

Developing Tools and Models for Evaluating Geospatial Data Integration of Official and VGI Data Sources



Maythm M Sharky Al-Bakri
B.Sc. Surveying Engineering
M.Sc. Surveying Engineering

**Thesis submitted for the Degree of
Doctor of Philosophy**

School of Civil Engineering and Geosciences

Newcastle University

October 2012

Abstract

In recent years, systems have been developed which enable users to produce, share and update information on the web effectively and freely as User Generated Content (UGC) data (including Volunteered Geographic Information (VGI)). Data quality assessment is a major concern for supporting the accurate and efficient spatial data integration required if VGI is to be used alongside official, formal, usually governmental datasets. This thesis aims to develop tools and models for the purpose of assessing such integration possibilities.

Initially, in order to undertake this task, geometrical similarity of formal and informal data was examined. Geometrical analyses were performed by developing specific programme interfaces to assess the positional, linear and polygon shape similarity among reference field survey data (FS); official datasets such as data from Ordnance Survey (OS), UK and General Directorate for Survey (GDS), Iraq agencies; and VGI information such as OpenStreetMap (OSM) datasets. A discussion of the design and implementation of these tools and interfaces is presented. A methodology has been developed to assess such positional and shape similarity by applying different metrics and standard indices such as the National Standard for Spatial Data Accuracy (NSSDA) for positional quality; techniques such as buffering overlays for linear similarity; and application of moments invariant for polygon shape similarity evaluations. The results suggested that difficulties exist for any geometrical integration of OSM data with both bench mark FS and formal datasets, but that formal data is very close to reference datasets. An investigation was carried out into contributing factors such as data sources, feature types and number of data collectors that may affect the geometrical quality of OSM data and consequently affect the integration process of OSM datasets with FS, OS and GDS. Factorial designs were undertaken in this study in order to develop and implement an experiment to discover the effect of these factors individually and the interaction between each of them. The analysis found that data source is the most significant factor that affects the geometrical quality of OSM datasets, and that there are interactions among all these factors at different levels of interaction.

This work also investigated the possibility of integrating feature classification of official datasets such as data from OS and GDS geospatial data agencies, and informal datasets such as OSM information. In this context, two different models were developed. The first set of analysis included the evaluation of semantic integration of corresponding feature

classifications of compared datasets. The second model was concerned with assessing the ability of XML schema matching of feature classifications of tested datasets. This initially involved a tokenization process in order to split up into single words classifications that were composed of multiple words. Subsequently, encoding feature classifications as XML schema trees was undertaken. The semantic similarity, data type similarity and structural similarity were measured between the nodes of compared schema trees. Once these three similarities had been computed, a weighted combination technique has been adopted in order to obtain the overall similarity.

The findings of both sets of analysis were not encouraging as far as the possibility of effectively integrating feature classifications of VGI datasets, such as OSM information, and formal datasets, such as OS and GDS datasets, is concerned.

Table of Contents

Abstract	i
Table of Contents	iii
List of Figures	ix
List of Tables	xiii
List of Abbreviations and Acronyms	xv
Publications from this Research	xvii
Acknowledgements	xviii
Chapter 1 Introduction	
1.1 Overview and background	1
1.2 Aim and Objectives of the research	4
1.3 Research methodology	6
1.4 Organisation of the thesis.....	7
Chapter 2 The Impact of Data Quality on Geospatial Data Integration	
2.1 Introduction.....	9
2.2 Definitions of multi-source geospatial data integration.....	10
2.3 Geospatial data integration issues	13
2.4 Applications for contemporary geographic dataset integration	16
2.5 The importance of data quality in geospatial data integration	18
2.6 Spatial data quality: concepts and issues	20
2.6.1 The uncertainty of spatial data.....	21
2.6.2 Spatial data variations and errors.....	23
2.7 Spatial data quality elements.....	25
2.7.1 Positional accuracy	25
2.7.2 Attribute accuracy.....	26
2.7.3 Temporal quality	26
2.7.4 Completeness.....	27
2.7.5 Logical consistency	27
2.8 Standards for reporting spatial data quality	28
2.8.1 FGDC	29
2.8.2 ANZLIC	29
2.8.3 CEN	30
2.8.4 ISO 19115.....	30

2.9	The assertion of data quality and credibility of alternative data sources	31
2.10	Practical testing of existing attempts of geospatial data integration.....	32
2.11	Chapter Summary	35
Chapter 3 Geospatial Mapping and VGI Databases		
3.1	Introduction.....	40
3.2	Official topographic mapping	41
3.2.1	Current Ordnance Survey (Great Britain) MasterMap data	41
3.2.2	The structure and characteristics of the MasterMap topography layer	42
3.2.3	Formal data in a contrasting agency –the Iraqi General Directorate for Survey.....	44
3.3	Contemporary aspects of informal data handling	45
3.3.1	Web 2.0 technologies and the development of geotagging	45
3.3.2	User Generated Content: concepts and applications	47
3.4	Development of volunteered geographic information and its features	50
3.5	VGI examples and initiatives.....	52
3.5.1	Exploration of the OpenStreetMap project.....	53
3.5.1.1	The analysis of the architecture components of OSM database	56
3.5.1.2	The structure of OSM datasets	58
3.5.1.3	Characteristics of OSM activities.....	62
3.5.2	Additional VGI examples - based Web 2.0 technologies.....	65
3.6	The differences between official and VGI data sources	67
3.7	Chapter summary	68
Chapter 4 Tool Development for Assessing the Similarity of Geometrical Entities		
4.1	Introduction.....	70
4.2	The study areas and data acquisition.....	71
4.3	Positional heterogeneity	76
4.3.1	Data accuracy standards	76
4.3.2	The US National Standard for Spatial Data Accuracy	77
4.3.3	Circular data for positional discrepancies analysis.....	80
4.3.3.1	The need for appropriate angular descriptive statistics.....	80
4.3.4	Positional similarity detection and implementation	83
4.3.4.1	Results and observations of formal and VGI datasets comparisons	84
4.4	Geometrical quality modelling for linear entities	95

4.4.1	Epsilon band accuracy models	96
4.4.2	An alternative method for buffering overlay.....	97
4.4.3	Linear similarity analysis tool	99
4.4.3.1	VGI and formal linear data integration assessment and output	100
4.5	Area shape similarity	108
4.5.1	Area moments invariant.....	110
4.5.2	Line (improved) moment invariant.....	112
4.5.3	Distances measurements.....	113
4.5.4	Polygon shape similarity analysis tool	114
4.5.4.1	Computing shape similarity of authoritative and VGI datasets	115
4.6	Chapter summary	119
Chapter 5 Factors Affecting Geometrical Integration of OSM Information with Official Datasets		
5.1	Introduction.....	121
5.2	Factorial design.....	122
5.2.1	An overview	122
5.2.2	Estimation of the effects in the 2 ³ factorial design.....	123
5.3	Stages of factorial design	126
5.3.1	Identifying the problem statement.....	126
5.3.2	Selection of factors and levels.....	127
5.3.3	Choosing the response variable	127
5.3.4	Selecting experimental design and performing the experiment	128
5.3.5	Experimental results analysis	128
5.4	Why these factors are significant for this experiment.....	129
5.5	Preparing tested datasets	131
5.6	Implementing the experiment: discussion and analysis.....	135
5.6.1	Experimental settings	135
5.6.2	The main and interactions effects.....	136
5.6.3	Testing significant effects.....	139
5.6.3.1	Numerical analysis	139
5.6.3.2	Graphical analysis	141
5.7	Chapter summary	143

Chapter 6 Semantic Similarity Models and Approaches

6.1	Introduction.....	145
6.2	Traditional models for semantic similarity analysis	146
6.2.1	Standard confusion matrices.....	146
6.3	Alternative similarity measurement models	149
6.3.1	Geometric models.....	150
6.3.2	Semantic network models.....	151
6.3.3	Transformational models.....	153
6.3.4	Feature models.....	153
6.3.5	Alignment models	155
6.3.6	Information content based models.....	156
6.4	Formal semantic lexical databases and their descriptions	156
6.4.1	WordNet database.....	156
6.4.2	Semantic similarity and relatedness	158
6.5	Methods for using semantic and structural models.....	159
6.5.1	Resnik (res).....	159
6.5.2	Lin (lin).....	160
6.5.3	Jiang and Conrath (jcn).....	161
6.5.4	Leacock and Chodorow (lch)	161
6.5.5	Wu and Palmer (wup).....	162
6.5.6	Path.....	163
6.5.7	Hirst and Onge (hso)	163
6.5.8	Banerjee and Pedersen (lesk).....	164
6.5.9	Vector	164
6.6	The achievement of WordNet::Similarity software	165
6.7	Chapter summary.....	167

Chapter 7 Finding Semantic and Structural Similarity for Classes and Instances

7.1	Introduction.....	169
7.2	Semantic similarity approaches	169
7.2.1	Feature based approach	170
7.2.1.1	Testing the semantic similarity suitability for feature classification matching purposes.....	170
7.2.2	Schema similarity approach.....	175

7.2.2.1	A categorization of schema relationships	175
7.3	Evaluating the similarity between different classes' features	177
7.3.1	Pre-processing phase	177
7.3.1.1	Pre-Processing of classes' names	177
7.3.1.2	Encoding feature classifications as an XML schema.....	180
7.3.2	Node similarity measurement phase.....	190
7.3.2.1	Label name similarity.....	190
7.3.2.2	Structural similarity.....	193
7.3.2.3	Data type similarity	196
7.3.2.4	Similarities combination	198
7.4	Chapter summary.....	200
Chapter 8 Conclusions and Recommendations for Further Work		
8.1	Thesis overview	203
8.2	Evaluation of the thesis objectives.....	204
8.3	Major conclusions of the thesis.....	206
8.3.1	The assessment of geometrical integration of formal and VGI datasets	206
8.3.2	The evaluation of various factors affect geometrical quality of VGI datasets	209
8.3.3	The assessment of semantic similarity of formal and VGI datasets for integration purposes.....	210
8.4	Data handling of VGI for integration with formal datasets	211
8.5	Utilising the distinct nature of VGI.....	213
8.6	Recommendations for future work (research)	215
References.....		218
Appendix A Maps of the Test Areas from OpenStreetMap Project.....		240
Appendix B The Programs of Geometrical Similarity Measurements Interfaces		
B.1	The positional similarity measurement interface's program (Pos.m)	243
B.2	The linear similarity measurement interface's program (Lin.m)	249
B.3	The area (polygon) similarity measurement interface's program (Mom.m).....	255
Appendix C XML Schema Codes		
C.1	XML schema code for OSM data in Cramlington-UK.....	263
C.2	XML schema code for OS data in Cramlington-UK	264
C.3	XML schema code for OSM data in Clara Vale-UK.....	266

C.4 XML schema code for OS data in Clara Vale-UK	268
C.5 XML schema code for OSM data in Baghdad-Iraq	270
C.6 XML schema code for GDS data in Baghdad-Iraq (in Arabic)	272
C.7 XML schema code for GDS data in Baghdad-Iraq (in English).....	273

List of Figures

Figure 2.1 Data flowline for multi-source geospatial data integration (Cai, 2002).....	11
Figure 2.2 Technical and non-technical issues of geospatial data integration (Mohammadi et al., 2008).....	14
Figure 2.3 A conceptual model of spatial data uncertainty (Fisher et al., 2006).....	23
Figure 2.4 The methods that have adopted by Girres and Touya (2010) to determine the linear differences between road features: (a) Hausdorff distance method (b) average distance method.....	35
Figure 2.5 The surface difference approach that has been used by Girres and Touya (2010) to calculate the differences between polygons.....	35
Figure 3.1 Comparison of the characteristics of Web 1.0 and Web 2.0 technologies (Barrett, 2012).....	47
Figure 3.2 The most popular types of mashups (Programmableweb, 2012)	51
Figure 3.3 The stages of creating a map for the OSM project (OpenStreetMap, 2011)..	54
Figure 3.4 A comparison of the details of maps for the centre of Newcastle upon Tyne – UK (images sampled on 09/05/2012, both rendered at equivalent zoom levels – 16/19 for OSM, 15/18 for Google maps). This comparison facility is now available at http://tools.geofabrik.de/mc/ . (a) OpenStreetMap data (http://www.openstreetmap.org/) (b) Google maps data (http://maps.google.co.uk/).....	56
Figure 3.5 The structure of the OSM project components (OpenStreetMap, 2012a)....	58
Figure 3.6 Examples of XML codes of OSM data types' structure.....	59
Figure 3.7 Statistics account graph reflecting the growth of OSM nodes, ways and relations data types (OSM-wiki, 2012).....	60
Figure 3.8 Wikimapia's growing performance (Wikimapia-statistics, 2012): (a) Statistics of users number (b) Statistics of marked places.....	66
Figure 4.1 Cramlington 1 and 2-UK site.....	72
Figure 4.2 Clara Vale-UK site.....	72
Figure 4.3 Baghdad-Iraq site.....	73
Figure 4.4 Field work using GPS to establish control points, and total station to survey the location of points.....	75
Figure 4.5 GPS Leica GX1230, employed to survey the location of features in open-space area using RTK technique.....	75
Figure 4.6 Distribution and intervals among tested points for NSSDA procedure (Givens, 1999).....	78

Figure 4.7 Example of unit vectors summation to calculate resultant length and direction (Fisher, 1993).....	82
Figure 4.8 The positional similarity measurement results for the comparison of FS and OS datasets in Cramlington1-UK.....	86
Figure 4.9 The positional similarity measurement results for the comparison of FS and OSM datasets in Cramlington1-UK.....	87
Figure 4.10 The positional similarity measurement results for the comparison of OS-OSM datasets in Cramlington1-UK.....	87
Figure 4.11 The positional similarity measurement results for the comparison of FS and OS datasets in Cramlington2-UK.....	89
Figure 4.12 The positional similarity measurement results for the comparison of FS and OSM datasets in Cramlington2-UK.....	89
Figure 4.13 The positional similarity measurement results for the comparison of OS-OSM datasets in Cramlington2-UK.....	90
Figure 4.14 The positional similarity measurement results for the comparison of FS-OS datasets in Clara Vale-UK.....	91
Figure 4.15 The positional similarity measurement results for the comparison of FS-OSM datasets in Clara Vale-UK.....	92
Figure 4.16 The positional similarity measurement results for the comparison of OS-OSM datasets in Clara Vale-UK.....	92
Figure 4.17 The positional similarity measurement results for the comparison of FS-GDS datasets in Baghdad-Iraq.....	94
Figure 4.18 The positional similarity measurement results for the comparison of FS-OSM datasets in Baghdad-Iraq.....	94
Figure 4.19 The positional similarity measurement results for the comparison of GDS-OSM datasets in Baghdad-Iraq.....	95
Figure 4.20 Perkal’s epsilon band approach (Perkal, 1965).....	97
Figure 4.21 Goodchild and Hunter method (Goodchild and Hunter, 1997).....	97
Figure 4.22 The BOS method elements (Tveite and Langaas, 1999).....	99
Figure 4.23 The results of the liner similarity measurement for the comparison of FS, OS and OSM in Cramlington2-UK.....	103
Figure 4.24 The results of the linear similarity measurement for the comparison of FS, GDS and OSM in Baghdad-Iraq.....	106
Figure 4.25 The interface of the output results of area shape similarity measurement for three datasets (FS, OS and OSM) in Cramlington1-UK.....	116

Figure 4.26 The interface of the output results of area shape similarity measurement for three datasets (FS, OS and OSM) in Cramlington2-UK.....	117
Figure 4.27 The interface of the output results of area shape similarity measurement for three datasets (FS, OS and OSM) in Clara Vale-UK.....	118
Figure 4.28 The interface of the output results of area shape similarity measurement for three datasets (FS, GDS and OSM) in Baghdad-Iraq.....	119
Figure 5.1 Graphical views of 2^3 factorial designs (Montgomery, 2001).....	123
Figure 5.2 Newcastle city centre site.....	133
Figure 5.3 Gosforth site.....	133
Figure 5.4 Cramlington site.....	134
Figure 5.5 Clara Vale site.....	134
Figure 5.6 Cube plot for response variable 'Euclidean distance'.....	136
Figure 5.7 Main effects plot for Euclidean distance.....	137
Figure 5.8 Interaction effects plot for Euclidean distance.....	138
Figure 5.9 Pareto chart of the standardized effects.....	142
Figure 5.10 Normal probability plot of the main effects and interactions among factors.....	143
Figure 6.1 An example of confusion matrix (Congalton and Green, 2009b).....	147
Figure 6.2 A hierarchical network structure.....	152
Figure 6.3 The relations between two sets of features (Tversky, 1977).....	154
Figure 6.4 The geometric configuration of alignment differences.....	156
Figure 6.5 A fragment of the WordNet is-a hierarchy (Giannis et al., 2005).....	158
Figure 6.6 The concept of similarity measure (Wu and Palmer, 1994).....	162
Figure 6.7 The correlation between human judgement results and WordNet::Similarity methods.....	167
Figure 7.1 One-sample t-test outcomes for the feature classifications semantic similarity scores of the comparison of FM and OSM datasets for three study areas: Cramlington-UK, Clara Vale-UK and Baghdad-Iraq.....	171
Figure 7.2 Box plot of the one-sample t-test results for the semantic similarity scores of feature classifications in Cramlington-UK site.....	174
Figure 7.3 Box plot of the one-sample t-test results for the semantic similarity scores of feature classifications in Clara Vale-UK site.....	174
Figure 7.4 Box plot of the one-sample t-test results for the semantic similarity scores of feature classifications in Baghdad-Iraq site.....	175

Figure 7.5 The comparison of tokenization rates of formal and informal datasets in three study areas: Cramlington-UK, Clara Vale-UK and Baghdad-Iraq.....	179
Figure 7.6 An example of XML document (W3Schools, 2012).....	181
Figure 7.7 An example of the respective XML schema for the document of figure 7.6 (W3Schools, 2012).....	181
Figure 7.8 XML Schema for feature classifications of OSM information in Cramlington-UK.....	184
Figure 7.9 XML Schema for feature classifications of OS datasets in Cramlington-UK.....	185
Figure 7.10 XML Schema for feature classifications of OSM information in Clara Vale-UK.....	186
Figure 7.11 XML Schema for feature classifications of OS datasets in Clara Vale-UK.....	187
Figure 7.12 XML Schema for feature classifications of OSM information in Baghdad-Iraq.....	188
Figure 7.13 XML Schema (in Arabic) for feature classifications of GDS datasets in Baghdad-Iraq.....	189
Figure 7.14 XML Schema (in English) for feature classifications of GDS datasets in Baghdad-Iraq.....	189
Figure 7.15 An example of structural similarity for part of OS and OSM schemas in Cramlington-UK.....	195
Figure 7.16 An example of structural similarity for part of OS and OSM schemas in Clara Vale-UK.....	195
Figure 7.17 An example of structural similarity for part of GDS and OSM schemas in Baghdad-Iraq.....	195
Figure 7.18 The hierarchy of XML schema data types (Hong-Minh and Smith, 2007).....	197

List of Tables

Table 2.1 A summary of the issues of geospatial data integration processing.....	36
Table 3.1 User generated content consumer percentage in the USA for the period between 2008 and 2013 (eMarketer, 2009).....	50
Table 4.1 The number of samples for positional similarity measurement tests.....	74
Table 4.2 The number of samples for linear similarity measurement tests.....	74
Table 4.3 The number of samples for area shape similarity measurement tests.....	74
Table 4.4 Comparisons of RMSE and NSSDA accuracy of compared datasets in Cramlington 1-UK.....	86
Table 4.5 Circular statistics of compared datasets in Cramlington 1-UK.....	86
Table 4.6 Comparisons of RMSE and NSSDA accuracy of compared datasets in Cramlington 2-UK.....	88
Table 4.7 Circular statistics of compared datasets in Cramlington 2-UK.....	88
Table 4.8 Comparisons of RMSE and NSSDA accuracy of compared datasets in Clara Vale-UK.....	91
Table 4.9 Circular statistics of compared datasets in Clara Vale-UK.....	91
Table 4.10 Comparisons of RMSE and NSSDA accuracy of compared datasets in Baghdad-Iraq.....	93
Table 4.11 Circular statistics of compared datasets in Baghdad-Iraq.....	93
Table 4.12 The overlap percentage of buffering area between FS and OS datasets in Cramlington2-UK study area.....	103
Table 4.13 The overlap percentage of buffering area between FS and OSM datasets in Cramlington2-UK study area.....	104
Table 4.14 The overlap percentage of buffering area between OS and OSM datasets in Cramlington2-UK study area.....	104
Table 4.15 The average displacement values for the comparisons of FS, OS and OSM datasets in Cramlington2-UK study area.....	105
Table 4.16 The overlap percentage of buffering area between FS and GDS datasets in Baghdad-Iraq study area.....	107
Table 4.17 The overlap percentage of buffering area between FS and OSM datasets in Baghdad-Iraq study area.....	107
Table 4.18 The overlap percentage of buffering area between GDS and OSM datasets in Baghdad-Iraq study area.....	108
Table 4.19 The average displacement values for the comparisons of FS, GDS and OSM datasets in Baghdad-Iraq study area.....	108
Table 5.1 Design matrix of 2 ³ factorial designs (Montgomery, 2001).....	124
Table 5.2 Design matrix and statistics of the developed experiment.....	135
Table 5.3 Minitab analysis for the experiment achieved in this project.....	141
Table 7.1 A summary of schema node relationships.....	177
Table 7.2 Results of schema relationships in Cramlington-UK.....	192

Table 7.3 Results of schema relationships in Clara Vale-UK.....	192
Table 7.4 Results of schema relationships in Baghdad-Iraq.....	192
Table 7.5 Portion of data type similarity.....	197
Table 7.6 Results of combined similarity between OS and OSM classification in Cramlington-UK.....	199
Table 7.7 Results of combined similarity between OS and OSM classification in Clara Vale-UK.....	199
Table 7.8 Results of combined similarity between GDS and OSM classification in Baghdad-Iraq.....	200

List of Abbreviations and Acronyms

AND	Automotive Navigation Data
ANOVA	Analysis of Variance
ANZLIC	Australian and New Zealand Land Information Council
API	Application Programming Interface
ASPRS	American Society of Photogrammetry and Remote Sensing
BOS	Buffer-Overlay-Statistics
CEN	European Committee Standardisation
CSDGM	Content Standards for Digital Geospatial Metadata
DCW	Digital Chart of the World
DMA	Defence Mapping Agency
FD	Formal Data
FGDC	Federal Geographic Data Committee
FS	Field Survey
f-VGI	facilitated-VGI
GDS	General Directorate for Survey
GI	Geographic Information
GIS	Geographic Information Science
GML	Geography Markup Language
GPS	Global Positioning System
GUI	Graphical User Interface
GUIDE	Graphical User Interface Development Environment
INSPIRE	Infrastructure for Spatial Information in Europe
ISO	International Organisation for Standardisation
ITN	Integrated Transport Network
JOSM	Java OpenStreetMap
LCS	Lowest Common Subsume
NMA	Norwegian Mapping Authority
NMAS	National Map Accuracy Standards
NNP	Normal Probability Plot
NSDI	National Spatial Data Infrastructures
NSSDA	National Standard for Spatial Data Accuracy
NTD	National Topographic Database

OS	Ordnance Survey
OSM	OpenStreetMap
POS	Parts of Speech
RMSE	Root Mean Square Error
SDI	Spatial Data Infrastructures
SDTS	Spatial Data Transfer Standard
TIGER	Topologically Integrated Geographic Encoding and Referencing
TOID	Topographic Identifier
UEFA	Union of European Football Associations
UGC	User Generated Content
UK	United Kingdom
UN	United Nation
US	United States
USGS	United States Geological Survey
VGI	Volunteered Geographic Information
WMS	Web Map Services
WSD	Word Sense Disambiguation
WVS	World Vector Shoreline
XML	Extensible Markup Language

Publications from this Research

During the period of this project, the following publications have been arise from the work presented in this thesis:

Al-Bakri, M. and Fairbairn, D. (2010) 'Assessing the accuracy of crowdsourced data and its integration with official spatial data sets', *The Ninth International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*. University of Leicester / UK, pp. 317-320.

Al-Bakri, M. and Fairbairn, D. (2011) 'User generated content and formal data sources for integrating geospatial data ', *25th International Cartographic Conference*. Paris, France, pp. 1-8.

Al-Bakri, M. and Fairbairn, D. (2012) 'Assessing similarity matching for possible integration of feature classifications of geospatial data from official and informal sources', *International Journal of Geographical Information Science*, 26, (8), pp. 1437–1456.

Acknowledgements

During the time of my PhD, there are several people who have supported and helped me to undertake this research. First, many thanks go to my supervisors Dr David Fairbairn and Mr Philip James for their guidance and patience. They have offered helpful advice and encouragement through the process of the academic development of this research and I thank them.

Additionally, I would like to thank the staff and the community of postgraduates in the Department of Geomatics / School of Civil Engineering and Geosciences at Newcastle University for the friendly atmosphere which made me enjoy the last four years. Thanks must also go to Martin Robertson and Wisam Hussein, who helped to record field survey data.

I am also very grateful to my family for their unconditional encouragement and inspiration. Finally, I would like to express my thanks to Ministry of Higher Education and Scientific Research / Republic of Iraq for providing me a scholarship and financial support to do this research.

Chapter 1 Introduction

1.1 Overview and background

During the last decade, the advancement of data acquisition technologies has led to a massive increase in the amount of digital spatial data available. There are now numerous institutions and individuals maintaining and storing digital spatial datasets at different levels of detail on the Internet (Chen et al., 2008). The development of the Web 2.0 technologies which enable users to produce and share data via the web has increased the availability of online data sources (Seeger, 2008). Thus, broadly speaking, spatial data available on the web can be categorised according to the community that collects and collates it. Data collection from the public can be generally categorised as User Generated Content (UGC), which is not limited to spatial data, and data collected by state sponsored companies and organisations can be considered as Formal Data (FD).

The spatial data which is collected and distributed as UGC has been termed as Volunteered Geographic Information (VGI) (Goodchild, 2007b). Currently, there is a wide variety of spatial VGI data sources available on the Internet, such as the OpenStreetMap (OSM) project, the Flickr service, the interactive Wikimapia website and Yahoo imagery service. The informal collaborative map data projects (e.g. OSM) seek to create free alternative maps which allow users to add or input new materials to the data of others. Basically, low cost Global Positioning System (GPS) receivers and the availability of GPS signals make it possible to acquire positional information about different locations and upload it to local or personal databases and to other VGI data sources. This enables the production of maps based on volunteered efforts, personal computers and the Internet (Goodchild, 2007b).

By contrast, maps for essential purposes such as land use maps, topographic mapping for military applications and cadastral maps have been produced by formal or governmental institutions over many centuries. In many countries, much of these data are protected by a data licence due to their high quality and costs of collection. Hence, they are a significant investment and their sale and use can be considered as a source of income for the economy (Perkins and Dodge, 2008). Nowadays, there are various government agencies that have provided a wide variety of maps; e.g. topographic maps.

One example of these national mapping agencies is the Great Britain Ordnance Survey (OS). This agency allows for the embedding of its own topographic mapping products in web applications in scales ranging from the whole of Great Britain down to street level through OS OpenSpace service. The US Census Bureau can be considered another example of a formal data source. The road vector data covering all of the United States is available through this agency (Chen et al., 2008).

The spatial data from diverse sources often has varying accuracy levels due to different data collection methods; most data accuracy does not meet the user requirements for the majority of applications across a range of different organizations. For example, the varied emergency services require comprehensive coverage of complex data to plan, manage and respond to demand on their services. It is unusual for the data sources used for these purposes to be completely accurate and compatible. Many efforts have been made to integrate multiple spatial datasets to improve accuracies (Omran and van Etten, 2007) and such data integration can produce more accurate results and more reliable information than that obtained from a single source. In addition, geospatial data integration processing may provide many other useful functions within the geographic information science (GIS) framework. For instance, conflating diverse datasets can serve to update the old version of a dataset by adding some up-to-date features from one dataset to the other. Another practical usage of data integration processing has been presented by Lv et al. (2008). They introduced a road network matching algorithm to overcome the low accuracy map matching problem in the intersections of roads. Their approach has reduced the number of errors that may occur in GPS navigation data. This can effectively improve the navigation location service that is used to provide information regarding the actual positions of users. Geospatial data integration may also assist in assessing the quality of compared datasets. In this context, Butenuth et al. (2007) pointed out that it is necessary first to assess the geometric and semantic quality of one dataset that is preparing for integration with another, rather than only incorporating those datasets into the GIS system by overlaying with no consideration of their quality. Therefore, gaining comprehensive valuable outcomes of multi geoinformation sources simultaneously can be achieved by integrating these datasets.

However, the integration process remains one of the main challenges facing spatial data users. The wide variety of geographic information creates more difficulties in the

integration of datasets (Ulubay and Altan, 2002). This is particularly so because of their diversity in collection time, purpose, scale or data quality elements in general. Data quality indicates the usefulness of any data for any particular purpose. Various components of data quality have been reported by many researchers, for example, Kumi-Boateng and Yakubu (2010), Devillers and Jeansoulin (2006), and Jakobsson (2002). The criteria for judging the quality of spatial data comes from the evaluation process, using such information as positional accuracy, attribute accuracy, temporal accuracy and many other parameters. Obviously, data quality assessment is one of the main challenges for supporting accurate and efficient geospatial data integration (Mustière and Devogele, 2008).

In the process of data quality measurements in order to support effective geospatial data integration, heterogeneity may occur in the geometrical or semantic level of individual datasets as they are being combined. Therefore, those elements should be considered and evaluated. A geometrical quality measurement is principally dealing with assessing the quality of the real world features such as points, line segments and polygon areas. The process of achieving such geometrical evaluation can involve positional accuracy assessment for point objects, and shape similarity measurements for examining the similarity between the linear properties and area shapes of compared objects. Any variations between the geometrical quality characteristics of integrated datasets may lead to non-alignment of corresponding features. For instance, the results of trying to integrate inconsistent linear features may superimpose different features together, such as buildings with rivers.

In addition, semantic similarity is another essential concept in GIS for performing beneficial exchanging and transferring of data among spatial databases. Semantic information can be regarded as people's descriptions attached to the kind of geometrical features inside databases and is usually presented in feature classes (Ziegler and Dittrich, 2004). For example, semantic information may define a polygon area as a 'building' not a 'park'; without such information it would be difficult to recognise with precision the type of geometrical features of such a dataset. The most important problem of heterogeneity in spatial datasets may occur in the semantic and structural similarity of classification data from different sources. The main issues are concerned with the meanings that the compared features may carry, and the relationships between the

structure of each dataset's feature classifications tree or schema. It is common to find the same concept for different names in two datasets. For example, a 'road' may reflect two different concepts in two different datasets e.g. referred to as a 'highway' in one and 'motorway' in the other. At the same time, comparing XML schema trees, which are usually created to organize feature classifications in ordered classes, may raise some problems. It is possible to find that a feature belongs to a 'sub-class' in one dataset and a 'super-class' in another. Therefore, integration of multiple spatial datasets remains one of the main challenges facing spatial data consumers.

Accordingly, this study is motivated by the above descriptions to develop and demonstrate practical mechanisms to investigate whether it is possible to effectively integrate official and VGI datasets. This introductory chapter of the thesis began by providing a background to spatial data sources and the issues and challenges of geometrical integration processing and semantic similarity measurements. The next section describes the overall aim and objectives of the work accomplished in this thesis. The methods that were used in order to achieve the objectives will be described afterwards. The chapter ends by presenting an outline of the thesis structure and content.

1.2 Aim and Objectives of the research

The aim of the research in this thesis is to assess the possibility of geospatial data integration from formal and VGI spatial data sources. Integrating VGI with formal datasets is a significant process in that it could make the updating of formal data more efficient and cheaper. Recently, the appearance of VGI data has provided new opportunities for GI communities to gain effective benefits from it. Why these kinds of dataset offer such opportunities is described in section 3.4. However, since data quality is an important part of spatial data components in general, it is often impossible to ignore when geospatial data integration is the ultimate target, as is explained in sections 2.5 and 2.9. In order to examine the assessment of such integration, the research objectives can be broken down into seven distinct tasks as follows:

1. To outline the history of the geospatial data integration process and recent research, and identify the effect of spatial data quality elements on it, especially the opportunities and problems related to the improvement of geospatial data integration in general, and integrating VGI with formal spatial datasets in

particular; showing the state of geospatial data integration development in the light of technical developments' addressing of spatial data interoperability;

2. To analyse how VGI data is, and could be, utilised for the purpose of updating formal datasets by integrating them together, the achievable description of such data and to what extent it is comparable to official datasets. This may provide a useful opportunity to understand the approaches that are followed for gathering, uploading, disseminating and sharing VGI data on the web, which may assist in VGI data handling;
3. To develop a system or a series of tools which assess, report and display geometrical similarity measurements such as positional, linear and polygon (area) shape measurements among tested datasets. The positional and shape descriptors of compared features are vital properties, as they can be utilised to determine the possibility of physical data integration;
4. To develop an experiment to investigate the effect of several factors, such as data sources, feature types and number of data collectors, on VGI geometrical data quality. The determining of the effect of each factor and the interaction values between all of them is important information that can be used to inform which factors need further development and consequently improve VGI data quality and integration;
5. To develop models for measuring semantic similarity between corresponding features and also between schema classifications for features of formal and VGI datasets. The developed models can be used to perform the evaluation of the ability of integrating semantic data from formal and VGI data sources, based on the results of semantic similarity processing;
6. To test the research flowline developed in Objectives 3 and 5 by using different study areas, data sources and feature types. This diversity is necessary in order to assess the possibility of geospatial data integration in differing situations;
7. From the results of the research, a conclusion may be drawn and analysis made of what further developments may be necessary to improve interoperability in

formal and VGI geospatial data integration, as well as suggesting more directions of related research.

