road names (*OpenStreetMap Main Page.*, 2013). There are many other tools, but some are not available in South Africa. Many of the tools may not have existed at the time when the data sets used for this investigation were extracted. In a recent investigation Elwood *et al.* (2012) stated that VGI initiatives have not established methods to enhance the level of trust of their data.

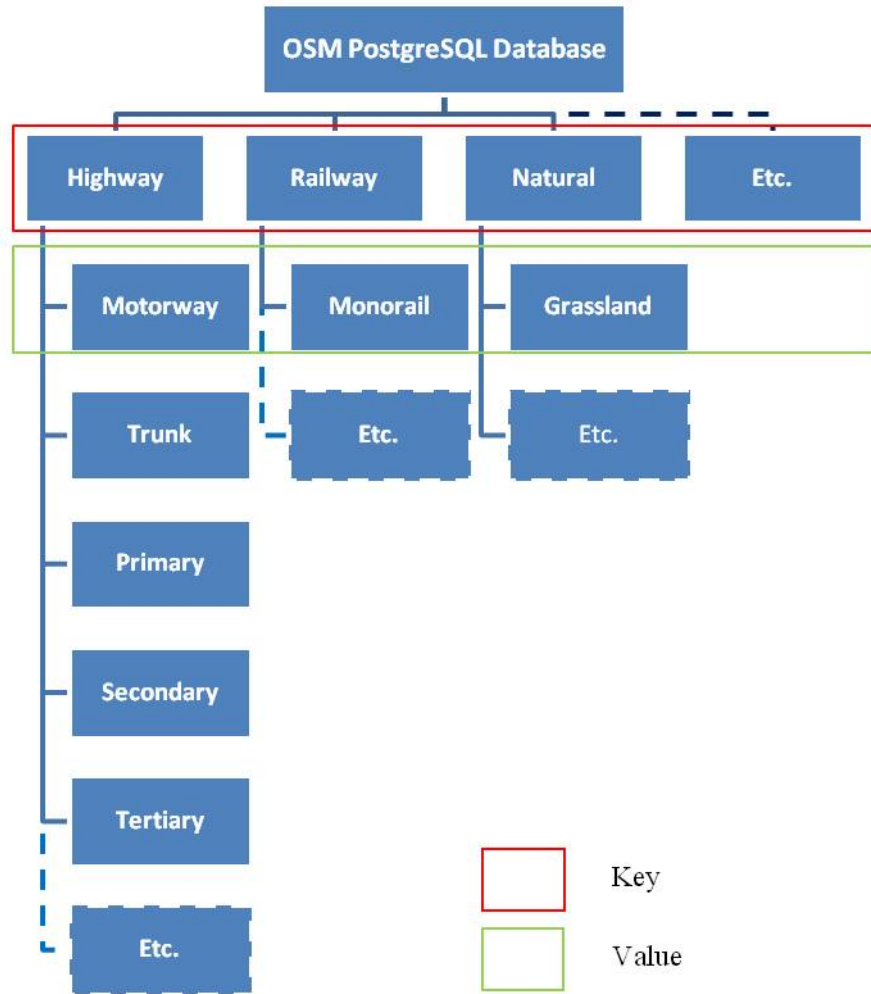


Figure 3.2: OSM feature classification

3.4.4 OpenStreetMap Topology

Ubeda and Egenhofer (1997) state that a database lacking topological integrity has a weak data structure. OSM does not enforce any topological rules, which may result in various topological errors. Every piece of information contributed is deemed correct until someone detects an error (Keler *et al.*, 2011). In 2008 Schmitz, Zipf and Neis (2008) investigated the OSM routing capabilities for three data sets. They found that the existence of topological errors resulted in failed routing requests. However, the number of routing failures did decrease over time. This may be an indication that the new error detection tools are increasing the OSM topological integrity. For this investigation, a topology test was performed for the commercial data set in Western Cape. Two examples of topological errors that were detected in the OSM data sets are overlaps and dangles.

Overlaps

A topology rule was used in ArcMap to find all the overlapping line features. There were many instances of overlapping features as can be in figure 3.3. The red lines in

figure 3.3 represent overlap errors. The error inspector revealed 445 overlap errors for the commercial data set. However, at a later stage it was discovered that the Esri shape-file format reads in relations (see section 3.4.1) as a feature and creates a duplicate of the map feature it is related to. When these were removed, there were no overlapping features.



Figure 3.3: Overlap errors generated for the Western Cape line commercial data set using ArcMap topology rules

The same rule was used to check the polygons of the same test area for overlaps. The result of the check is shown in figure 3.4. The type of overlap errors shown in the figure is specific to the CD: NGI. OSM users choose to digitise individual buildings as well as the property boundary, resulting in these types of errors. The CD: NGI will digitise either the buildings or the boundary depending on the size of the buildings, but not both.

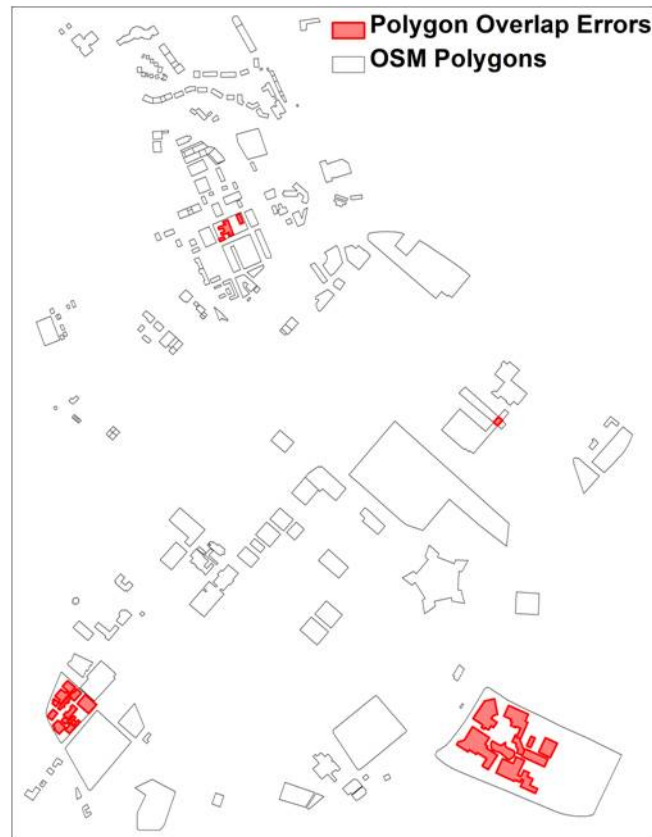


Figure 3.4: Overlap errors generated for the Western Cape polygon commercial data set using ArcMap topology rules

Dangles

Another topology rule was executed which checks the data for dangles (i.e. over-and-undershoots). A dangle can be defined “as the endpoints of lines that are not snapped to other lines in the feature class” (Environmental Systems Research Institute Inc., 2012). The results showed a few dangles, but not all dangles are because of incorrect digitising. In figure 3.5 three types of dangles are depicted. The blue lines represent the OSM road features for April 2012 and the orange lines represent the CD: NGI road features with a creation date August 2003. The first dangle shows incorrect digitising, as there is no road shown in the aerial image. For the second dangle, the entrance into a parking area is digitised as a road. Numbers three and four are not dangles, because they represent the end of the road (i.e. a cul de sac).



Figure 3.5: Examples of dangles generated for extract of Western Cape commercial line data set using ArcMap topology rules

3.4.5 Distribution of Data

Since September 2012, the OSM initiative operates under the Open Database License (ODbL) (*OpenStreetMap Foundation.*, 2013). This type of licensing states that the volunteer must be the owner of the contribution or that permission has been granted from the owner to contribute it. Users are allowed to use the data as they please, but the data may not be sold (Girres and Touya, 2010; *OpenStreetMap Main Page.*, 2013; *OpenStreetMap Foundation.*, 2013).

3.5 Analysis and Discussion

This chapter has shown that there are distinct differences between the data structures of authoritative and volunteer data in the way the data is acquired, stored and the policies and standards governing the data. The main differences are summarised in table 3.2. The CD: NGI follows the traditional top-down approach where it is the role of the expert to produce spatial data with specialised skills and specific software and non-experts are only the users of the data. In the case of OSM, as with most VGI initiatives, every citizen may be a producer of spatial data provided they have access to the necessary hardware and software.

Because the structures of the data elements between the two data sets are different, the data storage methods are also different. Fortunately, in the case of OSM, the data elements are convertible to a shapefile format, which mostly resembles that of the CD: NGI file structure. The OSM relation element causes some complication during the conversion process, but the GIS software allows this to be dealt with. It is therefore possible to compare the various topographical features between the data sets.

In terms of data usage, the OSM initiative operates under the ODbL, while CD: NGI distributes spatial data under the Copyrights Act (No. 98 of 1978). In both cases, the user may use and modify the data as they please, but they may not sell the data. The laws of the Copyright Act, although stricter, do make allowance for exceptions if the application for copyrights is found to be valid.

The comparison of the two data models has provided some insight into the level of complication for integrating volunteer and authoritative spatial data. The topology tests showed the lack of spatial data standards governing OSM data. As the NMA of South Africa, CD: NGI is continually aiming to add value to their products. Where value is measured by the usefulness the data offers to clients. The existence of topological errors results in erroneous information, which places a limit on spatial analyses resulting in de-valuation of the data.

Transforming the OSM data into the CD: NGI structure may prove to be a laborious task in an organisation where there is a staff shortage, which is currently the situation at CD: NGI. On the other hand, the benefit of gaining current spatial data at no cost may outweigh the cost of increasing the staff compliment in the end.

Table 3.2: Comparison of CD: NGI and OSM data models

Element	CD: NGI	OSM
Data Elements & Storage	Comprised of points, lines, polygons and attributes and stored in an Oracle 10g database.	Comprised of nodes, ways (open, closed, closed-filled), tags and relations and stored in a SQL database.
Acquisition & Compilation	Field and office annotation data, map deletion sheets, ortho-rectified images and compilation corrections are used during heads-up digitising from aerial imagery.	Contributions made via GPS track uploads, organisational mass uploads and heads-up digitising from aerial/satellite imagery.
Policies & Licensing	CD: NGI distributes spatial data under the Copyrights Act No. 98 of 1978. CD: NGI owns the data. Users may use and modify the data, but they may not sell the data. The laws of the Copyrights Act do make allowance for exceptions if the application for copyrights is found to be valid.	OSM operates under the ODbL. Volunteers own their contributions. Users may use and modify the data, but they may not sell the data.
Quality Control	The topographic compiler performs two checks on their own work and then the final quality control is done by a senior operator who ensures compliance to the standard for capture of topographic data.	Various tools are available for reporting bugs and detecting errors, automatically or manually (<i>OpenStreetMap Main Page.</i> , 2013).

Chapter 4

METHODOLOGY

4.1 Introduction

This chapter describes i) the methods used to determine the various quality elements and ii) the spatial analyses used to answer the qualitative research questions. Three quality measures were chosen: positional accuracy (geometric accuracy in the case of polygons), semantic accuracy and completeness. The CD: NGI considers these quantitative quality aspects most important in gaining an overall understanding of the quality. In addition, qualitative assessments were done in order to determine whether the OSM data is current and uniform across South Africa. The assessments undertaken are presented in table 4.1.

Section 4.4 describes the method for the quantitative assessments. It includes the method to determine the positional accuracy of OSM roads, the geometric accuracy of (polygon) buildings, the semantic accuracy of roads and the completeness of roads. These four assessments address the research question; does the OSM data meet the CD: NGI spatial data standards?

Sections 4.5 describes the method used for the qualitative assessments, that is the currency and uniformity in the acquisition of OSM data. The results will be used to answer the following research questions; i) is the data acquired evenly across the country?; ii) what is the rate of data generation and is it sufficient from a NMA perspective and iii) what is the most frequently generated data type and is this useful to the CD: NGI?

Table 4.1: Table showing the quality measures undertaken for each OSM data element

DATA ELEMENTS			ACCURACY TESTS				
OSM	Compared to	CD: NGI	Positional Accuracy	Semantic Accuracy	Completeness	Currency	Acquisition Uniformity
Node	Point				✓	✓
Open Way	Line	✓	✓	✓	✓	
Closed Way	Polygon	✓			✓	

4.2 Data

4.2.1 Data Sources

OSM data sets are freely available for download from third party websites.¹ The downloads are available in the raw OSM format (.osm) or it may be converted to shapefile format. The shapefile format was used for the OSM data in this investigation. The conversion process excludes certain feature categories. Therefore, data sets with the desired feature categories were requested directly from one of the OSM database administrators in the United Kingdom. The data was divided into point, line and polygon features and covered the period from October 2006 to April 2012. This provided a good study period to observe contribution trends across the country.

The reference vector and raster data were obtained from the CD: NGI. The vector data was in shapefile format. No feature manipulation was applied to the CD: NGI data because the purpose is to assess how much the test data (OSM) varies from the reference data set.

4.2.2 Co-ordinate System and Projection

The OSM data sets were provided on the Geographical Co-ordinate System, WGS84 reference ellipsoid. In order to overlay the OSM data onto the CD: NGI data, the OSM data was projected onto the Transverse Mercator projection type with reference to the respective central meridians.

4.2.3 Selection of Test Areas

Twenty-seven test sites were chosen throughout South Africa. For each of the nine provinces, three test areas were chosen. Each test area had to be representative of one of three settlement categories as defined by the CD: NGI, namely, residential land use (high urban density), residential land use (low urban density) and the combined land

¹www.cloudmade.com and www.geofabrik.com

use classes, commercial and industrial storage. For this study, they were named residential, low urban density and commercial respectively. Varying test areas will be used to determine how contributions differ between provinces and different settlement types within each province.

Residential Land Use - High urban density (called residential in this study)

“A residential land use (high urban density) is a built-up area where many buildings have been built close together, generally with spacing of less than 50 metres. Services like electricity, water and sewage disposal may be available, except in informal settlements.” (Chief Directorate: National Geo-Spatial Information, 2013*a*) (See figure 4.1)



Figure 4.1: Example of Residential land use - high urban density area

Residential Land Use - Low urban density (called low urban density in this study)

“A residential land use (low urban density) is a built up area, where buildings are close together, but not as close as a residential land use (high urban density). Services like electricity, water and sewage disposal may be available, except in informal settlements.” (Chief Directorate: National Geo-Spatial Information, 2013*a*) (See Figure 4.2)



Figure 4.2: Example of Residential land use - low urban density area

Commercial and Industrial Storage (called commercial in this study) The commercial class includes retail, financial institutions, restaurants and cafes, bars, taverns and night clubs, offices and informal trading. The industrial storage class includes light industries, heavy industrial, storage and wholesale distribution. (Chief Directorate: National Geo-Spatial Information, 2009) (See Figure 4.3)



Figure 4.3: Example of commercial and industrial storage area

The results from these test areas will be used to make generalised statements about the quality of OSM in South Africa. Analysis of the results will determine whether the inferences may be projected to a settlement type, province or the country.

The CD: NGI aerial imagery, which has a 0.5 m ground sample distance (GSD) in

conjunction with the definitions for settlement classifications, was used to discern the settlement categories when selecting the test sites. Settlement-patterns may vary significantly across an aerial image, therefore the size of the test areas were chosen in order to represent the settlement category appropriately. The size of each test area was $0^{\circ}01'30'' \times 0^{\circ}01'30''$. This equates to approximately 6.8 km^2 (i.e. the average area for the test areas). Because the test areas are spread across the country, the exact area will vary slightly for different latitudes. The bias in area was computed and applied to the test areas accordingly for the determination of the uniformity point acquisition only (see section 4.5.2).

4.3 Data Cleaning

4.3.1 Filtering

The OSM line vector data included all linear features such as roads, railway lines, some aero-way features (i.e. aircraft and air travel infrastructure (*OpenStreetMap Main Page.*, 2013)) and boundary lines. Only road features were assessed because it is considered the most important feature and the main focus of the OSM project (Ramm *et al.*, 2011, :64). Therefore, the highway key (or class) was filtered from the rest of the line features. Secondly, all highway features without a value (or attribute) were removed. The same process was applied to the polygon data sets, where only the buildings were extracted for comparison to the CD: NGI building (area) features. The OSM buildings were extracted because the CD: NGI classifies amenity buildings separately from other topographical features. Either the buildings are captured as polygons or points depending on its area as defined in the spatial data standards discussed in section 3.3.

Each OSM feature is assigned an OSM ID, which was used as the unique identifier during the processing phases. The OSM relation data element is not recognised in the ArcMap software. Thus, during the conversion from the .osm format to shapefile performed by the database administrator, the IDs for relations were converted to negative values (see section 3.4.4). Thus, the entire feature consisted of duplicate negative IDs representing the relations and the feature itself had a different unique ID. The negative IDs were filtered out and the unique feature ID maintained. An example is shown in figure 4.4, where the feature's unique ID is highlighted in yellow and five negative IDs representing the relations.

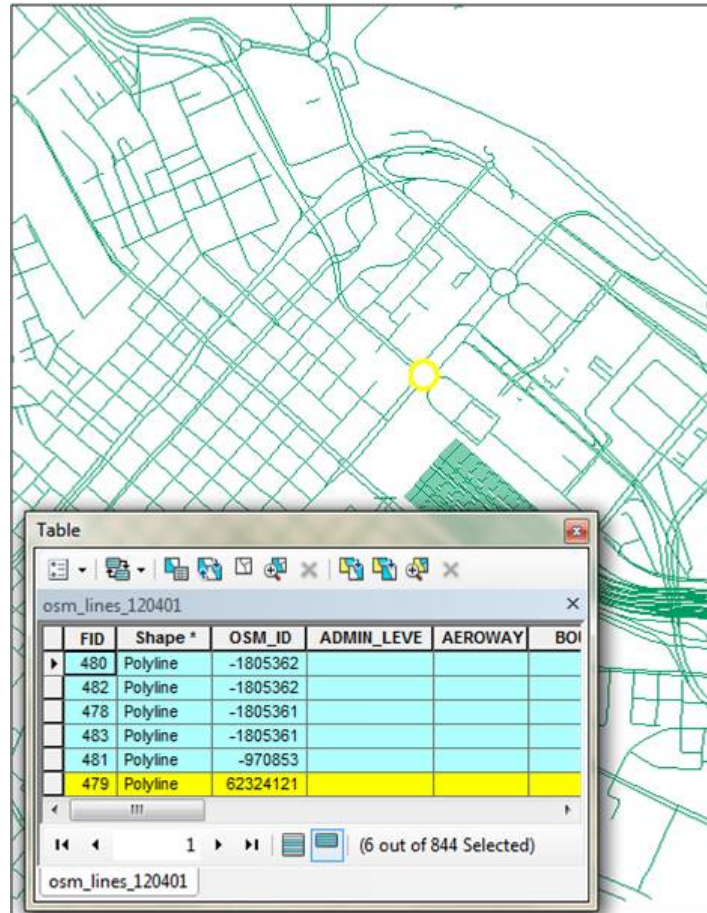


Figure 4.4: Example of features with multiple negative OSM IDs due to incorrect conversion from .osm to shapefile format

In terms of polygon features, all hole polygons were represented by negative IDs for the same reason. Hole polygons are similar to the example of the dual-lane road discussed in section 3.4.1. In this case however, there are two polygon boundaries —internal and external denoting the outer and inner walls of a building. Hole polygons were generalised by removing the internal boundary lines. The union and dissolve tools in ArcMap were used to generate simple polygons. Figure 4.5 (a) shows an example of a hole OSM polygon with a negative OSM ID overlaid onto the corresponding the CD: NGI polygon. Figure 4.5 (b) shows the same OSM polygon after generalisation was applied. This was necessary because the CD: NGI only generates the external wall boundary. This does place a limit on the positional accuracy of polygons.

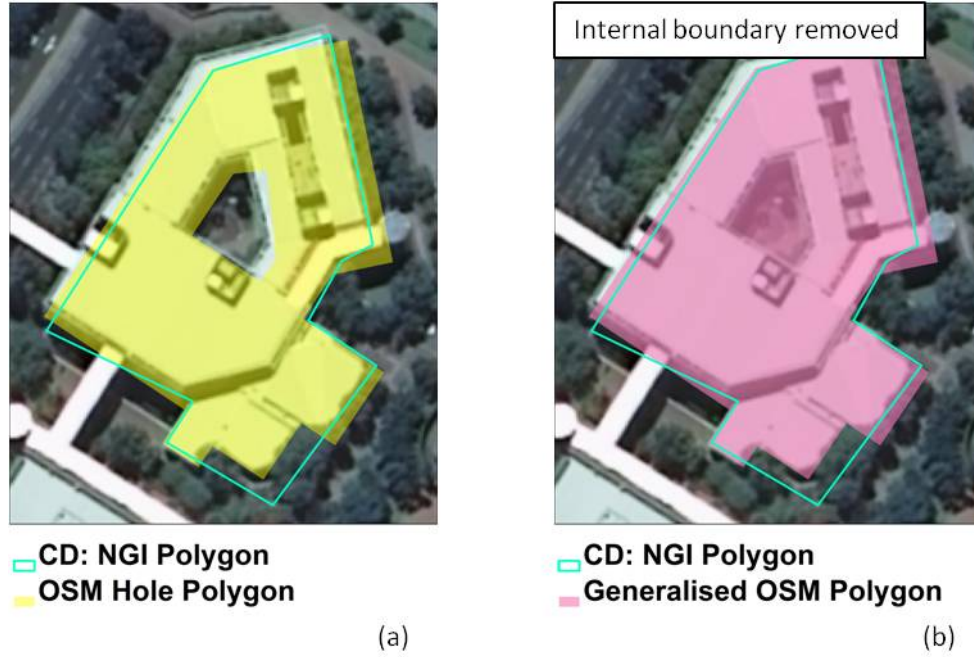


Figure 4.5: (a) shows an example of an OSM hole polygon with a negative OSM ID overlaid onto the corresponding CD: NGI polygon; (b) shows the same two polygons after removing the hole in the OSM polygon

4.4 Method for Quantitative Assessment

4.4.1 Positional Accuracy of Roads

Road Matching

Before any testing could be done, it was necessary to find corresponding roads between the CD: NGI and OSM data in order to eliminate any bias. Previous studies have used the road name to ensure that roads are correctly matched between the test and reference data (Ather, 2009; Kounadi, 2009) (see section 2.4.3). The CD: NGI does not capture road names; therefore, a different approach was required to match roads.

A 10 m buffer was generated for each of the CD: NGI road features. The buffer size was chosen based on the CD: NGI standard for capturing topographical features, which states that features captured by photogrammetric methods must have a positional accuracy not exceeding 10 metres at the 95% confidence level. This statement currently includes all roads and buildings. If an OSM road or part thereof was within any one of the buffers, it was considered a match. However, because of road intersections, some of the matches included parts of other roads.

Developing an automatic line-matching technique that would eliminate unwanted road sections is out of the scope of the project. Because of the size and number of data sets, it was also impractical to perform this task manually. A semi-automatic method was developed using the ArcMap geo-processing tools in Python scripting. The main stages of the process are listed below:

1. Separate each CD: NGI road feature according to its unique ID
2. Create 10 m buffers (B_w in figure 4.6) for each output feature in step 1
3. Clip the OSM data set to each output buffer in step 2
4. Compute the angle of intersection between the CD: NGI and OSM roads (angle a in figure 4.6). See appendix F1 for the Python script.
5. Remove all lengths shorter than $B_w(\cos(90-a))$ (see figure 4.6). See appendix F1 for the Python script.
6. Lines 2 and 3 in figure 4.6 will be removed
7. Split intersecting OSM roads, that is lines 1 and 4 in figure 4.6
8. Remove all lengths shorter than B_w
9. The shorter line segments of split lines 1 and 4 will be removed

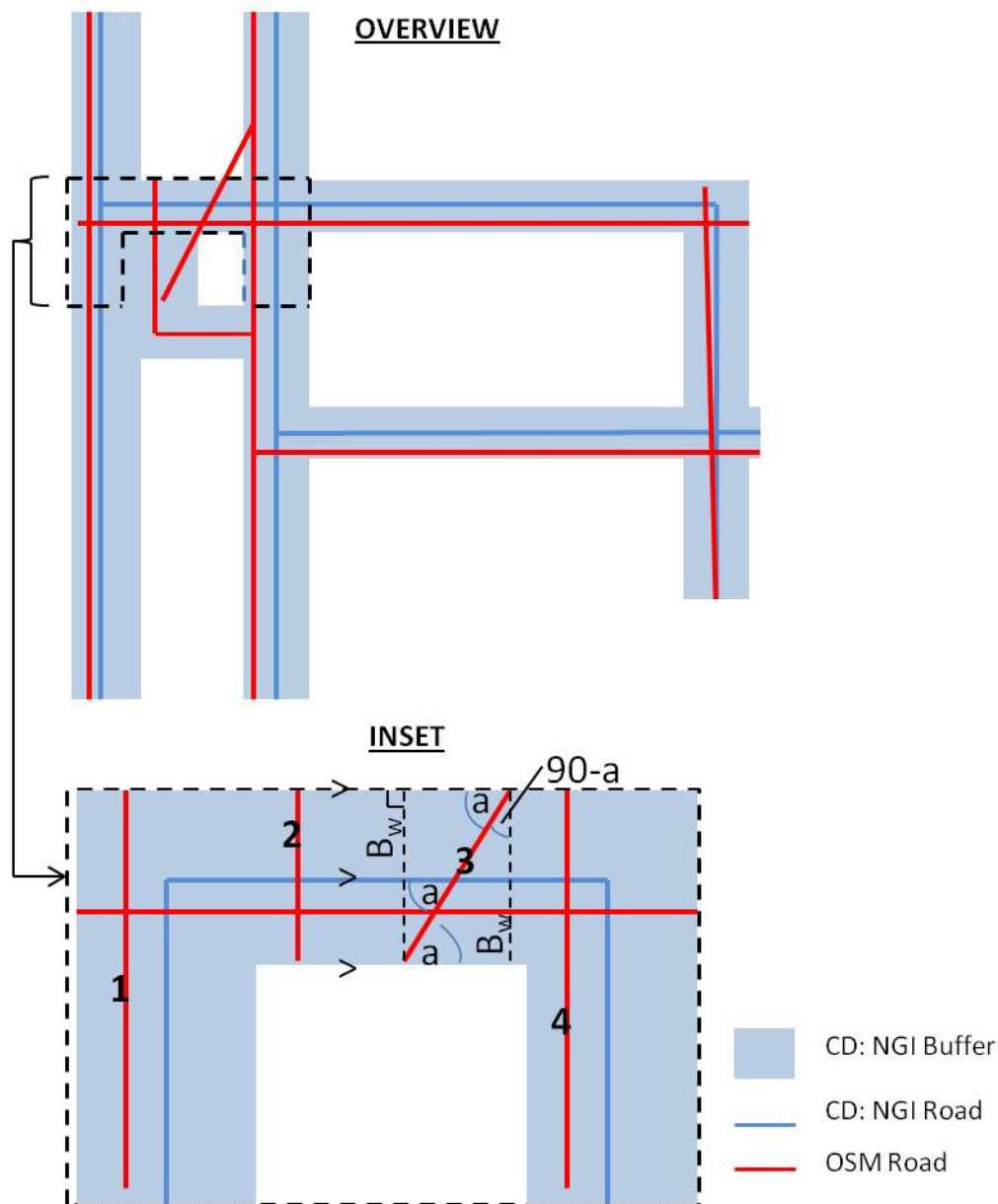


Figure 4.6: Method for removing unwanted OSM road sections within buffers of corresponding roads

Generating Centrelines

The CD: NGI generalises all multi-lane roads into a single centreline, whereas OSM allows for multi-lane digitising. It was necessary to generate singular centrelines for the OSM roads because it may cause the results to be better or worse than it actually is.