1.3 Research methodology

A methodology has been developed in order to accomplish the overall aim and objectives of this project. The research methodology has been fundamentally subdivided into several connected parts:

- Examination of the research field, by investigating data integration methods and opportunities, in addition to analysing the existing materials of formal and VGI data, to gain an understanding their characteristics, integration and interoperability issues;
- Application of positional similarity measurement methodologies, in particular the National Standard for Spatial Data Accuracy (NSSDA) and directional statistics techniques, to assess the similarity of positional and circular observations through developing a specific program interface;
- Execution of research-led procedures that can analyze shape similarity measurement, such as the curvature of linear features or the boundary of polygons, using the double buffering method and moment invariants models;
- Implementation of a factorial design experiment which involves a scientific approach to analyse the values of the factors influencing VGI geometrical data quality;
- Applying the semantic similarity analysis models developed in this research to assess the integration of corresponding feature classifications of compared datasets (this basically includes measuring the semantic similarity value between them and comparing with the threshold value to decide on the viability of integration);
- Evaluation of the significance of XML schema matching of feature classifications by developing practical models to determine semantic, data type, structural and overall similarity scores between the nodes of tested schemas.

1.4 Organisation of the thesis

The thesis consists of eight chapters. Chapter One introduces the research background which essentially includes a general overview, aims and objectives of the research and also a brief introduction to the methods of this study. A literature review of the relations between geospatial data integration and quality is presented in Chapter Two. Chapter Three discusses a comparison of formal and VGI data sources. The analysis of geometrical similarity measurements between formal and VGI datasets is demonstrated in Chapter Four, followed by an analysis of the factors that may affect the spatial data quality of VGI datasets. The next two chapters (Chapter Six and Chapter Seven) describe the models and results of semantic similarity measurements comparing official and VGI datasets, and a summary of the thesis, with recommendations for future work, is presented in the final Chapter. Individual chapters in this thesis address specific topics as follows:

- Chapter Two reviews the background of geospatial data integration processing in general and highlights the importance of data quality for it, this being the first objective of this project. This essentially involves considering sections of up-to-date research work being performed on the development and applications of geospatial data integration, and the barriers that may face multi-source geospatial data integration. The chapter then moves to consider in detail spatial data quality parameters, issues and standards. Finally, it concludes by describing concerns over the issue of data quality in contemporary spatial data sources with regard to the concept of geospatial data integration.
- Chapter Three provides an in-depth look into the creation and representation properties of spatial datasets from different data sources, with the intention of summarising the differences and similarities between formal and VGI datasets. It basically comprises two main parts: firstly, the formal data sources are described using different examples; secondly, the technologies and facilities of the VGI phenomenon are presented. Consequently, this chapter assists in achieving the second objective of this thesis.
- Chapter Four begins by introducing the study areas, which include urban and rural areas in Northumberland-UK (Cramlington and Clara Vale), and also an

urban area in Baghdad-Iraq. Different existing indices are examined and investigated, in particular, those which are related to positional, linear and area shape similarity analysis. The chapter addresses Objectives Three and Six of this project by developing and implementing three tools for measuring geometrical similarity between formal and VGI datasets for the varying sites by applying some of the metrics discussed in the chapter.

- Chapter Five investigates the techniques and approaches of a factorial design experiment. It analyses the results obtained from designing this experiment to determine the effect of different factors on VGI geometrical data quality, which addresses the fourth objective of this study.
- Chapter Six concentrates on semantic similarity assessment techniques and approaches, taking into account many particular models of this area of research. It discusses the mathematical models and information context of each model. The chapter ends with a comparison of different approaches of the WordNet::Similarity database to select an appropriate approach that can be used for the practical tests of Chapter Seven.
- Chapter Seven helps to address Objectives Five and Six of this thesis by developing and evaluating two models to show to what extent the semantic similarity of formal and VGI datasets can be integrated. The results and analysis from adopting the developed models within several study sites are illustrated and described.
- Chapter Eight presents the major conclusions of the thesis, discussing the findings achieved in Chapters Four, Five and Seven, in addition to describing the limitations of the study, in order to propose an outlook for future work in this field of research. The contents of this chapter basically address the last objective of this project.

Chapter 2 The Impact of Data Quality on Geospatial Data Integration

2.1 Introduction

Spatial datasets have fundamental roles to play in representing and managing a wide range of natural and/or constructed features in the real world. However, inconsistencies and errors are always inherent to these datasets. In order to determine the appropriate use of this information for such applications, it is essential to identify the different elements of spatial data quality that are characteristic of such datasets and their accuracy. The increasing demand for spatial data applications has led to a range of standards being developed that can manage data quality descriptions. Furthermore, the need for disseminating spatial datasets to potential users has also motivated a substantial amount of research on the spatial data quality field (Kumi-Boateng and Yakubu, 2010). Hence there is a necessity for details of the data quality information to be embedded as quality parameters into datasets.

In practice, many GIS outputs can be used to support decision making in areas such as geospatial data integration, environmental monitoring and evaluation of resources. In addition, it has become easier to use and distribute digital spatial data repeatedly for different applications. These trends have also emphasised the need for information regarding data quality parameters or the assessment of these elements. This chapter aims to provide a detailed overview of the issues related to data quality and to summarise researches that have explained the importance of spatial data quality information to spatial data users. In order to focus on the main objective of this research project, the emphasis will be on the aspects of the evaluation of spatial data quality elements that relate to geospatial data integration.

The sections of this chapter, therefore, illustrate some definitions that can be used to describe geospatial data integration concepts and also progress to presenting an explanation of spatial data quality concerns and standards. The first section discusses different terms that can be employed to define multi-source geospatial data integration processing. This is followed by a review of key issues concerning geospatial data integration, including both technical and non-technical issues. The importance of contemporary geospatial data integration will be discussed in section 2.4, with a focus

on the applications of the possibility of integration of formal and informal data sources, as well as the main challenges that may face the GI community in such an exercise. Before going into further details of spatial data quality issues and limitations, understanding the necessity of spatial data quality for data integration processing is essential and this is discussed in section 2.5. Subsequently, the chapter reviews the problems and concerns of spatial data quality. This initially includes discussion of some terminology related to spatial data quality such as 'uncertainty', 'errors' and 'accuracy'. The content of spatial data quality elements or parameters is described in section 2.7. This review focuses primarily on positional accuracy, attribute accuracy, temporal quality, logical consistency and completeness. Section 2.8 explains in detail the standards that can be applied to document data quality elements. This is followed by discussion of the importance of assessing the quality of alternative datasets. Some existing attempts of data integration research are considered in section 2.10, whilst the final section presents a summary of the preceding sections and concludes the chapter.

2.2 Definitions of multi-source geospatial data integration

In general, geospatial data integration processing has been defined by Rajabifard et al. (2003) as the process of making different datasets compatible to each other. They stated that time and effort expended by spatial data users can be reduced by managing integrated datasets. Their study also suggested that it is initially important to decide upon an agreement or standard for the amount, type and structure of spatial datasets that are to be integrated within different users' communities. This could contribute to the production of spatial datasets without redundancy or duplicating efforts. Recently, the main advantage of the availability of digital data in spatial data handling, compared to paper maps, is its ability to integrate or overlay spatial data from different sources. Such development is enhanced by the availability of digital data, the development of distributed web services and the ability to use online GI processing for decision making. Within this context, the management of spatial information may become more efficient from the perspective of geospatial data integration processing at various levels within nations. This includes, for example, local, regional, state and national and then it may proceed to global levels which could help to develop Spatial Data Infrastructures (SDI), a concept which will be described further in section 2.4.

One example of a real situation of multi-source geospatial data integration is shown in Figure 2.1. This figure displays the general information flowline for the assessment of the effectiveness of regional infrastructure processing. The connection between the information regarding the infrastructure facilities and the demand information, which can reveal a geospatial pattern of infrastructure, is a key target of this treatment. In Figure 2.1, Boxes 1 and 2 represent two sets of dataset layers: infrastructure surveying data from a number of companies and different types of users. Box 3 indicates the regional infrastructure dataset of the union of all layers of Box 1. The union of all layers of Box 2 forms a new dataset layer, 'regional demand', shown in Box 4. A new dataset can be created in Box 5 by overlaying 3 and 4, in which the value of each element is a function of the values of source datasets (i.e. the data of boxes 3 and 4) (Cai, 2002).

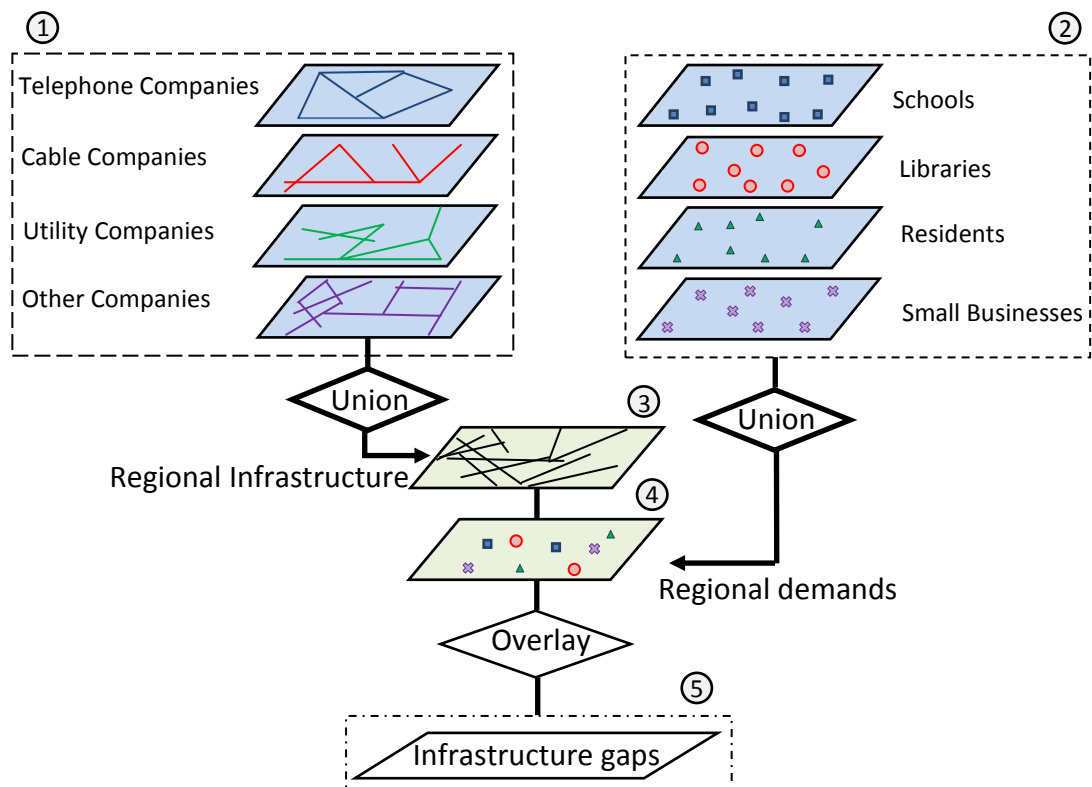


Figure 2.1 Data flowline for multi-source geospatial data integration (Cai, 2002).

In the example above the terms 'union' and 'overlay' are shown as possible data integration procedures. However, 'data integration' is a much-used phrase and there is no agreed definition. Within the field of spatial information, the terms 'integration', 'conflation', 'merging', 'fusion' and 'combining' have all been used to define the process of creating one database from two or more diverse datasets. Various studies have

attempted to specify the actual meaning of 'geospatial data integration' from a GIS perspective. For instance, Rhind et al (1984) defined data integration as the process of matching multiple datasets to each other in order to produce a master dataset, while Uitermark and Dutch (1996) described geographic dataset integration as a method of creating links among equivalent objects for different datasets for the same geographic region. Similarly, Cai (2002) believed that geospatial data integration includes the process of combining a variety of information from different sources. This can be carried out by establishing effective matching paradigms between corresponding entities across compared data sources. Likewise, Usery et al. (2005) suggested that a vital part of data integration is the matching of corresponding dataset properties such as topological, geometrical and attributes parameters. So far, however, the above discussion summarises several data integration concepts which are generally helpful with regard to processing in GIS technologies.

In some studies, researchers have used other phrases in their investigations into bringing different datasets together. For example, Saalfeld (1988) defined geospatial data 'conflation' as a composing or a collating of the overlapping regions of two different datasets. Saalfeld's pioneering work on conflation focused on improving the overall accuracy of the conflation of two or more datasets by eliminating inconsistency in spatial data using a coordinate modification procedure such as rubber sheeting. UCGIS (1996) argued that the word 'conflation' similarly refers to the integration of diverse datasets. It may be used to recognize the same features from different datasets for automatic registration processing. Furthermore, it can be applied to the modification of old versions of datasets by transferring the feature geometry of more accurate versions. Wald (1999) suggested the term 'fusion' as an equivalent to the concept of data integration. The same author also explained that there are other terms which could be used such as 'merging', 'combination' and 'synergy'. Samadzadegan (2004) pointed out that the term 'information fusion' refers to the process of combining entities from many information sources to produce a 'better' database. In this project, the term 'data integration' has been preferred to the other terms.

Geospatial data integration can be classified into several categories based on the characteristics of individual datasets (Jensen et al., 2005). For example, 'vector-to-vector integration' can be applied to the integration of corresponding objects from low

accuracy data sources into more precise datasets. 'Image-to-image fusion' may include the combination of, for example, any satellite image and a digital orthophoto. Furthermore, it can be used to detect the differences between two satellite images collected on two separate dates. Integration can also be performed between vector and image datasets, such as in the case of merging satellite images with a network road dataset. It can also be achieved between some measurements spreadsheets and vector datasets, such as in the integration of water data quality (stream gauging) and geodetically controlled base datasets. Another type of integration can be made between two measurements sets (measurement-to-measurement), such as creating a statistical link between vegetation biomass and its height.

Further distinctions in data integration were described by Jensen et al. (2005). This includes classifying data integration into three dimensions: horizontal, vertical and temporal combination. Horizontal integration involves side by side (adjacent) integration or combining spatially bordering datasets. Vertical integration refers to the superimposition of different datasets (e.g. by overlay) to produce one database. The integration of different datasets that have been obtained at different times is usually called temporal integration. The research described in this thesis is focused on assessing the possibilities of vector-to-vector integration and is also restricted to the sense of the vertical integration concept of the datasets, in this case from official and informal sources.

Although data integration can save time and money, there are numerous theoretical and technical issues which must be addressed when considering the challenges of geospatial data integration as will be discussed in the next section.

2.3 Geospatial data integration issues

Most geospatial data integration implementations include a number of processes to produce new databases. This involves collecting spatial data through human activities and performing spatial data combination tasks by means of technologies (e.g. computers, software and network services). These operations may be achieved without too many problems within one organisation; however, there may be many problems facing GI users, when applying this action across different companies or agencies (Usery et al., 2005). The most significant barrier to effective geospatial data integration

is that spatial data users commonly rely on more than one dataset from different sources. At the same time, spatial data sources are managed by different communities and backgrounds. Thus different standards, frameworks, policies and tools can be involved in such spatial data handling. For example, the concept of data integration for SDI initiatives means not only the superimposition of datasets from different sources, but also includes all institutional, legal, social and policy mechanisms which may affect data handling, together with technical tools to facilitate the integration of multi-sourced geospatial data (Piwowar and LeDrew, 1990). These issues have been identified and addressed comprehensively by Mohammadi et al. (2008) and Mohammadi et al. (2010) in an effort to ensure effective geospatial data integration. They have investigated these challenges and issues from the perspective of technical and non-technical concepts as shown in Figure 2.2.

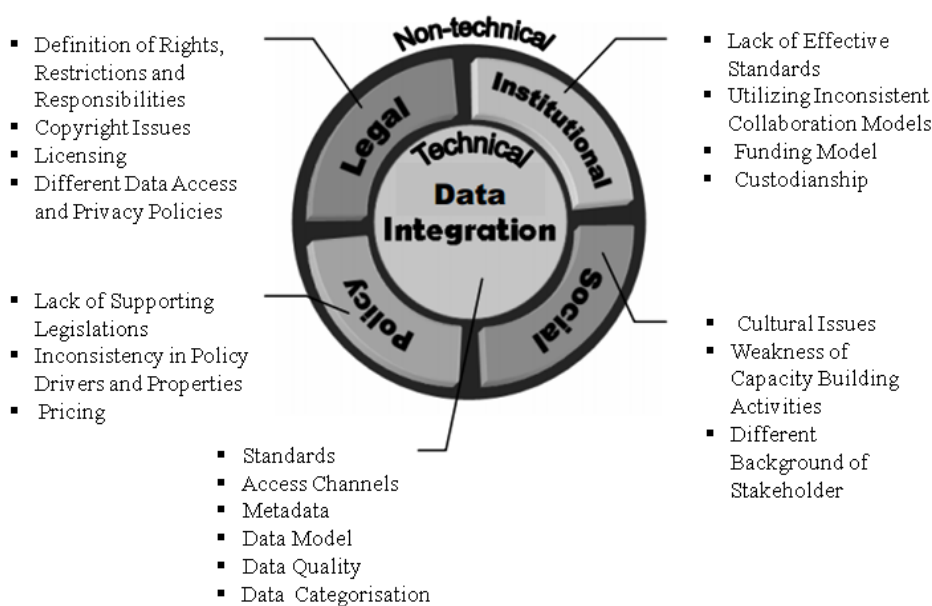


Figure 2.2 Technical and non-technical issues of geospatial data integration (Mohammadi et al., 2008)

There are several technical issues that can obstruct geospatial data integration. For instance, the lack of consistent data uniformity or specifications across different data sources often creates technical problems in attempts to integrate datasets. The technical issues of geospatial data integration and related concerns have been studied by many other researchers within the context of individual projects. For example, Edward and Simpson (2002) argued that the non-agreement between overlaying vector representations from different data sources can be due to variability in accuracy, errors, currency or temporality of an individual dataset. Finn et al. (2004) believed that the

major problems for national spatial dataset integration are topological consistency, geometric accuracy and resolution of data layers. Usery et al. (2005) confirmed some of the technical problems of geospatial data integration as being positional accuracy, resolution, formats and scales. They believed that data integration can improve the quality of involved datasets. Thus, providing detailed information about the consistency of the above-mentioned issues can play a key role in helping to establish effective geospatial data integration.

Mohammadi et al. (2008) also mentioned that non-technical problems such as policy, social, institutional and legal issues may also affect the geospatial data integration flow line. For instance, Thellufsen et al. (2009) described the social issues of geospatial data integration as one of the main constraints that should be taken into account when inter-organizational sharing of geospatial data is the target. They stated that data suppliers frequently oppose integration or sharing of geospatial data across different companies due to the related loss of independence, control and power. The operation of geospatial data integration has also many policy issues such as data access and pricing. For example, each geo-information agency may have applied different pricing models and licence conditions to transfer or share their datasets. Thus, it could be time consuming to obtain agreement to find the right information (Donker and Loenen, 2006).

Another prominent non-technical issue can be represented by institutional problems. These involve the diversity of the ways that datasets are stored in different organisations or companies. The differences in maintenance agreements and coordination, and weak collaboration between equivalent agencies are also other institutional difficulties for effective integration (Weaver, 2004; Ordnance Survey, 2003). According to Mohammadi et al. (2008), the final non-technical issue is legal. With the involvement of different spatial datasets from different organisations, it is necessary to specify the rights of spatial datasets and clarify the different licence conditions, including copyright (Donker and Loenen, 2006).

From the description above, it can be seen that the integration issues will differ in many different applications and even in different levels of the same application. However, it is necessary to remember that these issues are connected to each other and also have effects on each other.

For applications such as health, tourism, geology and disaster management, the need to combine diverse datasets has become very pressing. However, several datasets for the same area can be different in terms of scale, accuracy, currency and update dates. Casado (2006) classified the main types of spatial dataset conflation as geometric, semantic and topological conflation. The geometric type was defined as how to transfer the geometric properties of one dataset into another whilst minimizing the geometric differences and inconsistencies between them. The focus of this work was primarily on methods for geometric conflation of datasets covering the same area but at different scales. Semantic integration is described as the process of making feature classifications more uniform on the integrated dataset. Topological conflation is the third category suggested by Casado, the idea being that the integrated dataset topology can be regenerated, if necessary, by topological conflation of two datasets. Therefore, a consideration of all these categories in conflation processing would be useful in the case of disappearance, joining or merging of features.

In the same context Butenuth et al. (2007) reported that data integration is not only overlaying the data in a geographic information system, but also assessing how well the geometric and semantic properties of one dataset can be transferred to the other. However, in two different databases, the problem of heterogeneity may occur at geometric, semantic and structure of feature classifications tree levels which may lead to difficulty in attempting integration. For the research described in this thesis, some of the issues of assessing the possibility of data integration have been adopted. This will focus on geometric quality (positional and shape fidelity) initially. Also it will cover the semantic and structural similarity assessment of hierarchical ordering of categories in formal and VGI datasets.

2.4 Applications for contemporary geographic dataset integration

Recently, the need to allow users to access and identify geographical or spatial datasets from different levels and sources has become a priority of most of GI communities. For example, the emergence of Spatial Data Infrastructures (SDI) initiatives in many countries can be considered one of the most obvious geospatial data integration applications (Mohammadi et al., 2009). The Infrastructure for Spatial Information in Europe (INSPIRE) initiative, for instance, is one of the leading platforms for SDI. Over the past decade, several countries in Europe had developed their National Spatial Data

Infrastructures (NSDI). However, the legal framework of an SDI for all European mandates was initiated in May 2007 (Craglia, 2007). INSPIRE is based on the national SDI levels which are already produced in several European countries. INPIRE aims to supply a potential geospatial data integration service for overlaying and visualizing information from a wide range of data sources in the European Union (EU). The key goal of INPIRE is to enable all governmental levels to share and manipulate different datasets at both national and supra-national levels.

Other SDI activities include the US NSDI which was initiated to reduce the cost and improve the quality of sharing datasets. In addition, it aims to minimize the efforts of data collection among federal and state agencies and make spatial data more available to public users. The main objective of this infrastructure is to establish solution partnerships between states (Cho, 2005). The US NSDI has provided a data framework for different themes such as cadastral, orthoimagery, elevation, transportation, governmental units' boundaries, hydrographical data and geodetic control data. Therefore, it makes it easier for all people and institutions to search a wide range of spatial datasets in one user interface. However, both technical and non-technical issues can be considered major challenges for any spatial data infrastructure project, as indicated in the previous section.

Taking advantage of opportunities offered by contemporary technical developments such as web services and the Internet, new opportunities for geospatial data integration have emerged, such as integrating Volunteered Geographic Information (VGI) data into SDI. VGI can be essentially defined as geographic data that is usually collected and uploaded to Internet services by volunteers (Goodchild, 2007a). Further descriptions regarding VGI activity can be found in Chapter 3. McDougall (2009) observes that SDI has developed globally and also notes the increase of spatial data volume from private sources. Further, reliance on governmental data sources can result in problems but the integration of VGI to address such problems relies on new models for SDI. The integration of SDI and VGI involves substantial redesigning of institutional engagements and information flows. Coleman (2010) has examined such opportunities for integrating VGI information with SDI. The suggestion was that a successful integration can provide updates to data and improved datasets. This can become a

reality, if the limitations and constraints of VGI data are taken into account. These issues will be described and illustrated in the next chapter.

The free availability of data generated by volunteers can help during disasters, crisis and emergency management. For instance, the efficiency of the OpenStreetMap (OSM) project was shown when an intense earthquake hit Haiti in January 2010. Within a few days of the disaster occurring, the OSM community published emergency route service and damaged buildings maps. Digitizing the infrastructure and current situation was carried out based on contemporary satellite and aerial images. In order to obtain a more effective solution to disaster management, Neis et al. (2010) pointed to the necessity of the integration between actions carried out by official humanitarian organisations and VGI data; for example, integrating OSM data with the UN Spatial Data Infrastructure for Transportation. This may assist adding some missing information, such as up to date emergency routes or volume of damaged areas, to existing UN datasets when disasters happened and topographic changes take place.

Another example of VGI / formal data interaction was determined by Mooney et al. (2012) who integrated VGI datasets with pervasive health applications. VGI is now an active field of research and the topic of health care is of interest to many people, but although they found that this process may offer advantages such as the low cost of VGI datasets, they noticed that there were several disadvantages also. One of the main complications of this operation is the control of the VGI community. If there are problems with management of crowdsourced communities, issues regarding data consistency and data quality may appear.

From above description, it can be concluded that geospatial integration has the potential to reduce time, cost and efforts of collecting and disseminating spatial datasets. However, the integration of VGI information and authoritative datasets is still one of the main challenges for GI users.

2.5 The importance of data quality in geospatial data integration

For GIS processing such as geospatial data integration (including overlay), equivalent features may not geometrically match. Also the semantic relationships between corresponding objects may be mismatched. These differences can be due to the discrepancies between data quality characteristics. The assessment of data quality

becomes important for supporting accurate and efficient geospatial data integration. Thus it is essential for the GIS community to measure the quality of spatial data before decision making.

The interest in considering spatial data quality as a main parameter for supporting decision making has been an active area of concern for a long time. This has increased with the emergence of new spatial data collection technologies such as remote sensing images, laser scanning systems, Global Positioning Systems (GPS) and mobile GIS (Delavar and Devillers, 2010). The growth in the importance of spatial data quality may also be accounted for by the increasing amount of spatial data that has been created by private companies. However, a significant proportion of spatial datasets is still generated by government institutions; for example, Ordnance Survey (OS), the US Geological Survey (USGS), the Australian and New Zealand Land Information Council (ANZLIC) and other administrative agencies. Some of the spatial data that has been generated by these agencies is not mandatorily required to meet data quality standards. Therefore, the procedures for supporting geospatial data integration, for example, could be based on spatial information without taking into account the quality of the data's consistency. There is a distinct possibility of a subsequent interpretation or decision being wrong, if it is made depending on an integrated dataset without comprehensive quality considerations. Accordingly, data integration based on such decisions could be risky or dangerous; for instance, a building incorrectly integrated into an adjacent car park due to geometric errors, or a road integrated into a pathway class because of semantic errors. As a result of incorrect integration, there will be serious issues in using generated datasets for other GIS applications.

The topic of assurance of data quality and its potential implications on geospatial data integration have been highlighted by many authors. For example, Brimicombe (2003) reported that since 1987, the number of articles and symposia on spatial data quality have increased radically every year. The mainstream of GIS applications development is increasingly interested in data quality. Furthermore, there is a large volume of published studies describing the role of spatial data quality connected to data integration. For instance, Fonseca et al. (2002) addressed semantic heterogeneity as one aspect of geographic information integration. Edward and Simpson (2002) listed some data quality issues, such as source accuracy, errors, compilation standards and resolution that

may affect multi-source vector data integration. The influence of data quality problems was also introduced by Finn et al. (2004) as one of the most pertinent issues of the geospatial data integration process. The initial results of their investigation indicated that the integration of national datasets can only be achieved for spatial datasets that are similar in accuracy and resolution. In other words, data integration based on incompatible accuracies and resolutions of combined datasets may be difficult or impossible. Friis-Christensen et al. (2005) discussed the issue of classification name matching, in order to solve a schema heterogeneity problem across Europe. In particular, they presented an approach that included an examination of the use of ontologies to support the schema integration process. They found that applying ontologies can make the operations of schema integration more powerful; however, this required a special experience regarding data sharing and interoperability concepts which may be difficult to find in all areas of applications.

In recent years, a relatively newly-introduced difficulty of such data integration is the growth of crowdsourced spatial data sources on the Internet. In most cases of this kind of dataset, data collection properties and information (metadata) does not exist. The absence of these contents can lead to misunderstanding of the spatial dataset's quality. Without access to documented information of data quality, spatial data users have no ability to decide upon the suitability of the datasets for such integration. Further details regarding the concerns associated with crowdsourced spatial data are presented in Chapter 3. Therefore, data quality is an important factor for GIS products and it is also a major concern of the GIS community. The successful outcomes of geospatial data integration and analysis will be compromised by any inconsistency in the quality of the datasets that may be involved in the data integration process (Kumi-Boateng and Yakubu, 2010).

2.6 Spatial data quality: concepts and issues

Before discussing the problems of spatial data quality, it is necessary to understand the meaning of 'data quality'. Data quality as a concept may be defined differently, depending on the context in which it applies. There are many definitions of data quality in the literature. Each varies from organisation to organisation, application to application or person to person. For instance, the term 'quality' can be defined as an indication of high degree of craftsmanship or creativity (Veregin, 1999). In contrast,

Jakobsson (2002) regards data quality as a function of the difference between a dataset and the universe of discourse, when the universe of discourse is the actual objective world view and the dataset is the identifiable collection of any related dataset. In terms of spatial data, the notion of quality has been clarified by Korte (2001) as being the degree of how accurately the GIS data can be represented or meet a specific accuracy standard.

As mentioned in previous section, the notion of spatial data quality has been of increasing interest. The growth of spatial data exchange by means of different web technologies, such as the Internet, is a major reason for this. More generally, the development of GIS and availability of spatial data from satellites can be considered another reason for considering spatial data quality. In this context, Oort (2006) highlighted many reasons for spatial data quality concerns. For example, there is an increase in spatial data users who are less aware about data quality. Consequently, the spatial data has been used for any type of application, regardless of the suitability of their quality for a specific application. In addition, there is a potential gap in understanding between the people who produce the spatial data and have an idea about their quality and the users who use the spatial data. Thus, data quality should be an important factor in geographic information science research and data sharing between various organisations. Some GIS procedures and applications, including data integration, rely on spatial data which may be collected using different techniques, various sources, and may be in different levels of detail (Servigne et al., 2006). Hence, it is necessary to understand and consider issues and concepts related to data quality such as uncertainty, errors and accuracy. The concepts of uncertainty and errors will be described in this section, while the term accuracy will be discussed in section 2.7.

2.6.1 The uncertainty of spatial data

The topic of understanding and representing spatial uncertainty has been addressed by authors such as Griethe and Schumann (2005), Leyk et al. (2005), Foody and Atkinson (2002) and Pang (2001). One of the main challenges in geographical information science research is the conceptualizing or the definition of uncertainty. The use of the term 'uncertainty' with different meanings in different fields results in a fundamental misunderstanding of its real meaning. For example, in mathematics, 'uncertainty' describes the occurrence or lack of certain events as random. This concept is closely

related to probability approach. In psychology, 'uncertainty' can be defined as a subject of human state, such as anticipation and lack of confidence. In geographic information science, the definition of certainty is often a cloudy issue. The reason for this is that uncertainty forms an umbrella of many concepts such as error, ambiguity and vagueness (Drecki, 2007).

Many definitions of spatial uncertainty have been proposed. For example, Goodchild (2008a) defines spatial uncertainty as the discrepancy between a given value and its equivalent true value in the real world. He added that many other concepts are partially synonymous to spatial uncertainty, such as data quality or vagueness. Fisher et al. (2006) discussed uncertainty as a more general term. They frame the definition in the context of how objects or classes are defined (i.e. well and poorly defined objects). These classes can be further broken into three types, error, vagueness and ambiguity, and all of them have a degree of lacking clarity or precision. Figure 2.3 illustrates Fisher's uncertainty conceptual model and all terms are shown in a diagram. If an object is a well-defined feature such as a building which is usually created by human beings, then any errors in observations of spatial data collection will cause uncertainty in that feature. On the other hand, if the object is poorly-defined, for example where it is difficult to precisely identify its boundary (such as with woodland or vegetation), then the terms vagueness and ambiguity can be acknowledged.

Fisher et al. (2006) use the term 'vagueness' as a state of uncertainty that is associated with poorly-defined objects. Vagueness can be due to the method of observation or the nature of the object. An example is a lake whose water level depends on the amount of rainfall and the degree of evaporation. Information about the minimal and maximal extent of a lake can be obtained and portrayed. But the real size of a lake, which is somewhere between these two extreme limits is still vague (Pauly and Schneider, 2010). Similarly, human height defined as tall, medium and short reflects vague concepts and may be regarded poorly defined. Ambiguity is associated with the acuity of the specific phenomenon. It arises when there are many concepts that have the same name, but a different definition. It can be classified into two types: discord and non-specificity (Fisher et al., 2006). Discord occurs when the object is defined clearly, but may be assigned to more than one class or placed under different schemas. For instance, the definition 'soil' varies between several countries which put 'soil' under multiple classes

and schemas. On the other hand, if the assignment of the feature is unstable at all, non-specificity arises. For example, the relation of L is south of K can be considered as a clear instance of a non-specific case. This is because there are three cases in this meaning which can be represented as follows: L is south of K , but both of them lie on the same longitude; L lies south-east of K ; or L lies south-west of K .

In this thesis, uncertainty is considered to mean the lack of objective knowledge about accuracy in tested datasets. The work here with spatial uncertainty focuses on well and poorly defined objects observed in vector datasets. The processes applied here address spatial uncertainty, with an emphasis on point, line, and regional objects and are described in Chapter 4.

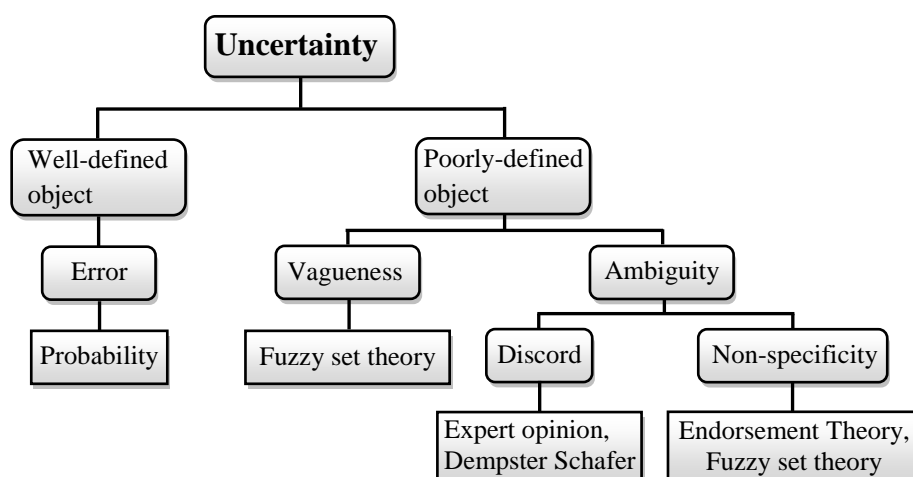


Figure 2.3 A conceptual model of spatial data uncertainty (Fisher et al., 2006).

2.6.2 Spatial data variations and errors

Although data integration can improve data representation or facilitate the processing of spatial analysis, the visual representation and the spatial analysis results will be affected by any variations or discrepancies between integrated datasets. The reasons for data errors or variations can be classified into two categories. The first category, such as the errors that occur as a result of differences of datum or projection, can be easily overcome. This kind of variation can be corrected by selecting the appropriate method of transformation. On the other hand, errors within the second group of reasons for variation are more difficult to correct. These can be represented by data quality elements which usually vary between different datasets, as will be described in the next section.

The various sciences that use and manipulate spatial datasets have usually measured and processed physical quantities. The utilisation of such measurements is often affected by different errors or variations. These discrepancies can appear in one of the three types of errors: blunders, systematic errors, and random errors (Chilani, 2010). Blunders are simply the mistakes that usually result from observer carelessness and can result in large errors; for example, by reporting inaccurate observations or observing incorrect samples. This kind of error gives extreme results which are clearly different than other results. Thus it is normally easy to recognise them as they are such large errors. Errors of this sort cannot be treated conventionally by statistics and therefore must be removed. Systematic errors have generally a constant magnitude occurring across a series of observations. These errors may be caused by mismatching between the ideal conditions and real conditions of observations, such as incorrect calibration of an instrument that has been used to collect spatial data.

The remaining discrepancies are usually known as random errors. This type of error may result from unknown reasons that are usually out of the control of the observer. In other words, these errors occur accidentally. The different technologies for acquiring spatial datasets have made a significant difference in the accuracy and errors in various datasets. For example, GPS is a widely used device to collect positioning data for many applications such as VGI datasets, as it is available in mobile phones or in-car navigation. However, the receivers of GPS can provide positions with varying degrees of accuracy. For example, some hand-held receivers (e.g. Garmin) can collect data that is accurate to within a range of 5 to 10 m. More advanced receivers can be linked to network reference stations to supply even more accurate results (i.e. some centimetres or millimetres). Similarly, in recent years freely available satellite and aerial images such as Yahoo imagery have been widely accessible. These images are being used for different applications such as digitizing and producing vector datasets. However, the positional accuracy of such product datasets will vary due to the variation in image resolution.

In general, the errors or the variations between datasets are fundamentally different and variable between different data sources. Thus, it is essential to take into account the accuracy of each dataset that may have been used for GIS processing such as geospatial data integration.

2.7 Spatial data quality elements

Spatial data quality usually provides information regarding the spatial database such as the flexibility and usability of data. The terms 'accuracy' and 'errors' provide one of the methods that can be employed to assess quality. Different types of quality can be measured by these two terms. For example, the errors concept measures the discrepancy between the measured and reality data, while accuracy can measure the differences from the modelled value, as discussed in the previous section. The quality measure can be represented by various components and many authors have highlighted these parameters; see for example, Lo and Yeung (2007), Devillers and Jeansoulin (2006), Shi et al. (2002), and Burrough and McDonnell (1998). In general, most researchers agree that the measure of data quality consists of the following components: positional accuracy, thematic accuracy, temporal accuracy, completeness and logical consistency.

Data quality components describe the ability and the purpose of using datasets for specific applications. For instance, most dimensions of geographical phenomena such as theme, space and time can be represented by spatial data quality parameters. The quality of such dataset can be specified by one or more quality parameters. Hence it is not a necessity that all components of spatial data quality are required for all GI processing. For example, Butenuth et al. (2007) reported that the discrepancies in geometric and thematic accuracy may hamper the integration of multi-sources spatial datasets.

Data quality can be measured by many diverse metrics and the project here describes several models for enabling assessment of the ability of integrating geospatial datasets from official and informal sources. The following subsections present an overview of the main components of spatial data quality.

2.7.1 Positional accuracy

Positional or spatial accuracy measures the quality values of geographic features' positions. It can be described by absolute and relative accuracy of the location of the object in spatial datasets. Absolute accuracy is the degree to which the coordinates' values of the feature in a dataset are close to the true or correct location with respect to the reference system. Relative accuracy refers to the accuracy of a data point in relation to other objects in the same coordinate reference (Choudhury et al., 2009). Furthermore, the data quality element 'positional accuracy' can be classified into horizontal and

vertical accuracy. These types of accuracies can deal with plan and height positions of objects in two or three dimensions with respect to specific datum. In addition to positional accuracy, there are other measures related to location accuracy such as geometric accuracy or accuracy of shape. These measures can be applied to other mapped features beyond positions such as linear or closed objects. There are various metrics that are used to determine positional or geometric accuracy, as described in chapter 4.

2.7.2 Attribute accuracy

The information that is usually assigned to features in spatial datasets is either qualitative (features' classifications or names) or quantitative attributes (statistical information, measurements). The attribute or thematic accuracy can be applied to refer to the accuracy of these categorical and quantitative attributes. Therefore, the metrics for measuring attribute accuracy depends on the nature of the information or attributes. For quantitative or continuous data, quality can be expressed in the same manner as measuring positional accuracy (e.g. RMSE) (Veregin, 1999). On the other hand, there are various metrics that can be used for the purpose of measuring nominal attributes' accuracy. The categorical data may be classified incorrectly; for example, as a road instead of a cycleway. Thus the values of attribute accuracy would be evaluated as right or wrong attributes, rather than absolute and relative accuracies as in positional accuracy. Furthermore, as the nominal classes are often used in vegetation or land use datasets, the accuracy of categorical attributes' values can be measured as a percentage of correctly classified data. For instance, the value of the nominal attributes may have fallen between two categories, such as 60% trees and 40% grass. Hence, if the attributes of any spatial feature are defined incorrectly, the effects may be enormous. For example, for geospatial data integration purposes, if a building has been defined as green area, the data integration process may yield wrong results. There are more complex models and techniques which can be used for measuring semantic or attribute accuracy, as shown in chapters 6 and 7.

2.7.3 Temporal quality

Temporal quality is often used to refer to the quality of the time that has usually been recorded within spatial datasets. Spatial data may have several different aspects of

temporality which can affect its quality. In general, there are three sub elements of temporal quality: temporal consistency, temporal accuracy and temporal validity (Stein, 2010). Temporal consistency refers to the appropriateness of the sequences of orders of events. For example, there is a temporal inconsistency if the date of the deletion of any feature from the database precedes the date of entering the same feature into the database. The temporal accuracy can be expressed as the correctness of the information at the time of reporting. For instance, if the time of the collection of data is recorded as 9:00am on a specific date, but the actual time was 11:00am, then the accuracy can be measured between the real and reported information. The temporal validity can be defined as the measuring of the validity of the dataset with respect to time, such as considering spatial data of one year ago to compare with data that has been collected more recently. Hence temporal element can take several forms and aspects that may affect the quality of spatial datasets.

2.7.4 Completeness

The data quality factor of 'completeness' deals with the correspondence relationship between the features in the real world and the features in digital format. It measures the omission and commission of spatial data with respect to data specification (Veregin, 1999). Omission shows the absence of data from the database. The error of omission can be considered when an object is not included in the data, but it already exists in reality. For example, 30% of the buildings omission means that 70 out of 100 buildings have been mapped and the rest might be missing as an omission error. Commission describes the exceeding of the data that has been presented in the database. For instance, 20% commission for the value of road features of such dataset refers to the fact that the roads are populated for 20 of 100 features on a dataset, while these features are not defined in the real world. If these road features are included in a dataset then this will be a commission error. Thus, data completeness can be described by checking whether the features in the database are completed or in progress.

2.7.5 Logical consistency

Logical consistency is a measure of how data conforms to the structure of the specifications that have already been defined in the database (Veregin, 1999). In other words, it is a measure of the percentage of the conformance of the rules that have been

defined by data users. The quality element of logical consistency is often separated into many sub elements. From this perspective, Harding (2006) illustrated that the Ordnance Survey usually check three elements of logical consistency: topological consistency, validity of recorded structure and validity of values. Examples of such rules are checking the connectivity of all polygons in order to make sure that all polygons are closed; all objects should be connected to each other in transportation networks; and the bridges should be signed at the intersection of the rivers with roads. The values of logical consistency can be reported as a percentage, number or ratio of the affected, for example.

2.8 Standards for reporting spatial data quality

As data quality has become a major concern with regard to most GIS processing, the importance of data quality documentation has increased. The main purpose of spatial data quality reporting is to provide effective mechanism to access and share datasets. The data quality information is usually reported in the contents of metadata. Metadata can be discussed as data about data or data that can be used to make the data more useful (Boin and Hunter, 2006). Data producers and consumers usually use metadata to reduce the risk of data misuse. They also use it to understand the limitations and the suitability of the datasets. In order to simplify the handling and disseminating of spatial datasets, standards should be applied to report data quality documentation.

From the mid 1980s onwards, many countries around the world with institutions handling spatial data have operated to establish such data quality standards. Several standards have been developed by international, national and sub national organisations in order to describe and document data quality. Although data quality elements may be defined by different standards and approaches, Moellering (1997) reported that most of the standards of spatial data quality adhere closely to the US Spatial Data Transfer Standard (SDTS).

Since 1992 the SDTS has been implemented by many government and private spatial data organisations in the USA. It defines five elements of spatial data quality: lineage, positional accuracy, attribute accuracy, logical consistency and completeness. Most of the spatial data community across the world accepted SDTS's five elements. Subsequently, this standard has been revised many times in the USA (ANSI, 1998). The

revised version suggested including cloud cover of remote sensing dataset as an optional sixth data quality element. The following sub-sections provide a brief overview of various typical data quality standards that have been suggested by several organisations.

2.8.1 FGDC

The increase in the need for efficient data distribution and dissemination has encouraged US agencies to develop US (NSDI). One part of NSDI's task was initiating and establishing a metadata content standard by the Federal Geographic Data Committee (FGDC). Initially, the standard entitled Content Standards for Digital Geospatial Metadata (CSDGM) was formed as a draft standard. After undergoing the process of many reviews, it was finally accepted in 1994 (FGDC, 1994). The data information is structured and documented at CSDGM in many sections such as identification, attributes and entities, information on spatial data institutions, distribution, information on data quality elements and spatial reference information. This standard provides a wide range of definitions and terminologies for metadata elements. The basic properties of a dataset can be described by utilising information such as the quality description, data format, resolution, reference system and the coverage extent. The main section of this standard is the data quality information part. In this context, the standard followed SDTS standard to report data quality components and suggested the same five data quality elements as those of SDTS (FGDC, 1998a).

2.8.2 ANZLIC

In Australia and New Zealand, the ANZLIC is the governmental spatial information agency for serving spatial datasets. It also supplies data standards and enables users to access online spatial information. To achieve these aims, metadata guidelines have been developed by ANZLIC. The first standard was established in 1996 with a view towards assisting the spatial information community to manipulate metadata elements, whereas the second version was developed in 2001 (ANZLIC, 2001). The metadata of ANZLIC standards are grouped into ten different sections. The components of the data quality subdivision are the same as FGDC elements. These include positional accuracy, attribute accuracy, lineage, completeness and logical consistency.

2.8.3 CEN

Another example of geographic information standards was developed by technical commission 287 of the European Committee Standardisation (CEN). The first attempt at establishing a European draft standard was instituted by CEN in 1991. The aim of this standard is to describe, define and structure spatial data in a standard way. In addition, it aims to create a standard system for updating and transferring geographic information. This can assist different users to access geographic information from various locations. The general principles for discussing spatial data quality have been established by this standard. For instance, a full description of the usage and lineage of a specific geographic dataset can be found in this standard. According to (CEN, 1998) the standard provides other spatial data quality components which go beyond FGDC and ANZLIC standards. This includes semantic accuracy and temporal accuracy. Semantic accuracy refers to the accuracy aspects of the semantics of spatial datasets, a feature which will be discussed in more detail in chapter six. The concept of temporal accuracy has been explained in section 2.6.3.

2.8.4 ISO 19115

In the last decade the need for the unification of standards to manipulate spatial data has increased considerably. One reason for this requirement is the movement towards universal spatial data interoperability. The International Organisation for Standardisation (ISO) responded to this necessity by forming the Technical Committee for Geographic Information. One of the main responsibilities of this team is establishing international standards for spatial datasets. Their initiative began by developing a group of spatial metadata standards named ISO 19100 series. One of the main standards of this series is the ISO 19115 standard. The design of ISO 19115 was influenced by many other standards such as FGDC, ANZLIC and CEN. This was to ensure that the standard could accommodate different international requirements.

This standard was released in 2003. It divided data quality aspects into three data quality overview elements and five data quality parameters (ISO/TC211, 2003). The group of overview elements includes purpose, usage and lineage which are usually used to describe the non-quantitative quality information. On the other hand, data quality elements contain quantitative information. These are the same as STDS elements, with

the exception of the inclusion of temporal accuracy as an additional quality element and changing the attribute accuracy into thematic accuracy.

2.9 The assertion of data quality and credibility of alternative data sources

In most cases, the VGI data on the web may not contain any information about their quality. From this perspective, Flanagan and Metzger (2008) supposed that the VGI data may improve spatial data content in general; however, the quality and accuracy of this data has still attracted the most attention to date. There are many reasons making VGI quality information extremely significant. For instance, the increasing of the decision making procedure based on the information of spatial data and the possibility of integrating different datasets which can be used for more GIS analysis and applications. The dependability of VGI data quality should be taken into consideration by people who have been collecting and disseminating this information. As will be described in the next chapter, VGI data is usually collected by volunteers; thus its quality will vary and nobody can guess or know the value of it. This drawback has been agreed upon by authors such as Haklay (2010) and Auer and Zipf (2009).

There are several legitimate criticisms that make the assessment of VGI quality difficult. For example, there is an enormous variety of people who contribute VGI data and there is no unified authority whose role is to assess the quality of spatial data. Additionally, because of the different perspectives of data developers, it is highly likely that heterogeneities will be found in resulting datasets (Elwood, 2009). This inspired Exel et al. (2010) to include crowdsourced dynamics as an indicator of crowdsourced spatial data quality determination. They aimed to establish spatial data quality operational indicators for both user and feature quality. Their proposed approach fundamentally considered different crowdsourced activities such as the number of editors or edits per feature and the historical (or temporal) information of the features which includes the development of such features over time. This suggested framework may assist in measuring the density of edits to an area of crowdsourced data and ultimately assessing its data quality.

Elsewhere, Goodchild and Li (2012) have argued that although VGI may offer numerous advantages such as the free availability and accessibility of spatial datasets, the quality of VGI data should be considered as a vital issue as VGI data does not

follow a standard structural design. Therefore, they investigated three different approaches to assess the quality of VGI data. Such quality assurance approaches are firstly, validation by crowdsourcing, secondly the social approach, relying on a hierarchy of 'trusted' individuals, and finally the geographic approach, which examines the probability of features being correctly located with reference to the surrounding context and geographical area. Subsequently, they compared these approaches with the quality assurance approach that is usually used by traditional mapping agencies. Some analysts (e.g. Hagenauer and Helbich (2012)) have pointed out that VGI data quality issues, especially completeness, can affect the fitness for use for such applications (e.g. urban planning). Therefore, they suggested a methodology to calculate through OSM data which urban areas in Europe are mapped or partially mapped. Their results found that the delineations of urban areas are based on the location.

The increase in the amount of data has also led to an increase in the heterogeneity between datasets. For instance, within different datasets, the features may be varying in accuracy due to the methods or skills that were employed for the purpose of collecting data. According to Haklay (2010), the distribution of errors in VGI data is usually based on the carefulness of each contributor. Therefore, the concern of trust of VGI data quality is the main issue facing the GI community. These heterogeneities may be especially problematic when the integration of multi-source spatial datasets is the target, as will be demonstrated in chapters 4 and 7 of this thesis.

2.10 Practical testing of existing attempts of geospatial data integration

In addition to the brief summary of VGI data quality issues and limitations, this section also presents the research literatures related to quantitative measures for evaluating VGI quality and discusses their outcomes. For example, Haklay (2010) examined the positional quality of OSM information by comparing it with OS-Meridian 2 datasets. The Meridian 2 dataset supplies detailed data of road networks in Great Britain such as motorways, minor and major roads. In addition to use more data sources in order to complete this investigation. These involved the 1:10,000 raster files from OS, and some data about the neighbourhood size which is based on Census from OS and national statistics office. The main focus of Haklay's work was limited to the measuring of the quality of roads or motorways of OSM datasets. The methodology which was applied to assess the quality of motorways of OSM data was based on approaches by Hunter

(1999) and Goodchild and Hunter (1997). The method of buffer was adopted to determine the accuracy of such lengthy objects by applying a certain distance of buffer size for this test. The results of the analysis showed that the average of overlap percentages when comparing OSM with OS datasets were approximately 80%, 88% and 77% for motorways, A-roads and B-roads respectively. The findings of Haklay's study also showed that the quality of OSM data is variable when compared to OS datasets within the average of 6m of positional accuracy.

Koukoletsos et al. (2012) provided an automated feature-based matching approach for assessing the completeness of VGI linear datasets through this matching processing. Their proposed method was fundamentally based on a multi-stages procedure that combines geometric and attributes elements. The OSM dataset was compared with the Ordnance Survey-Integrated Transport Network (ITN) layer dataset, as a formal or reference spatial data source, in a number of study areas around the UK. The results of their analysis found that matching errors were between 2.08% and 3.38% for urban and rural study areas respectively. Consequently, data matching processing proved its effectiveness and calculating OSM data completeness for small areas (tiles) can provide more heterogeneous and effective results.

The quality of VGI data has received more attention from the GI community; for example, the investigation that was carried out by Zielstra and Zipf (2010). They studied the quality of the routes and roads of OSM data in Germany. The OSM information has been compared with a commercial dataset known as Tele Atlas. They based their approach on that suggested by Goodchild and Hunter (1997) to measure the quality of linear features. Their results found that the overlap percentages between the roads of OSM data and Tele Atlas datasets were $\geq 80\%$ for most of the roads in major cities. In addition, they reported that the overlap percentage in towns of medium size was between 50% and 80%. Completeness has also been considered by Zielstra and Zipf (2010) as another factor of geospatial data quality. They argued that the results of their comparisons revealed that the positional accuracy seems quite good and the OSM data can be used for many routing applications. However, they found that there are still shortcomings in the completeness of the regional OSM datasets.

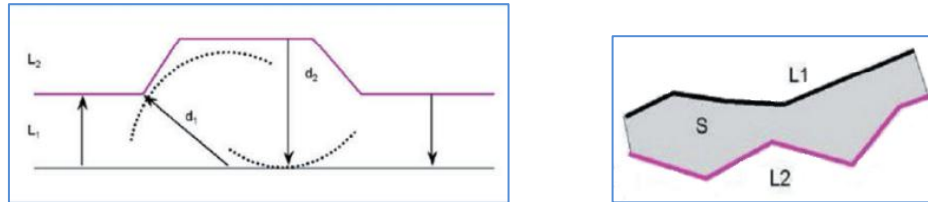
Another significant study in the comparison of OSM data quality with commercial spatial data sources was introduced by Ludwig et al. (2011). They stated that OSM data quality should be assured for commercial and geomatics purposes of the required area. Therefore, they compared OSM road data with Navteq data for the main and the most populated roads in Germany. The comparison technique was basically based on developing a fully automated matching methodology which could be directly applied to update both datasets. Their findings indicated that the relative completeness of features and their names, and relative precision of the quality of OSM data, were relatively high enough for maps at national level. However, there were considerable qualitative differences of OSM attributes and completeness at local level, as well as between the comparisons of towns or regions which are still relatively incomplete.

In Germany, further project research to analyse OSM datasets was conducted by Neis et al. (2011). In this case, the OSM data was compared with TomTom's commercial datasets for total network streets and route car navigation for the period from 2007 to 2011. They concluded that the total street network and pedestrians' route information of the OSM project in Germany had 27% data more than TomTom's commercial datasets. However, their analysis showed that about 9% of the OSM dataset related to car navigation routes is still missing when compared to those of commercial dataset.

Girres and Touya (2010) assessed the quality of OSM datasets in France. In their investigation, many quality elements for OSM data were evaluated. The more interesting parameters that will be illustrated and described in this thesis are geometric and semantic accuracy. Their analysis included the comparison of OSM data with the French National Mapping Agency geographic datasets. The results of their positional analysis of the road intersections indicated that the most frequent positional differences ranged between 2.5m to 10m, and the average value of the positional differences was nearly 6.65 m.

The differences between linear features were calculated by applying two techniques: the Hausdorf distance approach, which computes the maximum distance between the compared linear features, and the average distance approach, which takes the average distance between the compared polylines, a method suggested by McMaster (as cited in Girres and Touya, 2010). The principles of these methods can be seen in Figure 2.4. The

results of their study showed that the mean difference values between the compared roads were about 13.57m and about 2.19m for the Hausdorff distance and average distance methods respectively.



a- Hausdorff distance method

b- average distance method

Figure 2.4 The methods that have been adopted by Girres and Touya (2010) to determine the linear differences between road features

The same study also considered the differences between polygonal objects of lakes. The surface distance method which was proposed by Vauglin (as cited in Girres and Touya, 2010) was adopted to quantify polygon differences, as shown in Figure 2.5. The method is based on the common area of the two compared objects. The d_s value will be zero if polygon A is equal to polygon B , while it will be one if A is not equal to B . The results of this method showed that there is a small difference between the polygons of the comparison datasets. Tag names analysis was also included in their study. They found that nearly 100% of roads that were classified as motorway and primary were similar to the classes of the national datasets. However, only 49% of secondary class roads were correct between compared datasets.

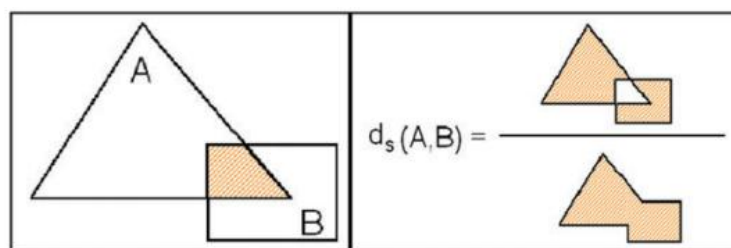


Figure 2.5 The surface difference approach that has been used by Girres and Touya (2010) to calculate the differences between polygons

As VGI is a relatively recent topic in the geographic data community, studies of their quality evaluation are fairly limited. In this work, the issues of assessing the possibility of data integration between formal and informal data sources have been investigated by developing different tools and models.

2.11 Chapter Summary

Although multi-source geospatial data integration can reduce time and effort for spatial data users and communities, spatial data quality remains a fundamental issue with regard to successful integration. The difficulty of spatial data integration becomes increasingly complex as more and more non-professional users have access to spatial data such as VGI datasets.

This chapter provided a review of different concepts and definitions that can be used to discuss multi-source geospatial data integration processing. The chapter then proceeded to describe the problems and issues that may influence geospatial data integration processing. The discussion included two types of main issue: technical and non-technical concerns. This was followed by highlighting the contemporary applications of geospatial data integration such as INSPIRE and US NSDI. With increasing numbers of people taking advantage of open source datasets, more applications have also become involved; for instance, integrating VGI data into SDI context or integrating VGI data into official humanitarian organisations' data to help during the course of disaster management. The possibility of integrating VGI datasets with pervasive health applications was also addressed. In addition, the chapter considered the necessity of data quality for geospatial data integration processing. A summary table was prepared in order to present the study findings for each document reviewed with regard to geospatial data integration issues, as shown in Table 2.1.

Table 2.1 A summary of the issues of geospatial data integration processing

Citation	Study description
Usery et al. (2005)	The geospatial data integration processing across different companies or agencies may be problematic because they are managed by different communities and people from different backgrounds.
Mohammadi et al. (2008)	The challenges and issues of geospatial data integration are discussed. They mention that successful integration not only includes technical data integration processing, but also highlighted non-technical issues such as social, legal, policy and institutional issues which should be taken into considered.

Continue

- Mohammadi et al. (2010) Their article initially investigated the technical and non-technical complexity and challenges of geospatial data integration. They proposed a tool to facilitate the integration of multi-source spatial datasets.
- Edward and Simpson (2002) The mismatching of overlaid datasets may be due to the discrepancy of currency, source accuracy, errors, temporality, sensor characteristics, resolution and scale of compared datasets.
- Finn et al. (2004) There are several significant problems that should be taken into account when the integrating data theme for National maps is the targets. This fundamentally includes differences in coordinate systems, precision and accuracy, projections, data models and datums for combined datasets.
- Thelluksen et al. (2009) They mentioned that the social issues of geospatial data integration are one of the main constraints that should be considered when inter-organizational sharing of geospatial data is the target.
- Weaver (2004) The diversity of the procedures that are followed for storing spatial datasets in different spatial data agencies, such as variations in maintaining coordination agreement and also the weak collaboration between such spatial data institutions, is a non-technical or institutional issue that may lead to difficulties of in achieving successful geospatial data integration processing.
- Casado (2006) The problem of geometric conflation was addressed. The geometric conflation algorithm was proposed in order to try to transfer geometric data for one dataset to another and minimise the geometric difference between them.
- Butenuth et al. (2007) The discrepancies in geometric and thematic accuracy and correctness may obstruct the combined geospatial data integration. This may occur because of the diversity of spatial data collection methods or the different methods of updating
-

Continue

McDougall (2009)	With the emergence of VGI, examination of the risks and the opportunities for using these data to update and enrich governmental spatial data sources by means of integrating processing, relying on new models of SDI.
Coleman (2010)	The integration of VGI within SDI can be a reality, if all VGI limitations and issues are taken into consideration and included in such a data integration framework.
Mooney et al. (2012)	Although using VGI data for processing such as integrating with health care data may advantageous, the uncontrolled management of VGI communities may lead to issues of spatial data quality and heterogeneity.
Fonseca et al. (2002)	Semantic inconsistency can be considered as one of the most significant factors that may affect the process of multi-source geospatial data integration.
Friis-Christensen et al. (2005)	Approaches presented in order to support effective schema matching processing to satisfy the requirements of exchanging and sharing spatial data across European applications.

It can be observed from the table above that several authors have examined various issues and problems that may make geospatial data integration more difficult and complex. In this research the focus will be on the evaluation of spatial data quality such as geometrical and semantics issues for multi-source geospatial data integration purposes.

Subsequently, several limitations and concepts related to spatial data quality research were also introduced in this chapter. The emphasis was on many different terminologies and terms such as certainty, error and accuracy. A detailed description of spatial data quality contents was also included. The main aim was to discuss the key elements of spatial data quality such as positional accuracy, thematic accuracy, temporal accuracy, completeness and logical consistency. The information on data quality is essential for valuable geospatial data integration. The exploration of data quality management also

involved examining several standards for reporting metadata and data quality. A brief exploration of the most popular national and international standards such as FGDC, ANZLIC, CEN and ISO 19115 was also carried out. The main purpose of spatial data quality reporting is to provide an effective mechanism to access and share datasets.

In order to understand the legitimate criticisms that the assessment of VGI quality is difficult, it is necessary to study previous researches and investigations which place emphasis on the assurance of VGI data quality. Furthermore, it is important to present research literatures related to quantitative measures for evaluating VGI quality, as was illustrated in the previous two sections. The next chapter will present an insight into the nature of authoritative and VGI datasets that have been used to assess the ability of integration of VGI and formal data sources.

Chapter 3 Geospatial Mapping and VGI Databases

3.1 Introduction

A map may be broadly defined as a visual representation of the phenomena on the earth's surface as they appear viewed from above. The recognition and distinguishing between different features on a map can be facilitated by using different colours, lines and symbols to represent map features. This thesis uses one particular kind of map, known as topographic maps. These types of maps are usually produced by official mapping organisations or agencies; however, nowadays they can be created and shared by non-professional people in the form of volunteered geographic information (VGI), as mentioned in Chapter 2 and will be described in further detail in the following sections. According to Hatzopoulos (2008), the term 'topographic mapping' is generally understood to mean the science of deriving a geometric representation of natural and human made topographic features such as hills, mountains, railways, roads and buildings by locating spatial points near the earth's surface. Historically, topographic maps and site plans were produced on paper sheets using traditional field or aerial surveys. Nowadays, with the emergence of new technologies such as the advent of computers and web service, the paper maps may not serve to satisfy all the requirements of GI users. Therefore, the concept of traditional topographic maps has changed dramatically into a digital concept by digitizing paper maps or producing direct digital maps. This digital flowline may make several geographic information science (GIS) processes, such as multi-source geospatial data integration, more efficient.

This chapter describes the current status of map development methodology of the national mapping agencies in Great Britain and Iraq, specifically the Ordnance Survey (OS) and General Directorate for Survey (GDS), as they are used as example of formal spatial data sources in this thesis. The chapter then presents alternative spatial data source (i.e. VGI) that contributes to free to use and faster to create topographic mapping activities. One aspect of the growth of VGI data is the possibility of being able to upload, share and change spatial datasets on the web. A detailed description of the technologies, such as Web 2.0, that enable the development of VGI data is included. Other concepts that are closely related to VGI, such as the user generated content (UGC) concept, which essentially include all 'free' data on the Internet, are also discussed.

Several VGI examples are demonstrated and discussed in order to understand the performance of VGI in further detail; however, the main focus is on OpenStreetMap (OSM) information, supplied as a spatial dataset which is the main interest in the comparisons made in this study. Finally, but before summarising the chapter, a comparison between official and informal datasets is addressed and viewed as an important initial step of this research to examine whether it is possible to integrate the official and VGI datasets together.

3.2 Official topographic mapping

The term official or formal spatial information has been applied to situations where spatial datasets have been provided by national governments which are in most cases recognised as the most trustworthy and accurate geographic datasets. In this thesis, two different formal data sources have been adopted: the data from the Ordnance Survey (Great Britain) and the data from the General Directorate for Survey (Iraq). In order to use these spatial datasets for this research effectively, it is necessary here to clarify exactly what is meant by the descriptions and the properties of each of these data sources.