Although ArcMap allows for automatic centreline generation, this process required human intervention. Two line segments indicating opposite traffic directions may represent part of an OSM road. The other parts of the same road may be represented by a single centreline. Figure 4.7 shows an example of a multi-lane road highlighting the differences in digitising between the CD: NGI and OSM. The discontinuities in the OSM line segments prevented automatic selection of centreline pairs. Therefore, the centreline

pairs were selected manually so that the maximum road lengths were maintained and centrelines generated where possible.



Figure 4.7: Example of discontinuities in OSM multi-lane road overlaid onto the CD: NGI single centreline road

Computing the Positional Accuracy of Roads

The Goodchild and Hunter (1997) method was chosen to compute the positional accuracy of OSM road features because of the advantages that no assumption is made about the accuracy of the test data and because it is insensitive to outliers. According to this method, the positional accuracy may be determined by the percentage of OSM road that falls within the buffer of the corresponding CD: NGI road feature (see figure 4.8) (Goodchild and Hunter, 1997). For this investigation, the buffer width was not generated iteratively, but set to the reference data's stated positional accuracy of 10 m.



Figure 4.8: (a) Single buffer for the CD: NGI road feature; (b) identifying the corresponding OSM road sections within the buffer

4.4.2 Geometric Accuracy Of Polygon Buildings

Identifying Corresponding Buildings

Unlike line features, polygon features require a method that assesses a single polygon feature as a whole and not in parts. Corresponding polygons between the CD: NGI and OSM data sets were identified first by comparing their centroids (see figure 4.9). If the centroid of an OSM polygon was found to be within a CD: NGI polygon, the two polygons most likely represented the same feature on the ground. The computation of the shape descriptors would verify this assumption.

Other incorrect matches were identified visually, for example in figure 4.10 two OSM polygons were identified as a match for this particular building. In this case, the larger more representative polygon was chosen for the computations.



Figure 4.9: Identifying corresponding polygons between the CD: NGI and OSM data using the centroids



Figure 4.10: Example of cases where more than one OSM polygon is identified as a match

Computing the Geometric Accuracy of Buildings

The Hausdorff distance was computed to find the positional accuracy of OSM polygons. Two shape descriptors (compactness and elongation) and the ratio of areas were used to compare the shape of the CD: NGI and OSM polygons. The disadvantages of the Hausdorff distance related to vertices are not applicable to polygons (section 2.4.1).

- The Hausdorff distance is the “maximum distance of one set to the closest point in another set” (Gregoire and Bouillot, 1998).
- The Ratio of the areas is a simple computation to measure how many times bigger or smaller an OSM polygon is than its corresponding CD: NGI polygon
- Compactness measures how regular a polygon is, in other words, how much a polygon deviates from a predefined shape (in this case, a circle) (Lee and Wan, 2004).
- Elongation is a ratio of the width and length of the smallest rectangle containing every point making up the polygon (Mingqiang *et al.*, 2008).

The equations for these measures are given by:

1. Hausdorff Distance:

$$DH(A, B) = \text{Max}(\sup_{x \in A} d(x, B), \sup_{x \in B} d(x, A))$$

(Hangouët, 1995)

2. Ratio of Areas:

$$\text{Area Ratio} = \frac{\text{Area of OSM Polygon}}{\text{Area of CD: NGI Polygon}}$$

3. Compactness:

$$C = \frac{\text{Area}}{(0.282 \times \text{perimeter}^2)}$$

(Maceachren, 1985)

4. Elongation:

$$E = 1 - \frac{W}{L}$$

(Mingqiang *et al.*, 2008)

Those Hausdorff distances that were significantly larger than average were identified as outliers. An outlier represented an incorrect polygon match as discussed in section 4.4.2. Each outlier was inspected manually to confirm that it was an incorrect match. In figure 4.11 the centroid of the OSM polygon was within the CD: NGI polygon, but it

is clear that they do not represent the same building.

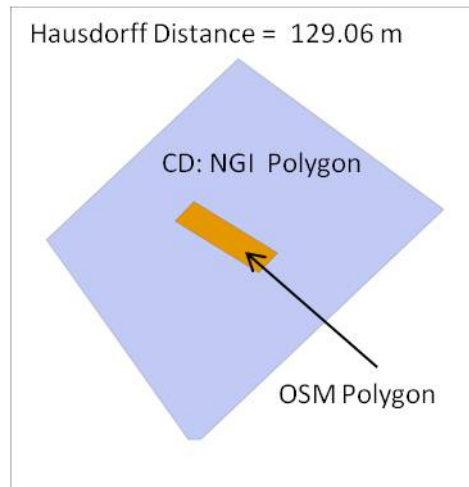


Figure 4.11: Example of incorrect polygon matching

4.4.3 Semantic Accuracy of Roads

A simple way to define the semantic accuracy of a data set is; compared to the reference feature classification, how often are features classed into their correct categories? In terms of the CD: NGI and OSM road classes, the semantic accuracy results will determine how often a feature is classified as the same type of feature in both data sets although the class names may differ. Therefore, it was necessary to determine which road classes defined the same type of roads between the two data sets.

The initial approach to assess the accuracy of the OSM data sets was to compare the predefined CD: NGI and OSM road classification definitions in order to establish matching road classes. A positive match could not be made between the road classes because some of the road class definitions were overlapping.

The Western Cape commercial data set was therefore chosen to determine how many times an OSM road class matched a certain CD: NGI road class for corresponding road features. The results are presented in the matrix in table 4.2, where the columns represent the CD: NGI road class and the rows represent the OSM road classes. Because the number of features was low for most of the road classes, only those road classes where there appeared to be possible matches were focused on. These were, main road, national freeway and street, as classed in the CD: NGI data set. From the results in table 4.2, it can be seen that there is not one definite match for these road classes between the data sets. Further investigation was done with respect to the respective road class definitions.

Comparing OSM and the CD: NGI Road Classes

Table 4.3 provides definitions for the road classes that could be possible matches between the CD: NGI and OSM roads for the Western Cape commercial data set. By comparison of the road class definitions, the three OSM classes were confidently matched to the CD: NGI classes. The results are shown in table 4.4. The column showing the percentage

Table 4.2: Sample matrix comparing the number of matches between the CD: NGI and OSM road classes for the Western Cape commercial data set

	CD: NGI ROAD CLASSES											
	Arterial Road	Footpath	Interchange	Main Road	National Freeway	National Road	On Ramp	Off Access	Secondary Road	Slipway	Street	Track
Crossing	0	0	0	0	0	0	0	0	0	0	0	0
Footway	0	0	0	0	0	0	0	0	0	0	2	0
Motorway	0	0	0	3	11	0	0	1	0	0	0	0
Motorway-Link	0	0	0	8	3	0	0	1	0	0	0	0
Path	0	0	0	0	0	0	0	0	0	0	0	0
Pedestrian	0	0	0	0	0	0	0	0	0	0	5	2
Primary	0	0	0	19	0	0	0	0	0	0	2	0
Primary-Link	0	0	0	0	0	0	0	0	0	0	1	0
Residential	0	0	0	0	0	0	0	0	0	0	65	0
Secondary	0	0	0	17	4	0	0	1	0	0	33	0
Service	0	0	0	3	0	0	0	1	0	0	8	0
Steps	0	0	0	0	0	0	0	0	0	0	0	0
Tertiary	0	0	0	0	0	0	0	0	0	0	9	0
Trunk	0	0	0	0	2	0	0	0	0	0	0	0
Trunk-Link	0	0	0	0	0	0	0	0	0	0	0	0
Unclassified	0	0	0	0	1	0	0	27	0	0	160	0

match represents how many times there was a match between the CD: NGI and OSM road classes. For example, the OSM primary road class had a 38% match to the CD: NGI main road class, while the OSM secondary road class had a 34% match. The positive road class matches were then used as a standard to assess the semantic accuracy of the other test areas.

Table 4.3: Comparing some of the CD: NGI and OSM road classes for the Western Cape commercial data set (Chief Directorate: National Geo-Spatial Information, 2013a; Ramm *et al.*, 2011, : 64)

CD: NGI ROAD CLASSES	OSM ROAD CLASSES	COUNT	MATCH %
MAIN ROAD - Main roads link the large towns, which are not on national or arterial routes, to the nearest major centre or city.	PRIMARY - Major long-distance (inter-city) road	19	38.0
	SECONDARY - Other major highways	17	34.0
NATIONAL FREEWAY - A national freeway is a dual carriageway (double road, each having two or more lanes) that is free of obstructions. No robots or intersections slow down the traffic and a minimum speed limit keeps slow vehicles off the freeway. Access is limited i.e. traffic must join and leave a freeway via an on or off ramp only.	MOTORWAY - Large, grade-separated, limited access freeway (or motorway). Each carriageway is drawn separately.	11	52.4
STREET - Streets make it possible to get access to the buildings in a town. They divide urban areas into blocks of houses or other buildings. Three basic patterns of streets can be described as grid, radial or irregular.	RESIDENTIAL - Residential street. Most inner-city roads use this type unless they are freeways.	65	62.8
	SECONDARY - Other major highways	33	11.6
TRACK - A track refers to a 'recreational track' and is more primitive than a motor sport track.	PEDESTRIAN - Pedestrian area or street.	2	66.7

Table 4.4: Positive road class matches between the CD: NGI and OSM

CD: NGI ROAD CLASSES		OSM ROAD CLASSES
Main Road	=	Primary Road
National Freeway	=	Motorway
Street	=	Residential

4.4.4 Completeness of Roads

The completeness of a data set can be defined as the omission of data and the presence of excess data (i.e. commission) (Girres and Touya, 2010) (see section 2.4.1). This definition applies to the attributes of the data set as well as the geometry. For this study, the completeness was investigated for road features only, based on the assumption that the CD: NGI data set is complete. The completeness was computed by dividing the total length of the April 2012 OSM road data set by the total length of the CD: NGI road data set as it was at 2012 (Haklay and Ellul, 2010). For example:

Total length OSM roads for the Gauteng commercial in April 2009 = 83.8 km

Total length CD: NGI roads for the Gauteng commercial as at April 2009 = 107.9 km

Thus, the OSM completeness at April 2009 = $100 \times (83.8 / 107.9) = 77.7\%$

4.5 Method for Qualitative Assessment

4.5.1 OSM Currency

The results from this part of the quality assessment were used to address the qualitative research questions regarding the uniformity and currency of OSM data across SA and how this compares to the CD: NGI requirements. The currency in this instance looks at the OSM growth rate and the data stability in comparison to the CD: NGI data.

Currency of Points

The currency was investigated by examining the evolution of OSM point, line and polygon data for 13 data sets covering the period from 2006 to 2012. Four classes of change were measured:

- Additions —where were new data added between consecutive data sets
- Deletions —where were existing data deleted between consecutive data sets
- Modifications —where were the geometric modifications to existing features between consecutive data sets
- No Change —where were the geometry of features unchanged between consecutive data sets

Consecutive data sets were compared by defining a buffer for each point feature. A 3 m buffer was chosen to compensate for the 5 m absolute positional accuracy of commercial

GPS devices. If a point from one data set is within the buffer width of a point from a consecutive data set, the two features are considered to represent the same topographical feature.

In figure 4.12, isolated points from data set 1 are treated as deletions; isolated points from data set 2 are treated as additions and points contained within the intersection of buffers from data set 1 and 2 are treated as either a modification or no change. Modified points are those points within the intersection of buffers, which exceed the minimum threshold value of 1 m. The threshold was decided upon by examination of the standard deviation of corresponding line features.

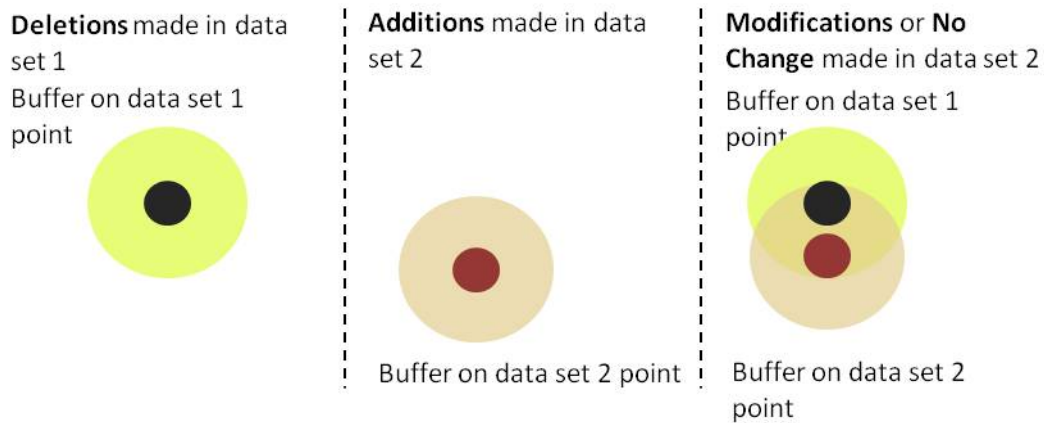


Figure 4.12: Point deletions, additions, modifications and no change

Currency of Roads

The same method was used to generate the additions, deletions, modifications and no change for the line data. Again, a 3 m buffer was created for every road feature and the intersections between consecutive data sets used to find the various categories of change. In figure 4.13, isolated line sections from data set 1 are treated as deleted roads. Isolated line sections from data set 2 are treated as additional roads. As before, modified lines are those lines within the intersection of buffers between consecutive data sets, which exceed the minimum threshold value of 1 m. Unchanged lines are those lines within the 1 m threshold value.

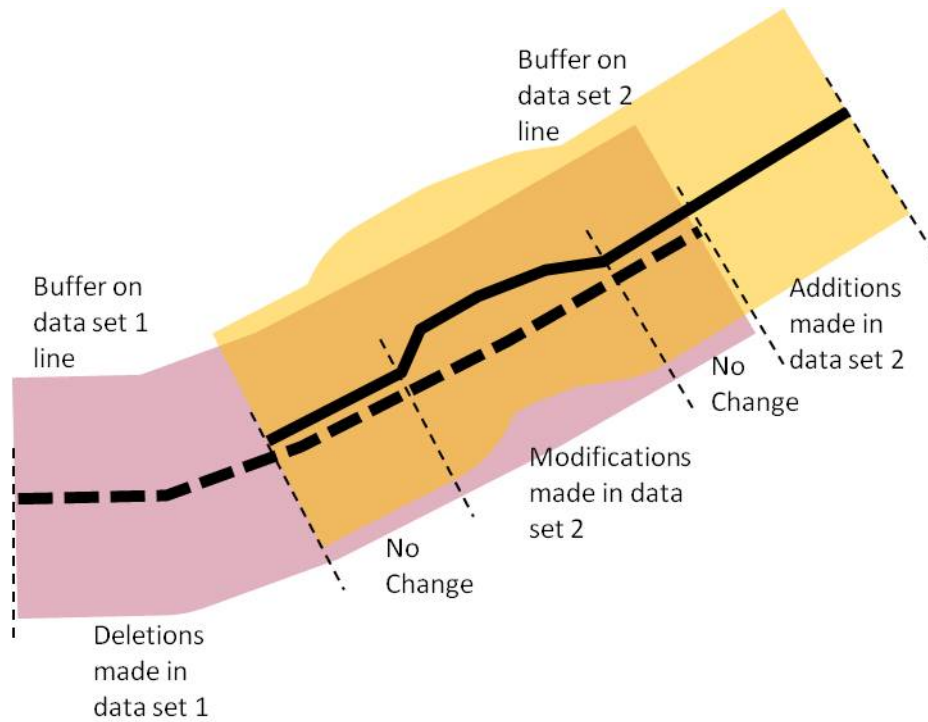


Figure 4.13: Line deletions, additions, modifications and no change

Currency of Buildings

As described in section 4.4.2, corresponding polygons between consecutive data sets were identified first by comparing their centroids (see figure 4.14). If the centroid of a polygon from data set 1 was found to be within a polygon in data set 2, the polygon in data set 2 was treated either as a modification or no change. Corresponding polygons with equal areas were categorised as no change, and the remaining polygons categorised as modifications.

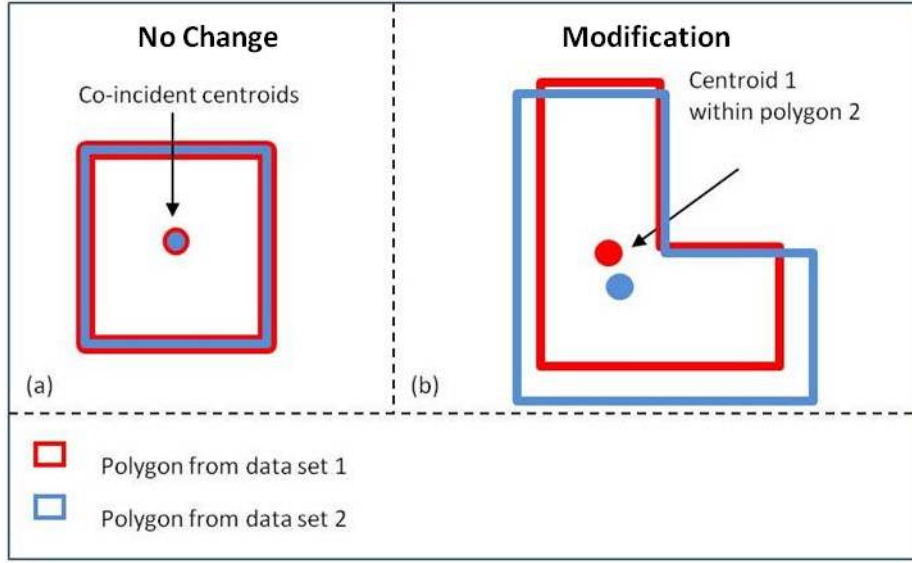


Figure 4.14: Identifying corresponding polygons between consecutive data sets by using their centroids; (a) shows corresponding polygons with the equal areas (no change); (b) shows corresponding polygons with different areas (modification)

Polygons from data set 1 that did not correspond with any polygons in data set 2 were treated as deletions. Polygons from data set 2 that did not correspond with any polygons in data set 2 were treated as additions.

4.5.2 OSM Uniformity in Acquisition

Uniformity of Point Acquisition

There are various areas of uniformity in topographical mapping, for example uniformity in distribution of data and uniformity in representation of data. Uniformity in acquisition of point data investigates whether contributions are made evenly the country. The uniformity was determined by investigation of point geometries with respect to their semantics. This investigation was an extension of a previous study by Siebritz *et al.* (2011). The authors state that the type of contributions made by volunteers will depend on factors such as culture and personal interests. For this part of the investigation, the OSM point data were classified into 8 amenity categories: Banking, Health, Education, Religious, Leisure, Safety, Postal and Transportation in order to determine how varied the contributions are across the country. In section 4.2.3, it was stated that the test areas differed slightly in size. This was accounted for in this specific assessment by ensuring all the areas were equal. Only those points within the specified area were included.

4.6 Analysis and Discussion

The OSM data required many stages of preparation in order to compare it to the CD: NGI data. Although most of the data preparation was done automatically, the process still required human intervention. Another disadvantage is that some of the Python scripts may take up to a few hours to run. Aside from the long preparation processing time, the semi-automatic method used to clean the data and to match corresponding

roads works well for the scope of this thesis. The Python scripts can easily be used to repeat the investigation for other areas.

After the data was in the desired format and corresponding roads matched, computing the positional accuracy was a straightforward task. The positional accuracy results showed that in some cases the cleaning techniques failed to eliminate all unwanted road sections. An example is shown in figure 4.15. The technique failed in this instance, because it relies on the intersection of either two OSM road sections or an OSM and CD: NGI intersection.

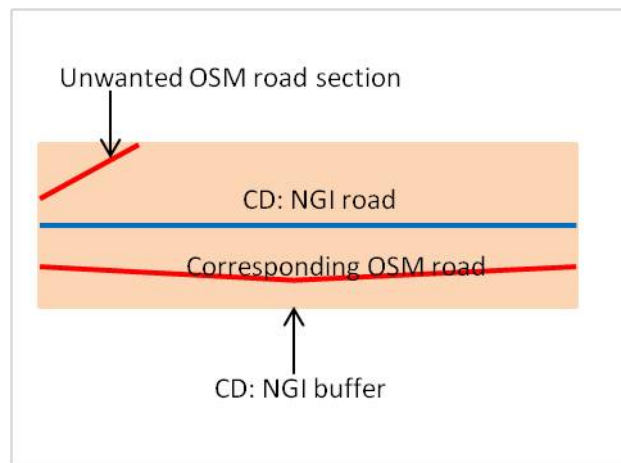


Figure 4.15: Example of the line matching technique failing

It was less difficult to identify outliers within the polygon positional accuracy results by observing the larger Hausdorff distances.

The process of matching corresponding road features made it easier to identify possible road class matches between CD: NGI and OSM. It cannot be said that the matched road classes were definitely representing the same features, but the match was confident enough to use as a standard for the rest of the test areas. Polygons were not tested for the semantic accuracy because, the OSM attributes were incomplete.

The positional and semantic accuracies of points were not tested because homologous points from OSM and CD: NGI could not be identified efficiently. Figure 4.16 shows the OSM and the CD: NGI point data for the Western Cape commercial test area, which is the data set with the most OSM point data. By visual inspection it can be seen that firstly, there is very little CD: NGI point data and secondly, very few of the points represent the same buildings between the data sets. Computing the positional and semantic accuracy would have yielded uncertain results.

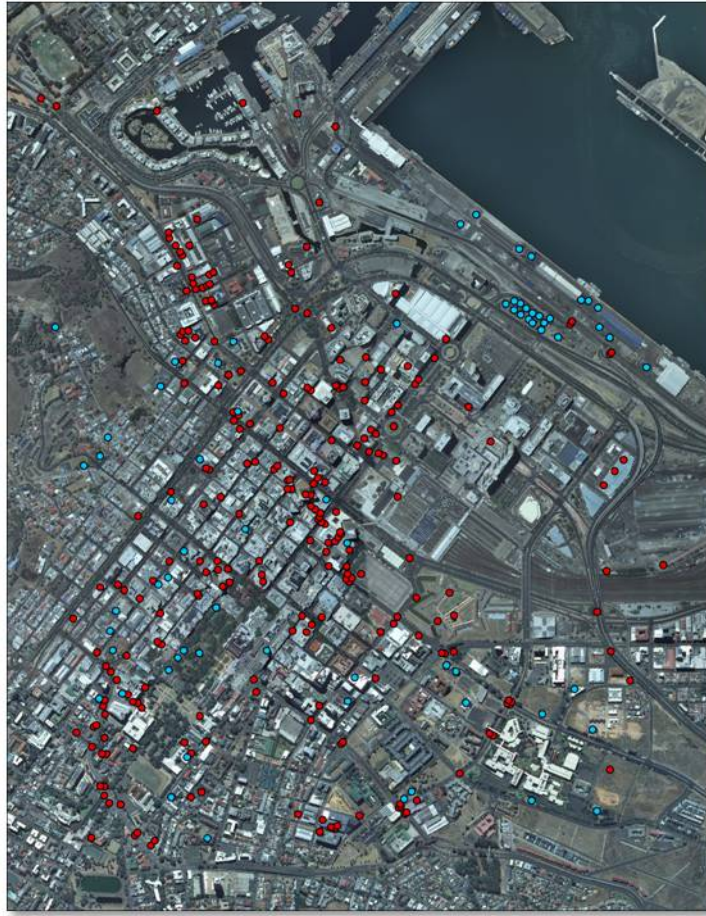


Figure 4.16: CD: NGI and OSM point data for Western Cape Commercial test area

In figure 4.17 (a) and (b) the blue and red points denote amenities in the CD: NGI and OSM data sets respectively. Figure 4.17 (a) and (b) shows two examples where points from both data sets denote the same building but represent different amenities according to the attribute data. This is not necessarily a semantic error in the OSM data set. Because the buildings being represented cover a large area in both (a) and (b), the building may be housing more than one amenity.

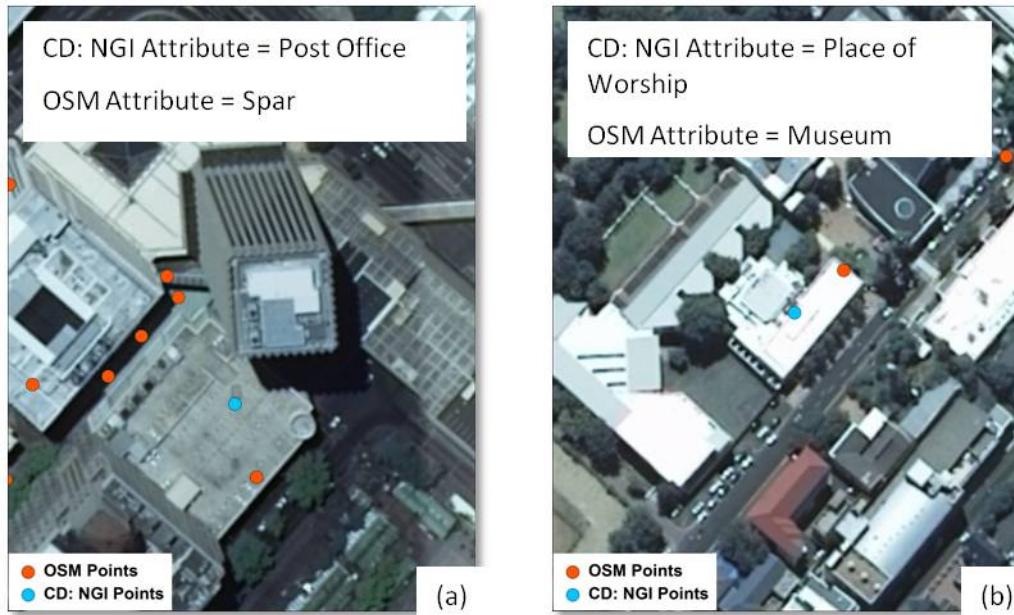


Figure 4.17: Comparing CD: NGI and OSM point data. (a) and (b) Examples of points denoting the same building but not the same amenity

The methodology used to address the qualitative research questions was executed without difficulty. However, the uniformity was limited to the point data.

The methodologies presented in this chapter has allowed all the research questions to be addressed, although only some were answered in this chapter, the remaining questions will be answered in the following chapters.

Chapter 5

RESULTS AND ANALYSIS

5.1 Introduction

This chapter presents the results of the various OSM quality assessments. The quantitative results are presented and discussed first section in 5.2. Thereafter the qualitative results are presented in section 5.3

Section 3.3.2 presented the CD: NGI standard for capturing topographical features. The standard provides the positional accuracy and semantic accuracy requirements. The standard states that features must have a positional accuracy not exceeding 10 metres at the 95% confidence interval. Because the method to determine the positional accuracy of roads is pre-set to 10 m, the confidence intervals were computed.

The CD: NGI maintains the same accuracy standard for polygon features. The positional accuracy determines how the average Hausdorff distance compares to the CD: NGI maximum deviation of 10 m. The shape accuracy results i.e. the compactness and elongation comparisons determine how well the OSM polygon shape compares with the CD: NGI shape. In addition, the areas of matching polygons were also compared.

The CD: NGI semantic accuracy standard requires that features be correctly classified with a confidence interval of 90%. The average weighted percentages (or confidence intervals) were computed for three OSM road classes within the three settlement categories.

The requirements for completeness and the qualitative measures (the currency and uniformity in data acquisition) are not specified in the standard. However, it is common knowledge that NMAs aspire to a data set that is complete, current and uniform. These quality elements were thus investigated in order to have a better understanding of the OSM quality.

5.2 Results for Quantitative Assessments

5.2.1 Positional Accuracy of Roads

Computing the Average Positional Accuracy of Roads

Figure 5.1 shows the average percentage overlap (or confidence level) for roads per test area. The percentages represent how much of the OSM roads are within 10 metres of

the corresponding CD: NGI roads. The Free State, Gauteng, Limpopo, North West and Western Cape data sets had similar results for the three settlement categories. The differences in results between commercial, residential and low urban density for these four provinces ranged from 3% to 14%. The Kwazulu Natal and Northern Cape test areas had greater differences between the three settlement categories. The Kwazulu Natal data set had a maximum difference of 22% and the Northern Cape 23%. The remaining two data sets showed the greatest differences in results, with a maximum difference of 40% for the Eastern Cape and 44% maximum difference for Mpumalanga.

It was seen that in some cases the positional accuracy of a data set is related to the number of features. Data sets with more features have a greater positional accuracy. This is the case for the Eastern Cape, Kwazulu Natal and Northern Cape data sets. For other provinces, like the Western Cape and Gauteng, fewer features did not have a significant impact on the positional accuracy. The percentage for the Mpumalanga's commercial data set is considerably lower than the other provinces. Further investigation showed that the OSM data for the commercial category has a shift in a north-westerly direction resulting in most of the OSM roads falling outside of the CD: NGI buffers (See figure 5.2). The surrounding areas were examined and it was found that there was a similar shift. The data contributed for this area was most likely set to a projection different to that of the base map data. This shift was not present for the residential and low urban density test areas.

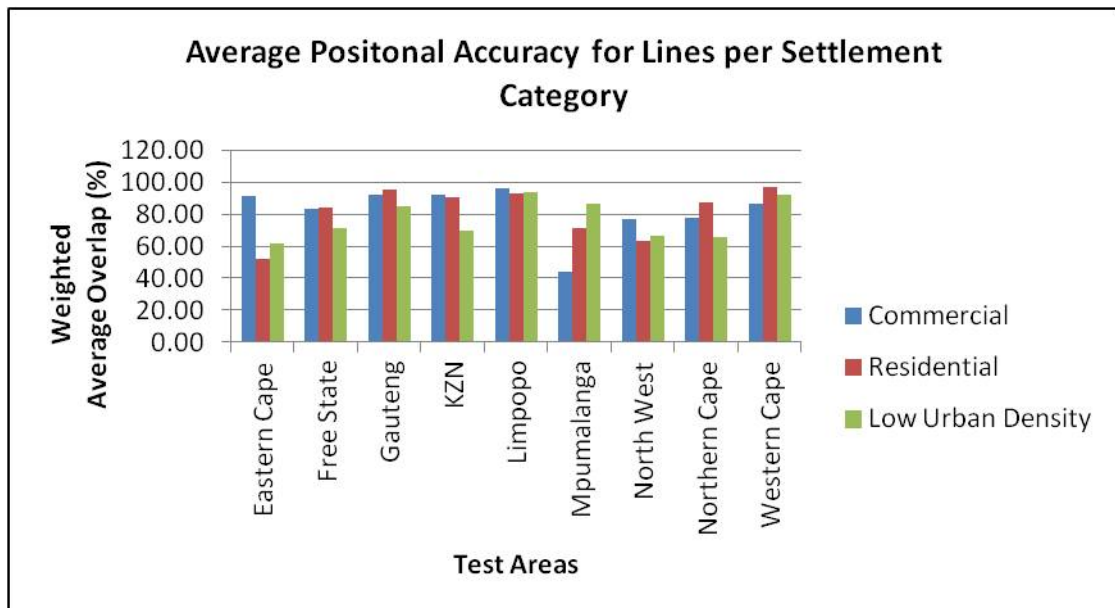


Figure 5.1: Graph showing the average percentage overlap for OSM roads per test area

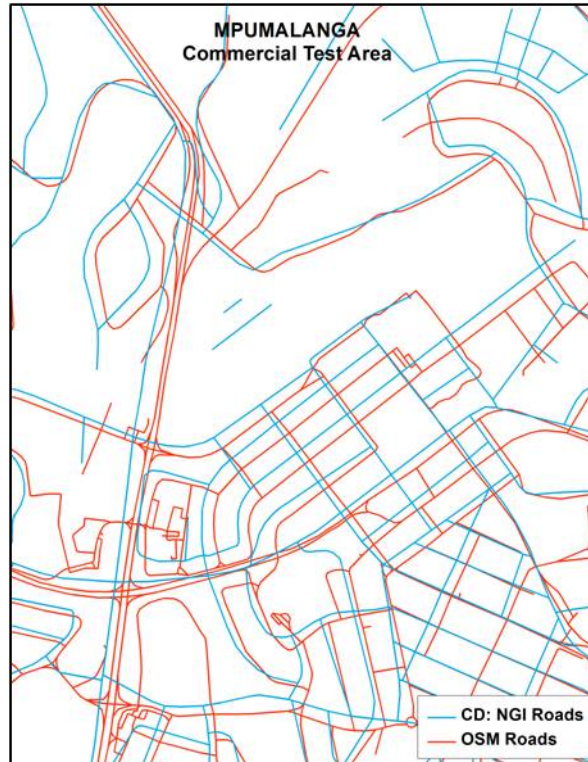


Figure 5.2: Depiction of the Mpumalanga data set with north-westerly shift in comparison to the CD: NGI data set

Because the number of features does influence the positional accuracy in some way, the count of features per test area was generated and used as the weight (and for all other computations where a weight was introduced) in order to compute the weighted averages. Firstly, it was computed for each province (see table 5.1) and then for each of the three settlement categories (see table 5.2). The weighted percentages range from 64.8% to 94.3% for the nine provinces and 74.1% to 85.7% for the three settlement categories. Five provinces are above 80% and four out of the five are within 10% of the CD: NGI requirement. These percentages are high considering that i) OSM does not have enforce accuracy, ii) the methods of data collection and iii) data is being generated by non-professionals, many of whom do not have a proper understanding of accuracy. The Gauteng and Limpopo provinces are very close to the 95% requirement. In fact, 5% of the OSM roads, which are not within 10 m could have been compared to the CD: NGI roads with an incorrect position. What this means is that the absolute positional accuracy of the OSM roads may be higher.

The North West province has a low percentage in comparison to the other provinces. The number of road features is not considerably low for any of the settlement categories in comparison to the other provinces. Without field survey data, it is difficult to determine whether the CD: NGI or OSM data set is incorrect. It could also be that many of the OSM roads exceed the 10 m buffer only slightly. Due to time constraints, further investigation was not possible.

Table 5.1: The weighted average percentage overlap for OSM roads per province

WEIGHTED AVERAGE POSITIONAL ACCURACY	
Province	Weighted Mean (%)
North West	64.8
Mpumalanga	71.1
Northern Cape	77.1
Eastern Cape	78.7
Free State	80.9
Western Cape	89.4
KZN	90.6
Gauteng	92.9
Limpopo	94.3

Table 5.2: The weighted average percentage overlap for OSM roads per settlement category

WEIGHTED AVERAGE POSITIONAL ACCURACY	
Settlement Type	Weighted Mean (%)
Commercial	84.9
Residential	85.7
Low Urban Density	74.1

Comparison of the CD: NGI and OSM Road Lengths

The positional accuracy results were used to perform further investigations. The CD: NGI and OSM road lengths were compared in a scatter plot for each test area. The scatter plots for the highest and lowest positional accuracies for each settlement category is shown in figure 5.3. The plots represent the length comparisons for: (a) Gauteng commercial, (b) Mpumalanga commercial, (c) Western Cape residential, (d) Eastern Cape residential, (e) Limpopo low urban density and (f) North West low urban density.

It is expected that the scatter plots are approximately linear if the CD: NGI and OSM roads match well. In this case, a good match is when the OSM length for a particular road is close to the length of the corresponding CD: NGI road. In other words, there is some similarity in the way CD: NGI experts and volunteers digitise roads. Longer roads are digitised as single features as far as possible and not as separate parts. This is in alignment with the CD: NGI topology rules discussed in section 3.3.2. The linear shape also demonstrates that the road matching technique has worked well.

The graphs to the left in figure 5.3 show those test areas with the highest positional accuracies for commercial, residential and low urban density, respectively, while those to the right show the lowest positional accuracies. The graphs to the left clearly have a more linear shape while those to the right have are more irregular, confirming what was expected. What is also noticeable from the graphs is that the graphs with higher accuracies have longer lengths, especially graphs (a) and (c). This again demonstrates

that better road matches results in higher accuracies.

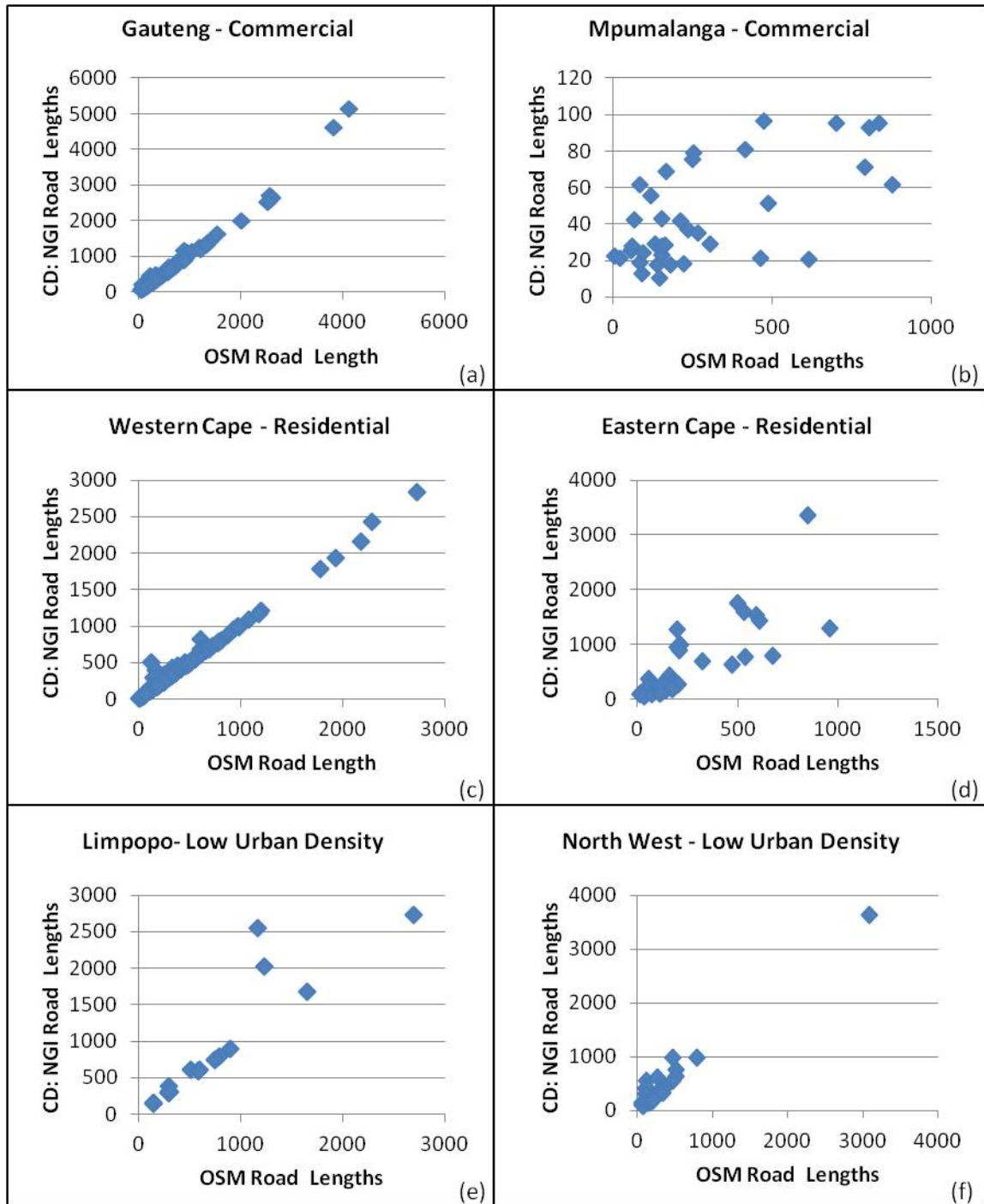


Figure 5.3: Scatter plots comparing the CD: NGI and OSM road lengths for (a) Gauteng commercial, (b) Mpumalanga commercial, (c) Western Cape residential, (d) Eastern Cape residential, (e) Limpopo low urban density and (f) North West low urban density test areas

5.2.2 Geometric Accuracy of Amenity Buildings

Comparison of the Hausdorff Distances

Only seven out of the twenty-seven test areas had polygon data representing amenities. There was no polygon data for the low urban density areas. As was discussed in section 4.4.2, incorrect polygon matches were identified by a large Hausdorff distance. In addition, for almost all the outliers (except one) the area ratio was either smaller than 0.5 or greater than 2.0. The outliers were removed and the compactness and elongation values computed.

Table 5.3 contains the average Hausdorff distances for each test area and the weighted averages for the commercial and residential categories. The dashes are for test areas with no data. The average Hausdorff distances for the commercial test areas range from 9.90 m to 22.03 m with a weighted average of 11.29 m, where the number of features is used as the weight. The averages for the residential areas range from 11.34 m to 17.36 m and a weighted average of 12.54 m. These deviations from the 10 m standard become insignificant when considering the size of a building. Nonetheless, the OSM polygons do not meet the CD: NGI accuracy requirements.

Table 5.3: Table comparing the Hausdorff distances for polygons in commercial and residential test areas

TABLE COMPARING HAUSDORFF DISTANCES				
Province	Commercial		Residential	
	DH(A,B) (m)	Std Dev (m)	DH(A,B) (m)	Std Dev (m)
Eastern Cape	10.01	1.81	17.36	9.69
Free State	—	—	11.34	12.06
Gauteng	22.03	14.38	—	—
Kwazulu Natal	13.47	5.14	—	—
Western Cape	9.90	5.74	13.75	7.20
Weighted Mean	11.29	6.01	12.54	11.31
Std Dev	5.70		3.03	

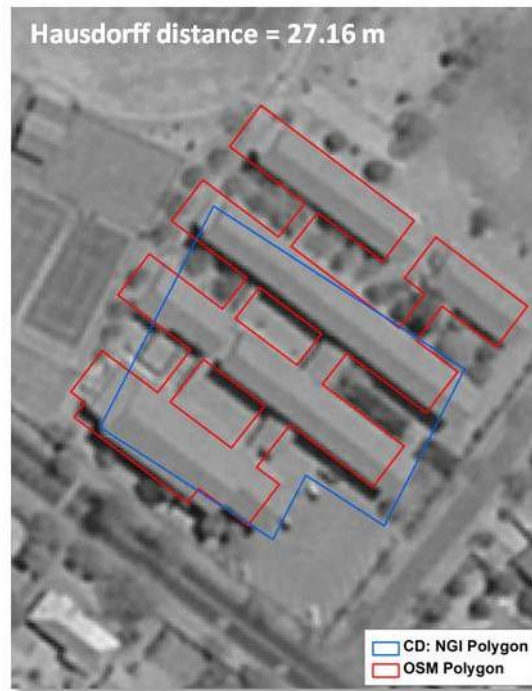


Figure 5.4: Example of generalisation applied to polygons at the CD: NGI

Comparison of the Area Ratio Results

Table 5.4 compares the area ratios for the commercial and residential areas. The ranges for commercial and residential areas are similar. The commercial area ratios range from 0.97 to 1.37 and from 0.97 to 1.34 for the residential category. This similarity in range is reflected by the equal standard deviations of 0.18.

The weighted average ratios are slightly different for the two categories with 1.06 for commercial areas and 1.27 for residential areas. For the commercial weighted average, which is close to one, it can be said that generally the CD: NGI and OSM polygons occupy similar areas for these test areas. Polygons in the tested residential areas tend to be slightly less similar in area. This may be because residential areas have houses, which tend to have irregular shapes. In commercial areas, buildings have a more regular shape.

Table 5.4: Table comparing the area ratios of matching polygons between the CD: NGI and OSM for commercial and residential test areas

TABLE COMPARING AREA RATIOS				
Province	Commercial		Residential	
	Area Ratio	Std Dev	Area Ratio	Std Dev
Eastern Cape	1.37	0.55	1.07	0.48
Free State	—	—	1.34	0.38
Gauteng	1.34	0.19	—	—
Kwazulu Natal	1.20	0.32	—	—
Western Cape	0.97	0.38	0.97	0.30
Weighted Mean	1.06		1.27	
Std Dev	0.18		0.18	

Comparison of the Compactness Results

The compactness values were computed for every pair of corresponding polygons. The averages per province are presented in tables 5.5 and 5.6 for commercial and residential test areas, respectively. The commercial compactness averages for both the CD: NGI and OSM polygons are mostly consistent between provinces. The consistency is shown in the graph in figure 5.5 and is reflected by the low standard deviations of 0.02 and 0.03 for the CD: NGI and OSM, respectively.

The ratio of the CD: NGI and OSM average compactness was computed for each province (see table 5.5). With a range of 0.99 to 1.12, the compactness between the two data sets generally compare well. The overall weighted average compactness is 1.01. The OSM polygons in the four commercial test areas therefore generally have a very similar compactness to the CD: NGI polygons.

The residential averages for the CD: NGI compactness vary between provinces, compared to the commercial compactness averages, reflected by the higher standard deviation of 0.12. The OSM residential averages are somewhat more consistent as can be seen in figure 5.6, reflected by the lower standard deviation of 0.04. As a result, the compactness ratios between the CD: NGI and OSM polygons are less consistent between provinces compared to the commercial results. Thus, as expected, the overall weighted average compactness is significantly greater than one. The CD: NGI polygons have a greater average compactness, thus the OSM polygons are less compact in the three residential test areas. As stated before, this may be because the residential areas have irregularly shaped houses.

Table 5.5: Table comparing compactness differences for commercial test areas

COMMERCIAL COMPACTNESS COMPARISON				
Province	CD: NGI Ave C	OSM Ave C	$\frac{CD:NGIC}{OSMC}$	Std Dev of C Diff
Eastern Cape	0.60	0.54	1.12	0.08
Gauteng	0.57	0.55	1.04	0.09
Kwazulu Natal	0.62	0.59	1.05	0.04
Western Cape	0.60	0.60	0.99	0.14
Weighted Mean	0.60	0.59	1.01	
Std Dev	0.02	0.03		

Table 5.6: Table comparing compactness differences for residential test areas

RESIDENTIAL COMPACTNESS COMPARISON				
Province	CD: NGI Ave C	OSM Ave C	$\frac{CD:NGIC}{OSMC}$	Std Dev of C Diff
Eastern Cape	0.53	0.50	1.06	0.18
Free State	0.65	0.41	1.57	0.18
Western Cape	0.41	0.44	0.92	0.03
Weighted Mean	0.61	0.43	1.42	
Std Dev	0.12	0.04		

The graph in figure 5.6 shows that the OSM compactness value for the Free State residential test area did not compare well to the CD: NGI value. Figure 5.7 shows an extract of the CD: NGI and OSM polygons for the Free State residential data set. The OSM polygons have more detail (polygons 1 and 2). In addition, adjacent buildings are digitised individually (polygon 3) unlike the CD: NGI, which combines adjacent buildings into one polygon. As discussed in section 4.4.2, in cases where separate buildings exist on a single property and were compiled as such by contributors, the polygon that was most representative of the property was chosen as the corresponding polygon. This provides an explanation for the large average compactness ratio of 1.57 for polygon 3.

The distribution of the compactness differences for the Free State residential data set is presented in figure 5.8. Ninety-five percent of the compactness values are above one, which means that for almost every matching polygon pair the CD: NGI compactness was greater than the OSM compactness, confirming that the CD: NGI buildings are more regular for this data set.