3.2.1 Current Ordnance Survey (Great Britain) MasterMap data

The Ordnance Survey (OS) is the national mapping agency of Great Britain. The main responsibility of this agency is to perform topographic, geodetic and photogrammetric surveying for the whole country and supply formal mapping to its customers who range from central government and defence to schools and recreational users. The OS was established in 1791 and since that date it has published mapping of Great Britain at different scales in relation to the varying requirements of each period. For example, up until the 1850s, maps at scales 1: 63,360 were produced for England and Wales. In 1858, the scales 1: 2500, 1: 10,560 and 1: 500 were adopted for the maps of cultivated, uncultivated and town areas respectively. The OS maps were published based on a Cassini projection until 1945. After World War II, one National Grid referencing system was decided upon, in addition to introducing a single Transverse Mercator Projection, following the recommendations of the Davidson Committee (Parry and Perkins, 1987). After that period the development of OS production has undergone notable changes. For instance, Ridley et al. (1997) reported that between the seventies

and nineties of the last century, all OS maps were changed into digital format to construct the National Topographic Database (NTD). This database contains the geospatial data that represents the essential OS products which are currently represented by several types of information such as the OS MasterMap datasets (Holland and Murray, 2000). These are the fundamental descriptions of Britain's geographical features at the largest scale possible.

The OS MasterMap is a reliable framework for the national topographic features database of Great Britain. It basically contains a wide range of information which is offered as four different, but complementary, layers. These consist of the Integrated Transport Network (ITN) layer, the topography layer, imagery layer and address layer. The OS MasterMap layers fundamentally cover the whole of Great Britain with about 450 million geographic features. In order to identify each feature inside the MasterMap database, a unique reference number has been located for every feature. This is known as a 'Topographic Identifier' (TOID) and consists of a unique 16 digit number for each feature. The layers have provided different types of datasets which are positioned on the British National Grid. For instance, the ITN layer can supply information about the structure of the road network and routing system which may assist drivers to select the most convenient route for a trip; the topography layer involves graphical representations for individual topographic features; the imagery layer includes the aerial coverage of Great Britain by orthorectified aerial images; and lastly, the postal addresses with geographic coordinates for residential and commercial properties have been united in the address layer (Ordnance Survey, 2012). The work described here adopted the topography layer as one source of formal spatial datasets, as the main focus of this study is to assess the integration of topographic features from different data sources.

3.2.2 The structure and characteristics of the MasterMap topography layer

The primary purpose of the topography layer is to offer detailed topographic data of Great Britain to the customers from different organisations. The current items in the OS MasterMap topography layer consist of physical objects of different geographical features such as roads, buildings and fences. The real world features are represented within the topography layer through basic units which usually include points, lines and polygons. The objects within the OS MasterMap topography layer have been structured into nine themes: administrative boundaries, buildings, heritage and antiquities, land,

rail, roads, tracks and paths, structures' terrain and height, and water. The main features that may be found in the data contained within the group of each theme can be logically identified from the name of the theme. For example, the theme of administrative boundaries shows the parliamentary boundaries and local government boundaries of Great Britain, while the buildings theme contains roofed constructions' features, and so on (Ordnance Survey, 2009).

Many new features have been designed and introduced to the OS MasterMap topography layer by taking advantages of the advancement of technologies in software development and spatial data management tools. One of the most important aspects of this is the storing of the OS MasterMap topography layer data along with its attributes in a standard form of database. This has enabled GIS or other databases' querying tools to manipulate the data of the topography layer in an easy and straightforward way. The sophisticated nature of the topography layer data can provide a solution to the needs of several organisations' requirements. For instance, the utilities companies may need to visit their infrastructure for maintenance, repairing or to change and add new parts. The crews can obtain assistance to locate the exact area that needs service by viewing their assets against topography layer features. The examination of the products of such utility organisations against topography layer data can provide a better customer service and reduce costs as well (Ordnance Survey, 2009).

The physical features in the topography layer are usually accompanied by attribution sets. These attributes can offer more information about each feature. For instance, information about the TOID that is suggested for each feature, meaning information that is related to the name of features. The customer can order the OS MasterMap topography layer data via an online service that has been developed by the OS. It supplies data in Geography Markup Language (GML) format as a seamless dataset. The spatial data of the topography layer are usually updated with the data from field or aerial surveys that are carried out by specialist teams employed in the OS agency (Ordnance Survey, 2009). As stated earlier, the spatial data from the topography layer have been adopted for the current work as one source of formal spatial data source in order to implement the main aim of this thesis, as will be seen in the following chapters.

3.2.3 *Formal data in a contrasting agency –the Iraqi General Directorate for Survey*

In Iraq the first surveying work was carried out at the beginning of the twentieth century, particularly at the end of World War I, with the assistance of the Survey of India. The first official Iraqi surveying office was established in 1917, known as the Office of the Director General of Surveys. Initially, the work of this office was limited to producing cadastral maps and large scale maps for the area between the two rivers of Iraq, the Tigris and Euphrates. Subsequently, between the thirties and forties of the last century, maps at a scale of 1:20,000 and 1:50,000 were produced for the central and southern parts of Iraq. The development of aerial photogrammetric surveying, especially after War World II, has affected map production procedures throughout the world, but especially in Iraq. At that time, the aerial photographic work was initiated in Iraq by foreign agencies; however, map production work was achieved at the survey department of the Ministry of Irrigation in Baghdad. The main output of these activities was agricultural maps and photomosaics at scales 1:10,000 and 1:20,000, in addition to producing a general map at scale of 1:250,000 (Bohme, 1993).

The modern requirements of accuracy led to the need for more surveying and triangulation work in Iraq to replace the original survey of India mapping. To satisfy these requirements, therefore, new triangulation was achieved in Iraq in the 1970s. In 1974 an agreement was signed between Iraq and Poland in order to perform surveying work for the whole of Iraq and produce maps at different scales. This essentially included the establishment of the first class of the horizontal and the vertical control points for the country, in addition to producing topographic maps from aerial surveying. In this context, Bohme (1993) stated that Poland assisted to produce 1500 map sheets at 1:25,000 between the period of 1974 and 1978. The Iraqi data was produced using Lambert projection until 1970 and Transverse Mercator projection since 1970. The Polish work also involved the intensifying of the networks of the horizontal and the vertical control points in Baghdad and also producing large scale maps of 1:500 for Baghdad.

The current national mapping agency in Iraq is called the General Directorate for Survey (GDS). The main responsibility of this agency is represented by all surveying works such as maintaining the Polish horizontal and vertical control points, setting the fundamental points using GPS and offering topographic mapping to customers.

Although the current maps of Iraq are based on Polish surveying works, the new development of the technology of collecting and producing spatial datasets has changed the traditional concept into a digital concept. Nowadays, there are new digital maps for Baghdad that can satisfy GIS requirements. These data are created by digitizing processing or tracing aerial images. For the work achieved in this thesis, spatial datasets from GDS have been used as other sources of formal datasets in addition to OS datasets. These spatial datasets have been used in order to test the possibility of integrating them with VGI datasets such as OSM information, as can be seen in Chapters 4 and 7.

3.3 Contemporary aspects of informal data handling

In broad contemporary data terms, informal or open source data can be defined as any data that is available free to any user without the limitations of copyrights. Before going into more detail and describing what open source data means and refers to, it would be better to have an understanding of the technology that is applied in order to offer an appropriate environment for developing and disseminating this kind of information, as will be explained in the following.

3.3.1 Web 2.0 technologies and the development of geotagging

The past 30 years has seen the start and the development of the Internet technologies which were initially used to obtain information only (Harris, 2008). However, in recent years, the advancement of web technologies has favoured the design of new patterns and practices models on the web extending beyond passive receiving of data. These cumulative developments are grouped under a common concept known as Web 2.0. The first official introduction of the term Web 2.0 was in the first conference of Web 2.0 (O'Reilly Media) by Tim O'Reilly in October 2004. Although the term Web 2.0 indicated a major change in the approaches of software developers on the web, it did not mean a new version of the World Wide Web. Web 2.0 has been described essentially as being a platform that can collect together different sites and software and make them easily available and useable to users (O'Reilly, 2005).

The emerging of Web 2.0 technologies has led to significant changes in the methods of producing, processing, sharing and spreading information through the Internet (Rinner et al., 2008). One consequence is enabling users to collaborate and interact among each other more easily and effectively. This is represented by the availability of a variety of

social networking and communicating sites such as Facebook, Twitter, Flickr and YouTube. By using these facilities, it is possible to upload pictures or videos for public sharing; at the same time, it is also possible to make comments on the postings of others. Both of these practices facilitate the sharing of huge amounts of information on the web. Hence, nowadays not only professional users, but also non-experts, can generate and publish information on the Internet.

Tagging can be considered one of the features or techniques that have been most typically used by Web 2.0 websites, a process that can be simply defined as single or multiple words which are usually created by users and attached to their contributions in order to describe the characteristics of their uploaded information. This process can also carry out functions such as helping in classifications and retrieving items on the web. For instance, a post showing a movie for the final of the UEFA Champion's League football match may be labelled as 'a super match' or 'a final match' when it is uploaded by the producer into the one of the social sites. This can give an idea about the content of the item, which makes it possible to reach the required information more quickly. The tagging process may also include adding geographical identifications (latitude and longitude coordinates) or time to any online item. This notion of the extension of the tagging procedure is commonly known as geotagging. The geographic coordinates' data can be either obtained from built-in GPS in mobile phones, for example, or added physically by users, while the time stamp is usually added automatically into the photos file by the camera itself. One important aspect of the geotagging concept that is different to the common tagging process is the need for preparing the location metadata (positional information) for the uploaded items on the web.

In general, the new Web 2.0 is different from Web 1.0 in a number of respects (Cormode and Krishnamurthy, 2008). Figure 3.1 displays a diagram for the comparison of the Web 1.0 and Web 2.0 technologies. It is clear from the figure that the numbers of sites and participants have increased noticeably for Web 2.0 models. This may be due to the fact that Web 2.0 has the ability to read-write information on the web, while the Web 1.0 realm is a static web or read only web. This is because Web 1.0 was designed to accommodate only a one-way information direction from producer to public. In contrast, Web 2.0 was designed to have two-way information flows, in order to share information between users and producers, and also between users themselves. This

concept has assisted in the growth of more interaction and participation between people, in contrast to a few years ago when they only read information on the websites. It became possible for any individual volunteer to modify the contributions of others and, in addition, led to the production and dissemination of free data on the Internet. This free sharing system was known as User Generated Content (UGC) (Krumm et al., 2008), which will be the focus of the subsequent section.

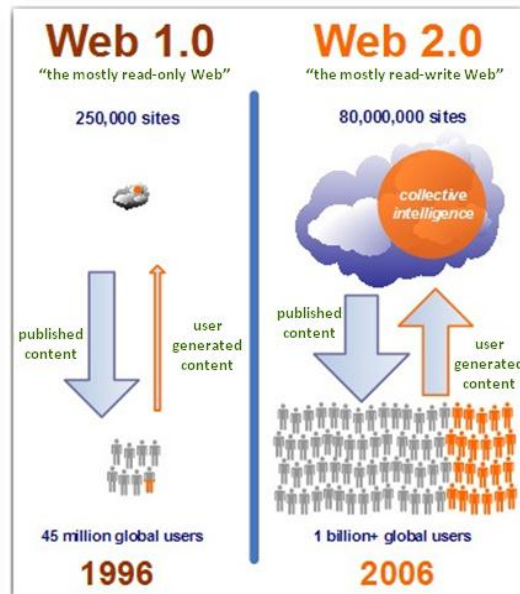


Figure 3.1 Comparison of the characteristics of Web 1.0 and Web 2.0 technologies (Barrett, 2012)

3.3.2 *User Generated Content: concepts and applications*

As described in the preceding section, one consequence of the development of Web 2.0 technologies was making the producing and sharing of information on the web through UGC more easy and effective. Since the term UGC refers to a variety of activities and applications, as will be discussed in this section, it is difficult to offer a standard definition for it. For instance, in their review of UGC, Vickery and Wunsch-Vincent (2007) identified the meaning of UGC as the ability of creating publicly, data on the Internet which can be achieved by amateurs or professionals with limited creative efforts. However, other authors did not accept this description as a common UGC definition; for example, Ochoa and Duval (2008) reported that the UGC concept may be considered local rather than universal, as uploaded data may be available only for a specific group and not for common usage, or it may be simply rearranging such information and not making new contributions. Nonetheless, generally speaking, UGC

can refer to information or media that may appear on the web which was contributed by volunteers without anticipation of any type of income (Krumm et al., 2008).

After the basic characteristics of UGC have been described, it would be useful to understand the various kinds of content that can be found on UGC websites. As UGC includes a wide variety of channels for gathering, viewing and transferring data to other people on the web, it may appear and be represented in various forms on the Internet. Difficulties arise, however, when an attempt is made to classify different kinds of UGC because users may group those websites into a variety of groups based on their area of interest. For instance, according to Steve Rosenbaum (2008) as cited in Balasubramaniam (2009), UGC has been classified into seven groups with respect to their utilisation. These include the personal information sharing websites such as Flickr where users can upload and share pictures, the media platforms such as You Tube for freely uploading videos, the Blog news for transferring personal news more quickly and specifically, and the social connection media such as Facebook and Twitter for chatting and connecting new people. Making money through UGC is also possible through e-commerce websites such as Ebay; there is a Meetup platform which offers a good opportunity for meeting people on the Internet and, lastly, the voices blogs are used to provide the economic, political and social views of people on the Internet. In addition to these classifications, a wiki is another form of UGC which refers to a website containing a collaborative work of multiple authors and can be used to share ideas and questions between groups of people (George and Skerri, 2007). This substantially highlights the fact that UGC has a multi-industry focus and supplies the basis of different innovative services on the websites.

A classic example of this kind of provision of information is Wikipedia (the free encyclopaedia). Wikipedia was originally established and founded by Larry Sanger and Jimmy Wales in 2001 (Miliard, 2008). As Wikipedia adopts an open model for uploading and editing the contributions of others, which are in most cases articles, the numbers of the registered users and the articles in Wikipedia have increased significantly according to its statistics. However, the question of accuracy will arise when comparing this freely available data with professional productions. Inaccurate structure, bad quality and wrongly edited articles might be expected. Nevertheless, this is not the case in some situations when the free data is created by a group of people

rather than single person, a factor that has been emphasised by Goodchild and Glennon (2010). They reported that the information obtained from a few contributors will be less accurate than that obtained from many people. Furthermore, the Wikipedia community has developed its regulations and rules through a specific section of wiki space. This offers an opportunity to members to contact each other and decide upon standards for documented data in Wikipedia.

Many analysts now argue that the strategy of UGC has been successful. (eMarketer, 2009), for example, reporting that the numbers of users of different UGC websites in the USA are growing dramatically, as shown in Table 3.1. From this table, it can be seen that there has been a marked rise in the percentage of total population who consume UGC in the USA since 2008. What is interesting in this data is that the blogs and the social networking users made up a higher rate than other websites. In 2008, the rates were 54.0% and 41.2% for consuming these two kinds of UGC websites respectively. The same year also saw less consumption for the user-generated video and wiki websites with approximately 36.0% and 33.9% respectively. This provides an idea about the increasing number of users with regard to UGC content. It is likely that these are the types of the UGC which are geotagged in some way. It is obvious that there is a movement towards favouring the more personal application websites, as those categories showed the highest rate of Internet users. In general, there is a clear trend of increasing of the rates of all UGC categories through the years between 2008 until 2013. The main reason behind this rapid development of UGC websites may be the short time between the producing and the accessing of them. In addition, different to traditional media, there is no need to obtain legal permission or pay fees for uploading or using UGC websites, factors which make these data more accessible, usable and transformable. This is why the concept of UGC expanded to several fields rapidly. For instance, the trend towards user-generated content which can be provided or shared online has had profound impacts on the geo-data scene, as will be described in the next section.

Table 3.1 User generated content consumer percentage in the USA for the period between 2008 and 2013 (eMarketer, 2009)

	2008	2009	2010	2011	2012	2013
User-generated video	36.0%	39.8%	42.5%	44.8%	47.2%	49.2%
Social networking	41.2%	44.2%	46.9%	49.1%	50.5%	51.8%
Blogs	54.0%	58.0%	61.0%	64.0%	67.0%	69.0%
Wikis	33.9%	36.6%	39.0%	41.0%	42.6%	43.9%
User-generated content consumers	60.0%	62.0%	64.0%	66.0%	68.0%	70.0%

3.4 Development of volunteered geographic information and its features

Recently, the advancement in geospatial data collection technologies, such as incorporating GPS technologies in mobile phones, has enabled users to gather their own geospatial data easily. By using Web 2.0 practices, an amateur can readily upload these data on the Internet. For example, nowadays any person can pick up the geographical information regarding their routes by using GPS in their driving or biking activities. Then, it is possible to contribute to updating and extending existing road databases on the web. It is also possible to add names or photographs to these datasets by means of the geotagging process, as mentioned earlier. As these data are typically produced by volunteers, they have been labelled by Goodchild (2007a) as volunteered geographic information (VGI).

There are many alternative names and definitions for the phenomenon of geospatial information on the web. For instance, the term 'geospatial information bottom-up' was used by Bishr and Kuhn (2007) to refer to geo-data on the Internet, whereas according to a definition provided by Turner (2006), this kind of information was coined as 'neogeography' and the same concept was also used by Haklay et al. (2008). On the other hand, for Sui (2008), VGI means 'geography without geographers'. Whatever concepts that have been used to describe the open spatial data enabled on Web 2.0, the term VGI is the most widely adopted by many authors; see for example, Mooney et al. (2010), Coleman et al. (2009) and Elwood (2008). The term VGI was generally used to refer to creating, disseminating and updating geospatial data voluntarily on websites. This basically means combining the efforts of individuals or collaborative communities in such a way as to supply this new kind of geospatial data. Similar to UGC applications,

VGI data can be effectively utilised by people other than the producers without any restrictions or rules.

One of the most significant current mapping facilities that may have an effect on VGI production is mapping 'mashup' technologies (Ho and Rajabifard, 2010). In general, a mashup means the ability to be able to combine several web services, such as Web Map Services (WMS), to produce one web application for displaying the combination of contents in a single interface. A simple example is the combination the address and photograph of a house can be combined on a Google map to generate a map mashup. The emergence of the application programming interface (API) provided by Google in 2005 has assisted the growth of mapping mashups. Although it is difficult to survey the exact number of map mashups that have been developed to date on the web, programmableweb.com has attempted to enumerate different kinds of mashups from a total of 6000 examined that have been created on the web (Figure 3.2). It can be seen that the top ten mashups types are calculated and compared to date (13-04-2012), and map mashups accounted for the highest proportion (28%). There are two main reasons for the widespread use of map mashup on the web. Initially, Google and Yahoo companies, for example, made some of their data resources free and available to everyone. Secondly, the increase in new mashup creator and editor tools reduced the need for a high level of programming skills in order to manage mashups on the web. This has enabled non-expert to create a map mashup without any complications.

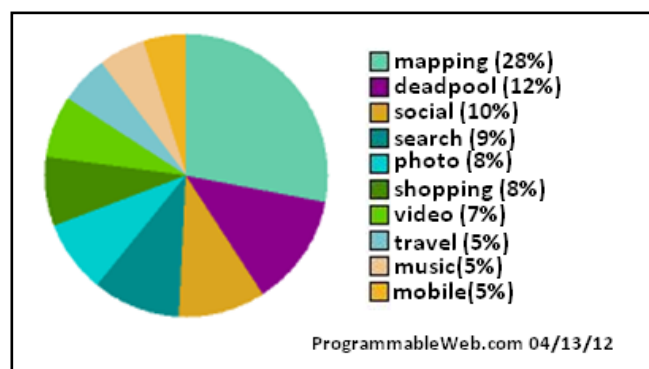


Figure 3.2 The most popular types of mashups (Programmableweb, 2012)

VGI is similar to other open source productions in that producers and users of VGI data come from a variety of backgrounds because any person can become involved in this activity. In addition, there are no standard methods for uploading this kind of data. Therefore, these databases are easily subject to heterogeneity and errors. However, VGI

does suggest a new and powerful approach that could be efficiently used to create up to date datasets. Real time geospatial data that can be obtained from VGI are necessary for several purposes such as emergency actions; for example, Zook et al. (2010) reported that the flexibility of free web-based mapping services played a major role in aiding and rescuing people when an earthquake hit Haiti in January 2010. After the earthquake free information technologies such as aerial photographs were used in order to supply the necessary information about the most devastated areas and to produce route maps to deploy resources. Therefore, VGI can offer the most interesting and the cheapest geographic information to users and sometimes it will be the only source of information, especially for remote areas (Goodchild, 2007a).

Since most VGI data are created by non-professionals, interest in integrating VGI with formal data, for instance, to develop and update formal datasets, may raise some concerns. From this point of view, Elwood (2008) highlighted the need to investigate the different types of VGI with specific emphasis on examining the impacts of VGI services, such as tools and procedures that were used to collect, create and share this data, on the accuracy and the validation of using VGI for multiple purposes. Because the research project described here presents an evaluation of the possibility of matching geospatial data from official and VGI data sources, the following pages focus on VGI data types and their nature and the issues related to the comparisons made with official datasets. The examining of VGI components and mechanisms can provide a useful impression regarding the challenges that may face the GI community when using VGI datasets for a variety of geospatial processing.

3.5 VGI examples and initiatives

VGI can generally be classified, on the basis of their characteristics and contribution purposes, into three categories (Deparday, 2010). The first group consists of geospatial data for public usage, and can involve downloading or obtaining free datasets and improving and updating the data of others. For this kind of database, GPS is usually used to capture different feature types such as road networks, pathways, buildings and green areas. An example of this category can be clearly seen in the case of the OpenStreetMap (OSM) project and a detailed description of it will be considered in the subsections below. In addition to the OSM scheme, commercial agencies such as Navteq and TeleAtlas have also found some benefits from this type of VGI

phenomenon. For instance, VGI might be exploited to generate or update some of the Tom Tom map share data, rather than following traditional expensive surveying and mapping procedures (Coleman, 2010).

Another group of VGI are typified by the ability of contributors to deliver free text data to identify their surrounding places. Different to the first VGI class, these data seek to gather people's discussions and opinions about a specific area of interest instead of the usual kind geographical information. It is openly subjective data which can usually be obtained in an open volunteered process. In this context, Seeger (2008) labelled the activity of using online interfaces by the public to add information on web maps as facilitated-VGI (f-VGI). Well known examples of this are the Wikimapia and Flickr initiatives.

There is a further class of VGI data which allows people to share their current geospatial locations with others (Elwood, 2008). Nowadays, the availability of smart phones has enabled users to achieve this task easily via the Internet connections. One advantage of this activity is giving an impression about who is nearby which can facilitate contact operations between users effectively. However, this kind of VGI data can be considered more private than the other two categories. Exchanging location information through mobile phone networks is usually kept private between the people involved. The general population are unable to access this information as it is only shared among friends or related persons. Examples of these are Loopt, Brightkite, and Plazes services.

In view of the fact that this thesis focuses on OSM data investigations in general, as OSM is the leading example of VGI projects that are concerned with geospatial data development around the world, the OSM project will be discussed in greater detail than others. The following pages are intended to give an overview and describe the main characteristics of VGI services, as presented above, with a particular consideration of the OSM project.

3.5.1 Exploration of the OpenStreetMap project

The OpenStreetMap (OSM) is an online geospatial database launched in England (London) by Steve Coast in 2004 (Chilton, 2009). In particular, it aims to produce and supply free editable geospatial datasets for a worldwide audience. The OSM

fundamentally relies upon the collaborative volunteers' contributions for collecting and uploading geographic data to the common data base on the Internet (Ciepluch et al., 2009). This mapping service can be categorised under the list of the first group of Deparday (2010) VGI classification groups. In general, making a map for OSM data includes five steps (Figure 3.3). Contributors can collect the OSM data by controlling handheld portable GPS devices (navigation mode) such as the Garmin series. Nowadays, it is also possible to use built-in GPS applications which are available in most mobile phones models such as iPhone. In order to map a certain area using GPS technique, for instance, the OSM community gathers volunteers through an activity called 'mapping parties'. An example of this can be found in the study carried out by Perkins and Dodge (2008) in which they illustrated a case study of a mapping party in Manchester, UK, in 2006. Although the GPS receivers may probably be considered as being the most important information source for the OSM project, there are also alternative data sources such as tracing data Yahoo imagery and /or Landsat images (Ramm et al., 2011). More analysis regarding the impact of the variability of data sources on the OSM geometrical data quality will be presented in Chapter Five.

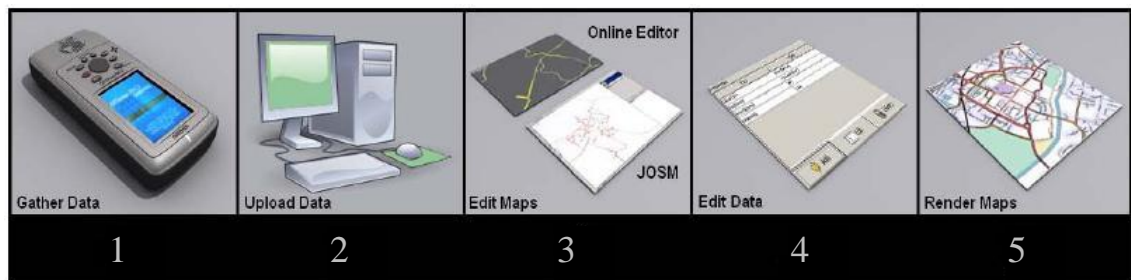


Figure 3.3 The stages of creating a map for the OSM project (OpenStreetMap, 2011)

The OSM database allows for every user to reproduce or edit its datasets without the necessity for any authorization although attributing the data to OSM is required. Thus, the OSM project is technically similar to the Wikipedia (free encyclopaedia) concept. The users of these systems are able to modify and add or even delete the contributions made by others. The underlying OSM map data can be uploaded by creating a user account and edited online through a wiki-like interface. There are many other services that provide mapping on the Internet freely. For example, Microsoft offers Bing Maps, and Yahoo Maps and Google Maps are readily available. However, the users of these alternative map sources have only been provided with a very limited right to use their datasets. It is not permitted for users of these services to edit or update their datasets.

Compared to the OSM data, there are several restrictions and conditions for using the Google Map service, as illustrated in GoogleMaps (2012). For example, the raw data of Google Maps is not available to the users at all; however, it can be used by commercial companies such as TeleAtlas and Navteq, as they pay for downloading these, while OSM data can be downloaded by any user. Consequently, the OSM project can be considered as being one of the most useful online mapping services in that it is suitable for education in schools and undergraduate studies. The survey conducted by Bartoschek and Keßler (2013) revealed that OSM is the most renowned online mapping service among students. The above mentioned positive aspects of OSM data, in addition to the possibility of using OSM as a base map for studies in cartography, make OSM a flexible tool in education.

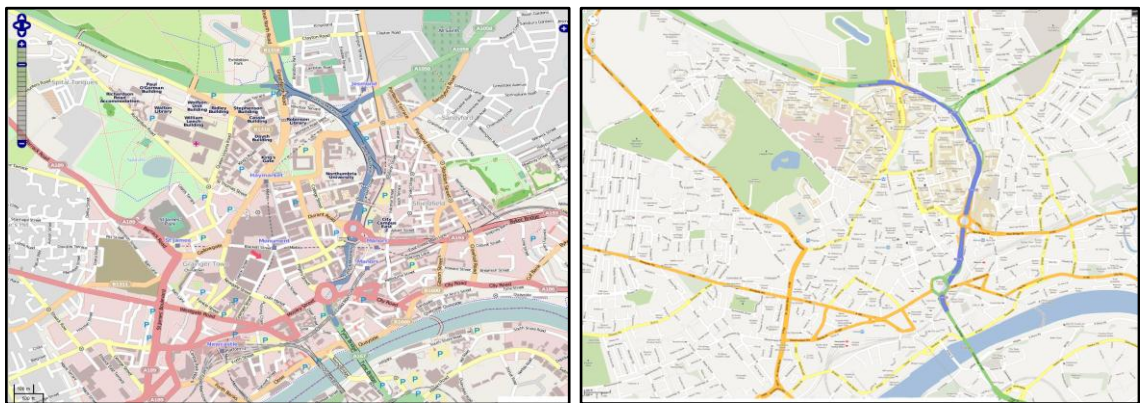
The main other difference that may be noticed when comparing OSM data to other public mapping services is the level of detail as far as features are concerned. It is clear from Figure 3.4 that the site of Newcastle upon Tyne appears more complete in the OSM map than in the Google Maps version. However, the levels of detail of OSM maps vary around the world (Ramm et al., 2011). There are some places, such as the UK, that are mapped very well, whereas there are other parts of the world, such as Iraq for example, that have a little coverage for the centres of the big cities only. In fact, there is no detailed data for the countryside or the suburban areas. The detail of OSM maps is fundamentally based on the number of the volunteers that are available in each place around the globe.

The amounts of OSM data are increasing every day on the Internet. The number of registered users of this project is also growing at a remarkable rate. For example, Haklay and Weber (2008) reported that in 2008 the number of registered users of the OSM project was approximately 33,000, while at the time of writing this thesis there are about 570,000 registered users (OSM-stats, 2012). It is clear that the number of registered people has increased by more than fifteen times during the last four years. However, the number who edit is a minority of these. This view is supported by Neis and Zipf (2012) who concluded that the rate of the registered users who achieved at least one edit of OSM data was only 38% of the total number of members. They also found that only 5% of the registered members have contributed more than 1000 nodes. Information can also be obtained from the OSM-stats (2012) with regard to the total number of uploaded GPS points, nodes, ways and relations for the real time OSM

database, as will be described in section 3.5.1.2. These statistics reflect the rapid growth of OSM data on the web.

There are various aspects to the most important motivations for these developments, for instance, gaining advantages from the free accessibility of OSM data (licence, cost, sharing) and opening up a new paradigm of 'geo-data-people's' SDI. Access to current OSM data without any charge is available to anybody with web connections. In addition, the wide-ranging coverage of OSM data sources (around the world) allows visitors to search a world map and download different portions from a distance for any part of the world. Although these are positive aspects, the problem of heterogeneous data quality has emerged (Al-Bakri and Fairbairn, 2011).

There are massive differences in the quality of geospatial data sources on the Internet. The evaluation of spatial data quality is an important issue in data integration research. Hence one part of the current project is to assess the capabilities of geospatial data integration from authoritative and VGI geospatial data sources, as will be established in Chapters Four and Seven.



a- OpenStreetMap data (<http://www.openstreetmap.org/>)

b- Google maps data (<http://maps.google.co.uk/>)

Figure 3.4 A comparison of the details of maps for the centre of Newcastle upon Tyne – UK (images sampled on 09/05/2012, both rendered at equivalent zoom levels – 16/19 for OSM, 15/18 for Google maps). This comparison facility is now available at <http://tools.geofabrik.de/mc/>.

3.5.1.1 *The analysis of the architecture components of OSM database*

On the homepage of the OSM project, a map appears through a javascript interface which allows users to navigate around the world efficiently. At the same time, it allows users to download data in different formats such as standard XML data or map images. There are ongoing operations for geospatial data processing, such as map editing,

structuring and storing spatial data into the OSM database, and map rendering, which are typically performed by specialist software in order to control these tasks. These can be represented by a noteworthy architecture of open source software that has been developed by the OSM community to generate, display and distribute OSM data on the web (OpenStreetMap, 2012a). The OSM software has the ability to create and render maps as images, which are usually called tiles, via the APIs. As the final objective of OSM project is to perform a universally complete coverage of the maps, this software is usually used to render and update existing tiles for the global production of OSM datasets.

The entire structure of OSM software is shown in Figure 3.5. From the diagram below, it can be seen that there are many applications which have been incorporated into the structure of the OSM software. The main database can be considered as the most important part of the OSM project's components since it represents the place where all OSM datasets are regularly kept (OpenStreetMap, 2012b). The OSM database is essentially managed by the PostgreSQL object-relational database management system as a distribution model of the OSM project. In addition, the OSM distribution model also includes a middle level called API 0.6 which is fundamentally developed in a Ruby on Rails free web application framework, as illustrated in Figure 3.5. Geospatial data can be uploaded into the OSM database as a GPX file, if it is originally produced from GPS tracking, or it can be traced directly from aerial imagery for the area with whole image coverage. Surveys such as that conducted by Wolf et al. (2011) revealed that different applications were developed in order to assist in the contribution of spatial data to the OSM database. This includes some of the iPhone applications, desktop applications such as Java OpenStreetMap (JOSM), Merkaartor and the web-browser based applications such as the Potlatch and Potlatch 2 interfaces.

In order to input a small amount of geospatial data into the OSM database or edit the data of an OSM project, tools such as browser-based Potlatch editor, JOSM and Merkaartor are typically used, whereas for exporting and importing a large amount of OSM datasets, tools such as osmosis are usually utilised. In general, the OSM software can be considered easy to use as it is open source software; however, the setup of the software requires some technical experience to understand the nature of the PostgreSQL and Ruby on Rails systems (Wolf et al., 2011).

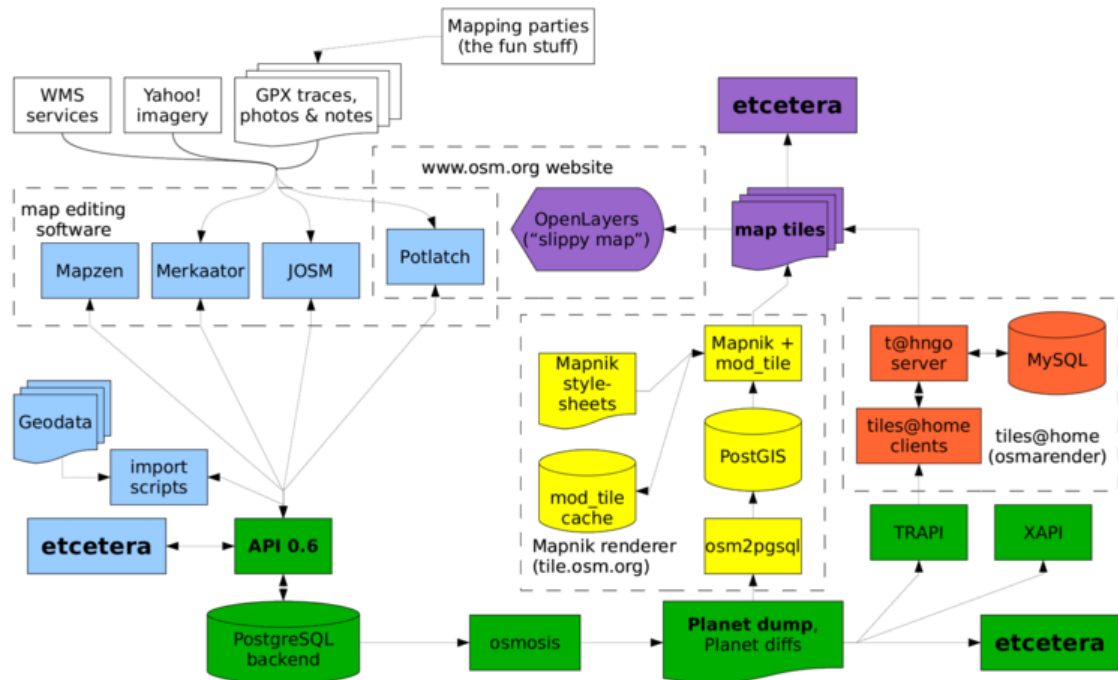


Figure 3.5 The structure of the OSM project components (OpenStreetMap, 2012a)

3.5.1.2 The structure of OSM datasets

In previous subsections, quite a few points of detail that relate to the development of the OSM project have been highlighted and identified. Questions may be raised concerning the strategies of storing and organising all the OSM information. The following paragraphs seek to answer these questions by addressing and describing the data models of the OSM project. From this perspective, Ramm et al. (2011) pointed out that the map features or the objects of an OSM map always have geographical coordinates and a type. These features are represented inside the OSM data model as nodes, ways and relations with attributes tags.

A node may be broadly defined as a point feature on the OSM map which can be considered the simplest type of OSM data. It may be used either as a single point to define a certain location such as a bus-stop and post box, or as vertices to represent ways and serving as linear features. Within the OSM database, a node consists of a number of key elements. These are a pair of lat/lon geographical coordinates, the last edit timestamp, the username and the ID of the latest editor, and tag names (Ramm et al., 2011). Figure 3.6a shows the XML code for a node that has been selected from the OSM database which represents the town of Cramlington-UK. Referring to the definition of the nodes, a way is a connecting of a list of nodes which can form a line

segment or can be closed to form a polygon. The configuration of ways data in the OSM database is similar to the nodes except there are many nodes rather than a single point in a node situation. The way data type is commonly used to depict road segments; however, it can also be used to refer to any other linear features such as rivers and railways. This kind of data is usually called non closed way, while the data that represent the features composed of areas such as buildings, land use and car parks are known as closed ways (Ramm et al., 2011). An example of ways data type is illustrated in Figure 3.6b. It is a sample of XML data from the OSM database which displays the way data of one 'road' in Cramlington-UK. The last kind of OSM data is 'relations' which are used to connect various kinds of objects. It is particularly used to represent complex areas, such as polygons with holes, or join segments of roads, to form a specific route on an OSM map. Again, the relations have similar elements of structure to the nodes and ways, except the relation data type is referenced by the ordered member tags list (Ramm et al., 2011). This is illustrated briefly by Figure 3.6c which includes the XML code for the relation data type of 'railway' in Cramlington-UK.

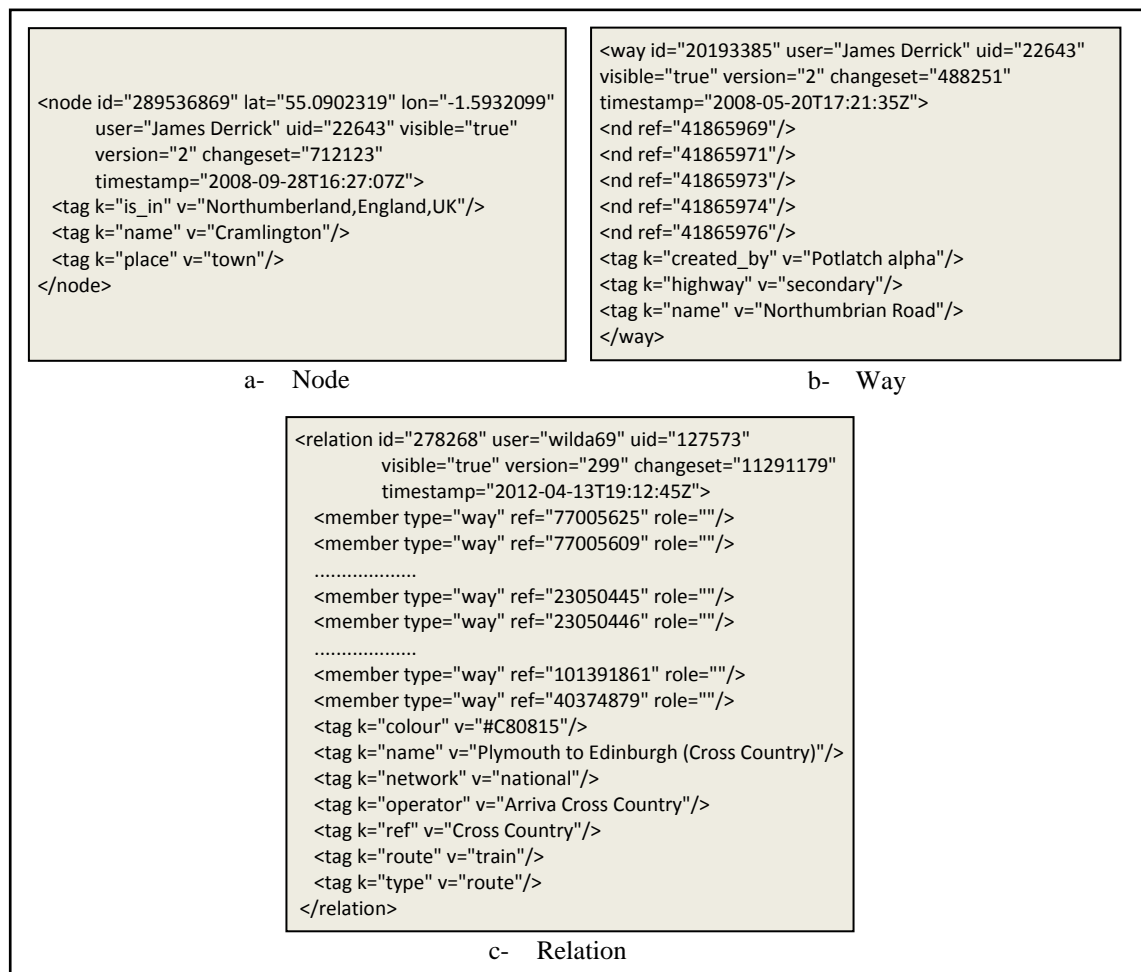


Figure 3.6 Examples of XML codes of OSM data types' structure

The number of nodes, ways and relations are dramatically increased every day, as shown in Figure 3.7. The growing numbers of nodes, ways and relations are represented on the vertical axis of this figure, while the horizontal axis includes various periods of date. It can be observed from the graphs in Figure 3.7 that there has been a gradual rise in the statistics account of the amount of OSM data types since the second half of 2007 (OSM-wiki, 2012). It is apparent from the figure below that the nodes resulted in the highest numbers of OSM data type. This is due to the fact that the nodes constitute the base unit of creating and producing the ways and relations data types. The up to date exact numbers of nodes, ways and relations can be obtained from OSM-stats (2012). It reflects the statistics of OSM data types at a real time. For instance, at the date of writing this thesis the statistics showed that there are more than 1.4 billion nodes, 133 million ways and 1.3 million relations. It seems possible that these huge amounts of OSM data are due to the popularity of the OSM project around the world in order to create a free features detailed worldwide map.

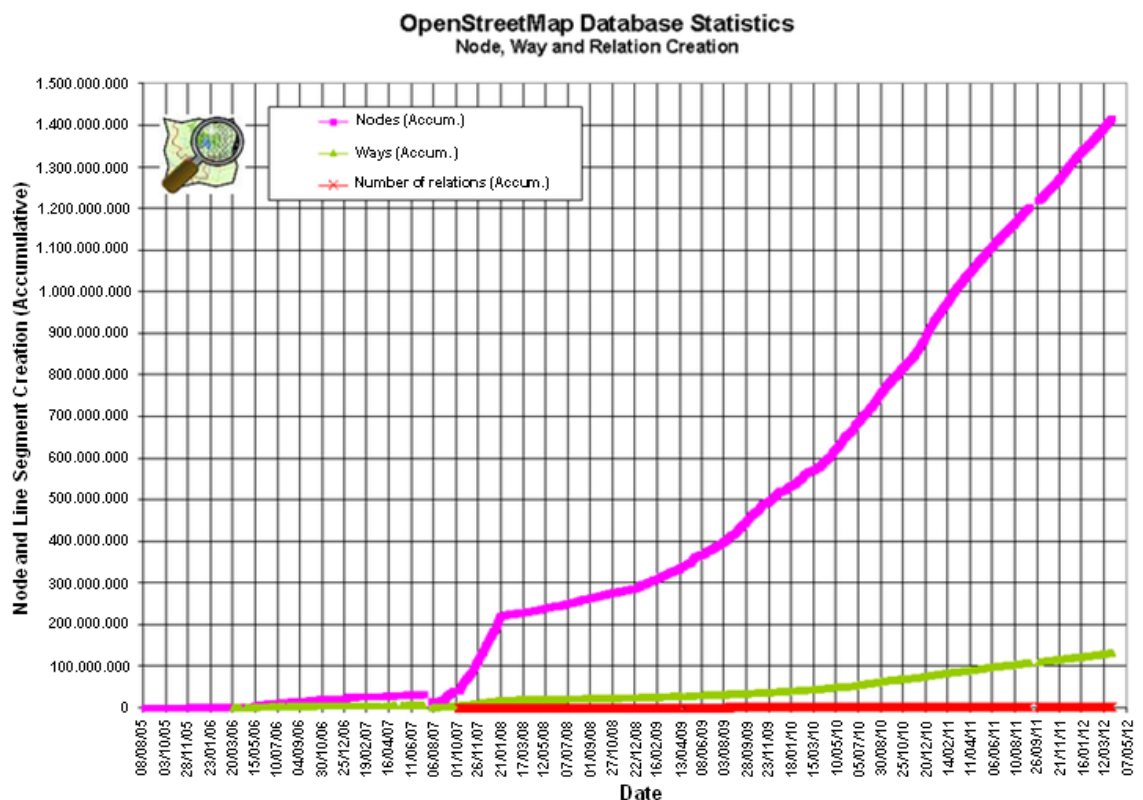


Figure 3.7 Statistics account graph reflecting the growth of OSM nodes, ways and relations data types (OSM-wiki, 2012)

As mentioned earlier, in addition to the three OSM data types, the attributes or the tags are also recorded for each feature in the OSM database. The tagging schema can be

considered as one of the most important parts of the OSM project as it provides the description of features. The OSM data tags usually consist of a key and a value pair to describe the nature of the OSM objects (Haklay and Weber, 2008). Both key and value can be string data, but it is not expected to be other data formats such as numeric or boolean. In the OSM initiative, tags are assigned for the whole object not a part of it. An example of OSM tags is the data displayed in Figure 3.7. From this figure, it can be seen that there are a number of tags for each of the nodes, ways or relations. All tags are in string formats to define related features and each tag uses a different key to the other tags. A detailed description and explanations regarding all types of OSM map features or tags can be found in OpenStreetMap (2012d).

The OSM tags are usually created by any member of the OSM community as there is no standard list of accepted tags of OSM projects. The proposing of a new OSM tag requires voting and discussion among the people who have an interest, but in general cases mappers create their tags without going through this process. They only seek advice regarding their suggestions from other OSM contributors by e-mail and their advice will be followed if it is sensible. These specifications may enable OSM users to provide and produce several different tag types without the difficulty of a complex authorisation process (Ramm et al., 2011). However, the lack of restrictions and standardisation may lead to some inaccuracies or errors when tagging OSM features. This will be investigated and examined in further detail in Chapter Seven where the evaluating of the ability of feature classification integration from formal and OSM data sources will be demonstrated.

Within the OSM environment, there are several methods and tools that can be employed to download OSM datasets. The OSM website or map viewer through the API can be considered as the main source to download the full OSM datasets. Tools such as Osmosis, pbftosm, osmconvert, or osmchange can also assist in extracting specific OSM data (OpenStreetMap, 2012c). Including data defined by area or data captured by individuals, in this research OSM data was selected and exported within OSM using a standard OSM interface which was then further manipulated in Microsoft Excel in order to prepare the required data for geometrical and semantic similarity analysis.

3.5.1.3 Characteristics of OSM activities

Nowadays, the OSM project has extended to cover most parts of the world as one of the main open source spatial data suppliers; however, there are several differences among the technologies for capturing spatial data, the experience of volunteers, and feature types and classes that may occur for various forms of universal coverage. Although in this research project the UK and Iraq will be demonstrated in Chapters 4 and 5 as local knowledge OSM study areas, there are numerous other OSM map activities in other countries around the world. For example, in developing countries such as Kenya, a country in East Africa, the OSM project has been essentially developed by creating spatial data using GPS tracking. On the other hand, the availability of free aerial images from the Bing service since the end of 2010 has made the creation of OSM easier for some of Kenya's regions such as Nairobi, Mombasa and Nakuru (Wiki-Project-Kenya, 2012).

An obvious example of the quick growth and well documented OSM data in Kenya can be seen in Kibera city. Kibera is the biggest informal settlement in Kenya; therefore spatial information is necessary to enable the government to improve the residential living conditions. However, up until October 2009, there was no online map data of Kibera at all. For this reason, the initiative of producing consistent data and making it available to the public, in order to show the distribution of residential features in this informal settlement, has been undertaken. A group of local young people were trained to use GPS and create maps from other sources such as free satellite or aerial images and then upload them into the OSM database (Veljanovski et al., 2012). Although there were many challenges and difficulties faced in collecting spatial datasets in Kibera, such as the lack of the familiarity with technology (i.e. computers and Internet browsing), economics (i.e. paying to be a volunteer) and community (i.e. difficulties in understanding the benefits of the project), the mission of developing the online map project in Kibera progressed well and it has now successfully contributed data to the OSM project (Map-Kibera, 2012).

Faraway from Africa, in Europe the OSM project received more attention and the amount of OSM data has increased massively. In addition to the UK, where the OSM project started, France can be considered to be another good example of volunteers successfully creating and contributing to the OSM global database. Even though the

OSM data in France has been basically developed by employing the traditional methods of creating OSM data, such as using GPS devices or tracing from free images, Mooney and Corcoran (2012) reported that recently in France spatial data can be imported directly into the OSM database from the CORINE Landcover databases for France. This service has provided free spatial data which can add more valuable and effective information into the OSM project. Mooney and Corcoran (2012) also mentioned that there are other countries following the same procedure to contribute spatial data into OSM global database. For example, in the Netherlands it is now possible to import the Automotive Navigation Data (AND) road network datasets for the whole country into the OSM database. In the same way, in the USA the TIGER (Topologically Integrated Geographic Encoding and Referencing system) road networks data has been also used in order to contribute to the OSM database. By initiating the activity of importing free available spatial datasets, volunteers have the ability to correct, update and complete the spatial data within the OSM platform in every area of these countries.

In other parts of the world such as Asia, and specifically in Japan, the development and growth of OSM activities has also been rapid. In Japan, the first OSM community was founded in March, 2008. Although it was a small group (about ten volunteers) at the beginning, the OSM community continued to be successful in their activities; for example, they established the OSM Foundation Japan, as well as obtaining support from the Geospatial Information Authority of Japan and Yahoo, Japan. Furthermore, the OSM has been introduced in academia, as some of the Japanese OSM community members had already worked as lecturers or were students in Japanese educational institutions. They are trying to use OSM data for practical work and some other activities in universities. Therefore, the OSM community in Japan is working on inviting over 20 academic institutions to promote OSM for educational and academic purposes. In March, 2011 when a tsunami swamped some Japanese towns, the number of OSM volunteers topped 1200 members. They successfully provided crisis mapping through digitizing spatial data from free satellite images and uploading into the OSM database (Wiki-OpenStreetMap, 2011). This approach is also supported by McDougall (2012) who examined the use of VGI data to develop a mapping system for crisis management. He noted that within several hours after the earthquake, the Japanese OSM community had started working hard to produce crisis maps for the area of the disaster. The same study also reported that within a few hours several communications

occurred between Japanese students at the Fletcher school who wished to volunteer and the OSM Tokyo team to provide up to date maps. The VGI produced maps were also used by several embassies in Japan in order to trace their citizens' locations. The implementation of the OSM project in Japan provides an insight into the effective processing of the cooperative work among the Japanese OSM community, as well as student volunteers from schools and universities, during around-the-clock efforts in order to rescue and save victims of the disaster.

In Thailand the contribution to and use of the OSM service is growing every day. The main OSM projects in Thailand are represented in three major areas: Bangkok (the capital of Thailand), Phang Nga (a province in the South West) and Chiang Mai (a province in the North). The OSM contributions are fundamentally based on GPS to collect spatial data. In addition, the free availability of aerial and satellite images, such as Bing and Landsat, means that they have also been used as additional spatial data sources of the OSM project. In Thailand, these services provide coverage for the whole country which makes creating OSM data for any part of this country more efficient (Wiki-Project-Thailand, 2012). Although the development of the OSM project has seen rapid growth in the developed and developing countries, as illustrated in previous paragraphs, in undeveloped countries such as Bangladesh, there are many challenges facing the development of OSM activities. Rifat et al. (2011) listed the main issues as the limitations of using GPS devices, the lack of technical skills, the high cost of using high speed Internet, the use of mapping services being unpopular with people, the smart phone not being introduced until 2011 and governmental support being very limited. However, in order to push the OSM project in Bangladesh forward, the same authors suggested some ideas for the enrichment of OSM. For instance, decentralization may be one of the factors that could help to promote the growth of the OSM project in Bangladesh. If each division or district in Bangladesh had its own OSM community, this will assist saving time and enrich OSM data. In addition, creating groups of volunteers and developing their technical mapping ability through map parties and frequent meetings may also help. Furthermore, the raising of awareness among citizens to encourage use of online mapping services will inspire companies to develop software for these purposes and assist in the growth OSM data.

The descriptions in previous paragraphs were for the purpose of reviewing different situations of OSM data in several countries around the world. The appraisal showed that the status of the OSM project is different and varies in each case. Differences may occur in spatial data sources, technologies of handling spatial datasets, the level of awareness among the volunteers and the governments' assistance. For example, the importing of spatial data directly into the OSM project in some countries such as France, the Netherlands and the USA has effectively corrected and improved the existing OSM data. Furthermore, intensive efforts to enrich the OSM data in some countries have been very helpful in a practical sense, as in the case of Kenya and Japan; however, for other countries, such as Bangladesh, OSM requires further efforts in order to be more useful. It is clear, therefore, that informal data collection projects such as OSM show significant variation in the nature, the completeness and the quality of their data. Data quality is dependent on a wide range of factors and it should not be a surprise if the quality of informal data does not match the formal data in the same way everywhere. The case studies presented in the next chapter, however, do attempt to examine some of the more important variables - type of feature collected, method of data capture, and nature of the volunteer effort.

3.5.2 Additional VGI examples – based on Web 2.0 technologies

In addition to the OSM dataset, there are many other VGI data sources on the Internet, as was explained in section 3.6. This subsection will cover some of them, especially those which have a similar concept to the OSM project, but less popularity than it has, such as Wikimapia and Flickr websites.

The phenomenon of Wikimapia has come to be used to facilitate individuals sharing descriptions about geospatial locations around the Planet Earth (Goodchild, 2010). It was originally launched in 2006 by Alexandre Koriakine and Evgeniy Saveliev. The Wikimapia service offers free and editable maps and satellite imaging resources on the web. Furthermore, it allows contributors to add notes or information on these maps and images in the same way as wiki technology. Therefore the Wikimapia activity combines the concepts of the systems of wiki and Google Maps (Moussa and Fritsch, 2010).

The numbers of users registered on the Wikimapia website are growing rapidly every year (Figure 3.8a). For instance, in the first quarter of the year 2012 the Wikimapia user account numbers reached 1.49 million, while in 2009 it only had about 400,000 users.

From these statistics, it can be observed that the user account numbers have multiplied by more than three and half times in only three years. Similarly, Figure 3.8b displays that there has been a noticeable rise in the number of places marked on Wikimapia. For example, at the time of writing (2012) Wikimapia had approximately 17.5 million places that had been described by the users. By comparing this value with the number of places in 2009 from the graph below, for instance, it can be noted that the number of places has increased by nearly 57%. These statistics may commonly reflect the increase in volunteers' interest in providing much richer descriptions of places with geographic locations on the web.

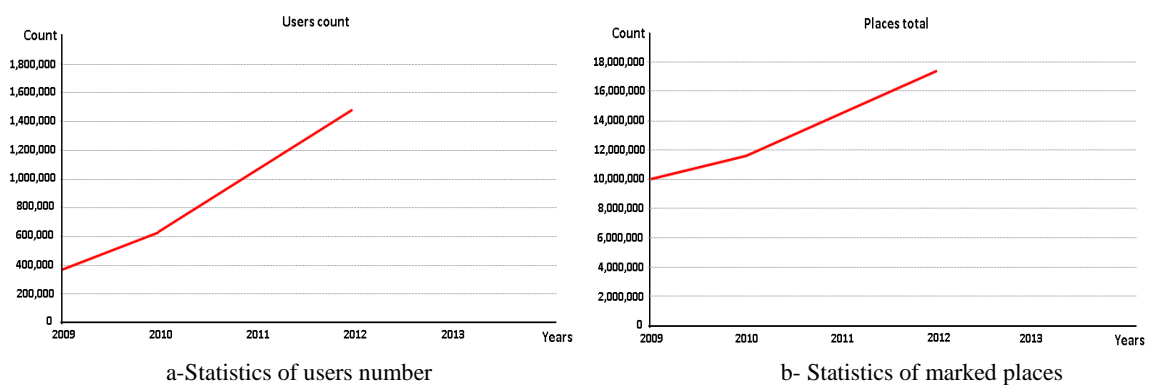


Figure 3.8 Wikimapia's growing performance (Wikimapia-statistics, 2012)

A further typical VGI example is the Flickr site which is basically a tool that enables citizens to share, publish and edit photographs on the web. Users can also add remarks or text tags for each picture to describe its content. The information attached to pictures includes different kinds of metadata such as the date and the time of the photograph, the place where the photograph was taken, in addition to the geographic location of the shot (georeferenced images). The geo-tagging process for Flickr photographs can be performed either manually by the authors themselves or different users or automatically by the GPS that is built-in to some types of camera (Newsam, 2010).

Flickr was created in 2004 by Lucicorp, and since it was launched the number of people who have become interested in this activity has been growing at a high rate. This is probably because Flickr can affect the daily life of individuals across the world by enabling them to manage and organize personal and common pictures on the web. For example, Yahoo reported that there are more than 50 million registered persons on the Flickr website to date (Yahoo, 2012). It is also interesting to note that the numbers of overall uploaded pictures into the Flickr site have exceeded 6 billion photographs. From

this amount, there are more than 187 million items classified as a geo-tagged photographs (Flickr, 2012). Despite the fact that the Flickr initiative was initially designed as a photo-based service rather than a geographic data resource, it can offer a massive amount of geographic positional information regarding the locations where photographs were taken. Therefore, as a consequence of the increase in the number of pictures that have been hosted by Flickr, the VGI sources on the web have undergone a relative expansion.

Similar to the OSM project, both the Wikimapia and Flickr services depend upon volunteer activity. The data of all three projects are gathered, distributed and updated on the Internet by individual people who are there as volunteers. However, in contrast to the Wikimapia and Flickr sites, OSM contributes purely geographic information. Therefore, it is adopted in this thesis to determine whether it is possible to integrate its geospatial data with official data sources, as will be described in the following chapters.

3.6 The differences between official and VGI data sources

As mentioned in Chapter Two, data quality issues have been a primary point concerning VGI data in general and the OSM project in particular. The VGI data may raise its own concerns regarding the value of information, its reliability and its quality, while official or governmental geo-data sources can provide more trust and reliable geospatial datasets (Rak et al., 2012). Thus, a quality check becomes critical, as assessing quality incorrectly can have serious social, scientific, personal and also political consequences. Furthermore, for most VGI projects no extensive metadata exists. In contrast, information obtained from traditional mapping agencies convey an immediate sense of care and attention to detail, based on experience, and it is immediately assumed to be of high quality. In the form of documented specifications, extensive metadata is available.

Another significant difference between authoritative and VGI data is the access network. Regarding this aspect, Castelein et al. (2010) pointed out that the geographic contents of VGI data can be presented by specific access network platforms which are prepared for this purpose. Furthermore, the VGI map viewer can be used to search for a location and the results appear immediately on the map. It is also possible to employ these platforms to download VGI data in order to use it for other applications. By contrast, downloading of formal data may be much more difficult after initial viewing using specific software.

In addition, Castelein et al. (2010) reported that in formal data sources, the standards are usually applied for metadata, while standards are applied only for data content in the VGI data sources. For example, OSM has pre-set map feature definitions as standards to define feature categories, such as recommending the map feature 'waterway' to be used to represent a stream or river. Further highlighting differences between formal and VGI data, Goodchild (2008b) uses examples as evidence that there are often detailed standards regarding the quality of authoritative gazetteers, whereas there are no standards concerning the relationships between a feature's actual footprint and a feature's footprint as entered into VGI data sources such as Wikimapia. In terms of policy, registration is the only requirement to be given rights to contribute to VGI projects. For instance, Wikimapia and OSM give the right only to registered users to contact other users or to change the contributions of others, whereas the authoritative data is secure and additions, deletions or modifications to data must be performed by an authorized person of the agency.

However, there are many positive aspects of VGI datasets compared to authoritative datasets. For example, Rak et al. (2012) pointed out that VGI can reduce the cost and the time for acquiring and maintaining geospatial datasets, as it is available and free to anyone, while governmental sources may be required to pay a specific fee and wait considerable time to obtain such spatial datasets. Although there are many limitations and differences between formal and VGI data sources, the facilities of alternative VGI data sources may encourage the GI community and organisations to take advantage of it. This may be represented by updating official datasets through their integration with VGI datasets. As mentioned in Chapter Two, the geospatial data integration process involves not only imposing two datasets or more together, but also includes the matching of geometrical and semantic elements between compared datasets. Therefore, the current project will investigate this issue in further detail.

3.7 Chapter summary

This chapter has highlighted the current status of official topographic mapping and informal spatial datasets. Traditionally, formal topographic mapping is usually produced by authoritative governmental mapping agencies, while an informal dataset is created and shared by non-professional people without any restrictions such as copyright control.

Additionally, the characteristics of each of the data sources have been described and the advantages and limits of these spatial data sources were also discussed. For this project, different formal spatial data sources such as OS (Great Britain) and GDS (Iraq), in addition to VGI datasets such as OSM information, have been used. Several types of VGI datasets, such as OSM, Wikimapia and Flickr have been illustrated and described; however, OSM information has been chosen for this research, as the OSM project provides geographic topographic data with feature descriptions or classifications which are appropriate for this study. The differences between formal and VGI datasets were discussed in the latter part of the chapter. In general, VGI datasets introduced challenges and new technologies for sharing and integrating multi-source geospatial datasets. Chapter 4, therefore, develops tools and demonstrates several tests in order to analyse the ability of geometrical integration of formal and VGI datasets. At the same time, feature classifications will be considered to carry out more integration that is useful and effective. The development of models to evaluate the integration of feature classifications from official and informal data sources is presented in Chapter 7.

Chapter 4 Tool Development for Assessing the Similarity of Geometrical Entities

4.1 Introduction

In the context of Geographic Information Science (GIS), two possible components of geometric accuracy can be considered: positional and shape (linear and area) accuracy. In order for such convergence, and ultimately data integration, to become useful, to assist in the development of Spatial Data Infrastructures (SDI) for example, the geometric correspondences have to be known. Geometrical quality is a dynamic problem in the domain of geographic scientific research because of the growth of data exchange through the web. The uncertainty with regard to physical quality is of crucial importance for geospatial data integration. Data integration is not only an overlaying of data in a geographic information system, but also involves assessing how well the geometric and semantic properties of one dataset can be transferred to the other (Butenuth et al., 2007). Hence, it is becoming increasingly difficult to ignore the assessment of such elements for the purpose of developing worthwhile multi-sources geospatial data integration.

The purpose of this chapter is to discuss the viability of geospatial data integration from field survey (FS) spatial datasets, official sources such as Ordnance Survey (OS) and General Directorate for Survey (GDS) datasets, as well as from volunteered geographic information (VGI) sources such as OpenStreetMap (OSM) data. It focuses particularly on the assessment of geometrical similarity measurements and their subsequent statistical analysis for the purpose of geospatial data integration. The research has tested different kinds of features of three study sites. The corresponding datasets for the three study areas were also obtained and processed through the use of specific tools. The tools were developed by Matlab for the purpose of measuring positional, linear and area (shape) similarity. At the end of this chapter, the final outcome will be the production of a range of interfaces that can be easily used for measuring and presenting geometrical similarity properties between tested datasets. The chapter will conclude with a description of an analysis of geometrical similarity results of formal and VGI datasets.

4.2 The study areas and data acquisition

Chapter 3 showed the variability of informal mapping worldwide, with differing data sources, volunteers, equipment, completeness, need and environment being evident in the output for the OSM project. This research, therefore, examined as its case study areas, varying environments (e.g. urban and rural), varying types of features (e.g. 'hard' and 'soft' or fuzzy), varying methods of data capture (e.g. GPS, aerial imagery), and varying personnel involved in measurement (e.g. amateurs, qualified engineers, overseas personnel unfamiliar with the location). This thesis used three different sites as study areas, each displaying a variety of different features, such as roads, buildings, car parks and woods, in order to test the methodology suggested for this part of the research. Two sites were sampled in the UK, one being an urban area where hard features such as kerb lines, roads and buildings were selected and tested, the other being a peri-rural area where the most common feature types were less distinct water edges and vegetation areas. Cramlington, UK, was chosen as the area with hard details, while the rural area for study was around the village of Clara Vale, UK. The third study area was selected as an urban area in the centre of Baghdad, Iraq. Figures 4.1 to 4.3 show these sites respectively as they appear in formal (OS and GDS) datasets. Similarly, these sites are illustrated in Appendix A (Figures A-1 to A-3) as fashioned in an informal (OSM) dataset.

Three different datasets were used for all three sites: self-generated reference field survey (FS) datasets, authoritative datasets, such as OS and GDS data, and informal VGI datasets such as OSM information. The idea behind this combination was to compare and test the capabilities of geospatial data integration between datasets from different authoritative topographic mapping agencies and VGI sources. This study also aims to examine whether there are any differences between 'hard' and 'soft' features, through data integration processing.