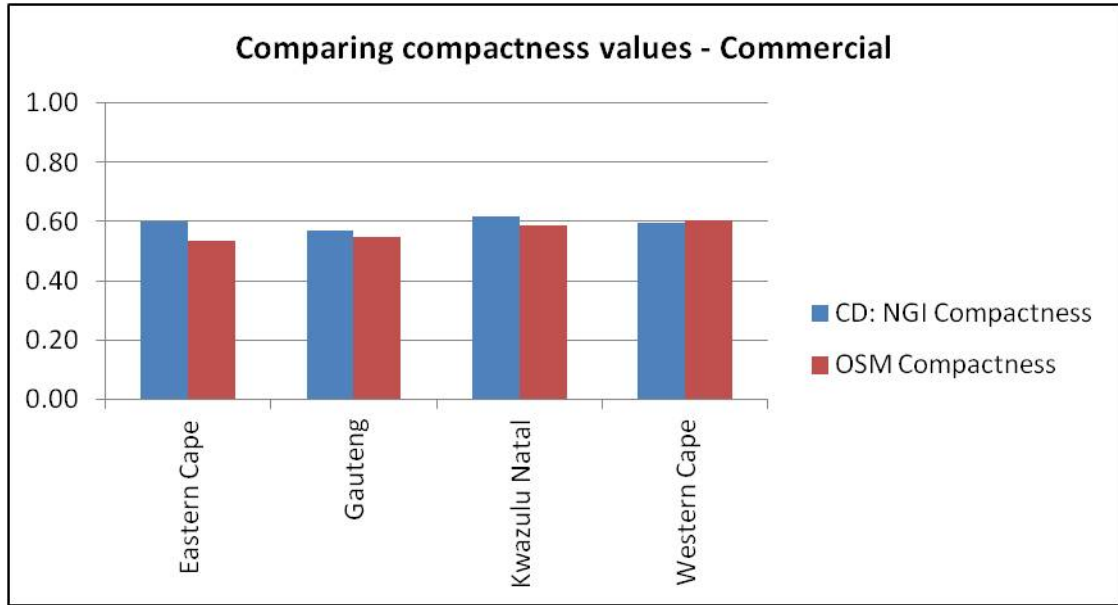


Figure 5.5: Chart comparing the CD: NGI and OSM compactness values for commercial test areas

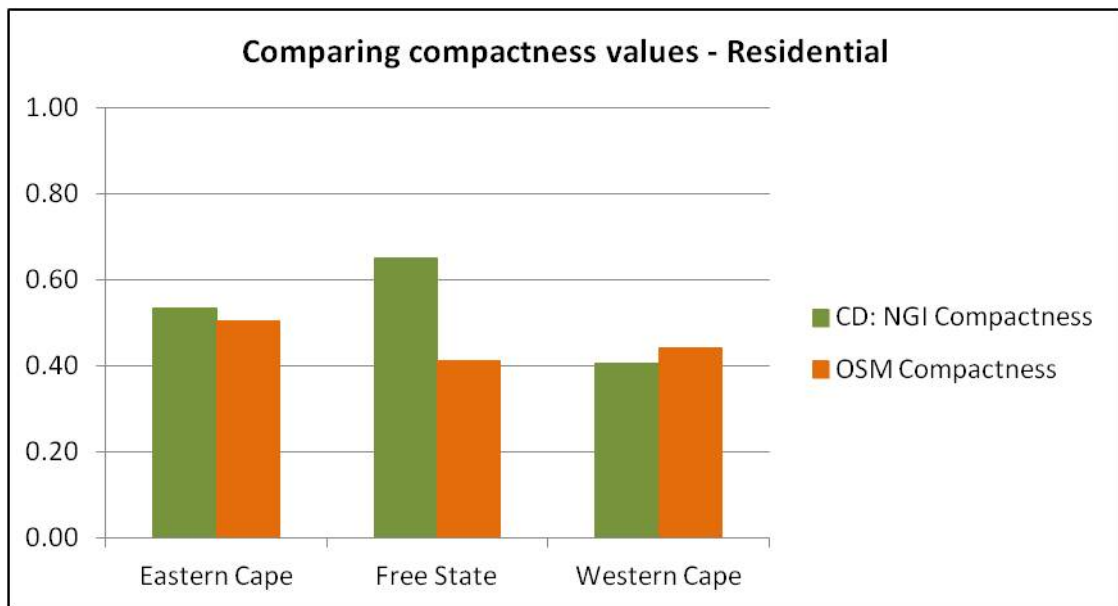


Figure 5.6: Chart comparing the CD: NGI and OSM compactness values for residential test areas

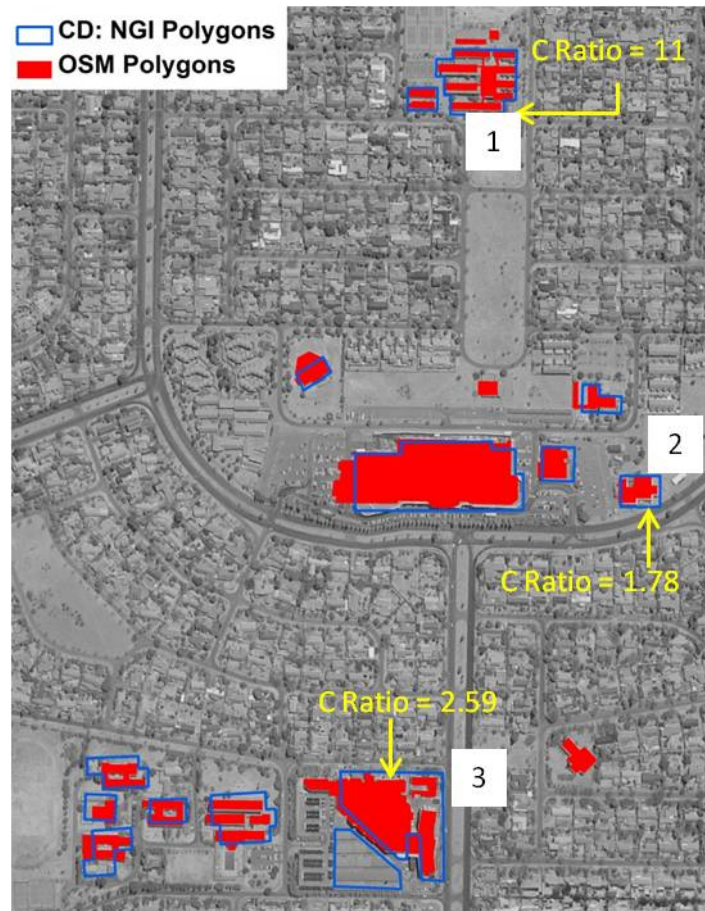


Figure 5.7: Comparing compactness between the CD: NGI and OSM polygons for an extract of the Free State residential data set

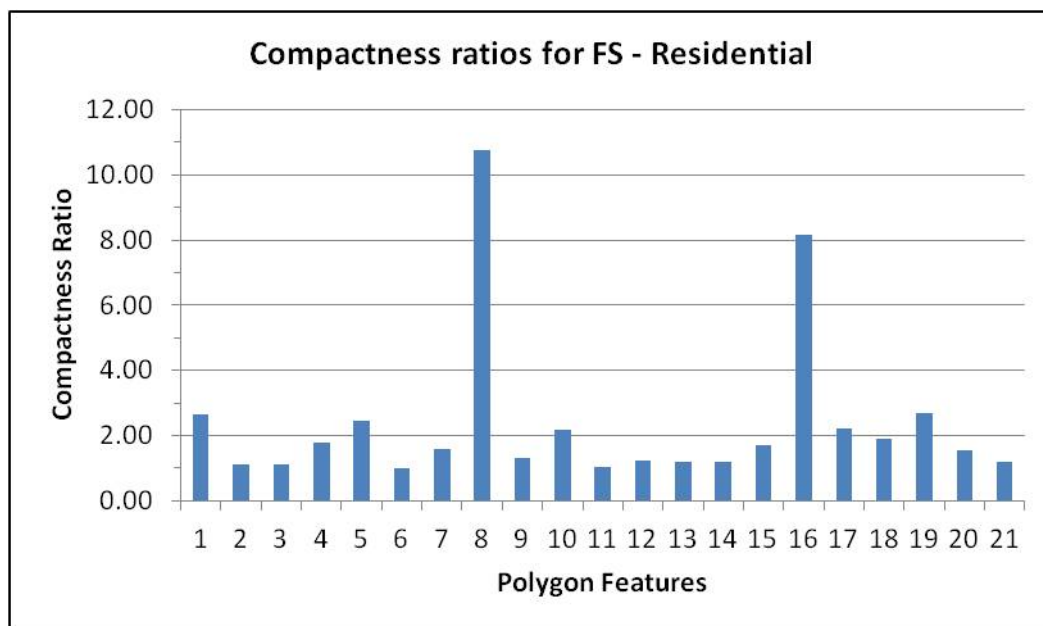


Figure 5.8: Chart showing the distribution of the compactness differences for the Free State residential test area

Comparison of the Elongation Results

The elongation was computed for every pair of corresponding polygons. The results for commercial and residential areas are presented in tables 5.7 and 5.8. The CD: NGI elongation averages per province are more consistent for commercial areas. The OSM data had a similar inconsistency for both commercial and residential areas.

The differences between the CD: NGI and OSM elongation values are small for most provinces in both the commercial and residential areas. The elongation differences for the Gauteng and Western Cape commercial test areas are slightly higher, resulting in the weighted average elongation being less similar to that of the CD: NGI. The OSM overall elongation was lower, which means that the polygons in these areas are elongated. The weighted average elongation for the residential areas was very similar to the CD:NGI elongation; they only differ by 0.01. Thus, polygons in these three areas have a similar elongation to the CD: NGI.

Table 5.7: Table comparing the elongation differences for commercial test areas

COMMERCIAL ELONGATION COMPARISON				
Province	CD: NGI Ave E	OSM Ave E	E Difference	Std Dev of E Diff
Eastern Cape	0.54	0.56	-0.02	0.07
Gauteng	0.45	0.56	-0.10	0.06
Kwazulu Natal	0.35	0.30	0.03	0.10
Western Cape	0.42	0.34	0.08	0.20
Weighted Mean	0.42	0.36		
Std Dev	0.08	0.14		

Table 5.8: Table comparing the elongation differences for residential test areas

RESIDENTIAL ELONGATION COMPARISON				
Province	CD: NGI Ave E	OSM Ave E	E Difference	Std Dev of E Diff
Eastern Cape	0.74	0.69	0.05	0.14
Free State	0.36	0.38	-0.03	0.22
Western Cape	0.46	0.41	0.05	0.14
Weighted Mean	0.43	0.44		
Std Dev	0.20	0.17		

5.2.3 Semantic Accuracy of Roads

The percentage match that is, how many times there was a match between the CD: NGI and OSM road classes, is presented in table 5.9. The dashes represent where the road feature did not exist in the CD: NGI data set.

The percentages for commercial areas are the lowest, which is unusual because in the previous assessments this category had better results compared to residential and low urban density areas. In all three settlement categories, the “street” class (or the residential class in OSM) had the highest count and is significantly higher than the other

two road classes. As was stated earlier, the number of features influences the accuracy results. However, only the residential and low urban density categories have a high percentage match for the “streets” road class and meet the CD: NGI requirement of 90% confidence level.

Considering the settlement pattern, it is expected that the commercial category have fewer roads in the “street” category. This explains why even though the count is high the percentage match is low. One explanation for the high percentage match for the “street” class in the other two settlement categories is that there is a default naming with this class against how many features are actually present in the data set. In other words, most of the time volunteers will classify a road as a residential road (or street) and for residential and low urban density areas, it just happens to be the correct classification.

The results for the commercial category thus provides a more correct indication of how often volunteers classify roads correctly. It is also the category with the most data available and thus provides a more balanced weighted average. The low percentages for this category, demonstrate that volunteers generally do not classify roads correctly. This could be either because they do not have sufficient understanding of road classes or simply because they are not motivated to ensure correct classification.

Table 5.9: Table showing the weighted average percentages for the CD: NGI and OSM road class matches

PERCENTAGE MATCH BETWEEN CD: NGI AND OSM ROAD CLASSES									
	Commercial (%)			Residential (%)			LU Density (%)		
Province	National Freeway	Main Road	Street	National Freeway	Main Road	Street	National Freeway	Main Road	Street
Eastern Cape	100.00	—	36.65	—	—	91.84	—	—	100.00
Free State	—	0	100.00	—	—	81.55	—	—	98.04
Gauteng	—	16.67	54.76	—	0	96.47	—	—	—
KZN	50	18.37	8.57	—	—	0	—	—	94.29
Limpopo	0	100.00	80.30	—	0	100.00	—	0	92.31
Mpumalanga	0	—	13.33	0	0	100.00	—	0	100.00
North West	—	0	100.00	—	—	100.00	—	—	100.00
Northern Cape	0	0	58.94	—	12.50	91.14	0	—	100.00
Western Cape	52.38	38.00	22.86	100.00	—	85.38	—	0	98.04
Weighted Mean	39.13	28.89	38.41	66.67	7.69	91.37	0	0	91.38
Std Deviation	41.03	36.01	35.22	70.71	6.25	31.80	0	0	2.98
Std Error	6.05	3.10	0.96	28.87	1.73	1.04	0	0	0.17

The graph in figure 5.9 shows the size of the standard error associated with the percentage for each road class. The standard error increases with smaller, more variable data sets (Culham, 2006). In comparison to the other two road classes, the error bars for

the “street” class are the smallest. Although, the percentage for this class is low for the commercial category, the count is high, resulting in a small standard error. The small standard errors confirm that “streets” are the most correctly classified. The “national freeway” class had the biggest standard errors. This may be because the class is the least correctly classified or because the number of features is low.

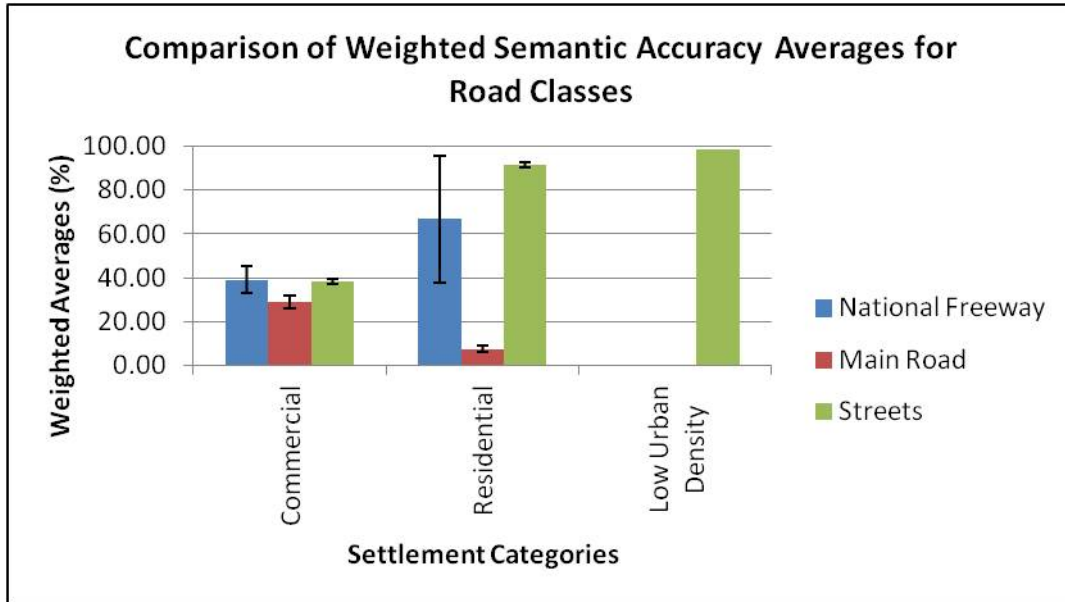


Figure 5.9: Graph comparing the standard errors associated with the weighted averages for the CD: NGI and OSM road class matches.

5.2.4 Completeness

The three graphs in figures 5.10, 5.11 and 5.12 show the completeness of OSM data from October 2006 to April 2012. The completeness is based on the initial assumption that the CD: NGI data set is complete. Thus, in cases where the completeness percentage exceeds 100%, the assumption no longer holds true.

In all three graphs, the OSM data reaches a peak and then evens out. Each settlement category has a different peak level and occurs at different dates. The commercial category had the most sites reaching their peaks during the periods 2009 to 2010 and 2011 to 2012. The residential category had the most sites reaching their peaks during the period 2011 to 2012 and for the low urban density category, during the 2010 to 2011 period. This shows that OSM South Africa received the most contributions from 2010 to 2012. Perhaps more people became aware of or were exposed to volunteer mapping during this period. Siebritz *et al.* (2011) state that specific events in time, like the 2010 FIFA Soccer World Cup, may have motivated citizens to participate in volunteers mapping in SA.

For commercial areas, three of the test areas did not reach a completeness of 100%, although two of the three areas had a maximum in the 93-96% range (i.e. omission). The residential category had five sites with a maximum completeness below 100%. In the low urban density category, only two sites reached a completeness of 100% and above. In the case where the completeness exceeds 100%, it means that at that point in time

the total length of OSM roads exceeds the total length of the CD: NGI roads (that is, commission). Thus, the OSM data sets reach absolute completeness, not at 100% completeness, but more likely at the point where the graphs even out after it reaches 100%.

The results show a clear difference in the level of completeness between the three categories, less developed areas have a lower level of completeness. The completeness graphs also show that commercial areas received contributions much quicker than the other two categories and therefore even out sooner.

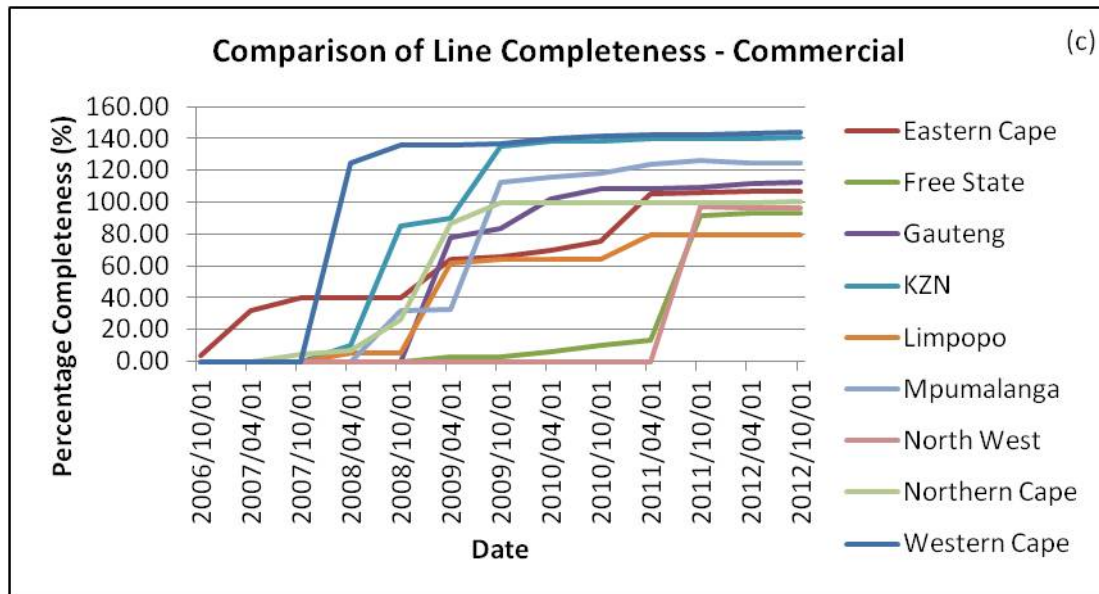


Figure 5.10: Comparison of the completeness of OSM line data for commercial areas

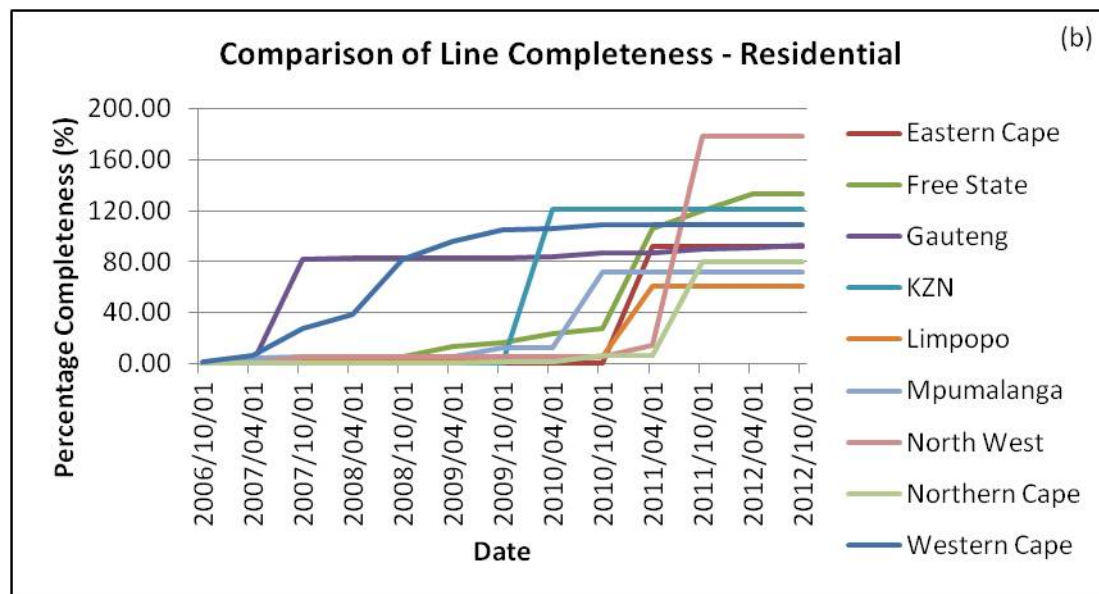


Figure 5.11: Comparison of the completeness of OSM line data for residential areas

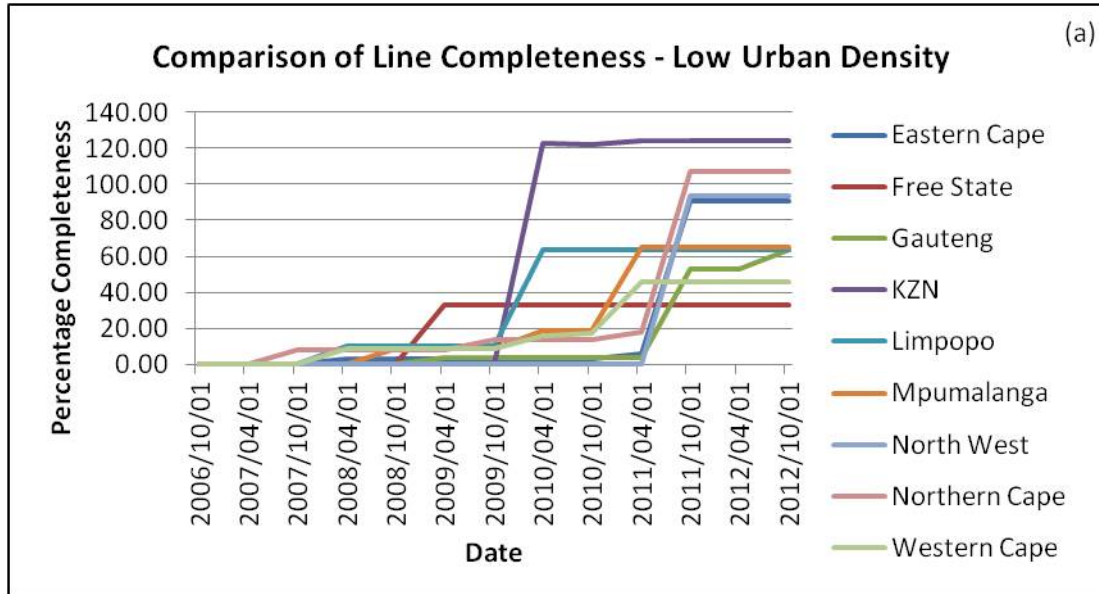


Figure 5.12: Comparison of the completeness of OSM line data for low urban density areas

5.3 Results for Quantitative Assessments

5.3.1 OSM Currency

The currency results consist of the total additions, deletions, modifications and unchanged data for points, lines and polygons. The total point and polygon additions to the commercial category were low for most provinces. There were almost no point and polygon additions for the residential category and no additions for the low urban density category. As a result, the number of deletions, modifications and unchanged features were very low and will thus not be discussed. Instead, the focus will be on the currency of roads. However, due to limited space, only some of the results will be presented.

The graph in figure 5.13 shows the total line additions for each test area. It confirms that less developed areas have fewer contributions. Gauteng, Kwazulu Natal and Western Cape had big amounts of data contributed for the commercial areas in comparison to the other provinces. There are three possibilities for the high amount of additions in commercial areas. The first of these being, that people in these areas may have a greater interest in volunteer mapping. Secondly, there may be more features to map. Thirdly, people in commercial areas may have better access to the resources needed to contribute data. Most likely, it is a combination of these factors.

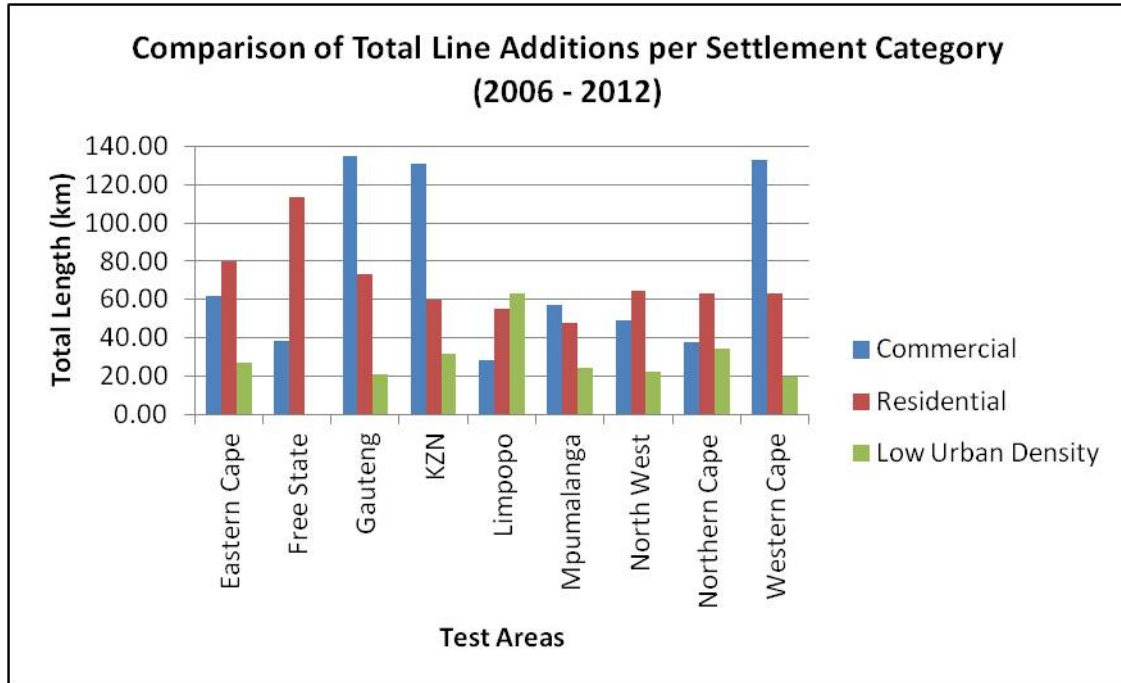


Figure 5.13: Comparison of OSM additions per province for the period 2006 to 2012

The graphs in figures 5.14, 5.15 and 5.16 show the total length of unchanged features between consecutive date sets for commercial, residential and low urban density, respectively. As with the completeness discussed in section 5.2.3, the OSM data sets reach a maximum and then the data evens out. The point where it evens out is generally earlier for commercial areas. In addition, the data peaks are the highest in commercial areas. Gauteng reached its peak much earlier than the other provinces for the residential and low urban density categories. This could mean that the initial base data was a mass upload from a single source.

The fact that most of the contributions were made during 2010-2012, means that the OSM data set can be seen as being current only after 2012 and only for some data sets. Further investigation, with later data sets is necessary to determine whether the currency seen in 2012, is being maintained.