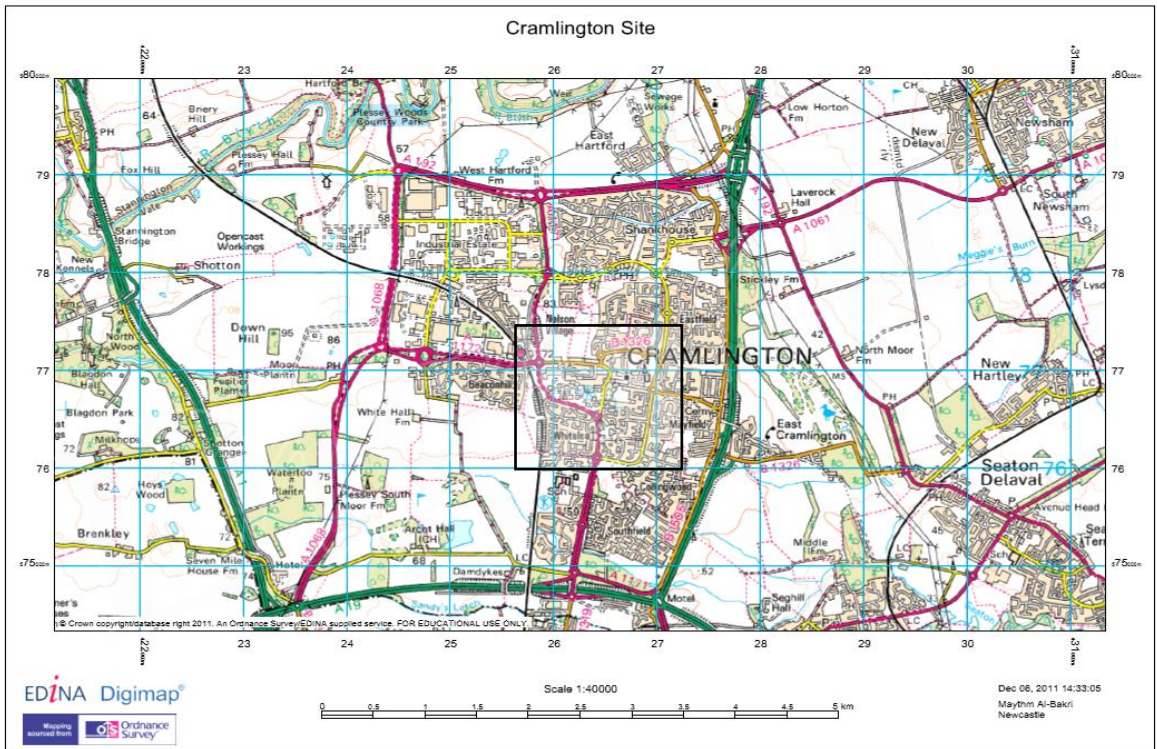


Figure 4.1 Cramlington 1 and 2-UK site

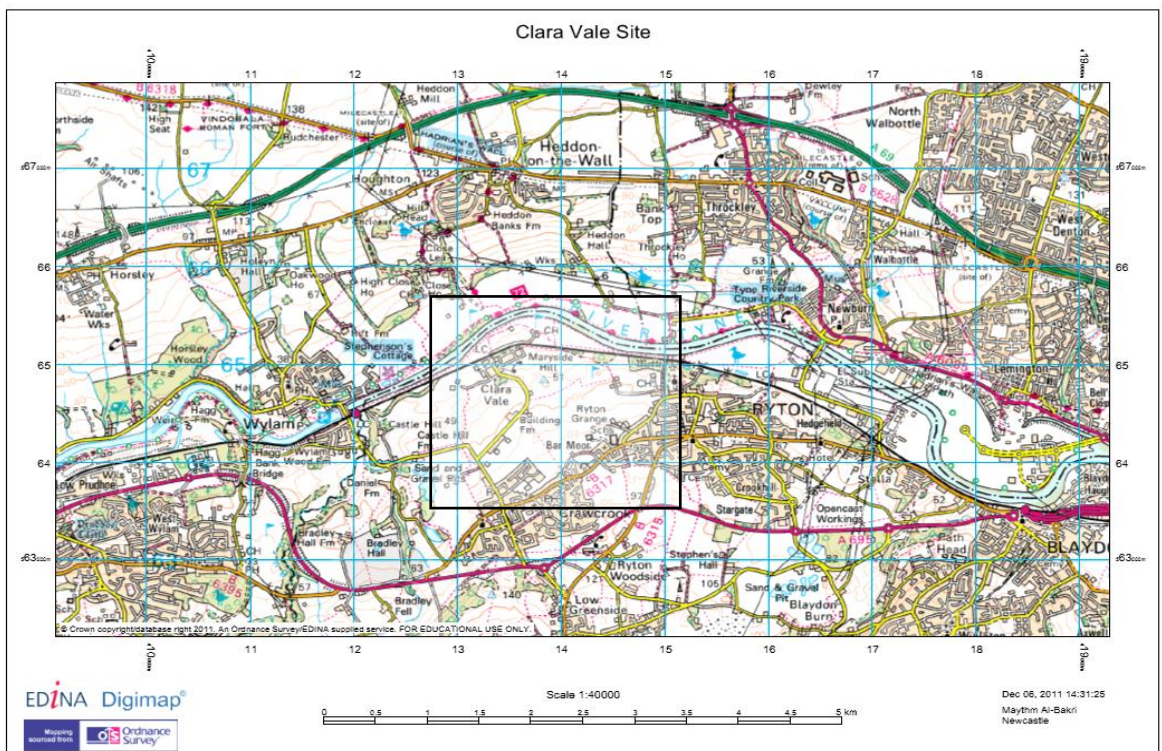


Figure 4.2 Clara Vale-UK site

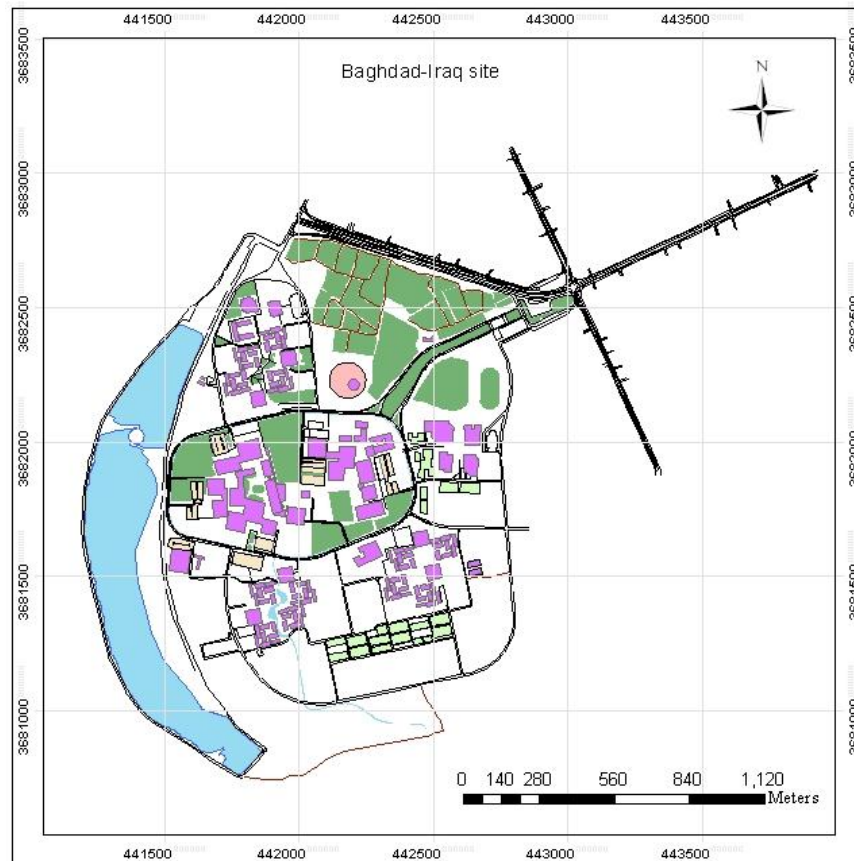


Figure 4.3 Baghdad-Iraq site

As mentioned in the previous section, this chapter aims to demonstrate the testing of the geometrical similarity measurement. In the case of positional evaluation, the planimetric coordinates of prominent features, such as corners of buildings, car parks, fences, vegetation verges and intersections of roads and pathways, were extracted, stored and processed. The numbers of the tested points for all study areas are illustrated in Table 4.1. The selection and distribution of tested points followed the recommended NSSDA approach which will be discussed in more detail in the next section. The linear features, such as roads and pathways, were used to provide the necessary datasets for linear similarity tests. The comparisons for linear features were carried out in the urban areas of Cramlington, UK and Baghdad, Iraq. This was basically concerned with pathways, main routes and small cul-de-sacs. The numbers and the average lengths of the linear tested samples are displayed in Table 4.2. The data used for the polygon shape similarity tests consists of vector datasets for the tested areas: urban and rural areas in the UK and an urban area in Iraq. The polygons typify the 'hard' details (buildings and car parks) in the Cramlington dataset, 'soft' features (ponds, woodland and golf courses) in the Clara Vale dataset, and urban features (car parks and buildings) in the Baghdad

dataset. The number and the average areas for each sample of datasets are shown in Table 4.3.

Table 4.1 The number of samples for positional similarity measurement tests

Test sites	Number of samples
Cramlington1-UK	32
Cramlington2-UK	24
Clara Vale-UK	20
Baghdad-Iraq	32

Table 4.2 The number of samples for linear similarity measurement tests

Test sites	Number of samples	Average length (m)
Cramlington2-UK	37	109.475
Baghdad-Iraq	30	411.628

Table 4.3 The number of samples for area shape similarity measurement tests

Test sites	Number of samples	Average area (m ²)
Cramlington1-UK	19	1178.368
Cramlington2-UK	29	169.980
Clara Vale-UK	10	15633.994
Baghdad-Iraq	20	4219.605

Field and ground data were created in several field surveys to serve as a benchmark of definitive reference datasets for geometrical similarity comparisons. Al-Bakri and Fairbairn (2010) stated that the application of VGI via the Internet introduced new spatial data sources that require assessment with regard to their quality. Field survey data therefore may be considered more appropriate than formal information sources for such geometrical accuracy comparison, because the topographic maps and data of most formal agencies are usually produced using indirect photogrammetric survey from aerial photographs.

A high quality field survey using high precision survey instruments was recorded for all study areas throughout 2009 and 2010. The ground reference data was collected using two techniques. For the building detailed areas, control points were established first with Leica GX1230 GPS and Topcon GR3 GPS for the UK and Iraq sites respectively,

and then the survey observations were performed by Leica TS02 total station and Topcon GTS-225 total station for the UK and Iraq study areas respectively, as shown Figure 4.4. The positional accuracy of the instrument essentially depends on angular precision and stated distance which were 3" and $\pm(1.5 \text{ mm}+2\text{ppm})$ for Leica TS02 total station and 5" and $\pm(2 \text{ mm}+2\text{ppm})$ for Topcon GTS-225 total station (Leica, 2012a; Topcon, 2012a). On the other hand, in open access areas (undeveloped land), GPS-RTK mode was used for land surveying, as shown in Figure 4.5. The expected planimetric coordinate accuracy of the instrument and method was $\pm(10\text{mm}+1\text{ppm})$ (Leica, 2012b; Topcon, 2012b). GPS surveying was not applied to the collection of data in detailed areas because the lack of visibility and multipath may affect the performance of the RTK method. The dataset of the sample points consisted of 3D coordinates (Easting, Northing and Elevation) and included a range of features which were collected to the highest possible degree of accuracy.



Figure 4.4 Field work using GPS to establish control points, and total station to survey the location of points



Figure 4.5 GPS Leica GX1230, employed to survey the location of features in open-space area using RTK technique

4.3 Positional heterogeneity

Assessing geospatial data integration in scenarios such as those introduced in Chapter 2 clearly requires testing of relative positional accuracy. Consider, for example, a suggested project requires combining the positions of certain public facilities from OSM data onto FM datasets. A distribution problem of overlaid features may occur, if there is an inconsistency between the locations of the compared datasets. For instance, some facilities may locate in the middle of roads or they may appear in the middle of rivers. The following subsection will, therefore, demonstrate details of metrics and indices for the assessment of the accuracy of positional and angular data. In addition, the processing steps that were applied to evaluate the possibility of positional integration of FS, FM and VGI data sources will be described. The section will conclude with a description of positional similarity comparisons, utilising the programs and coded interfaces that have been developed for this purpose.

4.3.1 Data accuracy standards

In Chapter 2, the term positional accuracy was broadly defined and discussed. Assessing such positional accuracy can be facilitated by adopting established standards for spatial data accuracy, thus providing an aid to the accuracy of measurement and reporting across a whole dataset. Despite the utility of this, spatial data accuracy standards were not developed until the middle of the last century. For instance, Marsden (1960) reported that during the period before 1900 and in the early part of the 20th century, in the United States evaluation of a map's positional accuracy was established by looking at the specifications of the field survey method. At that time, the accuracy evaluation of topographic maps was fundamentally based on the individual surveyor's skills or professional standing. The distribution of control points in survey networks was also considered as another element for such positional accuracy evaluation. Afterwards, in the 1930s, the rapid growth of the use of photogrammetry for the production of maps promoted the development of the US National Map Accuracy Standards (NMAS). Although the mapping and surveying professions would not accept that photogrammetry should replace traditional surveying methods, the development of standards did receive some support from many private institutions and government agencies such as the United States Geological Survey (USGS) (Marsden, 1960).

In 1939 the American Society of Photogrammetry published map accuracy specifications in order to develop national standards. After the following year, the committee of the Federal Board of Surveys and Maps developed an accuracy standard based on a map production method. They suggested different standards for maps produced from field surveys to those produced from photogrammetry. In 1941, the United States Director of the Bureau of the Budget issued the first version of the United States National Map Accuracy Standards (NMAPS). The standards were revised several times in 1943 and 1947. The NMAPS could apply to all federal maps agencies, as it was clear and easy. It could be used for measuring and reporting the accuracy of both horizontal and vertical locations (USBB, 1947). Afterwards, the 1947 standards were reviewed and updated by the American Society of Photogrammetry and Remote Sensing (ASPRS). The goal was to establish new standards for both hardcopy and digital maps. As a result there are many other standards for positional accuracy assessment that have been established and used in practice, such as Accuracy Standards for Large Scale Maps (ASPRS, 1989), or the more recent spatial data accuracy standards, such as the National Standard for Spatial Data Accuracy (NSSDA) (FGDC, 1998b).

4.3.2 The US National Standard for Spatial Data Accuracy

The standards described above are similar with regard to some properties; for example, data quality assessment can be performed by comparing a lower accuracy survey with a more accurate data source (e.g. a formal, official dataset). Also, the estimation of accuracy for all standards is essentially based on the analysis of the differences between the coordinates of compared points. However, existing standards such as the National Map Accuracy Standards (NMAPS) generally focus on testing paper maps rather than digital data. Furthermore, the standard measures the errors at the map scale instead of ground scale. This can be considered problematic when changing the mapping system into digital formats that can be output at varying scales. In contrast, in a case study approach NSSDA was chosen to specify and report the positional accuracy at ground scale rather than map scale. The obstacle regarding the estimation of positional accuracy on paper has been overcome by using NSSDA. It provides a positional evaluation method that can be used with both digital geographic data and graphic maps (Congalton and Green, 2009a).

The NSSDA provides a testing and statistical methodology for determining the horizontal and vertical accuracy of tested datasets. It also provides a formal approach to how the tested points should be identified, measured and distributed across the study area. It suggests that twenty or more test points are required to effectively test data accuracy. These points must be well defined, easy to measure and found in both tested and reference datasets. For horizontal accuracy assessment, the ideal distribution of tested points is even with at least 20 percent of the points in each quadrant when the dataset covers a rectangular area. The intervals between points should be at least 10 percent of the diagonal distance of the total area of the dataset (Zandbergen, 2008). Figure 4.6 shows the ideal distributions and intervals among NSSDA tested points.

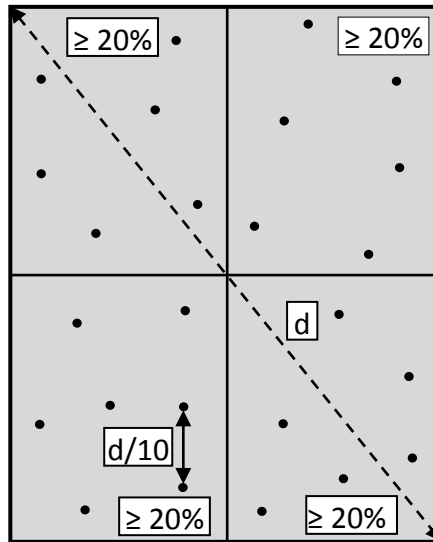


Figure 4.6 Distribution and intervals among tested points for NSSDA procedure (Givens, 1999)

FGDC (1998b) reported that the NSSDA estimates horizontal accuracy by using root mean square error (RMSE). The NSSDA records horizontal accuracy at the 95% confidence interval. The 95th percentile indicates that 95% of positions in the dataset will have an error compared to the true locations on the ground of equal or less than the reported accuracy value, and 5% of the errors will be larger than the reported accuracy value. Computing accuracy according to NSSDA can be summarised as follows:

$$RMSE_E = \sqrt{\frac{1}{n} \sum_{i=1}^n (E_c - E_r)^2}; \quad RMSE_N = \sqrt{\frac{1}{n} \sum_{i=1}^n (N_c - N_r)^2}; \quad RMSE = \sqrt{\frac{1}{n} \left(\sum_{i=1}^n \delta E_i^2 + \sum_{i=1}^n \delta N_i^2 \right)} \quad (4-1)$$

Where:

n : The number of check points

E_c, N_c : The coordinates of compared dataset

E_r, N_r : The coordinates of reference dataset

$\delta E_i^2, \delta N_i^2$: The direct linear distance mismatches for the i^{th} checked point in x and y directions

According to FGDC (1998b), the NSSDA accuracy can be computed for two cases:

If $RMSE_E = RMSE_N$ then:

$$\begin{aligned} RMSE &= \sqrt{2(RMSE_E)^2} = \sqrt{2(RMSE_N)^2} \\ &= 1.4142 \times RMSE_E = 1.4142 \times RMSE_N \end{aligned}$$

$$\begin{aligned} \text{NSSDA accuracy} &= 2.4477 RMSE_E = 2.4477 RMSE_N \\ &= 2.4477 \times (RMSE/1.414) \\ &= 1.7308 \times RMSE \end{aligned}$$

If $RMSE_E \neq RMSE_N$ then:

$$\text{NSSDA accuracy} = 2.4477 \times 0.5 \times (RMSE_E + RMSE_N)$$

Greenwalt and Shultz (1962) showed that the factor 2.4477 can be used to calculate the horizontal accuracy at the 95% confidence interval. The value of this factor can be calculated as following:

$$k = \sqrt{-2 \ln(1 - p)} \quad (4-2)$$

Where:

p : is the desired probability (confidence interval) value which is in this case equal to 0.95

For this study, the NSSDA methodology has been adopted, and a special code in Matlab has been developed, as will be discussed in subsection 4.3.4. This is basically performed in order to test and report the ability of integration of the locations of point's data from FS, FM and VGI data sources.

4.3.3 Circular data for positional discrepancies analysis

The method of measuring circular or directional discrepancies is usually applied in order to investigate whether there are any systematic errors present in the directions of point discrepancies. Circular observations can be summarized as locations on a unit circle or as angles over a 360° or 2π radians range. Each angular observation can be specified by a direction or unit radius on circumference of a unit circle centered at the origin. Directional data is utilized in a number of disciplines of science: computer science, for image analysis to represent the orientation of the texture on fingerprints for example (Hanbury, 2003); earth science and geology, for the estimation of relative movements of the tectonic plates and inferring the directions of earthquake effects (Dey and Ghosh, 2008); meteorology, to estimate the wind direction and speed (Bowers et al., 2000); agricultural sciences, to analyze the effect of the wind on trees and forest growth (Aradóttir et al., 1997); environmental science, to study the movement direction of ice and variation in the direction of birds' migration (Arnold and SenGupta, 2006); in geography, where Corcoran et al. (2009) investigated another application of circular data by analyzing "journey to work" data, a study that involved determining the relationship between the residential passengers' flow zones and destination zones. Circular data is also used in spatial data analysis for assessing the direction of the error vector of points that connect the correct and measured locations (Polo and Felicísimo, 2010). Although these examples come from a wide range of scientific fields, they all use circular data and they need special statistics for calculation and analysis, as will be discussed in the following subsection.

4.3.3.1 The need for appropriate angular descriptive statistics

Distance and direction are the implicit concepts of all spatial relationships. However, for some instances in geography, the observations of directions are only of interest. Statistical problems may occur when the data are in the form of directional or angular measurements. This is because there is a difference between directional and linear data when it comes to the analysis of certain elements. For example, linear data is usually measured with respect to a common specific point or origin. In angular data the observations are measured from an arbitrary reference direction or axis. In addition, the data in the form of angles or directions are treated in a different way to linear data. For instance, the mean direction of two observations at 358° and 2° is not 180° , as would be

obtained from linear arithmetic mean. Thus, the standard statistical calculation methods for linear variables are not appropriate for directional analysis. On the other hand, there has been an increasing amount of literature on mathematical statistics of circular or directional data; see for example, Jammalamadaka and SenGupta (2001), Mardia and Jupp (2000), and Fisher (1993). These all provided a useful description of the calculation and development of circular data analysis by applying vector addition technique.

For instance, if there are unit vectors v_1, v_2, \dots, v_n with their directions $\beta_1, \beta_2, \dots, \beta_n$, the mean direction $\bar{\beta}$ of the unit vector's directions would be the direction of the resultant vector R . It is also known as the vector of the centre of mass. There are two main components for the resultant vector: the mean direction $\bar{\beta}$, and the mean resultant length \bar{R} . They often form a useful starting point for any directional analysis. A simple example of vector addition calculation for the directions of four unit vectors is illustrated in Figure 4.7. In general, angular value is based on the rotation angle (clockwise or counter clockwise) and the zero direction (azimuth). In this project, the North is considered as the origin of direction and the sense of rotation is increasing clockwise. The mean directional angle and the length of the combined vectors (resultant) can be calculated as follows:

$$S = \sum_{i=1}^n \sin(\beta_i) \quad (4-3)$$

$$C = \sum_{i=1}^n \cos(\beta_i) \quad (4-4)$$

$$R = \sqrt{C^2 + S^2} \quad (4-5)$$

$$\bar{R} = R/n \quad (4-6)$$

$$\bar{\beta} = \begin{cases} \tan^{-1}\left(\frac{S}{C}\right) & S > 0, C > 0 \\ \tan^{-1}\left(\frac{S}{C}\right) + \pi & C < 0 \\ \tan^{-1}\left(\frac{S}{C}\right) + 2\pi & S < 0, C > 0 \end{cases} \quad (4-7)$$

$$V = 1 - \bar{R} \quad (4-8)$$

Where:

β_i : The direction of (units) vectors

R : The resultant length

\bar{R} : The mean resultant length associated with the mean direction $\bar{\beta}$

$\bar{\beta}$: The direction of the mean resultant vector

V : The circular variance

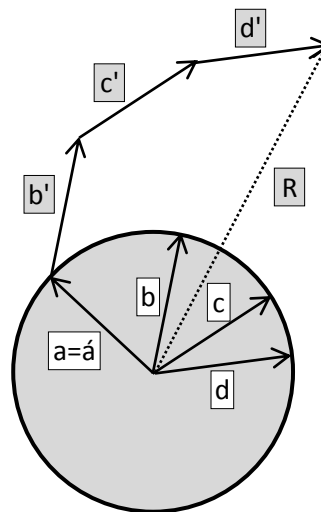


Figure 4.7 Example of unit vectors summation to calculate resultant length and direction (Fisher, 1993)

Polo and Felicísimo (2010) stated that \bar{R} is a natural measure of the spread of the unit vectors. The interpretation of mean resultant length is that if the data are closely clustered around the mean, then \bar{R} is close to 1. This indicates that a set of positional discrepancies have a uniform direction. However, for varying direction of positional errors, \bar{R} will be near zero. Circular variance is another indicator of dispersion of circular data. Like \bar{R} , the circular variance is bounded in the interval $[0, 1]$. If the circular variance is small, then all the point discrepancies are in the same direction. If the discrepancies are spread out evenly around the circle, the circular variance will be close to 1. For the current project, circular statistics were tried for further analysis of the directions of positional errors. This was undertaken through developing a program as a part of the location errors measurement code, a process which will be described in the next subsection.

4.3.4 Positional similarity detection and implementation

The positional discrepancies between any tested and accepted reference datasets can be represented as vectors. In 2D space, a vector can be defined by scalar and angular values. Therefore, this provides an appropriate opportunity to analyse the magnitude and the directions of errors between compared datasets simultaneously (Cuartero et al., 2010). In this section, ways of analysing the deviation between the reference data (FS) and the corresponding data with FM and VGI sources, and hence between each other, have been undertaken. The statistical operations suggested in previous subsections were applied by developing and designing a positional data analysis interface. A Graphical User Interface (GUI) was developed through the Matlab environment. Matlab can be considered as one of the most effective programming languages for solving technical computing problems. In this context, Tahir and Pareja (2010) mentioned that Matlab has these properties because it has the ability for efficient mathematical computations, especially matrices operations, development of algorithms, data manipulating and visualising, producing engineering and scientific graphs, in addition to including tools for programming in an easy to use system for constructing GUIs.

A GUI or graphical user interface was generally defined by Nordin (2008) as a convenient system to represent the program output information, while providing a suitable way to understand and control the manipulating of data through creating specific management components such as menus, pushbuttons and list boxes. In addition, GUIs can offer an easier way to specify what is required as input data for such programs because of their visual approach that allows computer users immediate interactivity with this task. Tahir and Pareja (2010) reported that from version 5.0, Matlab allowed users to produce their own GUIs. The Matlab GUI can be created by means of the Matlab Graphical User Interface Development Environment (GUIDE) which can provide tools that make GUIs programming and processing easier (MathWorks, 2012). In general, when producing a GUI by Matlab there are two files that will be generated with different file extensions: the first file will contain the different GUIs' figure components, which are (.fig) files, while the second one includes the code that can be used to load the GUI components, which are (.m) files.

For the current project, the positional similarity measurements' interface encoding (.m) file for tested datasets is presented in Appendix B, whereas the figures for the GUI (.fig)

file can be seen in Figures 4.8 to 4.19. The data properties (E, N) of tested points should be extracted to files in order to determine the positional similarity analysis. The data required for processing can be imported to the interface in any convenient data format (e.g. .txt and .xlsx) by clicking a button which has been created at the top of the developed interface. The interface basically consists of several parts. Firstly, the linear statistics were employed in order to analyse the linear error components among compared tested points. This was carried out by applying the equation 4.1 to calculate RMSE and consequently compute NSSDA accuracy, in addition to calculating the descriptive statistics such as mean, standard deviation, median, maximum, minimum and interquartile range along the differences of (E, N). Secondly, the directional statistics were adopted to calculate the components of angular errors, as in equations 4.3 to 4.8. The graphical analysis was included as a third part of this interface in order to obtain a full location data analysis. The positional similarity interface also offers the ability to export and save the results. The developed positional similarity measurement tool was tested for different datasets of different sites, as shown in Figures 4.8 to 4.19.

4.3.4.1 Results and observations of formal and VGI datasets comparisons

Figures 4.8 to 4.10 show the outcome interfaces of the numerical and graphical analysis results of the comparison of FS, OS and OSM datasets in the Cramlington1-UK study area. It can be observed from Table 4.4 that the RMSE and NSSDA accuracy values for the first comparison (FS/OS) are relatively lower than the other two comparisons. The findings indicate that there is a wide range of discrepancies between FS/OSM datasets and also between OS/OSM datasets. The same interfaces have also provided the magnitudes and graphical display of the values of the deviations between the locations in the tested datasets. In addition, the descriptive statistics such as mean, standard deviation, median, maximum, minimum and interquartile range along the differences of (E, N) of the compared data were calculated. The analysis revealed that the components of the errors for the comparison of FS/OS are very small or less than one metre for most of sample points. In contrast, the tests of FS/OSM and OS/OSM datasets showed that there are significant differences between the discrepancies of these datasets. For instance, the maximum values of the deviations in the (E, N) for the comparing of FS with OS datasets are (0.452m, 0.518m); whereas, they are (6.898m, 5.946m) and (6.749m, 5.815m) for the comparisons of FS/OSM and OS/OSM respectively. The

initial results of positional comparisons confirm that it would be difficult to match OSM data with reference or formal datasets.

The assessment of the directional accuracy was calculated for the sets of the sample points using angular statistics. In this part of the project, the parameters of the directional statistics, such as mean direction, mean resultant length and circular variance, were considered with the same interface tool as the positional similarity measurement, as can be seen at the bottom of Figures 4.8 to 4.10. The initial explorations of the angular error components for the compared datasets, FS/OS, FS/OSM and OS/OSM, are summarised in Table 4.5 respectively. The discrepancies between FS and OS data indicate less variability than is found in the other comparison. A relatively higher value of the mean of the resultant length of the first case (FS/OS) shows that its data was slightly more concentrated around the mean direction than other situations. As the other basic parameter for angular error analysis, the circular variance is a measure of the concentration of data around the mean direction. For the tests undertaken here, it is obvious that the value of circular variance is much lower when comparing the reference data with formal data directly than when comparing each of the reference and formal data with OSM information. Therefore, the data of the FS/OS test show a more uniform distribution with relation to the mean direction than other comparisons.

Although numerical analyses of the positional and directional error components are included, the analysis can be considered incomplete without graphical representation of the discrepancies of the compared datasets. This can provide information regarding significant characteristics of error elements, such as the length and the distribution of error vectors. Measures are graphically presented in the right part of Figures 4.8 to 4.10. The distribution of the tested data and the sample size can be assessed through these figures. For instance, it is apparent from the figures below that the discrepancies vectors are not following any regular direction, as the errors have a wide range and multi-directional distribution. Figures 4.8 to 4.10 also show a vector diagram containing all vector errors which radiate from the centre of the circle. The mean direction of the compared datasets is plotted with a red line. Visual interpretation of these polar plots refers to varied directional discrepancies of compared datasets. However, the comparison of FS with OS datasets showed more concentration of the error directions around the mean direction than the comparisons of FS/OSM and OS/OSM datasets.

From the above description, the results analyses indicated that there is a convergence between the FS and OS datasets, while there are significant differences between FS, OS and OSM datasets which make the using of OSM data for the purpose of geospatial data integration a difficult task.