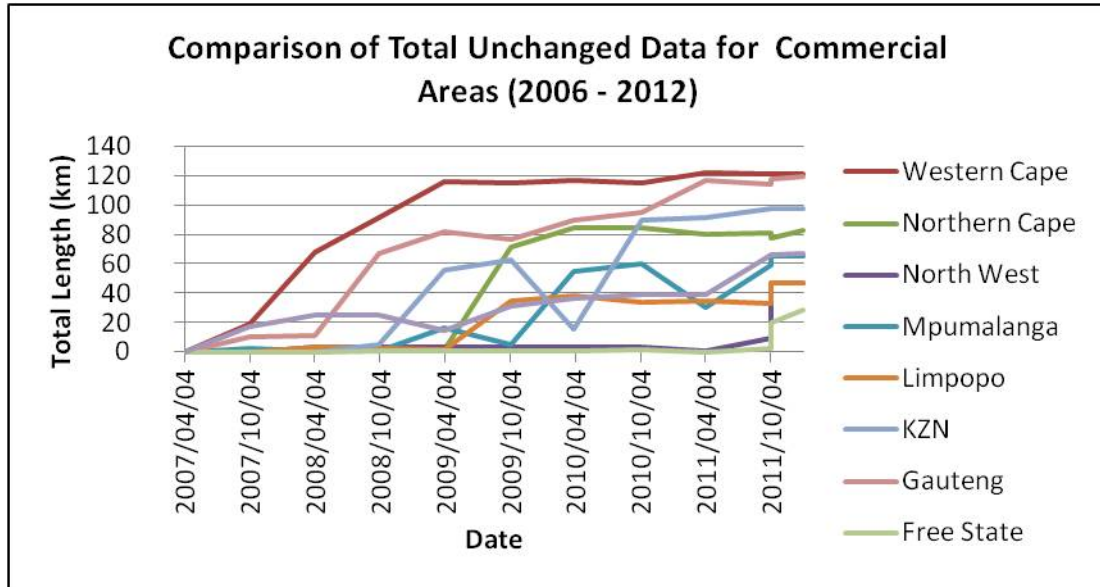


Figure 5.14: Comparison of OSM unchanged data per province in commercial areas for the period 2006 to 2012

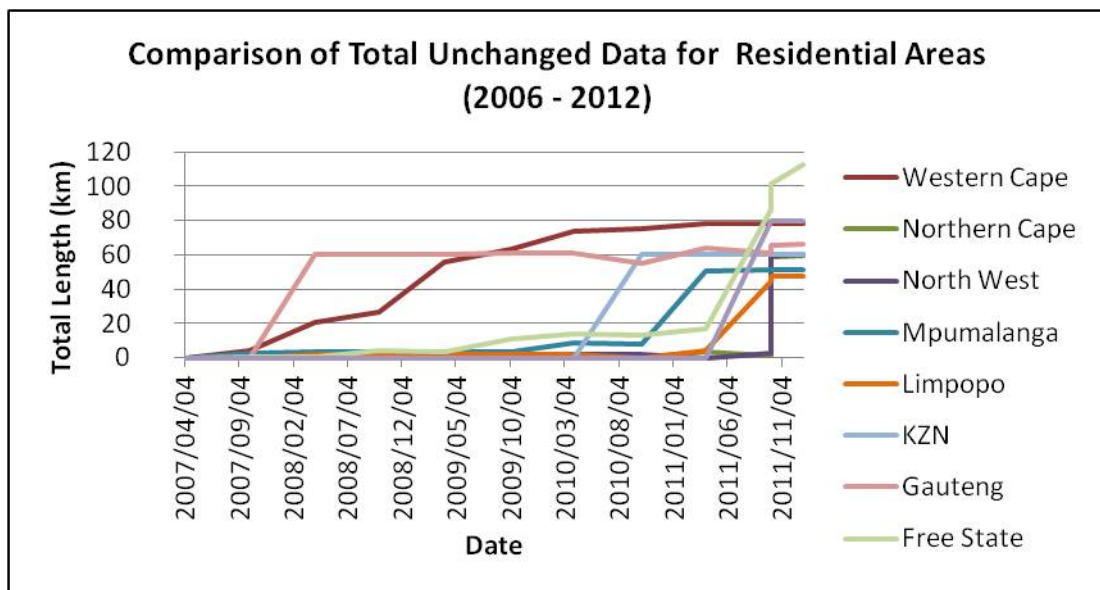


Figure 5.15: Comparison of OSM unchanged data per province in residential areas for the period 2006 to 2012

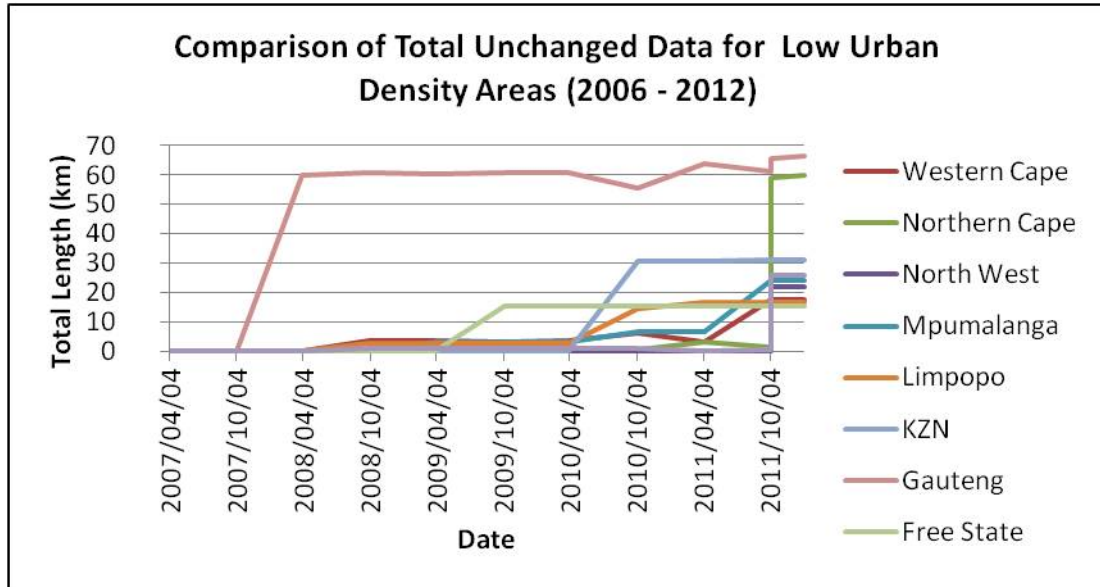


Figure 5.16: Comparison of OSM unchanged data per province in low urban density areas for the period 2006 to 2012

5.3.2 OSM Uniformity of Point Acquisition

The count for each of the eight amenity categories are presented in figures 5.17 and 5.18 for commercial and residential areas respectively. The commercial category had six test areas, while the residential category only had three test areas. For the commercial category the “leisure” category had the highest number of contributions for four provinces. Mpumalanga had the most contributions in the “transportation” category and Eastern Cape has the same amount of contributions for the “leisure” and “transportation” categories. The other amenity categories had only a few contributions and not all of the provinces had contributions for all of the amenity categories.

The point contributions for residential areas were very low as stated in section 5.2.4. “Banking” is the only category that had contributions for all three provinces. There were no contributions for the postal category.

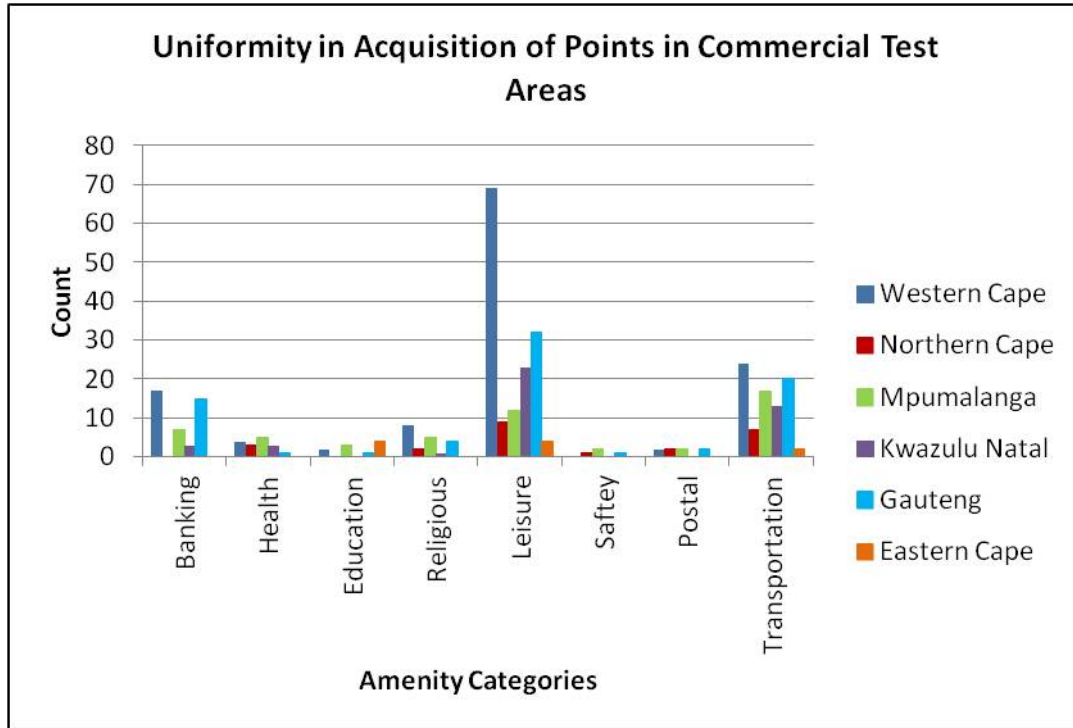


Figure 5.17: Chart showing the distribution of contributions for different point categories in commercial areas

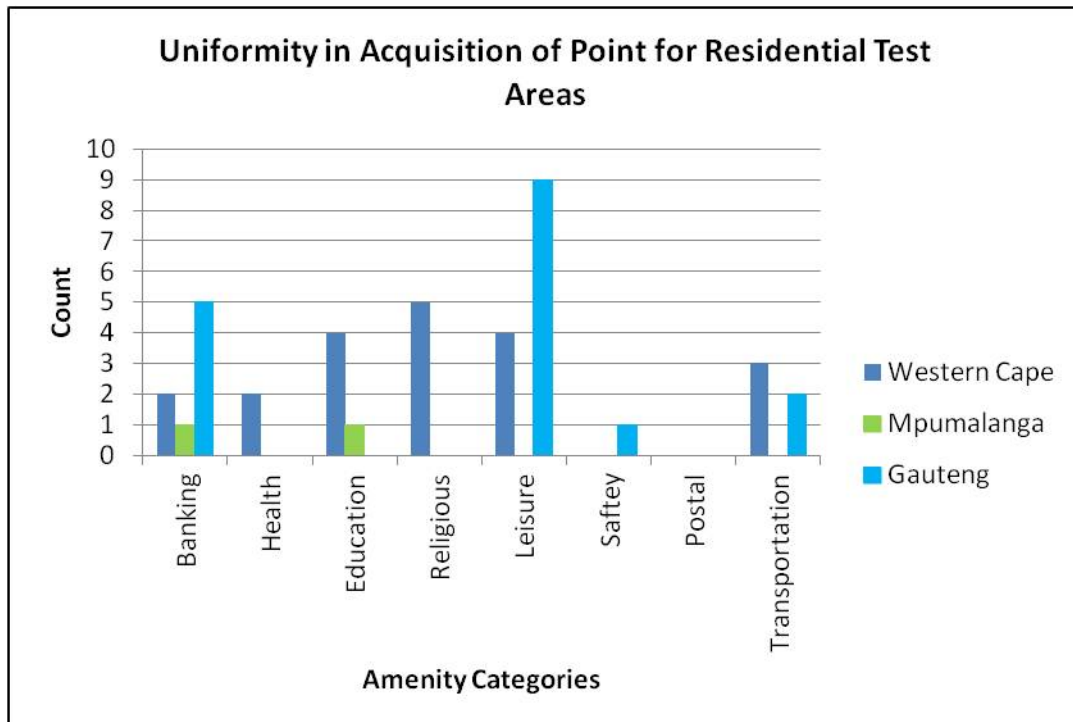


Figure 5.18: Chart showing the distribution of contributions for different point categories in residential areas

Figure 5.19 compares the total contributions from all test areas for commercial and

residential. Overall, the commercial areas had the most contributions in the “leisure” category and the second most in the “transportation” category. The residential test areas had the most contributions in the “banking” category. The contributions made to residential test areas are not varying and the number of contributions is low, thus it is difficult to draw conclusions about the type of contributions made by people in residential areas. The commercial results provide a better indication about the type of contributions made by volunteers. People in this area consider leisure amenities and transportation facilities to be the most important. This result is not unexpected, as these areas have more leisure amenities. Also, people have to travel to these areas daily for work, thus information on the availability of transportation would be of importance. What can be said with certainty is that OSM contributions are not uniform across the country, whether it is the type of contributions or the amount of contributions.

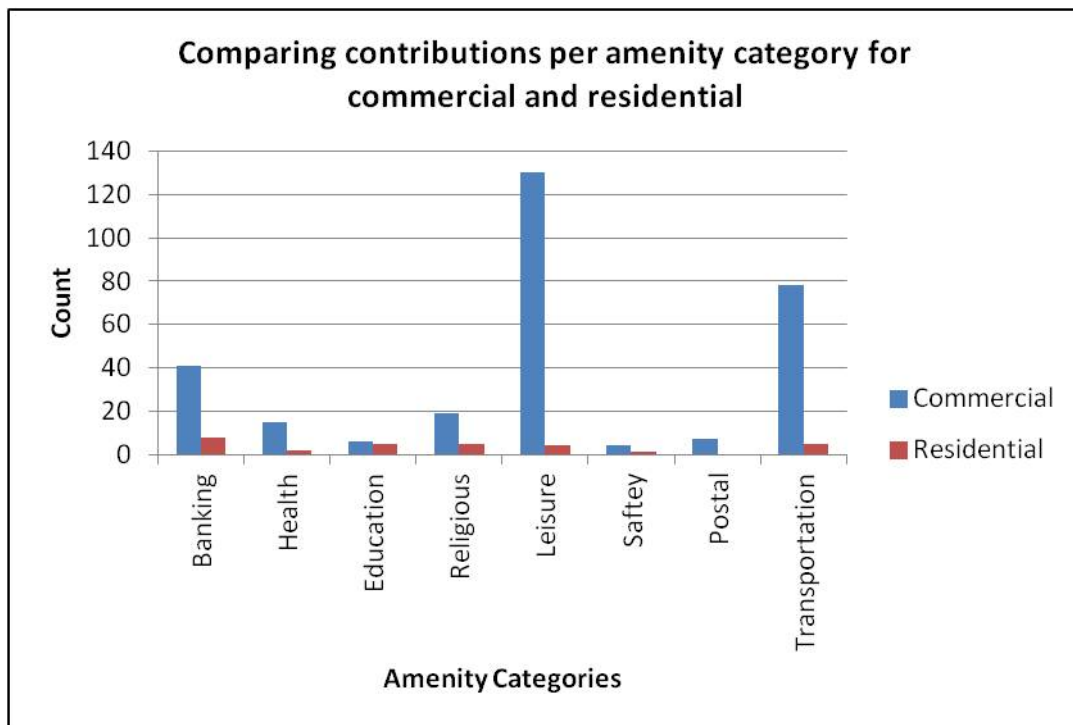


Figure 5.19: Chart showing the distribution of contributions for different point categories in residential areas

5.3.3 Analysis and Discussion

The aim of presenting the results in this manner was to determine whether the OSM data meets the CD: NGI minimum accuracy requirements. The CD: NGI requirements are set out for the positional accuracy and semantic accuracy, thus the quality assessments were aligned to the predefined standards as far as possible. The positional accuracy results for roads were directly applicable to the CD: NGI standard. For polygon buildings however, the method did not allow the 10 m threshold to be set, instead, the average positional accuracy was computed. A computation of the positional accuracy for polygons would be insufficient; therefore, shape comparisons between the two data sets were also necessary. Two well-known shape descriptors, the compactness and elongation were used to compare the shape of matching the CD: NGI and OSM polygons. An additional

comparison of the polygon areas was also performed.

For five of the provinces the overlap percentages for roads are high, with greater than 80% overlap. Considering the 5% error in the CD: NGI data, four provinces may have an absolute accuracy of 95% and greater. In terms of the positional accuracy of OSM building polygons, the average Hausdorff distances in commercial and residential areas compare well with the CD: NGI's stated positional accuracy of 10 m. The commercial areas have a higher positional accuracy, with a deviation of 1.29 m and 2.54 m in commercial and residential areas, respectively. However, there was very little polygon data available, therefore, the results cannot be generalised to all commercial and residential areas in the country. In addition, although the deviations are not large, the positional accuracy of polygons does not meet the CD: NGI standard.

The second part of the geometric accuracy results concerning the shape comparisons for polygons showed that OSM building polygons in the tested commercial areas compare well with the CD: NGI polygons in area and compactness. The OSM elongation values in these commercial areas are less consistent with the CD: NGI polygons. Polygons in residential areas compare well in terms of the elongation values, but the area and compactness results are less consistent with the CD: NGI polygons. Generalisation and operator/contributor interpretation are two factors that influence the positional and shape accuracies, more so for polygons as there is more room for interpretation.

The OSM semantic accuracy of roads is only high for one of the road classes in only two settlement categories. Only the "streets" road class in the tested residential and low urban density areas meets the CD: NGI standard of a 90% confidence level. Overall, the semantic accuracy of OSM data is low and do not come close to the 90% requirement. The results give a bad impression of the non-expert's understanding of feature classification, as well as the importance of correctly classifying features. It also confirms the statement by Baglatzi *et al.* (2012) discussed in chapter 2, that freedom in OSM tagging results in semantic interoperability problems.

For the other quality elements that were presented here, that is, the completeness, currency and uniformity in data acquisition there was no predefined standard. The results obtained for these measures were therefore gauged logically.

Roads in commercial and residential areas generally have a high level of completeness and in many cases exceed the CD: NGI completeness level. Low urban density areas are in most cases not complete. It may however increase as more people become involved with volunteer mapping. The OSM point and polygon data is still minimal. The amount of data generated across the country is therefore not evenly distributed.

In a previous investigation by Siebritz *et al.* (2011), the results showed that different communities will contribute different types of data depending on their motivation and interest. This study was hoping to build on from those conclusions, but the lack in point data for so many of the test areas did not allow for further conclusions.

Roads are unquestionably the most frequently generated data type. The CD: NGI is not the data custodian of roads and because the data sharing culture amongst governmental departments is still developing, roads are definitely an important feature to the CD:

NGI. The CD: NGI has identified new or modified roads as one of the major indicators of where topographical changes may have occurred to the landscape.

Chapter 6

INTEGRATION

6.1 Introduction

Integrating data sets from various sources may prove to be a difficult task, especially when integrating data from volunteer and authoritative sources. This dissertation was focused on issues associated with integrating data with different levels of accuracy for the various quality aspects. The discussions in the previous chapters have however shown that there are numerous other factors to consider even before assessing the accuracy. The technical factors include different reference systems, different representation of features, duplication of features, omission of attribute and metadata, different file formats, the physical and structural differences of spatial databases and different naming conventions. While these factors can be mitigated, some of the processes involved can become laborious for bigger data sets. There are also those factors concerned with policies, licensing and spatial data standards. Lastly, there are the differences in the quality control and quality assurance procedures for the CD: NGI and OSM. The degree to which these obstacles can be overcome and the effort involved to do so will determine the level of integration.

This chapter firstly summarises all the technical issues encountered in this investigation and the possible solutions for transforming the OSM data into the CD: NGI format. The next part makes recommendations as to how the legalities of integrating data distributed under different licenses or policies may be aligned. Comparison of the quality assurance and quality control is discussed thereafter and finally, a conceptual workflow for ingesting OSM data into the CD: NGI iTIS is presented.

6.2 Previous Investigations into Integrating Authoritative Data and OSM Data

There are very few documented cases of VGI officially being integrated into authoritative data models. Three different projects of VGI and authoritative data integration were discussed in section 2.3.3, including the on-line map editing service by the USGS. In addition, the organisation has also completed their first phase of its collaborative project with OSM, the OpenStreetMap Collaborative Prototype (OSMCP) (Wolf, Matthews, Mcninch and Poore, 2011). The project aims to determine whether the software developed for OSM may be used for data contributions that meet the USGS spatial data

standards (Wolf *et al.*, 2011). The first phase did not include contributions by volunteers, instead road data from the chosen partner was used (Wolf *et al.*, 2011). The road data was similar in accuracy to the requirements of the USGS, but it lacked in standard geometric representations (Wolf *et al.*, 2011). The customised OSM software was used for co-editing the roads data (Wolf *et al.*, 2011). Although it was not the OSM data being edited, the discussion on pre-processing mentions many of the tasks described in this investigation. The tasks included: “evaluating self-intersecting lines; running a geometry filter to check feature type; running a matching process to fix geometry problems, such as duplicate lines; testing to remove line slivers; running a snapping process to ensure connectivity in the new network; and, finally, converting the KS roads geometry into the USGS schema” (Wolf *et al.*, 2011). The project report also states that staff members required training in the software and that the training opportunities were scarce (Wolf *et al.*, 2011). The advantages of the system do however motivate further investigation. The project outcome showed that the customised OSM software does allow for collaborative editing (Wolf *et al.*, 2011). The inclusion of volunteer information is planned for the second phase of the project (Wolf *et al.*, 2011).

Fairbairn and Al-Bakri (2013) investigated the quality of OSM data with the intention of assessing possible integration. The OSM quality was assessed by comparison to two authoritative data sets, the Ordnance Survey UK data and General Directorate for Survey (GDS) Iraq data. From the results, the authors concluded that integrating the OSM data is not feasible. The authors state that VGI should not be used as a means to update authoritative databases, that NMAs should rather have no data than inaccurate VGI (Fairbairn and Al-Bakri, 2013).

Instead of enriching the authoritative data set, Pourabdollah, Morley, Feldman and Jackson (2013) focused their study on enriching the OSM data set using authoritative data from Ordnance Survey’s Open Data. This was done by either adding or correcting the road name and reference attributes and by flagging those roads in the OSM data set that were missing (Pourabdollah *et al.*, 2013). The enriched OSM data set is then served over open Web services (Pourabdollah *et al.*, 2013). The purpose for this process is to assist volunteers with contributing correct information (Pourabdollah *et al.*, 2013). The authors state that they encountered conflation issues such as, differences in the reference systems and file formats, heterogeneous road classifying and missing road names in the OSM data set (Pourabdollah *et al.*, 2013). While the OSM data set had more features in terms of the completeness, the geometric and attribute accuracies were lower than the OS data set (Pourabdollah *et al.*, 2013).

6.3 Technical Considerations

6.3.1 Different Reference Systems

Both the CD: NGI and OSM data sets are referenced to the WGS84 ellipsoid. The OSM data is however set to the Geographical Co-ordinate System (GCS), while the CD: NGI data is projected using the Transverse Mercator projection. Projected OSM data is not directly offered to users (Slater, personal communication, 2014 January 21). Transforming from the GCS to the Transverse Mercator projection is easily performed in the ArcGIS software. Apart from the Mpumalanga commercial data set, which displayed a

north-westerly shift, the OSM reference system appears to be consistently correct. This can only be said with certainty with further investigation.

6.3.2 Different Representations of Topographical Features

It is inevitable that features will be represented differently from one person to the next. The user and expert interpretation of features may differ vastly. In addition, the motivation for capturing features may also differ greatly. The CD: NGI experts are guided by the rules and standards employed by the organisation while, the users of OSM capture features as they believe would be the best representation. The CD: NGI experts capture features as accurately as possible because they are employed to do so. This means that because consistent data acquisition rules are being applied, the CD: NGI data is uniform. Volunteers, however mainly capture features out of interest and thus have no obligation to correctness, leading to heterogeneous data. The disadvantage of integrating the CD: NGI data with OSM data is that the CD: NGI data uniformity will be degraded.

Buildings

The preceding chapters have shown that OSM users capture more detail for buildings compared to the CD: NGI (see section 5.3). Automatic generalisation process may be applied to simplify OSM polygon features. For example, the tool that generates the minimum bounding rectangle may be used.

The CD: NGI standards specify that buildings that exceed a certain extent (M.D.C.U.) are not captured as point features, but as polygons. Figure 6.1 shows an example where a building is represented by both a point and polygon feature in the OSM data set. A method for automatically detecting these types of errors or other suspicious features is needed. This is especially necessary, as manual corrections are inevitable.

Figure 4.10 highlighted that OSM users also capture adjacent buildings as individual polygons, whereas the CD: NGI may generate a single polygon depending on the size of the buildings. The method presented in this investigation may be used (see section 4.5.1) or the generalisation tool, which aggregates polygons within a specified extent.

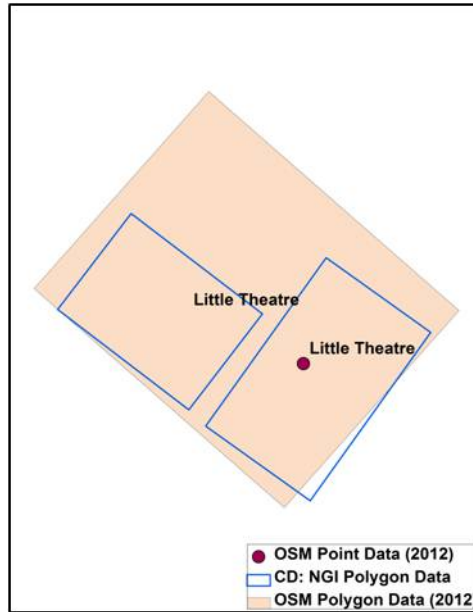


Figure 6.1: Example of different feature types representing the same feature in the OSM data set.

Roads

Figure 4.7 showed that multi-lane roads are represented by multiple centrelines in the OSM data set, but only a single centreline for the same road in the CD: NGI data set. The tool for automatically generating centrelines may be used, although in some cases human intervention is needed.

The results in figure 5.3 showed that in some areas the OSM roads are digitised correctly (i.e. a single road as a single feature), while in other areas not. The method developed to identify corresponding roads between the two data sets makes allowance for this issue (see section 4.4.1). The method may be taken a step further, where corresponding OSM road sections are consolidated into a single road feature.

6.3.3 Duplication of Features

The conversion to shapefile format renders duplicate OSM features (see section 4.3.1). For roads, identifying and removing the duplicate features is a simple task. Duplicate features are represented by negative OSM IDs, which makes the filtering process easy. For polygon features, duplication is rendered in the form of hole polygons. Removing the hole requires slightly more effort, but the process may be automated (see section 4.3.1).