Table 4.4 Comparisons of RMSE and NSSDA accuracy of compared datasets in Cramlington 1-UK

Datasets	RMSE _(m)	NSSDA accuracy _(m)
FS/OS	0.492	0.846
FS/OSM	5.429	9.143
OS/OSM	5.331	8.989

Table 4.5 Circular statistics of compared datasets in Cramlington 1-UK

Datasets	Mean direction of positional discrepancy ($\bar{\theta}$)	Mean resultant length (\bar{R})	Circular variance (V)
FS/OS	248.194°	0.455	0.545
FS/OSM	299.956°	0.199	0.801
OS/OSM	313.743°	0.133	0.867

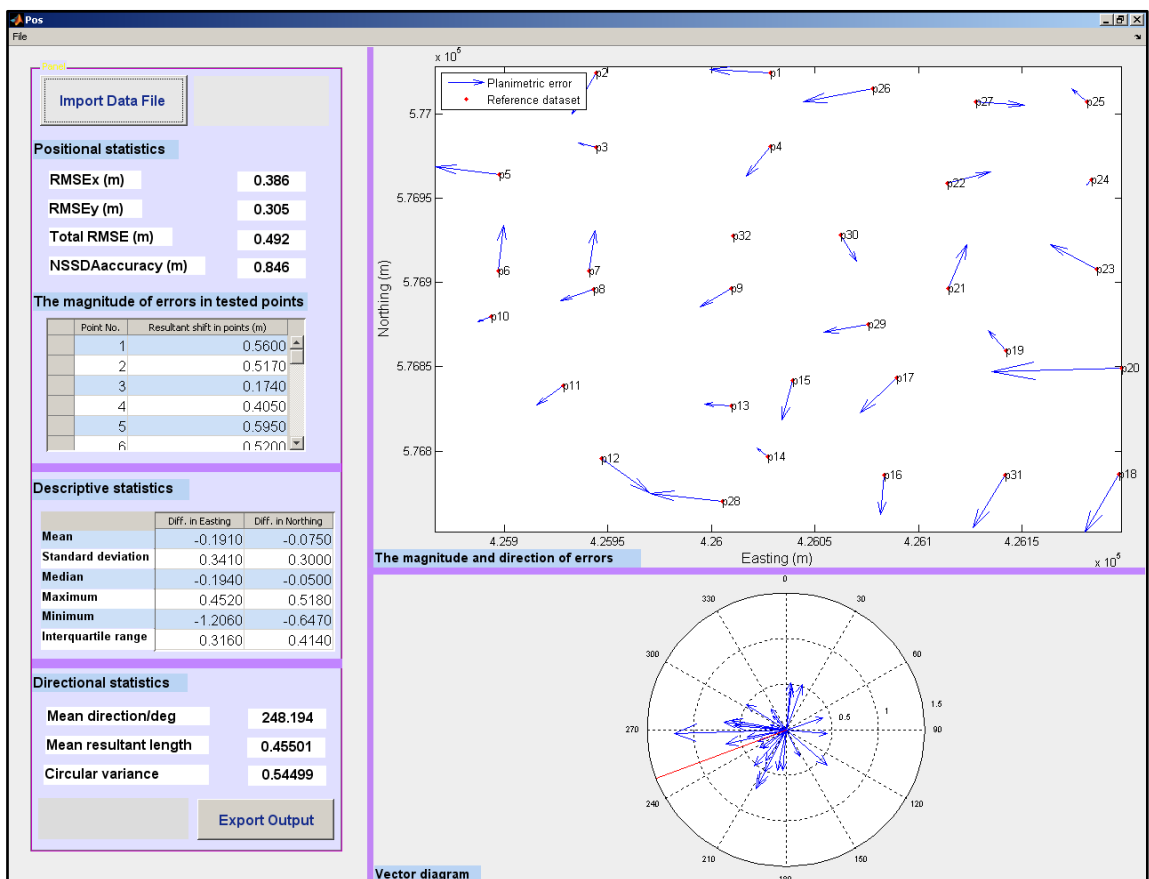


Figure 4.8 The positional similarity measurement results for the comparison of FS and OS datasets in Cramlington1-UK

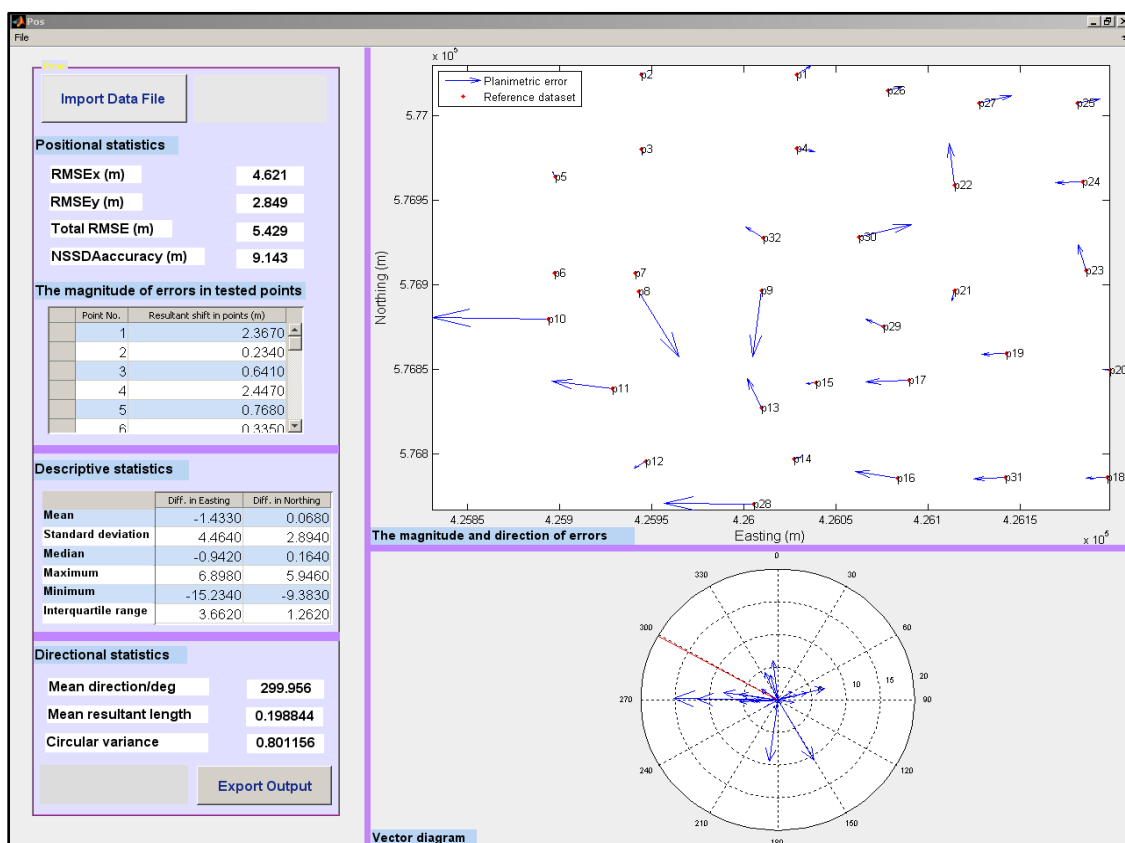


Figure 4.9 The positional similarity measurement results for the comparison of FS and OSM datasets in Cramlington1-UK

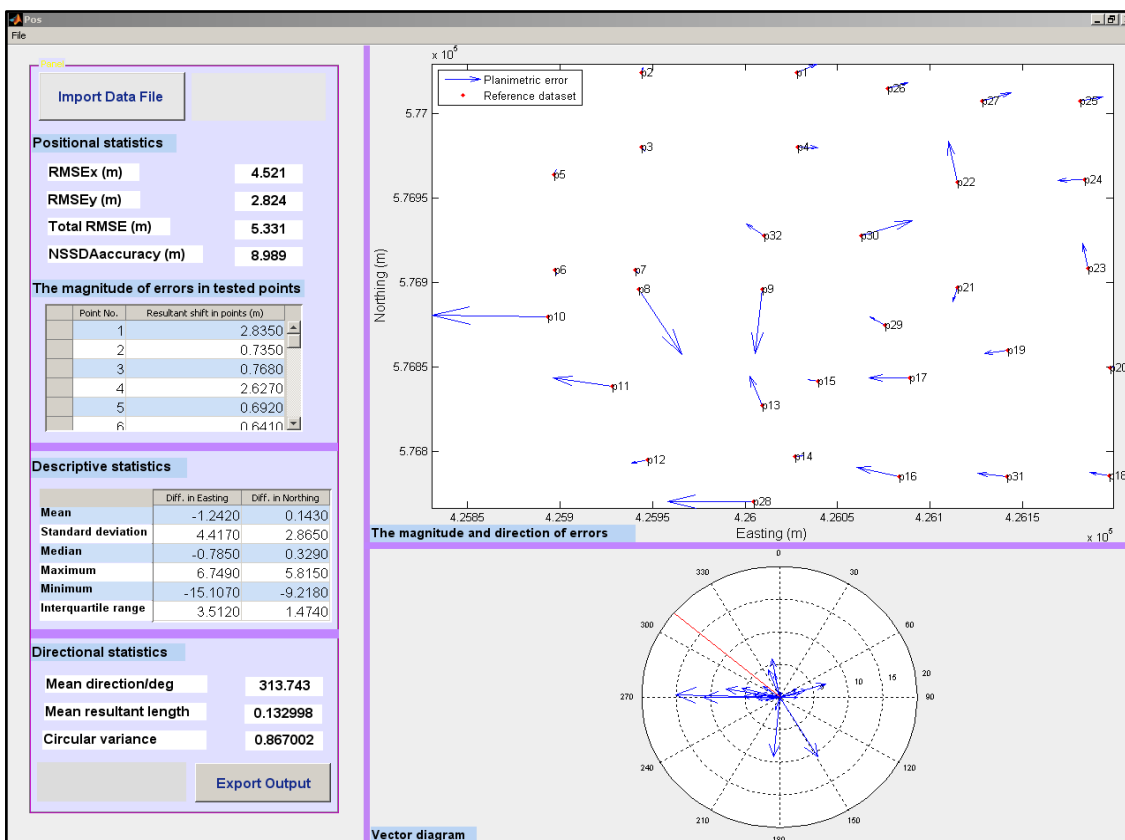


Figure 4.10 The positional similarity measurement results for the comparison of OS and OSM datasets in Cramlington1-UK

Figures 4.11 to 4.13 illustrate the results of positional analysis of the Cramlington2-UK dataset. Table 4.6 reported the RMSE and NSSDA accuracy values for the comparisons of FS/OS, FS/OSM and OS/OSM datasets respectively. The findings of this part of the study indicated that the RMSE and NSSDA accuracy values when comparing the FS with OS data were very low compared to FS and OS data against OSM information. The descriptive statistics such as mean, standard deviation, median, maximum, minimum and interquartile range along the differences of (E, N) of the compared datasets can also be seen in the same figures. The results revealed that there is a significant difference between the E and N components of errors for the FS/OSM and OS/OSM tests. However, the results showed that the discrepancies between the FS and OS datasets are very close to each other. For example, the maximum values of the differences in (E, N) were (0.261m, 0.693m), (7.317m, 5.556m) and (7.526m, 5.672m) for the FS/OS, FS/OSM and OS/OSM comparisons respectively. The directional statistical analysis found that the distribution of error directions varies in all comparisons, as shown in Table 4.7. This can be confirmed by looking at the graphical representations for the discrepancies of compared datasets. These plots show multi-directional distribution errors for all tests undertaken.

It can be initially concluded that there is a possibility of integrating the positional data from reference field survey and formal data sources as they have relatively small discrepancies. However, there is a significant difference in the positional similarity between the comparison of FS and OS against OSM datasets.

Table 4.6 Comparisons of RMSE and NSSDA accuracy of compared datasets in Cramlington 2-UK

Datasets	RMSE _(m)	NSSDA accuracy _(m)
FS/OS	0.342	0.590
FS/OSM	4.500	7.714
OS/OSM	4.564	7.796

Table 4.7 Circular statistics of compared datasets in Cramlington 2-UK

Datasets	Mean direction of positional discrepancy ($\bar{\theta}$)	Mean resultant length (\bar{R})	Circular variance (V)
FS/OS	136.415°	0.046	0.954
FS/OSM	212.771°	0.275	0.725
OS/OSM	211.313°	0.257	0.743

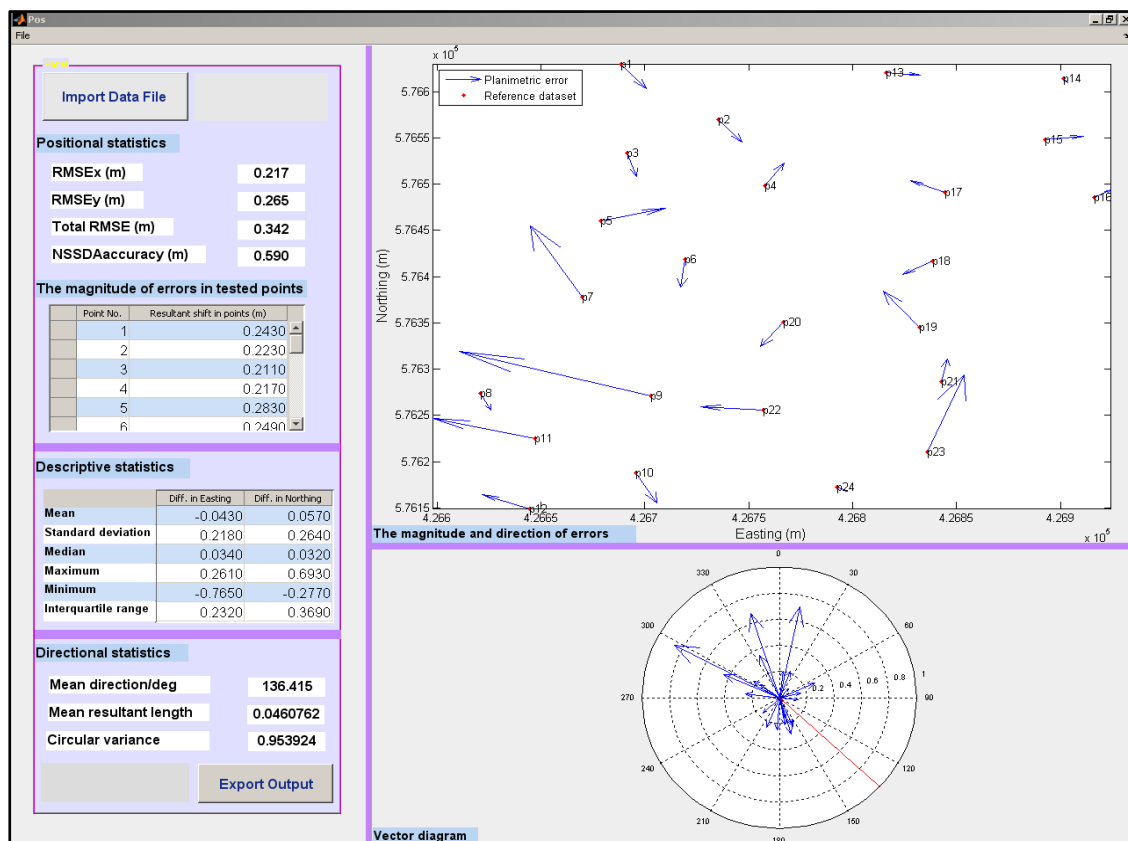


Figure 4.11 The positional similarity measurement results for the comparison of FS and OS datasets in Cramlington2-UK

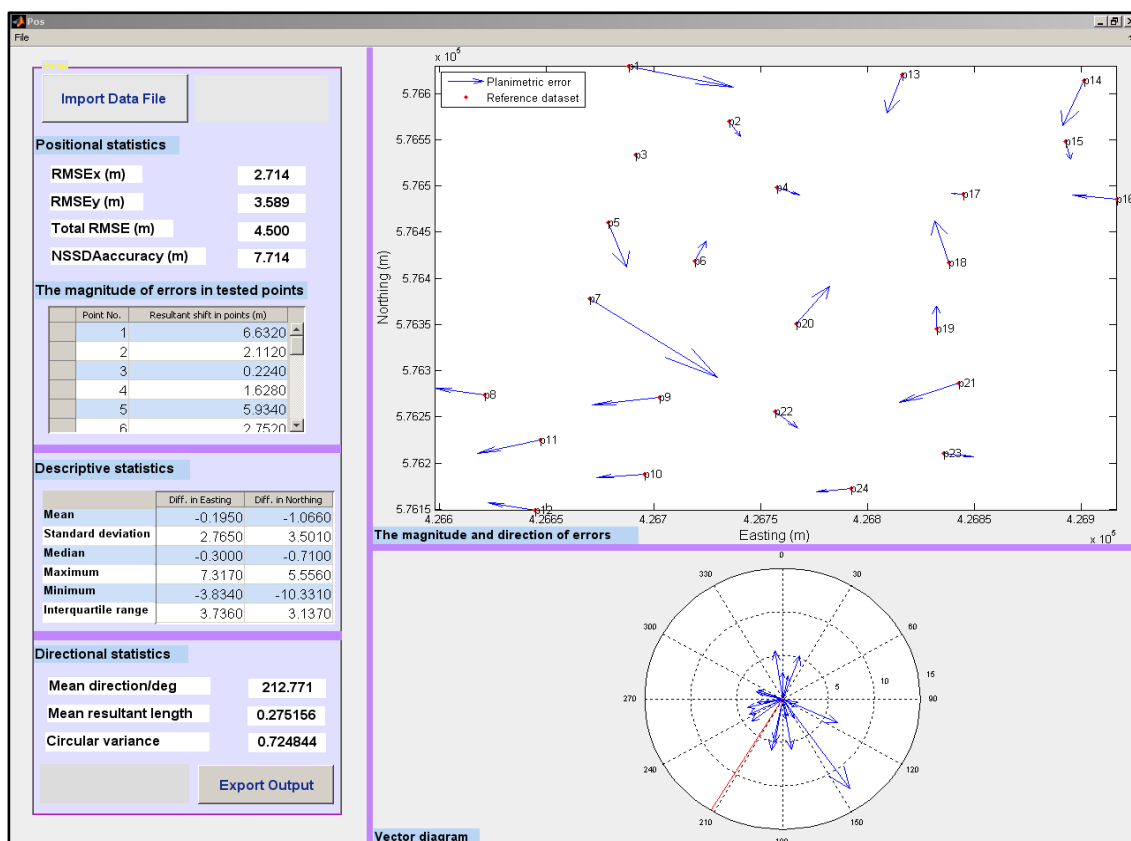


Figure 4.12 The positional similarity measurement results for the comparison of FS and OSM datasets in Cramlington2-UK

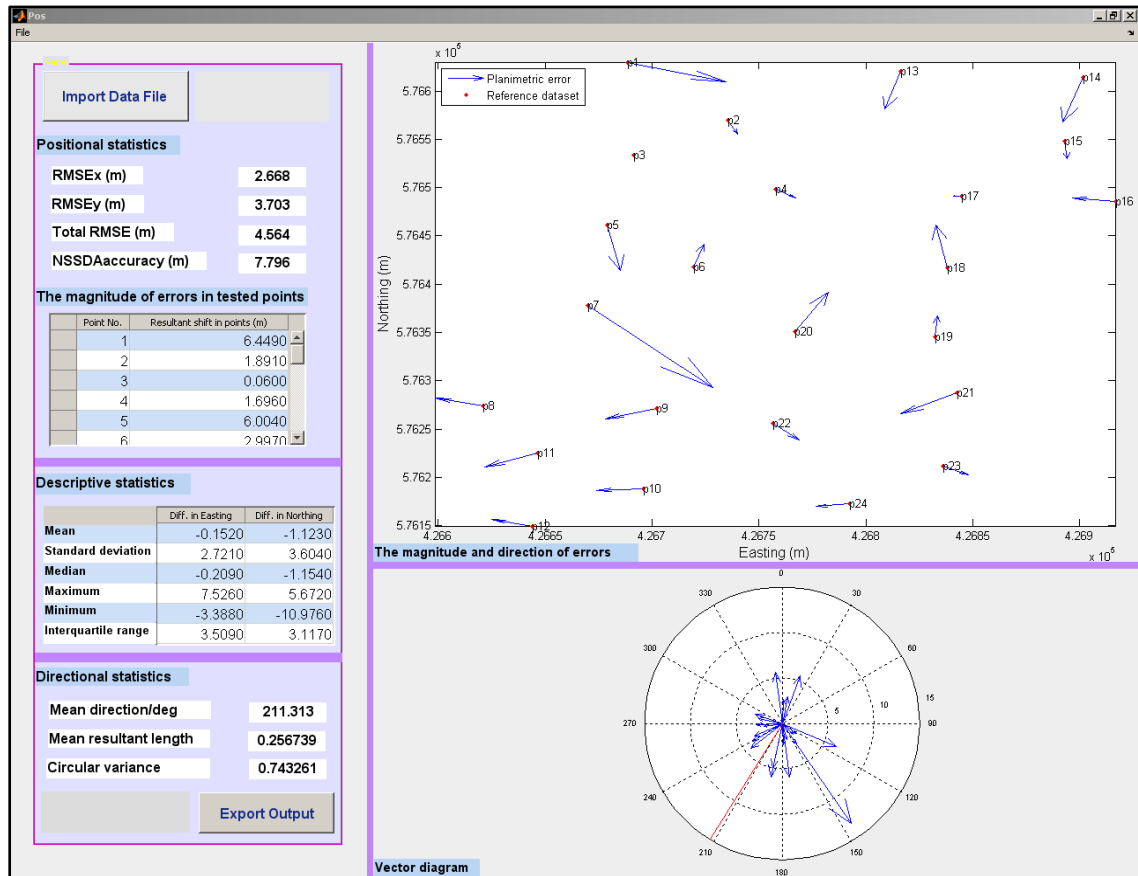


Figure 4.13 The positional similarity measurement results for the comparison of OS-OSM datasets in Cramlington2-UK

In order to assess the possibility of positional integration among datasets, both VGI and formal data, in the rural area, the same procedure for analysis as employed in the previous tests was adopted (Figures 4.14 to 4.16). The results obtained from the preliminary analysis of the total RMSE and NSSDA accuracy of FS/OS, FS/OSM and OS/OSM comparisons are presented in Table 4.8 respectively. The data from this table showed that the first situation (FS/OS) reported significantly different results from the other two groups. The linear statistical analysis showed that the values of the RMSE and NSSDA accuracy when comparing FS/OS are less than the values of comparing FS/OSM or OS/OSM. The descriptive statistics calculations for the component of errors (E, N) suggested the same observations of the linear statistics analysis. It is clear from the figures below that FS/OS resulted in the lowest values of all the descriptive statistics calculations such as mean, standard deviation, median, maximum, minimum and interquartile range for the discrepancies of (E, N) of the compared datasets. For example, the maximum values of the differences between the components coordinates (E, N) were (4.401m, 2.062m), (17.796m, 14.169) and (17.778m, 13.164m) for FS/OS, FS/OSM and OS/OSM comparisons respectively. In the uniformity of directional error

analysis, the numerical and graphical representations revealed that the discrepancies between the tested datasets are not homogeneous for all comparisons, as illustrated in Table 4.9 and figures 4.14 to 4.16. This part of analysis concluded that there are such similarities that exist between the FS and OS datasets, whereas OSM information does not match well with FS or OS datasets.

Table 4.8 Comparisons of RMSE and NSSDA accuracy of compared datasets in Clara Vale-UK

Datasets	RMSE _(m)	NSSDA accuracy _(m)
FS/OS	1.843	3.189
FS/OSM	11.650	20.161
OS/OSM	10.887	18.832

Table 4.9 Circular statistics of compared datasets in Clara Vale-UK

Datasets	Mean direction of positional discrepancy ($\bar{\theta}$)	Mean resultant length (\bar{R})	Circular variance (V)
FS/OS	52.055°	0.346	0.654
FS/OSM	45.478°	0.341	0.659
OS/OSM	43.395°	0.277	0.723

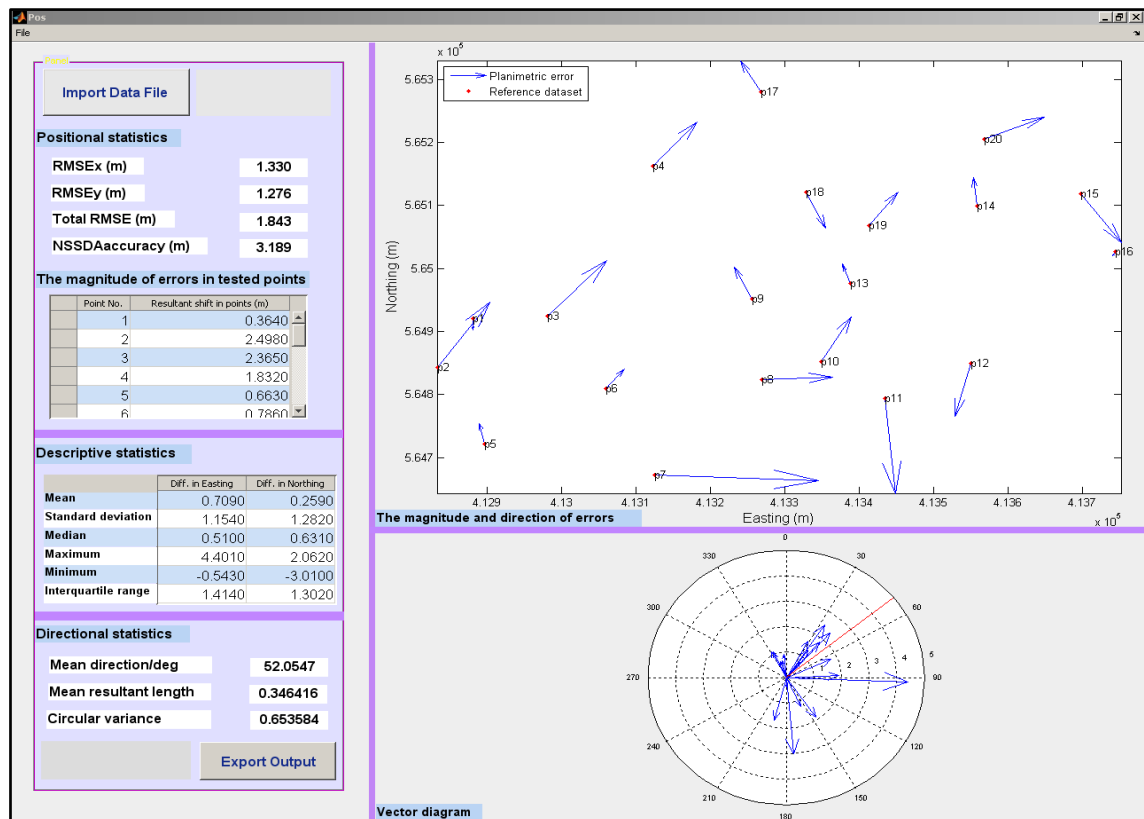


Figure 4.14 The positional similarity measurement results for the comparison of FS-OS datasets in Clara Vale-UK

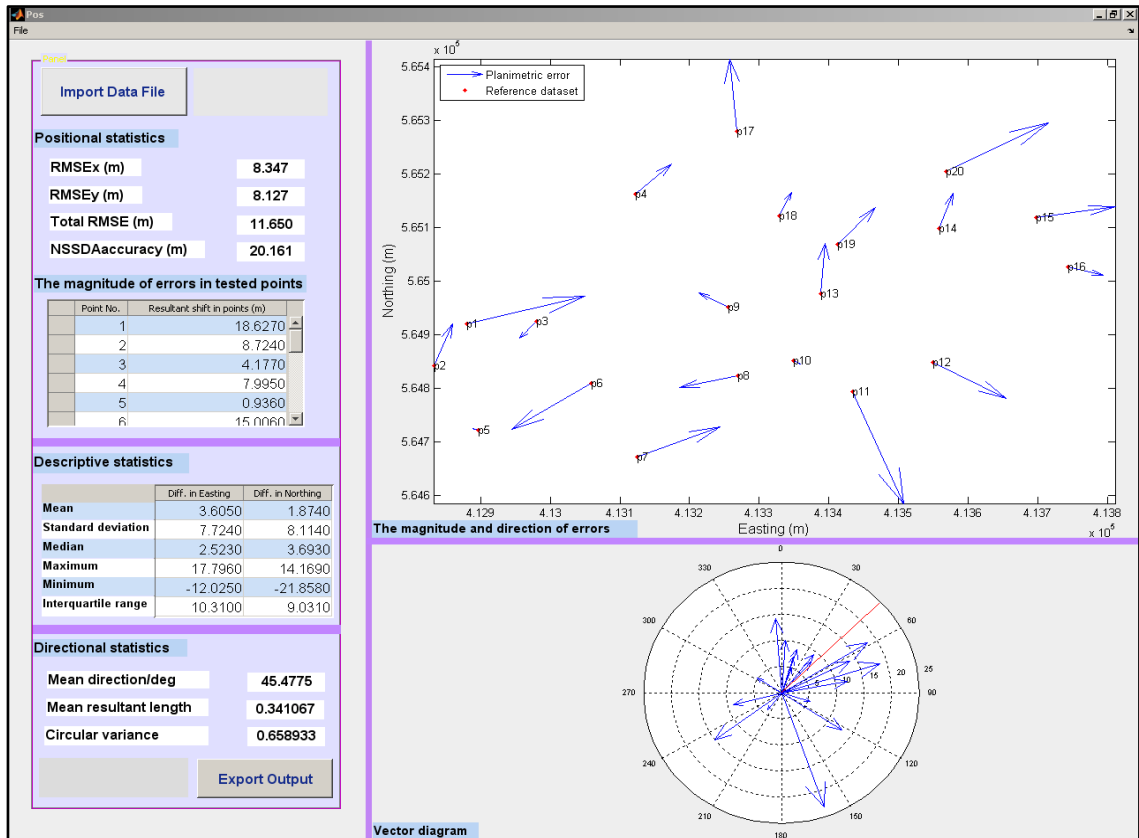


Figure 4.15 The positional similarity measurement results for the comparison of FS-OSM datasets in Clara Vale-UK

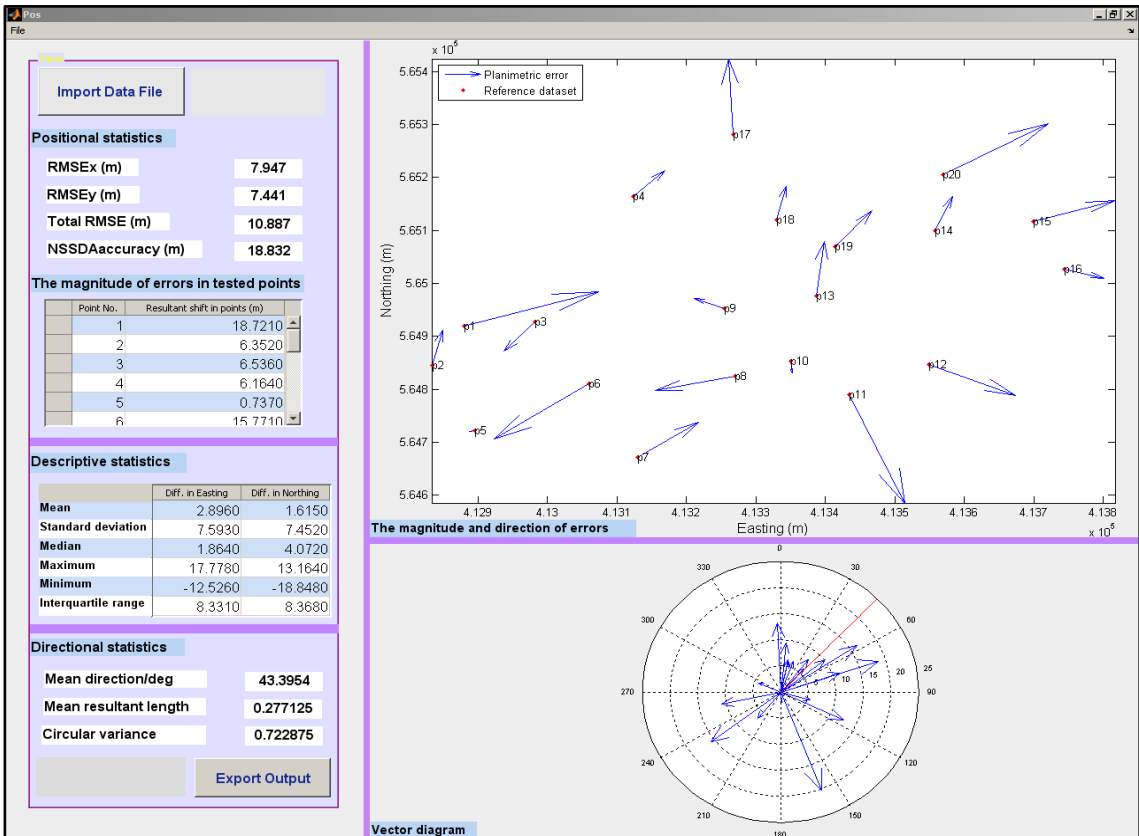


Figure 4.16 The positional similarity measurement results for the comparison of OS-OSM datasets in Clara Vale-UK

The study also aims to derive a comparative accuracy assessment of reference, formal and VGI datasets for the Baghdad-Iraq study area, with the intention of assessing the possibility of geospatial data integration in a different setting. Figures 4.17 to 4.19 and tables 4.10 to 4.11 illustrate some of the main characteristics of the linear and angular statistical analysis. In general, the findings of the Baghdad-Iraq study are somewhat similar to the results of the UK urban sites. At the same time, the results can be considered better than the results of the UK rural area. It is interesting to note that in the Baghdad study, the RMSE values of the comparisons of the reference and formal data with VGI data are only slightly different to the RMSE values of the same comparisons in the Cramlington1-UK and Cramlington2-UK (urban) sites. The RMSE values were different by only approximately half a metre between this study area and Cramlington1-UK, and about one metre for the other urban comparison. On the other hand, the findings show that the difference was around five meters for the comparison with the rural area. The same observations can be made for the comparisons of NSSDA accuracy values. Although the RMSE and NSSDA accuracy values in urban areas are better than the rural area, there is still quite a wide discrepancy between the comparisons of OSM data with field survey or formal datasets.

The mismatches between FS and OSM datasets and in the comparisons of OSM with OS and GDS datasets mean that using OSM data to initiate or revise the OS or GDS dataset would be extremely difficult. In order to obtain valuable geometrical similarity assessment, the linear features are also included in the research analytical steps. The next section will focus on the assessing of the linear similarity measurement of formal and VGI datasets by comparing them with benchmark field survey data.