6.3.4 Omission of Attribute and Metadata

The attribute accuracy was not investigated, but by visually inspecting the OSM data, it is seen that there are many features lacking attribute information. Determining whether this has improved over time requires further investigation. Attribute information may be useful to the CD: NGI but it not a necessity. Therefore, the differences in naming

conventions do not present a major problem for the type of integration being proposed in section 6.6.2. This is discussed in more detail in the following section on the processing of ancillary data. In terms of the metadata, users may add tags to the “source” field, which specifies how the data was created (Fairbairn and Al-Bakri, 2013). The data sets used for this investigation did however not contain those tag fields. It may be acquired by inspecting the JOSM file format as seen in figure 6.2. The information is contained in the “generator” and “origin” tags (indicated by the red and blue arrows).


```

<?xml version='1.0' encoding='UTF-8'?>
<osm version='0.5' generator='JOSM'>
  <bounds minlat='51.5076478723889' minlon='-0.127899783553507' maxlat='51.5077445145483' maxlon='-0.127774884645096' origin='OpenStreetMap server'
  <node id='26821100' timestamp='2009-02-16T21:34:57+00:00' user='dankarran' visible='true' lat='51.5077286' lon='-0.1279688'>
    <tag k='created_by' v='Potlatch 0.10f' />
    <tag k='name' v='Nelson&apos;s Column' />
    <tag k='tourism' v='attraction' />
    <tag k='monument' v='statue' />
    <tag k='historic' v='monument' />
  </node>
  <node id='-1' visible='true' lat='51.507661490456606' lon='-0.1278000843634869' />
  <node id='346364767' action='delete' timestamp='2009-02-16T21:34:44+00:00' user='dankarran' visible='true' lat='51.5076698' lon='-0.1278143' />
</osm>

```

Figure 6.2: Example of JOSM file format containing the source information (*OpenStreetMap Main Page*, 2013)

6.3.5 Different File Formats

OSM data is stored in the .osm file format. This is not a familiar file format, but it is convertible to other formats, like shapefile, which was used for this investigation. As mentioned earlier, some minor complications occur during the conversion process. This investigation did not commit a lot of time to OSM file conversions and perhaps a better option is available for the CD: NGI.

6.3.6 The Physical and Structural Differences of the Databases

Sections 1.2.1 and 1.2.3 summarised the structures of the CD: NGI database and the main OSM database, respectively. The two databases are vastly different in the way they function and store data. For example, the main OSM database stores features in nodes, tags, ways and relations, but the CD: NGI database stores points, lines and polygons. In addition, while the CD: NGI has one database; the OSM model employs various databases, which serve various functions (*OpenStreetMap Main Page.*, 2013). There are databases for storing user information, changesets and many other data (*OpenStreetMap Main Page.*, 2013). The proposed integration of the OSM data into the CD: NGI data model is not at the database level. The CD: NGI is nowhere near integration of this sort. Instead, integration is considered for small areas and even individual features.

6.4 Policies, Licensing and Spatial Data Standards

6.4.1 The CD: NGI Policy and OSM Licensing Concerning Data Distribution

The CD: NGI's policy for integrating ancillary data is outlined in a procedure document. Ancillary data is data that is produced outside of the organisation. A procedure document describes the workflow for a particular task. There is no policy restricting the CD: NGI to certain ancillary data sets, thus, the CD: NGI is free to use VGI. The CD: NGI is however very conservative about the ancillary data sets that are chosen. The data accuracy and liability associated with incorrect information is too great a concern. Considering that even experts are capable of producing erroneous information, perhaps the CD: NGI and NMAs in general should reconsider their view (Foody, See, Fritz, Velde, Perger, Schill and Boyd, 2013).

OSM operates under the Open Database License (ODbL), which allows users freedom of use (*OpenStreetMap Foundation.*, 2013). The license does not permit any user to sell the data (*OpenStreetMap Foundation.*, 2013). If however a user does use the OSM data as part of a derived product, the product including all data may be sold (*OpenStreetMap Foundation.*, 2013). In this instance, the OSM license requires that the user attributes OSM as one of the sources as follows: “©OpenStreetMap contributors” (*OpenStreetMap Foundation.*, 2013). In addition, the OSM derived data must be distributed under the same ODbL (*OpenStreetMap Foundation.*, 2013). This means that the client is allowed to distribute the data or derived product freely (*OpenStreetMap Foundation.*, 2013). If the derived product is not distributed, the user is not obliged to share the data (*OpenStreetMap Foundation.*, 2013).

The CD: NGI freely distributes all its products under the Copyright Act, No 98 of 1978, but charges a fee for hard copy maps to cover the direct costs (e.g. ink and printing). In this case, the CD: NGI policy regarding copyright does not infringe the requirements of the ODbL. Clients also have freedom of use, but no selling of the data or any derived product is allowed. However, data vendors may arrange an agreement with the CD: NGI, which allows them to sell the CD: NGI product or any derivative thereof (Du Plessis, personal communication 2014, January 17). This presents a problem. The CD: NGI may set up the agreement so that vendors are prohibited from selling the OSM derived data, but administering this will be challenging.

Attributing ancillary data sources is something that the CD: NGI already practices. The CD: NGI will however not agree to change its distribution rights to an ODbL. For this part, the CD: NGI and OSM must set up an agreement, which benefits and satisfies both parties. Currently, there is agreement between the CD: NGI and OSM, which allows OSM to upload the CD: NGI imagery. The CD: NGI was supposed to benefit by receiving regular change layers from OSM. Thus far, no change layers have been provided. The agreement does not make provision for ingesting OSM vector data. Perhaps the existing agreement can be modified to include data ingestion.

These legal issues require further investigation, with consideration to legislation like the SDI Act and the Promotion of Access to Information Act (PAIA), 2000 (Act 2 of 2000) (Du Plessis, personal communication 2014, January 17).

6.4.2 Adherence to the CD: NGI Spatial Data Standards

In terms of accuracy, the OSM data generally does not meet the CD: NGI requirements. One of the research questions posed in section 1.5.2 was; is it necessary to meet all the accuracy requirements in order to use the OSM data? The answer to this question is no. According to the process for ancillary data (discussed in the following section), the CD: NGI requires a high positional accuracy. The attribute information, although it would be helpful, is not a necessity. The question, which then follows on from this is; (not a research question) is it necessary that the OSM positional accuracy meet the CD: NGI requirements for the entire country in order to use the OSM data? The answer to this question is also no. The results showed that in some areas the OSM data was very close to the CD: NGI positional accuracy requirements. Thus, the OSM data is still useful for updating these areas. It may even be useful in areas where the CD: NGI data set is lacking, although the positional accuracy is lower. Otherwise, the CD: NGI may use the OSM data as a tool for detecting where and how the landscape has changed only. These two possibilities are presented in the following sections.

6.5 Differences in Quality Assurance and Quality Control Processes

Section 3.2 discussed the methods employed by the CD: NGI to ensure high quality data and products. The implementation of spatial data standards and rules for capturing topographic features are used to achieve the desired quality. While the QA division is responsible for ensuring that these standards are maintained, every operator is respon-

sible for adhering to the standards. The workflow in appendix B1 indicated that the operators perform three visual inspections during a compilation job (the size of a 1:50 000 map sheet). The checks are put in place to identify inaccurate features. Other than topological tests within the Geomedia environment, the CD: NGI has no tools for automatically detecting errors.

The OSM initiative employs no official quality assurance procedures. Section 3.4.3 did however summarise the various quality control techniques. The tools for detecting bugs may be used automatically, semi-automatically or manually (*OpenStreetMap Main Page.*, 2013). The error detection tools run automatically, identifying potential data errors (*OpenStreetMap Main Page.*, 2013). There is a variety of error detection tools available, although not all of them are available in South Africa.

6.6 The Process for Acquiring and Processing Ancillary Data

The Survey Services Division at the CD: NGI is currently responsible for acquiring and processing ancillary data. The process for sourcing and processing ancillary vector data is presented in appendix E1. The CD: NGI has an annual production plan, which is used to identify areas where topographic vector data is needed. Possible data sources are then identified and the necessary data obtained. The task descriptions for each stage in the processing of ancillary data are:

1. Refine ancillary data —The data is visually inspected for consistency and accuracy. Spatial attributes and projection parameters are verified. The operator may perform some accuracy analysis. The data is formatted if necessary. Geographical names are extracted for simulation into the iTIS and boundary information is made available to the Topographical Compilation Division. (Chief Directorate: National Geo-Spatial Information, 2009).
2. Quality control of ad hoc processed ancillary data sets —During this process, sample sets are chosen and the positional accuracy is tested (Jansen, personal communication 2014, January 8). In the case of road and point features, a simple buffering technique may be used or roads may be inspected visually against the ortho-rectified imagery (Jansen, personal communication 2014, January 8). Polygon features are inspected visually in conjunction with the imagery, cadastral data and other surrounding topographical features (e.g. rivers) (Jansen, personal communication 2014, January 8).
3. Archive ad hoc ancillary data sets —The data is archived (not into the iTIS) and made available on the network to other mapping divisions.

The CD: NGI never ingests an entire data set, even after exhaustive accuracy testing (Jansen, personal communication 2014, January 8), (McLachlan, personal communication, 2014 January 8). Instead, what has been done in the past, the ancillary data set is used to create a base layer separate to the iTIS base layer, which is then used as a reference in the topographical compilation process (Jansen, personal communication

2014, January 8). If an area has poor data coverage and the ancillary data is proved accurate enough, small amounts of data may be ingested. However, even this does not occur often (McLachlan, personal communication, 2014 January 8).

6.6.1 Using OSM Data for Change Detection

Currently, the change detection process at the CD: NGI is accomplished by differencing the source and target vector data sets (Jansen, personal communication 2014, January 8). The change constitutes those features that are not present in the CD: NGI data set. For the change detection process, the ancillary data sets undergo the refining process only (Jansen, personal communication 2014, January 8). The change differences are generated, analysed, reformatted if necessary and made available to the Topographical Compilation Division. This aids in prioritising areas where data capture is most needed (Jansen, personal communication 2014, January 8). The OSM data can easily be incorporated to detect landscape changes following this process, but in an automated technique is needed (see figure 6.3).

The change layer, which is produced, is also useful in providing the CD: NGI with other indicators. It provides hints about the type of features and the type of areas that people are interested in and the level of detail people prefer (e.g. building polygons with no generalisation applied). Again, this statistical information aids with prioritising areas for data collection.

6.6.2 Proposed Integration of the CD: NGI and OSM Data

The level of integration between data from different sources is determined by the ability to overcome those factors, which hinder integration. Considering the CD: NGI's current manual quality control methods, a high-level integration may not be feasible. For this type of integration, the targeted database is transformed, mainly automatically, into the required format. The desired features may then be extracted and merged with the (in this case) authoritative database. At a lower integration level, the databases are not merged; instead the features are ingested in part. For example, only the geometry or only some of the attribute fields are ingested. At an even lower level of integration, the data from the different sources remain separable. In this case, maintaining the topological integrity becomes challenging. At all three levels, but especially for the lowest level, the metadata becomes crucial. A proposal for a level two-type integration for the CD: NGI and OSM data is given below.

The proposed work flow for ingesting OSM data into the iTIS is presented in figure 6.3. It is based on the existing task descriptions discussed section 6.6. However, the proposed workflow incorporates a transformation process whereby the OSM data is manipulated to resemble the CD: NGI geometry. The transformation process (stage 3 in figure 6.3) includes the technical issues discussed in 6.3.1. Also, the geographical names are extracted at this stage and not at the refining stage. Thereafter, the quality assessment is done (stage 4 in figure 6.3). The CD: NGI does not use the attribute information from ancillary data, only the geometry. The existing methods for assessing the positional accuracy must be modified to better suit the OSM data.

The next step would be to generate a base layer (stage 5 in figure 6.3) as described in section 6.6. Features will be ingested into the iTIS (stage 6 in figure 6.3) if the positional accuracy is high or if the CD: NGI has poor coverage in an area. The completeness assessment has revealed that the CD: NGI data is lacking for some areas, at least for roads. Attribute information will be added in accordance with the CD: NGI spatial data standards.

Unlike other ancillary data sets, OSM allows for frequent updating, because volunteers contribute data daily. Thus, the CD: NGI could generate several base layers per year. The metadata is crucial when updated OSM base layers are frequently generated.

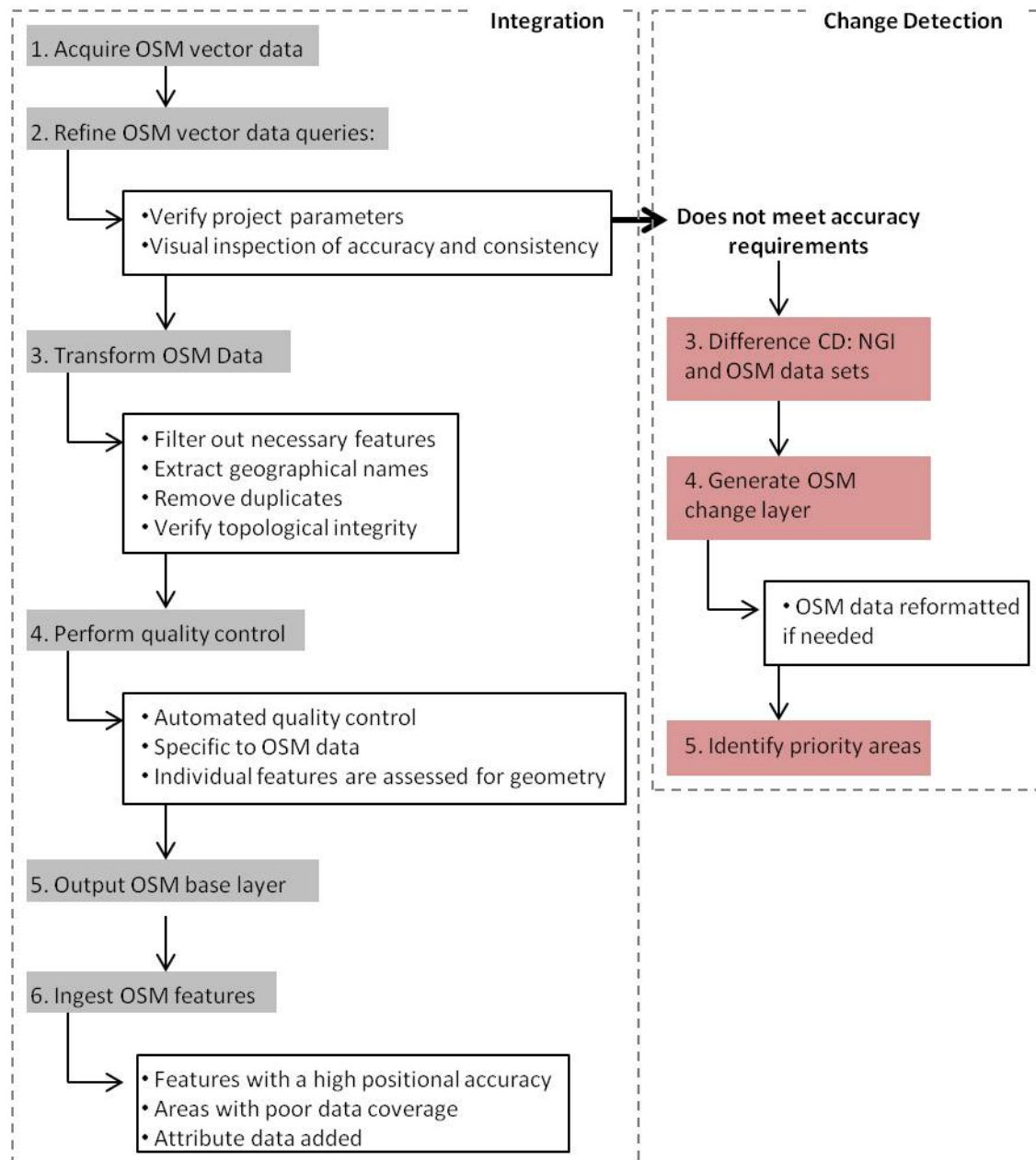


Figure 6.3: Process for ingesting OSM data into the CD: NGI iTIS

6.6.3 Institutional Reorganisation Needs

The current verification process is time consuming because there is a shortage of staff. Currently, only the senior operators are allowed to perform the quality control and understandably so (Jansen, personal communication 2014, January 8). Correctly identifying features and their positions is a task that requires not only knowledge but also experience. There are but a few senior operators available for the task. The proposed integration requires a rigorous quality control technique, which poses a problem with the current staff situation.

In 2012 the CD: NGI set out the following strategic objective:

By 2017 geo-spatial information in the iTIS will have a currency of no more than three years with 99% of the changes to all main features updated within 12 months of the change occurring, in compliance with the prescribed standards (Chief Directorate: National Geo-Spatial Information Management, 2013)

Before then, there was a five year revision cycle for all features for the entire country. This target proved to be impossible. In the light of this, generating consecutive change or base layers cannot occur often and thus the updating of features will be delayed. The solution for this is to invest in an automated technique for extracting, processing and verifying OSM data.

6.7 Analysis and Discussion

This chapter has highlighted the various factors concerned with integrating the CD: NGI and OSM topographical data. Performing an investigation on the accuracy of the OSM data has assisted in identifying the many factors that must be considered. Many of the technical issues are common to previous investigations (Wolf *et al.*, 2011), (Fairbairn and Al-Bakri, 2013) and (Pourabdollah *et al.*, 2013). This dissertation has demonstrated that these issues may be mitigated; some by automatic methods and others by manual methods. The CD: NGI has not invested in the development of in-house automated tools, whether it is for data transformations or quality control. OSM, has with the help of users with various expertise made significant progress in the this area. Most of the tools for detecting OSM data errors are automated.

The legal issues involved with ingesting OSM data need further investigation. Although, the Copyright Act No 98 of 1978 supports the ODbL requirements, complications arise with the existing and future agreements between the CD: NGI and data vendors. In depth, knowledge of the act and the ODbL is needed to come to a sound agreement between the CD: NGI and OSM.

Because the the accuracy of OSM data is heterogeneous across the country, the CD: NGI cannot ingest the entire OSM data set, but the data should not be disregarded either. Instead, ingestion may be limited to those areas with a high positional accuracy or areas where there is poor data coverage. Another option is to strictly use the OSM data to detect changes to the landscape by performing a differencing between the CD: NGI and OSM vector data sets.

A conceptual workflow for integrating OSM data into the CD: NGI iTIS was presented. Because, the CD: NGI already has procedures in place for sourcing and processing ancillary data, the conceptual workflow designed for the OSM data was not a difficult task. Aside from the lack in human resources, the CD: NGI does require other changes to the current workflow to account for the specifics of the OSM data. For example, the method of change detection and quality control requires automation.

Chapter 7

CONCLUSIONS AND RECOMMENDATIONS

7.1 Introduction

NMAs worldwide are struggling to maintain current spatial data, because the costs involved are high. In an attempt to reduce production costs, NMAs have considered VGI as a possible data source. Although VGI may be a valuable source of current spatial data, it must be proved that the quality meets the NMA mapping standards. This study was focused on firstly determining the quality of the OSM data. The second objective was to investigate how the integration of OSM data with the CD: NGI data would work and the factors to consider.

The working hypothesis was that VGI meets the CD: NGI national mapping standards and can thus be used for updating the CD: NGI topographical database. The methods used to assess the OSM accuracy aspects were based on the assumption that the CD: NGI reference data set is of a higher accuracy. The first part of this chapter concludes on the quality of the OSM data. The second part provides a summary on the proposed integration of the CD: NGI and OSM data.

7.2 OSM Positional Accuracy

In terms of the three quantitative assessments, the positional accuracy results were identified as the most important quality indicator for the CD: NGI. The CD: NGI has an existing workflow for acquiring and processing ancillary data. During this process, except for geographical names, none of the attributes are extracted. Attribute information is thus not crucial. Even so, it was a good exercise to see how well the OSM data compares with authoritative data for the other quality elements.

The positional accuracy of the OSM road and building data is heterogeneous across the country. The average percentages of roads that are within 10 m of the CD: NGI roads, range from 65% to 94% for the nine provinces. The percentages per settlement category are less varied; however, it is still clear that low urban density areas tend to have a lower positional accuracy. It was seen that for many test areas, the number of features influenced the positional accuracy. Test areas with the lowest feature count (which in most

cases was low urban density areas) resulted in the lowest weighted average percentages.

For road features, none of the test areas met the CD: NGI's 95% confidence level requirement. Although, four provinces had average percentages very close to 95%. Considering the 5% error within the CD: NGI data, these four data sets may even have a higher absolute positional accuracy. Further investigation is needed to verify this conjecture.

It was important to compare both the position and shape of the OSM polygon features. There were very few test areas containing polygon features, so the results were specific to those that had polygon features. For both the commercial and residential test areas, the weighted average Hausdorff distances exceeded the 10 m threshold by 1.29 m and 2.54 m, respectively. There were no pre-defined standards for the polygon shape assessments. The results however compare well for commercial areas but are less consistent for residential areas.

Generally, the CD: NGI applies generalisation techniques when capturing topographic features. Adjacent buildings may be grouped into a single polygon and multi-sided buildings may be represented by a simplified polygon. Until 2 years ago, the CD: NGI's focus was on cartographic representation, which means that features may be modified to better suit the map layout. This process obviously detracts from both the true position and shape of polygon features. Thus, the OSM geometric accuracy may be more comparable to the features as it exists on the ground.

7.3 Qualitative Aspects of the OSM Data

It was not sufficient to only report on the quantitative measures. The qualitative aspects were focused on the rate of data generation, uniformity of data acquisition and the usefulness of the OSM data. This section summarises the various forms of heterogeneity that exist in the OSM data. There is heterogeneity in: i) the volume of data contributed across SA, ii) the type of features contributed and iii) the way people compile topographical features. This is because of different user motivations, different interests and backgrounds, different levels of spatial knowledge and the availability of resources. As long as these factors exist and OSM remains an open initiative without any feature compilation rules, the data will continue to be heterogeneous. OSM has made no mention of changing their guidelines regarding feature compilation.

7.3.1 Heterogeneity in OSM Data Volumes

The results in section 5.3.1 showed that road contributions in low urban density areas are consistently low across the country compared to residential and commercial areas. The completeness investigation in section 5.2.4 confirmed that it is the number of contributions that are low and not the number of features existing in the area. Although the CD: NGI is on a mission to source current spatial data, the focus has specifically been on low urban density areas. However, in most of the commercial test areas and some of the residential areas, where the OSM data exceeds the CD: NGI data, the OSM data may still be useful.

Growth Rate

Assessment of the OSM currency comprised two aspects, the growth rate (i.e. is the data repository growing?) and the data stability (i.e. does the data reach a state of completion?). The repository has definitely grown significantly, more so in commercial and residential areas. This was deduced from the graphs in section 5.3.1.

The completeness results in section 5.2.4 showed that most of the test areas reached a point where the data evens out before or during 2011. It was seen that a 100% completeness did not mean that the OSM data set had reached its maximum contributions. The method for computing the completeness makes it impossible to determine the absolute completeness because; the completeness percentages exceed 100% (i.e. commission). When a test area reached its peak at a completeness percentage below 100% (that is, omission), it meant that users had stopped contributing data. This was the case for 14 of the 27 data sets. What this demonstrates, is that OSM has the potential to provide CD: NGI with sufficient data for commercial and some residential areas. The data volumes in low urban density and some residential areas are insufficient. The question may arise; will the volunteers continue to contribute data in the future? Further investigation will determine if those features in the OSM low urban density data sets exist in the CD: NGI data set. If they do not and those features are valuable to the CD: NGI, the contributions may still serve as complementary data.

OSM Data Stability

The fact that the amount of unchanged data within a test area evens out (see section 5.3.1), indicates that very few or even no modifications and deletions are being made. This implies that the data is considered correct (Siebritz *et al.*, 2011). Either users are making contributions that are more accurate or the existing data has been improved because more error detection tools are available (see section 3.4.3).

7.3.2 Heterogeneity in OSM Data Acquisition

Users will contribute different features depending on culture, motivation and education (Siebritz *et al.*, 2011). The results for point feature acquisition in section 5.3.2 were inconclusive for residential areas and could not support this statement. In commercial areas, volunteers had the greatest interest in the “leisure” category.

The semantic accuracy results indicated that for roads, the OSM “residential” class had the most contributions. This does not however mean that volunteers prefer to contribute to this road class, but that they prefer it as the default classification. The result is incorrect road categorising. Thus, even in this, it can be said that volunteers preferences are influenced by their knowledge and level of interest, resulting in heterogeneity in data acquisition.

In terms of the three data types (points, lines and polygons), line features or roads have definitely had the most contributions. This is expected as OSM was originally intended for road networks. There is therefore also heterogeneity in data acquisition between data types.

7.4 Integration Opportunities

The processes needed to transform the OSM data into the most usable format for the CD: NGI can become tedious if manual methods are used. The methods used in this investigation were semi-automatic. It is possible to modify the techniques so that they are automatic. Verification of the transformed data will inevitably include manual checking and editing. This is mainly because the CD: NGI currently performs manual checks, but also because there will always be the concern about the credibility of VGI.

Section 6.6 presented two possible uses for the OSM data at the CD: NGI. The data may be used for flagging changes to the landscape that have not yet been captured in the iTIS. This is a simple process, which does not require much pre-processing, but does require an automated technique. The other possibility is to ingest OSM feature geometry for those features with sufficient positional accuracy (geometric accuracy for polygons) and where the CD: NGI has poor data coverage. The challenges associated with this proposal include a shortage of staff and expertise.

7.5 Analysis and Discussion

The investigation has proved that the hypothesis is false in part (see table 7.1). For some of the assessments, like the semantic accuracy for two of the road classes, the completeness in low urban density areas and the currency of point and polygon features, the accuracies are low. However, for the other assessments, like the semantic accuracy of the “residential” class, the completeness for most of the commercial test areas and some of the residential test areas, as well as the currency of roads in commercial areas, the OSM data does meet the CD: NGI requirements. In terms of the positional accuracy, none of the provinces met the requirements, but some of the accuracies were very close. Thus, the data may still be good for updating the CD: NGI data, but that decision lies with the organisation.

The assumption that the CD: NGI reference data set is more accurate is not true for every accuracy aspect. This was clearly not the case for the completeness. Further investigation is needed to determine other instances where this assumption does not hold true.

Table 7.1: Table showing the level of compatibility of OSM data with the CD: NGI data

	OSM COMPATIBILITY			
	Yes	No	Mostly	Partly
Reference System	✓			
File Format	✓			
Topology			✓	
Relevant Features			✓	
Semantics		✓		
Positional Accuracy —Roads				✓
Positional Accuracy —Polygons				✓
Shape Accuracy —Polygons			✓	
Completeness			✓	
Licensing			✓	

7.6 Recommendations and Future Work

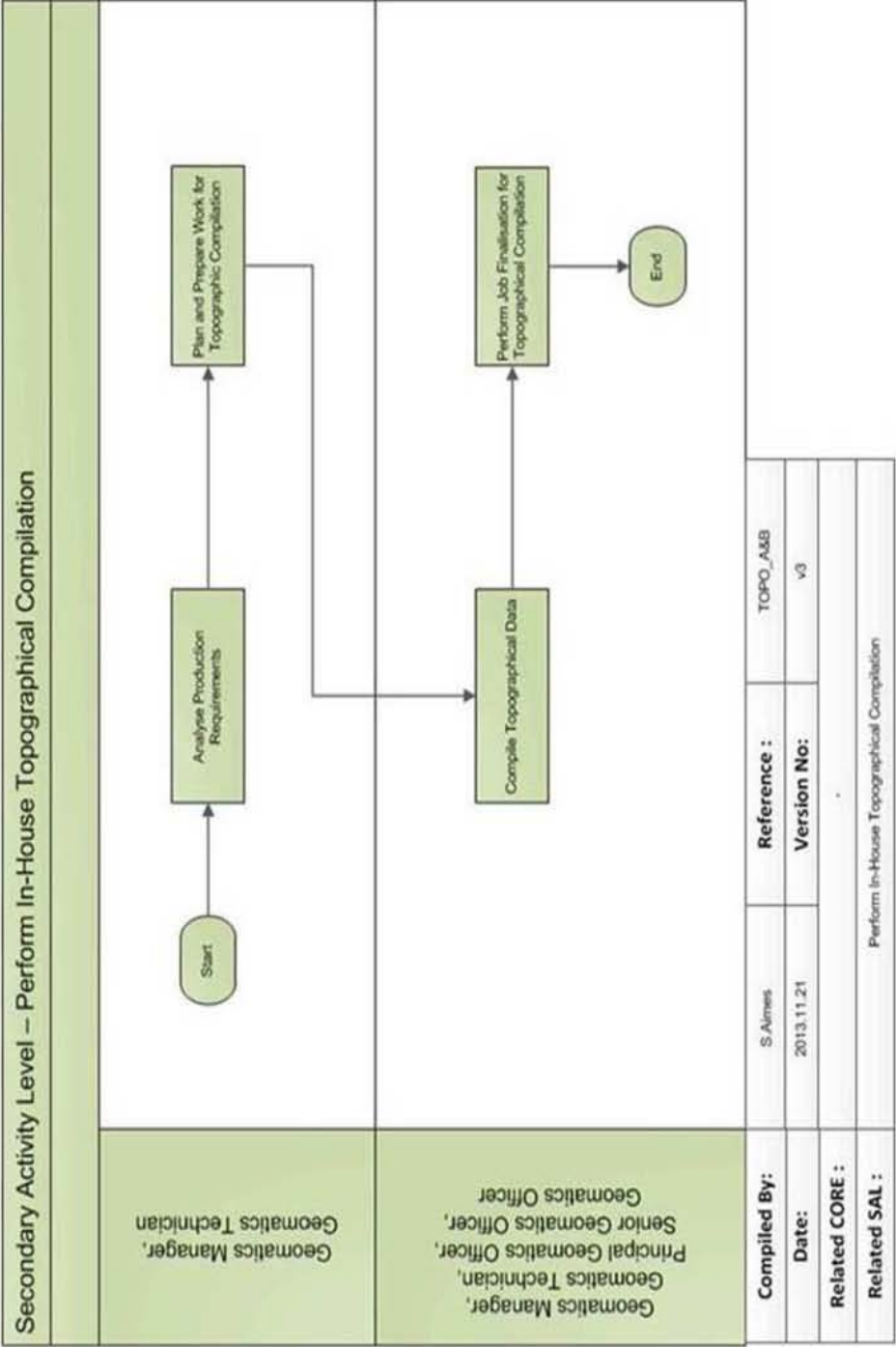
This investigation has revealed the need for two automatic techniques within the CD: NGI's current workflow. One technique is needed for a dynamic change detection process and another for the transformation and verification of ancillary data. This is necessary because the current methods are not feasible for large data sets. Both the change detection and verification process are time consuming. As a result, the updating process is delayed. The development and implementation of these techniques may take some time. It is thus recommended that during this developmental stage, that junior operators in the Ancillary Data Division be trained in the current change detection and verification methods. In this way, senior operators are available to assist with the development of the automated techniques. After successful implementation of the techniques, senior operators may perform other tasks like administering and maintaining various OSM base layers, instead of verification.

Another option is to establish partnerships with OSM contributors where they are trained to make contributions that satisfy the CD: NGI requirements. This may be achieved by offering free workshops in the various provinces. Only those contributions made by the partners will then be considered. These contributions can easily be identified and separated by the user log-in details. However, both parties must benefit from the partnership. Volunteers must be motivated to contribute data in alignment with the CD: NGI requirements. They may be motivated if they are being publicly acknowledged as data contributors. Providing small monetary incentives is another option. It would thus be good to conduct a study on the type of people who contribute data to OSM in SA and the motivations behind the contributions.

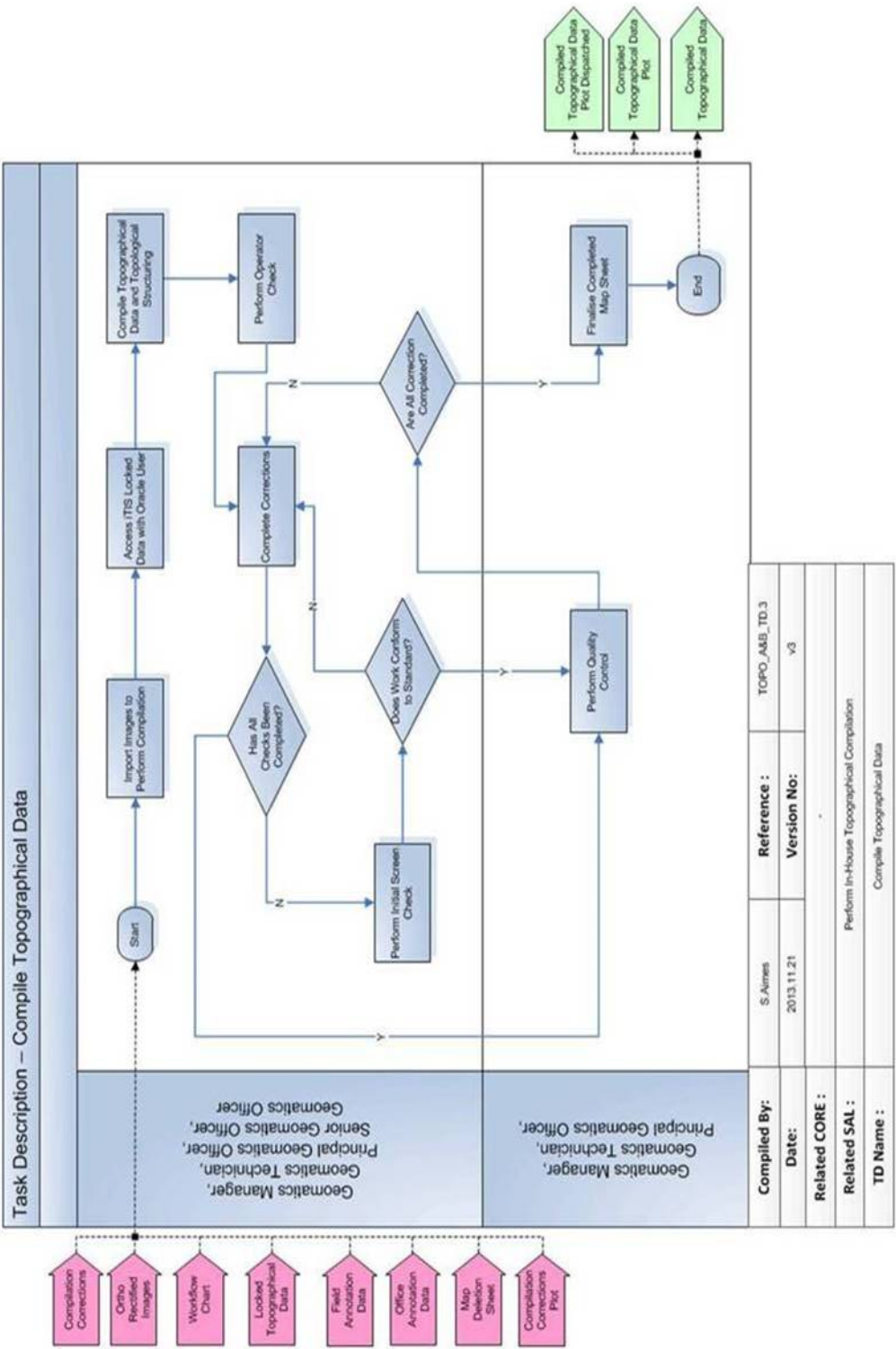
OSM has implemented various mechanisms for detecting data errors, but perhaps quality assurance techniques should be considered. For example, incorporating attribute domain lists, which force volunteers to choose attributes from a pre-defined list. This is just one of the mechanisms that could be used to decrease heterogeneity.

APPENDICES

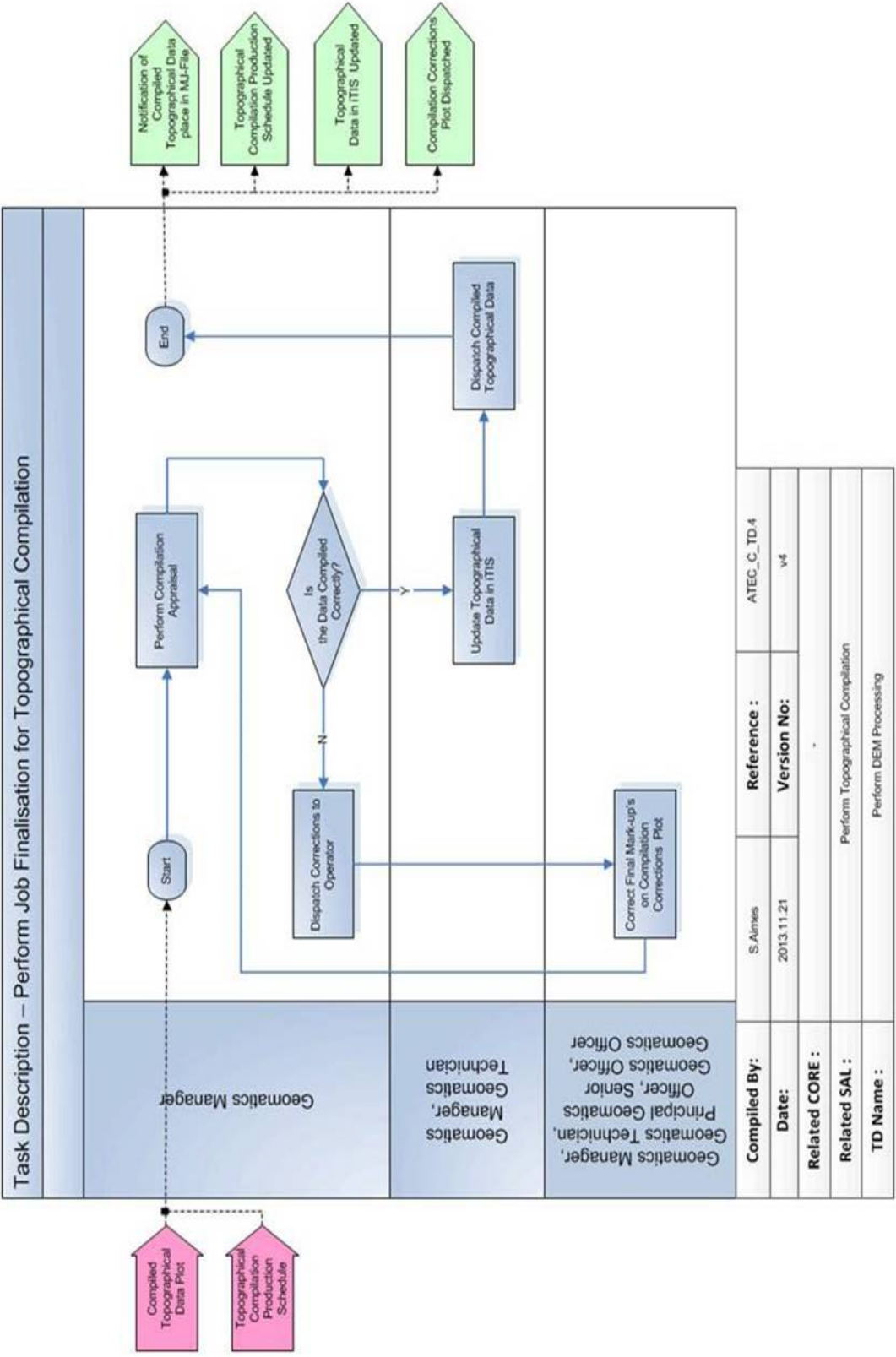
A1. Appendix A: Secondary Activity Level (Chief Directorate: National Geo-Spatial Information, 2013a)



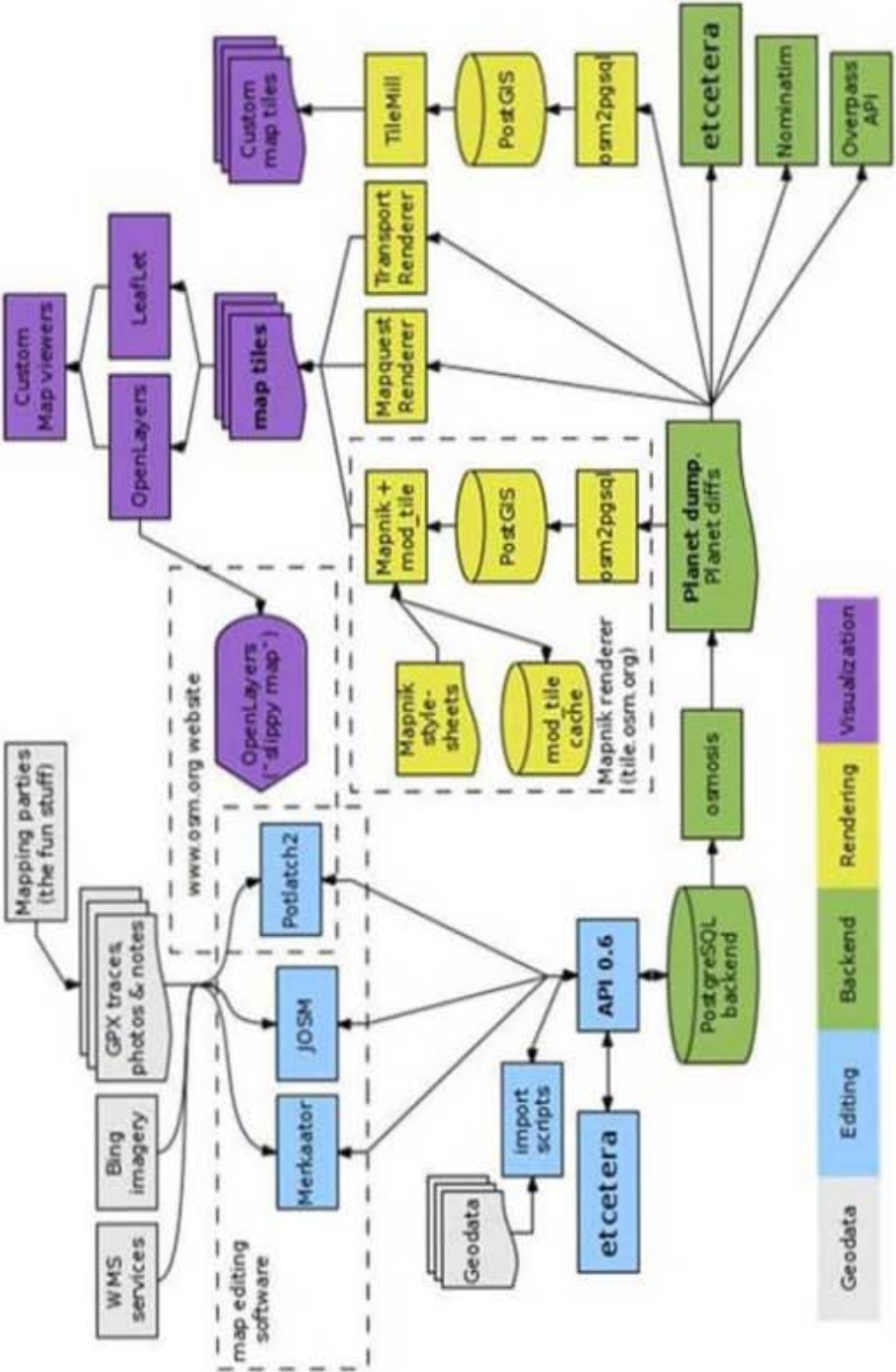
B1. Appendix B: CD: NGI Process for Compilation of Topographical Features (Chief Directorate: National Geo-Spatial Information, 2013b)



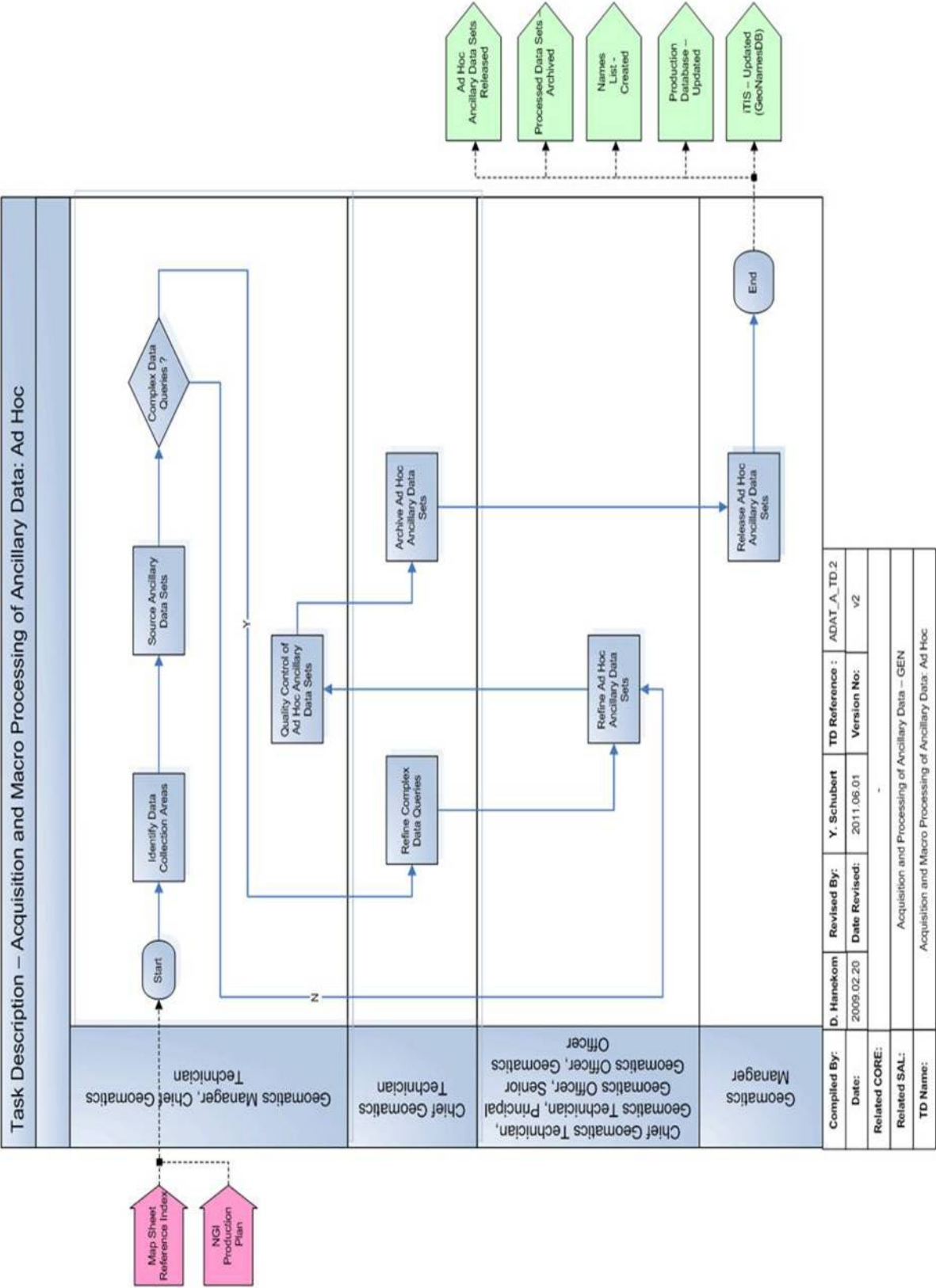
C1. Appendix C: Data flow for after approved compilation task (Chief Directorate: National Geo-Spatial Information, 2013b)



D1. Appendix D: OSM Data Model Components (*OpenStreetMap Main Page*, 2013)



E1. Appendix E: Data flow for acquisition and macro processing of ancillary data: ad hoc (Chief Directorate: National Geo-Spatial Information, 2009)



F1. Sample Python scripts as discussed in section 4.4.1

Script to compute the angle of intersection between the CD: NGI and OSM roads (angle a in figure 4.6)

```
import arcpy, sys, string, os, arcgisscripting, glob, math
from arcpy import env

arcpy.env.workspace = r"G:\Local Disk (D)\Projects\MSc Thesis\Vector Data\Freds_Full
Historical Dataset\October Test Areas\Western Cape\3318\Remove Angle\Commercial\Joined
NGI and OSM"
fclist = arcpy.ListFeatureClasses()

for fc in fclist:
    arcpy.AddField_management(fc, "Angle", "DOUBLE", "", "", "", "", "NON_NULLABLE",
        "NON_REQUIRED", "")

    NX0 = ("NGL_X0")
    NY0 = ("NGL_Y0")
    NY1 = ("NGL_Y1")
    NX1 = ("NGL_X1")
    OX0 = ("OSM_X0")
    OY0 = ("OSM_Y0")
    OY1 = ("OSM_Y1")
    OX1 = ("OSM_X1")
    Angle = ("Angle")

    rows = arcpy.SearchCursor(fc)

    for row in rows:

        if (row.getValue(NY0)) > (row.getValue(OY0)):

            dx = math.fabs(row.getValue(NX1)) - math.fabs(row.getValue(NX0))
            dy = math.fabs(row.getValue(NY1)) - math.fabs(row.getValue(NY0))

            dxN = math.fabs(row.getValue(OX0)) - math.fabs(row.getValue(OX1))
            dyN = math.fabs(row.getValue(OY0)) - math.fabs(row.getValue(OY1))

            if dx > 0 and dy > 0:
                directionN = (math.atan(math.fabs(dy)/math.fabs(dx)))*180/math.pi
            elif dx > 0 and dy < 0:
                directionN = 180 - ((math.atan(math.fabs(dy)/math.fabs(dx)))*180/math.pi)
            elif dx < 0 and dy < 0:
                directionN = 180 + ((math.atan(math.fabs(dy)/math.fabs(dx)))*180/math.pi)
            elif dx < 0 and dy > 0:
                directionN = 360 - ((math.atan(math.fabs(dy)/math.fabs(dx)))*180/math.pi)

            if dxN > 0 and dyN > 0:
```

```

        directionO = (math.atan(math.fabs(dyN)/math.fabs(dxN)))*180/math.pi
    elif dxN > 0 and dyN < 0:
        directionO = 180 - ((math.atan(math.fabs(dyN)/math.fabs(dxN)))*180/math.pi)
    elif dxN < 0 and dyN < 0:
        directionO = 180 + ((math.atan(math.fabs(dyN)/math.fabs(dxN)))*180/math.pi)
    elif dxN < 0 and dyN > 0:
        directionO = 360 - ((math.atan(math.fabs(dyN)/math.fabs(dxN)))*180/math.pi)

    if directionO > directionN:
        angle = directionO - directionN
        if 90 < angle < 180:
            angle = 180 - angle
            print fc, angle

    elif 180 < angle < 270:
        angle = angle - 180
        print fc, angle

    elif 270 < angle < 360:
        angle = 360 - angle
        print fc, angle

    elif angle < 90:
        angle = directionO - directionN
        print fc, angle

    elif directionO < directionN:
        angle = directionN - directionO
        if 90 < angle < 180:
            angle = 180 - angle
        elif 180 < angle < 270:
            angle = angle - 180
            print fc, angle

        elif 270 < angle < 360:
            angle = 360 - angle
            print fc, angle

        elif angle < 90:
            angle = directionN - directionO
            print fc, angle

    elif (row.getValue(NY0)) < (row.getValue(OY0)):
        dx = math.fabs(row.getValue(NX1)) - math.fabs(row.getValue(NX0))
        dy = math.fabs(row.getValue(NY1)) - math.fabs(row.getValue(NY0))

        dxN = math.fabs(row.getValue(OX1)) - math.fabs(row.getValue(OX0))
        dyN = math.fabs(row.getValue(OY1)) - math.fabs(row.getValue(OY0))

```

```

if dx > 0 and dy > 0:
    directionN = (math.atan(math.fabs(dy)/math.fabs(dx)))*180/math.pi
elif dx > 0 and dy < 0:
    directionN = 180 - ((math.atan(math.fabs(dy)/math.fabs(dx)))*180/math.pi)
elif dx < 0 and dy < 0:
    directionN = 180 + ((math.atan(math.fabs(dy)/math.fabs(dx)))*180/math.pi)
elif dx < 0 and dy > 0:
    directionN = 360 - ((math.atan(math.fabs(dy)/math.fabs(dx)))*180/math.pi)

if dxN > 0 and dyN > 0:
    directionO = (math.atan(math.fabs(dyN)/math.fabs(dxN)))*180/math.pi
elif dxN > 0 and dyN < 0:
    directionO = 180 - ((math.atan(math.fabs(dyN)/math.fabs(dxN)))*180/math.pi)
elif dxN < 0 and dyN < 0:
    directionO = 180 + ((math.atan(math.fabs(dyN)/math.fabs(dxN)))*180/math.pi)
elif dxN < 0 and dyN > 0:
    directionO = 360 - ((math.atan(math.fabs(dyN)/math.fabs(dxN)))*180/math.pi)
directionN = 0

if directionO > directionN:
    angle = directionO - directionN
    if 90 < angle < 180:
        angle = 180 - angle
        print fc, angle

    elif 180 < angle < 270:
        angle = angle - 180
        print fc, angle

    elif 270 < angle < 360:
        angle = 360 - angle
        print fc, angle

    elif angle < 90:
        angle = directionO - directionN
        print fc, angle

    elif directionO < directionN:
        angle = directionN - directionO
        if 90 < angle < 180:
            angle = 180 - angle
            print fc, angle

        elif 180 < angle < 270:
            angle = angle - 180
            print fc, angle

        elif 270 < angle < 360:

```

```

        angle = 360 - angle
        print fc, angle

    elif angle < 90:
        angle = directionN - directionO
        print fc, angle

rows2 = arcpy.UpdateCursor(fc)
for row2 in rows2:
    row2.Angle = angle
    rows2.updateRow(row2)

```

Script to remove all road lengths shorter than the lowerbound (shorter than Bw(cos(90-a)) in figure 4.6)

```

arcpy.env.workspace = r"G:\Local Disk (D)\Projects\MSc Thesis\Vector Data\Freds.Full
Historical Dataset\October Test Areas\Western Cape\3318\Remove Angle\Commercial\Angles
to Remove-Lengths"

fclist2 = arcpy.ListFeatureClasses()

for featureclass2 in fclist2:

    arcpy.AddField_management(featureclass2, "Len_Segs", "DOUBLE", "", "", "", "",
    "NON_NULLABLE", "NON_REQUIRED", "")
    arcpy.CalculateField_management(featureclass2, 'Len_Segs', '!shape.length@meters!', 'PYTHON')
    desc = arcpy.Describe(featureclass2)

    fc_copied = r"G:\Local Disk (D)\Projects\MSc Thesis\Vector Data\Freds.Full
    Historical Dataset\October Test Areas\Western Cape\3318\Remove Angle\Commercial\Short
    Lengths Removed" + "\\\" + str(featureclass2)
    arcpy.CopyFeatures_management(featureclass2, fc_copied)

    lyr = r"G:\Local Disk (D)\Projects\MSc Thesis\Vector Data\Freds.Full
    Historical Dataset\October Test Areas\Western Cape\3318\Remove Angle\Commercial\Short
    Lengths Removed" + "\\\" + desc.name.replace(".shp", "") + ".lyr"
    arcpy.MakeFeatureLayer_management(fc_copied, lyr)

    rows = arcpy.SearchCursor(fc_copied)

    for row in rows:
        row = row.getValue("Len_Segs")
        rads = 65*math.pi/180
        lowerbound = 20/(math.cos(rads))
        if row < lowerbound:

            selection = arcpy.SelectLayerByAttribute_management (lyr, "NEW_SELECTION",
            "Len_Segs <" + str(lowerbound))
            output = r"G:\Local Disk (D)\Projects\MSc Thesis\Vector Data\Freds.Full

```

```
Historical Dataset\October Test Areas\Western Cape\3318\Remove Angle\Commercial\Short  
Lengths Removed" + "\\\" + str(fc_copied)  
arcpy.DeleteFeatures_management(lyr)
```


Bibliography

- Al-Bakri & Fairbairn. (2010). Assessing the accuracy of crowdsourced data and its integration. *in* 'Accuracy 2010 Symposium'. Leicester.
- Al-Bakri & Fairbairn (2011). User generated content and formal data sources for integrating geospatial data. *in* 'Proceedings of 25th International Cartographic Conference'. Paris. p. 8.
- Anand, Morley, Jiang, Heshan & Hart (2010). When worlds collide: combining Ordnance Survey and Open Street Map data. *AGI Geocommunity 10* p. 7.
- Antoniou (2011). User Generated Spatial Content: An Analysis of the Phenomenon and its Challenges for Mapping Agencies. PhD thesis. University College London.
- Ather (2009). A quality analysis of OpenStreetMap data. Masters thesis. University College of London.
- Baglatzi, Kokla & Kavouras (2012). Semantifying OpenStreetMap. *in* 'The 11th International Semantic Web Conference'. Boston. pp. 39–48.
- Basiouka & Potsiou (2012). VGI in Cadastre: a Greek experiment to investigate the potential of crowd sourcing techniques in Cadastral Mapping. *Survey Review* **44**(325). 153–161.
- Baskaran & Muchie (2006). *Bridging the Digital Divide: Innovation Systems for ICT in Brazil, China, India and South Africa*. 1st edn. Adonis & Abbey Publisher Ltd. London.
- Behrens (2011). Segmentation of OpenStreetMap Data-Generating, Merging, and Distributing Tiles. PhD thesis. University of Bremen.
- Bernard (2002). Experiences from an implementation Testbed to set up a national SDI. *in* '5th AGILE Conference on Geographic Information Science'. Spain. p. 9.
*Available: <http://ifgi.uni-muenster.de/bernard/publications/AGILE2002.Bernard>
- Bishop, Escobar, Karuppannan, Suwarnarat, Williamson, Yates & Yaqub (2000). Spatial Data Infrastructure for cities in developing countries: Lessons from the Bangkok Experience. *Cities* **17**(2). 85–96.
- Boin & Hunter (2006). What communicates quality to the spatial data consumer. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* **34**. 8.
- Budhathoki, Bruce & Nedovic-Budic (2008). Reconceptualizing the role of the user of spatial data infrastructure. *GeoJournal* **72**(3-4). 149–160.
*Available: <http://link.springer.com/10.1007/s10708-008-9189-x>. [2013.04.08]

- Budhathoki, Nedovic-Budic & Bruce (2010). An interdisciplinary frame for understanding volunteered geographic information. *Geomatica* **64**(1). 11–26.
- Carrera & Ferreira (2007). The future of Spatial Data Infrastructures: Capacity- building for the emergence of municipal SDIs. *International Journal of Spatial Data Infrastructure Research* **2**. 49–68.
- Chief Directorate: National Geo-Spatial Information (2009). Task Description: Acquisition and Processing of Ancillary Data: Ad Hoc. Technical report. Chief Directorate: National Geo-Spatial Information. Cape Town.
- Chief Directorate: National Geo-Spatial Information (2013a). Secondary Activity Level: Perform In-house Topographical Compilation. Technical report. Chief Directorate: National Geo-Spatial Information. Cape Town.
- Chief Directorate: National Geo-Spatial Information (2013b). Task Description: Compile Topographic Data. Technical report. Chief Directorate: National Geo-Spatial Information. Cape Town.
- Chief Directorate: National Geo-Spatial Information Management (2013). Chief Directorate: National Geo-Spatial Information Strategic Planning 2014-2018. Technical report. Chief Directorate: National Geo-Spatial Information. Cape Town.
- Clarke (2011). Initiatives and challenges of spatial data infrastructure in South Africa. in ‘AfricaGeo Conference 2011’. Cape Town. p. 12.
- Coetzee, Cooper & Strydom (2007). Spatial standards make your life easier. *PositionIT* pp. 37–39.
- Coleman (2010). Volunteered geographic information in spatial data infrastructure: An early look at opportunities and constraints. *Proceedings of GSDI 12 Conference, Singapore* p. 18.
*Available: <http://217.219.78.9/upload/file/amoozeshi/4/4.10.pdf>. [2011.12.21]
- Coleman (2013). Potential Contributions and Challenges of VGI for Conventional Topographic Base-Mapping programs. *Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice* .
- Coleman, Georgiadou & Labonte (2009). Volunteered Geographic Information: the nature and motivation of producers. *International Journal of Spatial Data Infrastructures Research* **4**. 20.
- Coleman, Nkhwanana & Sabone (2010). Volunteering Geographic Information to authoritative databases: Linking contributor motivations to program characteristics. *Geomatica* **64**(1). 383–396.
- Committee for Spatial Information (2012). ‘Committee for Spatial Information (In terms of the SDI Act, 2003): Data Custodianship Policy’.
- Cooper (2010). Perceptions of virtual globes and volunteered geographic information and spatial data infrastructures. *Geomatica* **64**(1). 73–88.
- Cooper, Coetzee, Kaczmarek, Kourie, Iwaniak & Kubik (2011). Challenges for quality in volunteered geographical information. in ‘AfricaGeo Conference 2011’. Cape Town. p. 13.

- Cooper & Eloff (2011). 'Spatial Data Infrastructure to be established in South Africa'.
*Available: www.csir.co.za/enews/2011_mar/12.html. [2012.03.12]
- Cooper, Kourie & Coetzee (2010). Thoughts on exploiting instability in lattices for assessing the discrimination adequacy of a taxonomy. pp. 1–12.
- Cooper, Rapant, Hjelmager, Laurent, Iwaniak, Coetzee, Moellering & Duren (2011). Extending the formal model of a Spatial Data Infrastructure to include Volunteered Geographic Information. *in* 'Proceedings of 25th International Cartographic Conference'. Paris. p. 10.
- Craglia (2007). Volunteered Geographic Information and Spatial Data Infrastructures: When do parallel lines converge?. *in* 'Position Paper for VGI Specialist Meeting'. European Commission Joint Research Centre. Santa Barbara. p. 3.
- Culham (2006). Error Bars - What they tell you and what they don't. Technical report. The Brian and Mind Institute - The University of Ontario. Ontario.
*Available: culhamlab.ssc.uwo.ca/JodyCulham/Courses/ErrorBars_Lecture.ppt.
- Devilleers, Jeansoulin & Moulin (2007). Towards spatial data quality information analysis tools for experts assessing the fitness for use of spatial data. *International Journal of Geographical Information Science* **21**(3). 261–282.
- Elwood (2008). Volunteered Information: Geographic key questions , concepts and methods to guide emerging and research practice. **72**(3/4). 133–135.
- Elwood, Goodchild & Sui (2012). Researching Volunteered Geographic Information: Spatial Data, Geographic Research, and New Social Practice. *Annals of the Association of Geographers* **102**(3). 571–590.
- Environmental Systems Research Institute Inc. (2012). 'ArcGIS Help: Exercise 4b: Using geodatabase topology to fix line errors'.
*Available: <http://resources.arcgis.com>. [2013.02.01]
- Fairbairn & Al-Bakri (2013). Using Geometric Properties to Evaluate Possible Integration of Authoritative and Volunteered Geographic Information. *ISPRS International Journal of Geo-Information* **2**(2). 349–370.
*Available: <http://www.mdpi.com/2220-9964/2/2/349/>. [2014.01.06]
- Flanagin & Metzger (2008). The credibility of Volunteered Geographic Information. *GeoJournal* **72**(3-4). 137–148.
*Available: <http://www.springerlink.com/index/10.1007/s10708-008-9188-y>
- Foody, See, Fritz, Velde, Perger, Schill & Boyd (2013). Assessing the Accuracy of Volunteered Geographic Information arising from Multiple Contributors to an Internet Based Collaborative Project. *Transactions in GIS* **17**(6). 847–860.
- Fuchs & Horak (2008). Africa and the digital divide. *Telematics and Informatics* **25**(2). 99–116.
*Available: <http://linkinghub.elsevier.com/retrieve/pii/S0736585306000359>
- Genovese & Roche (2010). Potential of VGI as a resource for SDIs in the North/South context. *in* 'Global Spatial Data Infrastructure 12'. Singapore. p. 15.

- Girres & Touya (2010). Quality assessment of the French OpenStreetMap dataset. *Transactions in GIS* **14**(4). 435–459.
 *Available: <http://doi.wiley.com/10.1111/j.1467-9671.2010.01203.x>. [2011.10.10]
- Goodchild (2007). Citizens as sensors: The world of volunteered geography. *International Journal of Spatial Data Infrastructure Research* **69**(4). 211–221.
- Goodchild (2008a). Commentary: Whither VGI?. *GeoJournal* **72**(3-4). 239–244.
 *Available: <http://www.springerlink.com/index/10.1007/s10708-008-9190-4>
- Goodchild (2008b). Spatial Accuracy 2.0. in ‘Proceedings of the 8th international symposium on spatial accuracy assessment in natural resources and environmental sciences’. Shanghai. pp. 1–7.
- Goodchild (2009). NeoGeography and the nature of geographic expertise. *Journal of Location Based Services* **3**(2). 82–96.
 *Available: <http://www.tandfonline.com/doi/abs/10.1080/17489720902950374>
- Goodchild & Glennon (2010). Crowdsourcing geographic information for disaster response: a research frontier. *International Journal of Digital Earth* **3**(3). 231–241.
 *Available: <http://www.tandfonline.com/doi/abs/10.1080/17538941003759255>
- Goodchild & Hunter (1997). A simple positional accuracy for linear features. *International Journal of Geographical Information Science* **11**(3). 299–306.
- Goodchild & Li (2012). Assuring the Quality of Volunteered Geographic Information. *spatial Statistics* **1**. 110–120.
- Gould (2007). Vertically interoperable geo-infrastructures and scalability. Technical report. Santa Barbara.
- Gouveia & Fonseca (2008). New approaches to environmental monitoring: Information of ICT to explore Volunteered Geographic Information. *GeoJournal* **72**(3/4). 185–197.
- Government Gazette (1978). ‘Copyright Act No.98 of 1978’.
- Government Gazette (2004). ‘Spatial Data Infrastructure Act No.54 of 2003’.
- Government Gazette (2008). ‘Standards Act No. 8 of 2008’.
- Gregoire & Bouillot (1998). ‘Hausdorff distance between convex polygons’.
 *<http://cgm.cs.mcgill.ca/~godfried/teaching/cg-projects/98/normand/main.html>
- Grobbe (2012). *TomTom Africa*. Prepared for meeting at the Chief Directorate: National Geo-Spatial Information[2013.02.07].
- Guelat (2009). Integration of user generated content into national databases - Revision workflow at Swisstopo. in ‘1st EuroSDR Workshop on Crowdsourcing for Updating National Databases’. Waben, Switzerland.
- Haklay (2010). How good is Volunteered Geographical Information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and Planning B: Planning and Design* **37**(1). 682–703.

- Haklay, Basiouka, Antoniou & Ather (2010). How many volunteers does it take to map an area well? The Validity of Linus Law to Volunteered Geographic Information. *The Cartographic Journal* **47**(4). 315–322.
 *Available: <http://openurl.ingenta.com/content/xref?genre=article&issn=0008-7041&volume=47&issue=4&spage=315>. [2013.04.16]
- Haklay & Ellul (2010). Completeness in volunteered geographical information - the evolution of OpenStreetMap coverage in England (2008-2009). *Journal of Spatial Information Science* **2**. 18.
- Haklay, Singleton & Parker (2008). Web Mapping 2.0: The Neogeography of the GeoWeb. *Geography Compass* **2**(8). 2011–2039.
- Hangouët (1995). Computation of the Hausdorff distance between plane vector polylines. in ‘Twelfth International Symposium on Computer-Assisted Cartography’. North Carolina. p. 10.
 *Available: <http://mapcontext.com/autocarto/proceedings/auto-carto-12/pdf/computation-of-the-hausdorff-distance-between-plane.pdf>. [2011.10.06]
- Hovland, Janis & Kelley (1953). *Communication and persuasion*. CT: Yale University Press. New Haven.
- ISO/TC 211 Presentations. (2012). Available: <http://www.isotc211.org>. [2013.02.01].
- Johnson & Sieber (2011). Motivations driving government adoption of the Geoweb. *GeoJournal* **77**(5). 667–680.
 *Available: <http://link.springer.com/10.1007/s10708-011-9416-8>. [2013.04.19]
- Johnson & Sieber (2013). Situating the Adoption of VGI by Government. in Sui, Elwood & Goodchild, eds, ‘Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice’. Springer Dordrecht Heidelberg. New York London. pp. 65–82.
- Keler, Trame & Kauppinen (2011). Tracking Editing Processes in Volunteered Geographic Information: The Case of OpenStreetMap. in ‘Identifying Objects. Processes and Events in Spatio-Temporally Distributed Data (IOPE). workshop at Conference on Spatial Information Theory 2011 (COSIT’11)’. Maine, USA. p. 7.
- Keniston (2003). *The four digital divides*. 1 edn. Sage Publishers. Delhi.
- Kounadi (2009). Assessing the quality of OpenStreetMap data. Masters thesis. University College of London.
- Lee & Wan (2004). Compactness measure of digital shapes. in ‘2004 IEEE Region 5 Conference: Annual Technical & Leadership Workshop’. Oklahoma. pp. 103–105.
 *Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1300173. [2012.04.23]
- Li, Goodchild & Church (2013). An efficient measure of compactness for two-dimensional shapes and its application in regionalization problems. *International Journal of Geographical Information Science* p. 24.
 *Available: <http://www.tandfonline.com/doi/abs/10.1080/13658816.2012.752093>
- Maceachren (1985). Compactness of geographic shape: Comparison and evaluation of measures. *Geografiska Annaler* **67**(1). 53–67.

- Makanga & Smit (2010). A review of the status of Spatial Data Infrastructure implementation in Africa. *South African Computer Journal* (45). 18–25.
- Masser, Rajabifard & Williamson (2008). Spatially enabling governments through SDI implementation. *International Journal of Geographical Information Science* **22**(1). 5–20.
- Mcdougall (2009). The potential of citizen Volunteered Spatial Information for building SDI. in ‘GSDI-11 World Conference’. Rotterdam. p. 10.
- Mcdougall (2010). From silos to networks —Will users drive Spatial Data Infrastructures in the future?. in ‘FIG Congress 2010. Facing the Challenges —Building the Capacity’. Sydney Australia. p. 13.
- McLaren (2012). Between a rock and a hard location. *Geospatial World* **2**(12). 38–40.
- McMaster (1986). A statistical analysis of mathematical measures for linear simplification. *The American Cartographer* **23**. 103–117.
- Mingqiang, Kidiyo & Joseph (2008). A survey of shape feature extraction techniques. *Pattern Recognition* pp. 43–90.
- Mooney, Corcoran & Winstanley (2010). A study of data representation of natural features in OpenStreetMap. in ‘Sixth International Conference on Geographic Information Science’. Zurich.
- Musinguzi, Bax & Tickodri-Togboa (2004). Opportunities and challenges for SDI development in developing countries - A case study of Uganda. in ‘Proceedings 12th International Conference on Geoinformatics Geospatial Information Research: Bridging the Pacific and Atlantic’. Sweden. pp. 789–796.
- Neis, Zielstra & Zipf (2012). The street network evolution of crowdsourced maps: OpenStreetMap in Germany 2007 —2011. *Future Internet* **4**. 1–21.
*Available: <http://www.mdpi.com/1999-5903/4/1/1/>. [2012.02.03]
- Norris (2000). The worldwide digital divide: Information poverty, the Internet and development. in ‘Annual Meeting of the Political Studies Association of the UK, London School of Economics and Political Science’. London. p. 10.
- OpenStreetMap Foundation*. (2013). Available: <http://wiki.osmfoundation.org>. [2014.01.17].
- OpenStreetMap Main Page*. (2013). Available: <http://wiki.openstreetmap.org>. [2013.01.21].
- Ostermann & Spinsanti (2010). A Conceptual Workflow For Automatically Assessing The Quality Of Volunteered Geographic Information For Crisis Management. in ‘The 14th AGILE Conference on Geographic Information Science’. Utrecht. p. 6.
- Patwardhan, Banerjee & Pedersen (2003). Using measures of semantic relatedness for word sense disambiguation. in ‘Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics’. Mexico. pp. 224–257.
- Pedersen, Patwardhan & Michelizzi (2004). WordNet::Similarity-measuring the relatedness of concepts. in ‘The Nineteenth National Conference on Artificial Intelligence’. California. pp. 1024–1025.

- Perkal (1966). On the Length of Empirical Curves. *in* ‘Discussion Paper No. 10’.
- Pourabdollah, Morley, Feldman & Jackson (2013). Towards an Authoritative Open-StreetMap: Conflating OSM and OS OpenData National Maps Road Network. *ISPRS International Journal of Geo-Information* **2**(3). 704–728.
- Rajabifard, Binns, Masser & Williamson (2006). The role of sub-national government and the private sector in future spatial data infrastructures. *International Journal of Geographical Information Science* **20**(7). 727–741.
 *Available: <http://www.tandfonline.com/doi/abs/10.1080/13658810500432224>
- Ramm, Topf & Chilton (2011). *OpenStreetMap*. UIT Cambridge Ltd.. England.
- Riecken, Bernard, Portele & Remke (2003). North-Rhine Westphalia: Building a regional SDI in a cross-border environment/ad-hoc integration of SDIs: Lessons learnt. *in* ‘9th EC-GI & GIS Workshop’. Spain. p. 7.
 *Available: http://www.ec-gis.org/Workshops/9ec-gis/papers/pd_portele.pdf
- Roche, Sureau & Caron (2003). How to improve the social utility value of geographic information systems for French local governments? A Delphi study. *Environment and Planning B: Planning and Design* **30**(3). 429–447.
 *Available: <http://www.envplan.com/abstract.cgi?id=b12964>. [2013.02.06]
- Rotter, Skulimowski & Kotropoulos (2005). Fast shape matching using the Hausdorff distance. *Science and Technology* pp. 429–447.
- Ryttersgaard (2001). Spatial Data Infrastructure - Developing trends and challenges. *in* ‘Committee on Development Information’. Addis Ababa. p. 8.
- Sabone (2009). Assessing alternative technologies for use of Volunteered Geographic Information in authoritative databases. Masters thesis. University of Brunswick.
- Schenk & Guittard (2009). Crowdsourcing: What can be outsourced to the crowd, and why?. *in* ‘Workshop on Open Source Innovation, Strasbourg, France’. p. 29.
- Schmitz, Zipf & Neis (2008). New Applications based on collaborative geodata —the case of Routing. *in* ‘XXVII INCA International congress on collaborative mapping space technology’. Gandhinaga, India.
- Shekhar (2010). Contributors of volunteered geographic world: Motivation behind contribution. *in* ‘GSDI 12 World Conference’. Singapore. p. 2.
- Siebritz, Sithole & Zlatanova (2011). Assessment of the homogeneity of Volunteered Geographic Information in South Africa. *in* ‘XXII International Society for Photogrammetry & Remote Sensing Congress’. Melbourne. p. 6.
- Sui (2004). Toblers first law of geography: A big idea for a small world?. *Annals of the Association of American Geographers* **94**(2). 269–277.
- TeleAtlas. (2013). Available: <http://mapinsight.teleatlas.com/mapfeedback/index.php>. [2013.02.07].
- Theobald (2001). ‘ArcUser Online - Understanding Topology and Shapefiles’.
 *Available: <http://www.esri.com/news/arcuser/0401/topo.html>. [2013.01.24]

- Thomas, Hedberg, Thompson & Rajabifard (2009). A Strategy Framework to Facilitate Spatially Enabled Victoria. *in* 'GSDI 11 World Conference'. Rotterdam, Netherlands. p. 19.
 *Available: <http://www.gsdi.org/gsdi11/papers/pdf/167.pdf>. [2013.04.16]
- Tveite & Langaas (1999). An accuracy assessment method for geographical line data sets based on buffering. *International Journal of Geographical Information Science* **13**(1). 27–47.
- Ubeda & Egenhofer (1997). Topological error correcting in GIS. *in* 'Advances in Spatial Databases - Fifth International Symposium on Large Spatial Databases, SSD '97'. Vol. 1262. Berlin. pp. 283–297.
- United States Geological Survey (2011). 'History of volunteer mapping at the USGS'.
 *Available: <http://nationalmap.gov/TheNationalMapCorps/history.html>
- Vauglin (1997). Modèles statistiques des imprécisions géométriques des objets géographiques linéaires. Unpublished phd thesis. Marne-la-Vallée University.
- Vorster (2009). The south african address standard: The national geo-spatial information perspective. *in* 'SABS Workshop on the South African Address Standard, Pretoria, South Africa'. p. 30.
- Vorster & Duesimi (2010). Geospatial data management within SA's national mapping organisation. *PositionIT* pp. 31–35.
- Williams (2001). *What is the digital divide?* University of Michigan (Unpublished).
- Wilson (2006). *The Information Revolution and Developing Countries*. MIT Press. Cambridge.
- Wolf, Matthews, Mcninch & Poore (2011). OpenStreetMap Collaborative Prototype , Phase One. Technical report.
 *Available: <http://pubs.usgs.gov/of/2011/1136/pdf/OF11-1136.pdf>. [2014.01.14]
- Wunsch-Vincent & Vickery (2007). Participative web: User-created content. Technical report. Organisation for Economic Co-operation and Development.
- Wytzisk & Sliwinski (2004). Quo Vadis SDI?. *in* '7th AGILE Conference on Geographic Information Science'. Heraklion, Greece. pp. 43–49.
- Zielstra & Zipf (2010a). A comparative study of proprietary geodata and Volunteered Geographic Information for Germany. *in* '13th AGILE International Conference on Geographic Information Sciences 2010'. Vol. 1. Portugal. p. 15.
- Zielstra & Zipf (2010b). Quantitative studies on the data quality of OpenStreetMap in Germany. *in* 'Sixth International Conference on Geographic Information Science'. Zurich. pp. 20–26.