Table 4.10 Comparisons of RMSE and NSSDA accuracy of compared datasets in Baghdad-Iraq

Datasets	RMSE _(m)	NSSDA accuracy _(m)
FS/GDS	1.246	2.149
FS/OSM	5.903	10.190
GDS/OSM	5.806	10.012

Table 4.11 Circular statistics of compared datasets in Baghdad-Iraq

Datasets	Mean direction of positional discrepancy ($\bar{\theta}$)	Mean resultant length (\bar{R})	Circular variance (V)
FS/GDS	252.535°	0.330	0.670
FS/OSM	277.473°	0.291	0.709
GDS/OSM	282.595°	0.236	0.764

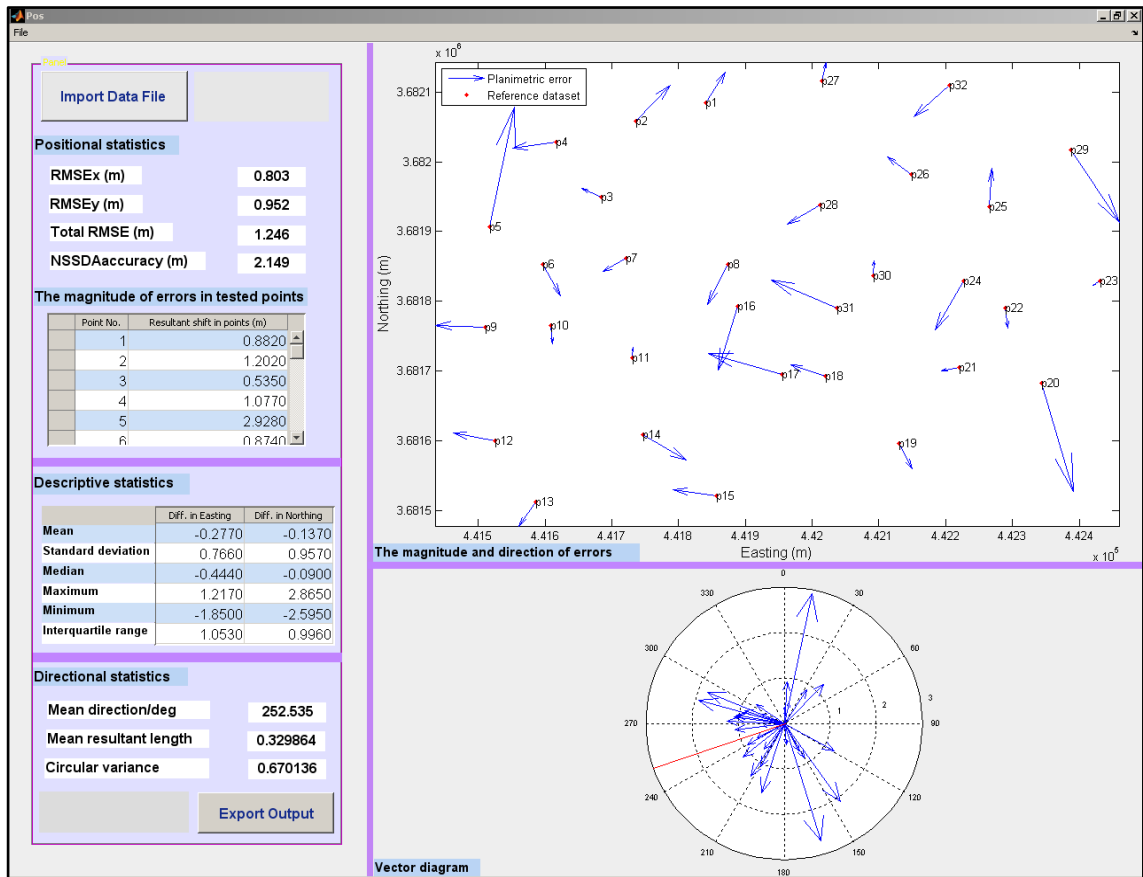


Figure 4.17 The positional similarity measurement results for the comparison of FS-GDS datasets in Baghdad-Iraq

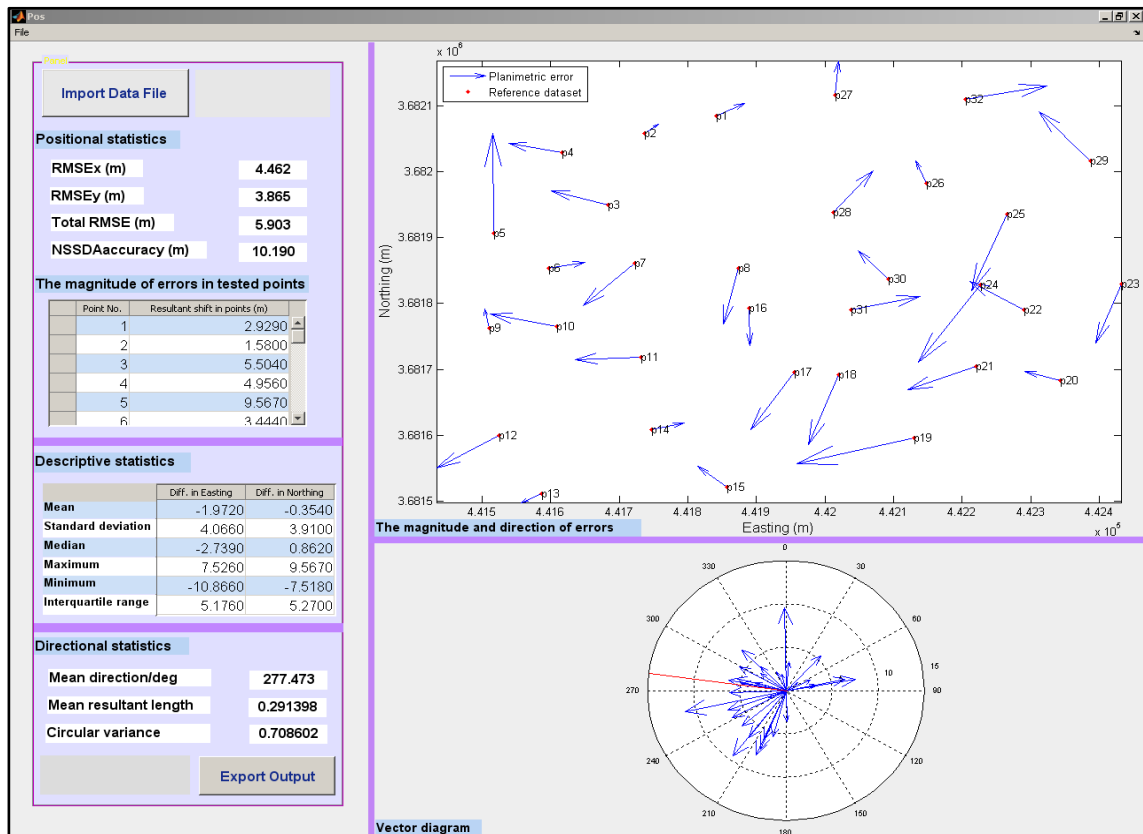


Figure 4.18 The positional similarity measurement results for the comparison of FS-OSM datasets in Baghdad-Iraq

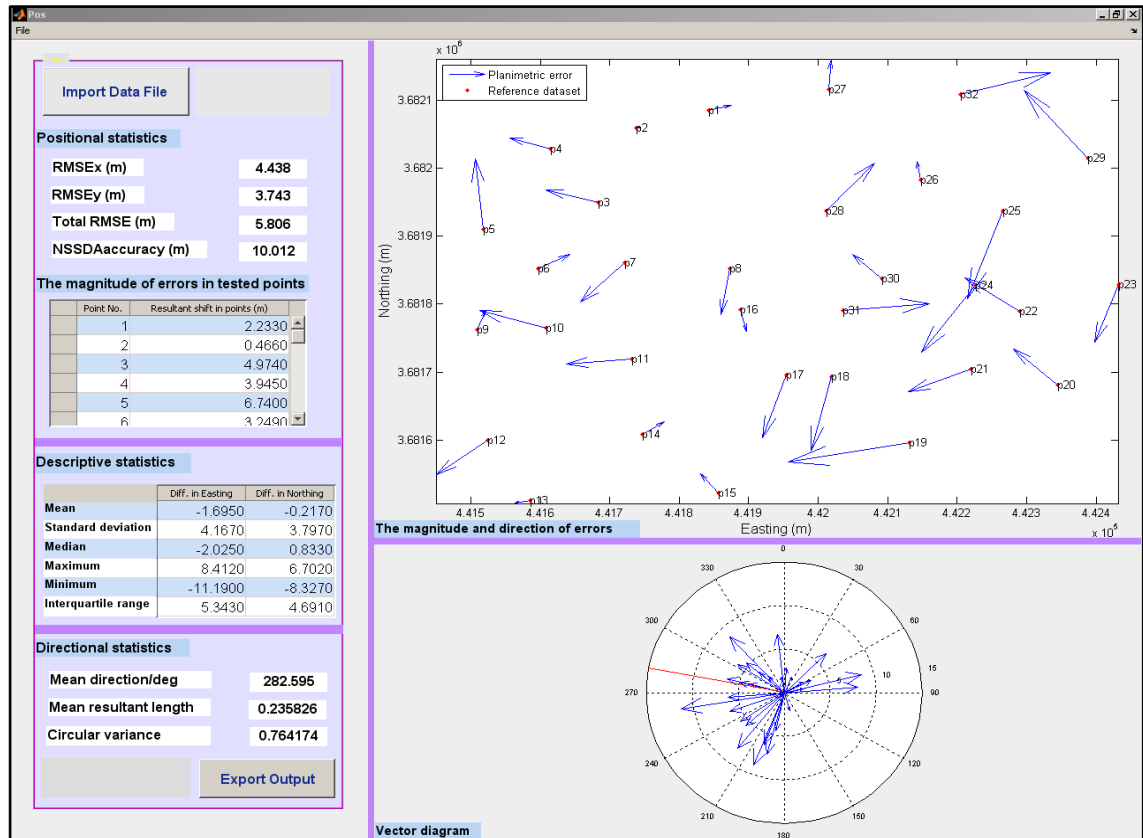


Figure 4.19 The positional similarity measurement results for the comparison of GDS-OSM datasets in Baghdad-Iraq

4.4 Geometrical quality modelling for linear entities

Although the evaluation of the possibility of positional integration from different spatial data sources has been considered and analysed in previous section, geographical phenomena formalised in GIS may be in a form such as linear features which is more complex than simple points. These are usually not straight segments and they may be complex or sinuous features. Linear phenomena can be natural or human made features such as roads, pathways, railways, coasts, streams, and canals. For the process of geospatial data integration, for example, matching of real world and digital representations of linear entities may be successful for some parts of a feature, but overall differences may occur in lengths or the smoothness of compared features. For this research, procedure for the evaluation of the similarity of linear features has been adopted which extends the point comparison already described. The variety of metrics used in linear similarity measurements will be described in the following subsections. The results of the comparison of linear features from formal and informal spatial data sources, and their comparison with respect to reference data, will also be included and

described. This has been undertaken by developing a code interface in order to test spatial datasets prior to successful geospatial data integration.

4.4.1 Epsilon band accuracy models

Linear accuracy can be examined as an indication of the shape or curvature similarity between two lines, as well as the positional accuracy of points along the line. When line features such as roads or railways are considered, comparison using point accuracy of locations on the line is insufficient to capture the geospatial complexity of linear features. While point measures may be straightforward to understand and calculate, they do not capture all aspects of line accuracy. Thus, there are several methods for modelling geometric errors in linear features. For instance, the Perkal epsilon band model is one commonly used method for analysing errors of cartographic line segments (Perkal, 1956, 1965). This technique involves creating a constant width buffer around a line when a circle of diameter epsilon (ϵ) is rolled along both sides of the line. An example of this method can be seen in Figure 4.20 in which the arc (x) is enclosed by a 2ϵ buffer area width and two semi-circular areas of radius ϵ . The model assumes that the true location of the cartographic line lies within the buffer band and it never deviates outside it.

The potential of using the epsilon band approach inspired several researchers in their measuring of the physical characteristics of linear features. For instance, Chrisman (1982) applied the epsilon band model to measure the error in the USGS data of the city of Pittsburgh. The conclusion was that the epsilon band method is an appropriate approach for measuring the error of all features on a map, including the complex shapes of linear features. In addition, Blakemore (1984) illustrated how the epsilon band theory can be used in order to analyse the accuracy of point-in-polygon procedures. In Blakemore's test, the band width was based on the error values existing in the digitized employment zones of a study area in North West England. The suggestion was that the points would be within the epsilon band of the polygon in four categories: possibly in, possibly out, definitely within, and definitely outside the polygon. These can explicitly be utilised to define the error in the location of a polygon boundary.

In another major study, Goodchild and Hunter (1997) established a buffer for describing and analysing the accuracy of linear features. Their method relies on buffering the reference datasets of higher accuracy only. The procedure determines the proportion of

the length of the feature of tested datasets that lies within the buffer of reference datasets (Figure 4.21). The model was tested on a sample of data from the Digital Chart of the World (DCW) dataset and the southern coastline of Victoria, near Melbourne, Australia, was selected as a study area. The approach was to increase the buffer width gradually and measure the percentage of the tested line that lies within the buffer area for each buffer size. The evidence from this study reported that this technique is simple and it can be applied for assessing linear matching to long and complex linear features. However, by using this method it is not possible to measure the significant relationships of compared datasets such as curvature similarity, a fundamental comparison of tested lines. The next subsection will introduce another approach for measuring further geometrical similarity characteristics between linear features. It is also based on a buffering technique, but follows a different procedure.

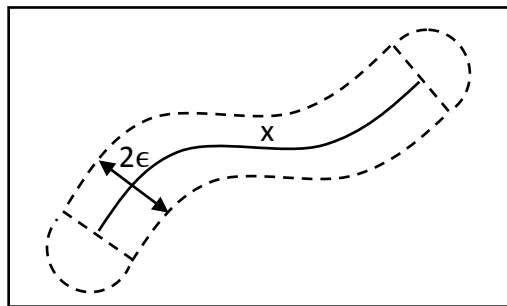


Figure 4.20 Perkal's epsilon band approach (Perkal, 1965)

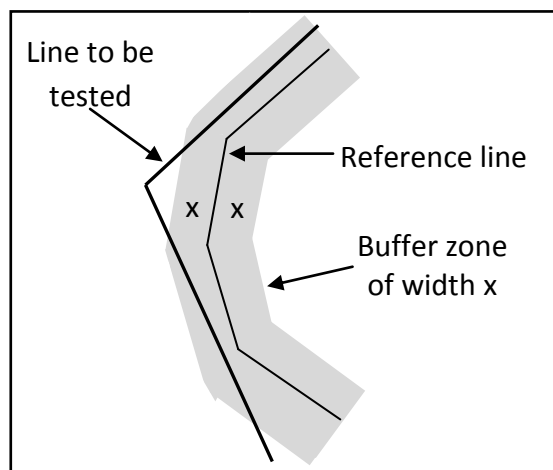


Figure 4.21 Goodchild and Hunter method (Goodchild and Hunter, 1997)

4.4.2 An alternative method for buffering overlay

In (1999) Tveite and Langaas proposed the buffer-overlay-statistics (BOS) method as an alternative approach to assessing geographical line datasets accuracy. Similar to the Goodchild and Hunter approach (1997), this method is based on a comparison of the

lines of unknown data quality to reference or higher quality datasets. However, Tveite and Langaas (1999) suggested generating a buffer around both tested and reference datasets to measure the degree of overlap of each line, rather than creating a buffer around reference datasets only, as was proposed by Goodchild and Hunter (1997). This method was implemented by Tveite and Langaas (1999) using different real-world datasets such as digital geographic data from the Norwegian Mapping Authority (NMA), DCW (Digital Chart of the World) and WVS (World Vector Shoreline). DCW and WVS were produced by the Defence Mapping Agency (DMA), USA. The results of their investigations showed that the procedure of the double buffer method is a suitable and effective means for measuring linear geometric accuracy. Other authors such as Gruber et al. (2008) have pointed out that this method can be useful for different applications. For instance, the double buffer technique was involved in the research workflow that was suggested by Gruber et al. (2008). Their investigation included creating a three-dimensional city model from aerial images and evaluation of the positional accuracy of the building footprints by applying the buffer overlay method.

The elements of this method are shown in Figure 4.22. Lines in each dataset are denominated as (X) for the unknown quality datasets, and (Q) for higher quality reference datasets. These lines get a number of buffers of various sizes (e.g. XB, QB). It is an iterative approach because it is impossible to estimate the appropriate buffer size in advance. The buffer size is iteratively increased with different matching accuracy results and a point is reached where increased buffer size gives limited improvements in the accuracy. After performing a buffer operation on each of the two lines, statistical calculations should be carried out. In order to apply the overlapping operations on XB and QB areas, it is necessary first determining the buffering area for the following situations (Tveite and Langaas, 1999):

$$\begin{aligned}
 &\text{Area inside XB and inside QB: } \text{Area}(XB \cap QB) && \text{(a)} \\
 &\text{Area inside XB and outside QB: } \text{Area}(XB \cap \overline{QB}) && \text{(b)} \\
 &\text{Area outside XB and inside QB: } \text{Area}(\overline{XB} \cap QB) && \text{(c)} \\
 &\text{Area inside XB or inside QB : } \text{Area}(XB \cup QB) && \text{(d)}
 \end{aligned}
 \tag{4-9}$$

The statistical calculations in the set of equations above can be used to obtain the measures of the deviation of unknown quality line from the line of known quality. Interpretation of the results is facilitated by presenting them graphically. This will be described in further detail in the next subsections. The comparison can be carried out by

using buffers to determine the area inside both buffers, the area inside X and outside Q and the area outside X and inside Q. In addition, the average displacement information between two lines can be calculated as shown in the following equation:

$$DE = \pi \times bs_i \frac{Area(\overline{XB}_i \cap \overline{QB}_i)}{Area(XB_i)} \quad (4-10)$$

Where:

bs_i : The width of buffering size.

The assessment of the linear matching process can be summarized as follows: identify tested and reference datasets and create buffer zones with different sizes for each dataset; then process the data by performing overlay and statistics operations as discussed in equations 4.9 and 4.10.; the final step can be represented by analyzing the results and making a decision regarding the ability of geospatial data integration. For the current research, in addition to previous calculations, the overlap percentage between datasets (calculated as the ratio between 4.9(a) and Area XB - the buffer from the unknown quality dataset) has been taken into account, as in equation (4.11). To perform such an operation, the creation of a professional tool is necessary. The use of the tool developed for this research is therefore illustrated in the next subsection.

$$Overlap\ percentage = \frac{Area(XB \cap QB)}{Area\ XB} \times 100 \quad (4-11)$$

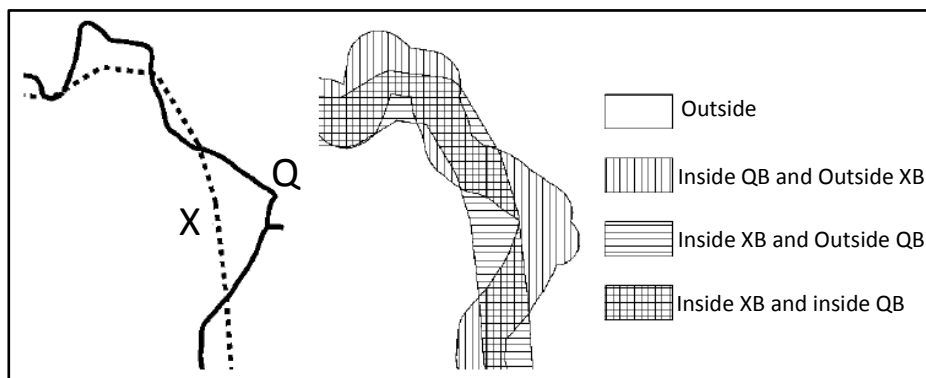


Figure 4.22 The BOS method elements (Tveite and Langaas, 1999)

4.4.3 Linear similarity analysis tool

It should be noted that this subsection is only concerned with developing a tool that can compare a linear geometrical aspect of a variety of spatial datasets such as reference field survey, authoritative and informal datasets. In order to understand how and why certain aspects of the linear similarity measurement tool were created, it is first

necessary to have some idea about the tool structure. It is essentially encoded within a Matlab environment, as documented in Appendix B and illustrated in Figures 4.23 and 4.24. Three distinct tasks can be identified within the developed linear similarity measurement interface code: input of data; analysis/statistics; and results output (i.e. numerical similarity graphs and data overlay drawings). The first stage is fairly straightforward, as input data can be imported as a text-based file of planimetric coordinates and buffer areas of compared datasets. During analysis, the methodology development described in the preceding subsection is undertaken in order to enable checking for linear accuracy. The interface also provides facilities to export the output results as a text-based report. All three of these tasks have been tried to create reusable and flexible solutions for evaluating the ability of multi-source linear geospatial data integration.

The developed interface in Figure 4.23 basically consists of three main parts. It is similar to the positional similarity assessment which has been shown in Figures 4.8 to 4.19. This essentially involved the numerical analysis of results, as displayed on the left side of Figure 4.23. The graphical analysis is also included in this interface, represented in the middle of the same figure. The drawing of the overlay of compared datasets occupies the right of the developing linear similarity measurement interface. The interface is also specifically designed to import data for linear similarity measurement processing and to export output results through two buttons at the top and the end of the numerical analysis part. The practical testing of this tool was tried and implemented for different study areas, as will be illustrated in the following paragraphs.

4.4.3.1 VGI and formal linear data integration assessment and output

The method described in subsection 4.4.2, has been implemented here for the comparison of benchmark field survey and formal data with VGI datasets, and between each other, for the Cramlington2-UK study area. This has been achieved for linear features such as roads and pathways, as explained in section 4.2. The buffer operations were performed using the ArcGIS10 software package, while the metrics of measures, statistics, graphs and overlay drawing were calculated and produced through the specific developed interface, as illustrated in the previous paragraphs. The method was applied for buffer sizes from 0.5m to 12.5m for comparisons involving OSM data, and for the comparison of FS with formal data the buffer size ranged from 0.2m to 5m with interval 0.2m. This can be observed in the graph plots of Figure 4.23.

As mentioned in subsection 4.4.2, the results of BOS can be visualised more easily by means of graphs. For the current study, the graphical representations of the results have included the overlap percentage and the average displacement (equations 4.10 and 4.11) between the compared datasets with respect to increasing buffer size, as shown in Figure 4.23. The horizontal axes of the plots correspond to the buffer size, and the vertical axes show the overlap percentage of the common buffer area for one set of calculations, and the average displacement values for the other set of calculations. The first three graphs represent the relationships between the overlap percentage values and the increased buffer width, while the fourth plot illustrates the average displacement comparison with increased buffer area. The analysis included the comparisons of FS/OSM, OS/OSM and FS/OS datasets respectively. Each overlap percentage chart displays three curves, buffer area inside both reference (Q) and tested (X) linear datasets, buffer area inside (Q) and outside (X) linear datasets, and buffer area outside (Q) and inside (X) linear datasets. On the other hand, the average displacement plot involves three curves for the comparisons of the FS/OSM, OS/OSM and FS/OS datasets. The numerical data calculations of all plots are illustrated in Tables (4.12 to 4.15). These tables represent the overlap percentage and the average displacement values for the corresponding buffer size iteration for each of the compared datasets.

From Figure 4.23, it can be observed that the overlap percentage graphs (blue lines) flatten out as the buffer size increases, revealing the overlap percentage of the pairs of lines. There is a significant difference in the buffer overlay results of OS and OSM with the OS data being closer to the reference FS dataset. For example, Table 4.12 shows that the overlap percentage for the comparison of FS/OS is nearly 96% when the buffer width is 5m; whereas Tables 4.13 and 4.14 show that the overlap percentage was about 84% when the buffer size was 12.5m for both FS/OSM and OS/OSM comparisons. The same figures and tables also illustrate the overlap percentage of the buffer areas inside the reference data and outside the compared data: it is approximately 2% when the buffer size is 4.2m for the comparison of FS/OS (where the graph flattens), while it is 9% for the comparisons of FS/OSM and OS/OSM when the buffer size is 12.5m. The same observations can be made for the overlap percentage values of the buffer area outside the reference data and inside the tested data (red lines).

On the question of the average displacement comparisons, this study found that the FS data is very close to OS datasets. Figure 4.23 and table 4.15 show that the average

displacement is about 0.29m for a buffer width of 1m after which the plot remains steady. The differences between FS/OSM and OS/OSM, by comparison, were approximately 3.35m for both of them when the buffer size was 12.5m. It is clear that the OSM data shows higher average displacement values than the formal datasets when compared with FS and OS datasets. The slope of the graphs can also be used for results analysis. The greater variability of the average displacement values are obtained with steeper graph gradients. The attaining of a flattening curve can confirm higher accuracy of the compared datasets. As can be seen from the figure below, the slope of the average displacement comparison of FS/OS is apparently steady with no sharp slope. On the other hand, the other comparisons show steep and variable measurement graphs.

The numerical analysis, illustrated in the left part of the same interface, also emphasised that there is a clearer divergence between the comparisons of FS, OS datasets with OSM data than between FS and OS datasets. The visual inspection from the overlay drawing at the right of Figure 4.23 also indicated that the OSM data does not match the reference or formal datasets. Therefore, a similar conclusion to that obtained from the comparison of point position can be drawn for the current part of the project. The linear features in the OS dataset are more convergent with reference FS data than the comparison of linear features in OSM data versus OS and FS datasets. Both the overlap percentages and the displacement calculations caution against obtaining successful integration of the benchmark dataset or the formal data with the OSM data; however the findings favour the possible integration of the formal data with the reference data.

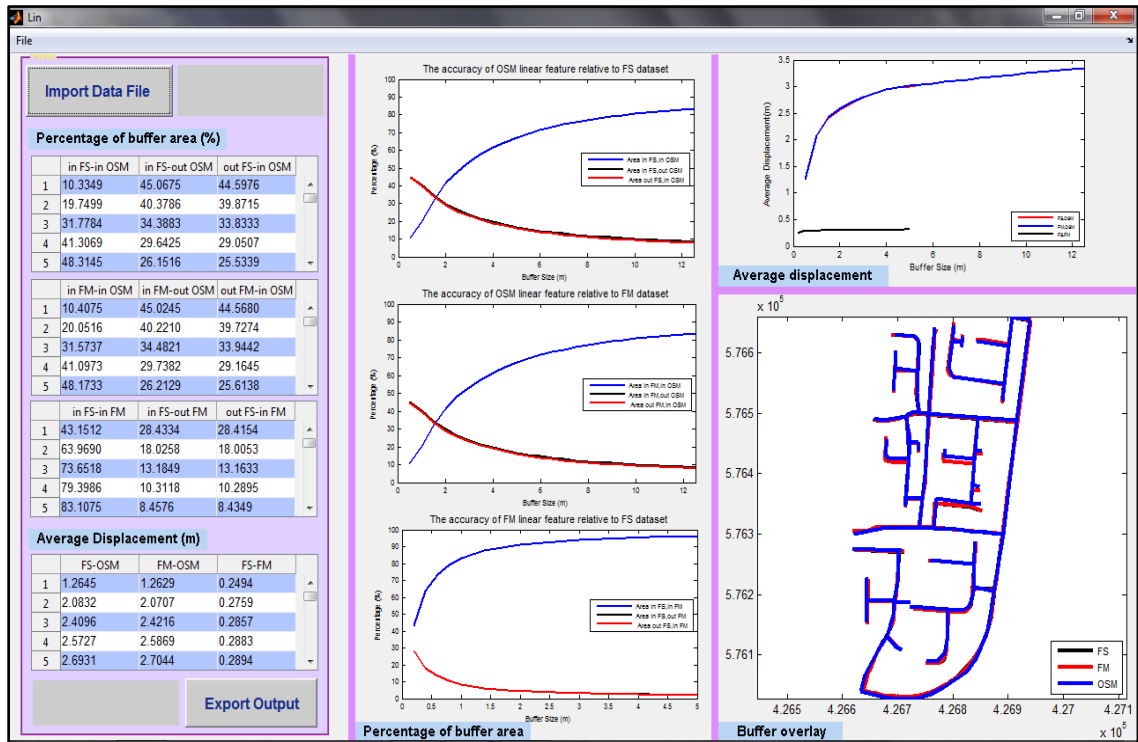


Figure 4.23 The results of the liner similarity measurement for the comparison of FS, OS and OSM in Cramlington2-UK

Table 4.12 The overlap percentage of buffering area between FS and OS datasets in Cramlington2-UK study area

Buffer size iteration (m)	The percentage of buffer area in FS-in OS (%)	The percentage of buffer area in FS-out OS (%)	The percentage of buffer area out FS-in OS (%)
0.2	43	28	28
0.6	74	13	13
1.0	83	9	8
1.4	88	6	6
1.8	90	5	5
2.2	92	4	4
2.6	93	4	4
3.0	94	3	3
3.4	95	3	3
3.8	95	3	3
4.2	95	2	2
4.6	96	2	2
5.0	96	2	2

Table 4.13 The overlap percentage of buffering area between FS and OSM datasets in Cramlington2-UK study area

Buffer size iteration (m)	The percentage of buffer area in FS-in OSM (%)	The percentage of buffer area in FS-out OSM (%)	The percentage of buffer area out FS-in OSM (%)
0.5	10	45	45
1.5	32	34	34
2.5	48	26	26
3.5	58	21	21
4.5	65	18	17
5.5	69	16	15
6.5	73	14	13
7.5	76	12	12
8.5	78	11	11
9.5	80	11	10
10.5	81	10	9
11.5	83	9	8
12.5	84	9	8

Table 4.14 The overlap percentage of buffering area between OS and OSM datasets in Cramlington2-UK study area

Buffer size iteration (m)	The percentage of buffer area in OS-in OSM (%)	The percentage of buffer area in OS-out OSM (%)	The percentage of buffer area out OS-in OSM (%)
0.5	10	45	45
1.5	32	35	34
2.5	48	26	26
3.5	58	21	21
4.5	65	18	17
5.5	69	16	15
6.5	73	14	13
7.5	76	12	12
8.5	78	11	11
9.5	80	10	10
10.5	81	10	9
11.5	83	9	8
12.5	84	9	8

Table 4.15 The average displacement values for the comparisons of FS, OS and OSM datasets in Cramlington2-UK study area

Buffer size iteration (m)	The average displacement between FS and OSM datasets (m)	The average displacement between OS and OSM datasets (m)	Buffer size iteration (m)	The average displacement between FS and OS datasets (m)
0.5	1.265	1.263	0.2	0.249
1.5	2.410	2.422	0.6	0.286
2.5	2.693	2.704	1.0	0.289
3.5	2.8766	2.883	1.4	0.292
4.5	2.979	2.983	1.8	0.294
5.5	3.035	3.038	2.2	0.296
6.5	3.086	3.089	2.6	0.298
7.5	3.134	3.136	3.0	0.300
8.5	3.181	3.182	3.4	0.302
9.5	3.225	3.226	3.8	0.304
10.5	3.269	3.269	4.2	0.306
11.5	3.311	3.311	4.6	0.308
12.5	3.352	3.351	5.0	0.310

Similarly, Figure 4.24 shows the linear similarity assessment for the Baghdad-Iraq study area, with the same buffer size and number of iterations that were applied for the Cramlington2-UK site. The quantitative calculations of all plots of Figure 4.24 are illustrated in Tables (4.16 to 4.19). It can be observed from Figure 4.24 and Table 4.16 that the percentage value of the buffering overlay was approximately 91% when the buffer size was 5m for the comparison of FS and GDS datasets. The same graph and table also illustrate the overlap percentage for the area of the buffer inside the reference data and outside the GDS data, as well as for outside the reference data and inside the GDS datasets. These were nearly 5% when the buffer size was 5m for each of the analyses. On the other hand, the other comparisons, the informal with the reference and the formal datasets, are quite revealing in several ways (Tables 4.17 and 4.18). Firstly, unlike the previous comparisons, the overlap percentages were about 80% when the buffer size was 12.5m for both the FS/OSM and GDS/OSM analyses. Secondly, the percentage of the buffer area inside the FS and GDS data and out of the OSM datasets was nearly 10% when the buffer size was 12.5m. The findings also revealed that the buffering area outside the FS and GDS data and inside the OSM datasets was approximately 9% when the buffer size was 12.5m.

From the same interface and table 4.19, the plots and values of the average displacement measurement illustrated that the slope of the graph for the comparison of the FS/GDS

was stable when the average displacement about 0.75m and the buffer size 2.5m, while it was approximately 3.70m when the buffer size was 12.5m for both FS/OSM and GDS/OSM analyses. The close proximity between the reference and the formal datasets on the one hand, and the mismatching between the OSM against FS and GDS on the other hand is also emphasised in the part of data overlay drawing of Figure 4.24. The same observations can be seen from the numerical calculations of the overlap percentages and the average displacement of the tested datasets in the left part of the same figure.

There are similarities between the data expressed by Figure 4.24 and those described in Figure 4.23. It is interesting to note that in both cases of this part of the study, UK and Iraq, there is a significant linear similarity between the reference and the formal datasets, while there are significant differences between the comparisons of the informal data against both the reference and the formal datasets. In order to include the variation of all kinds of feature, the evaluation of the ability of polygon shape integration from formal and informal data sources will also be included in this research, as will be demonstrated in the following section.

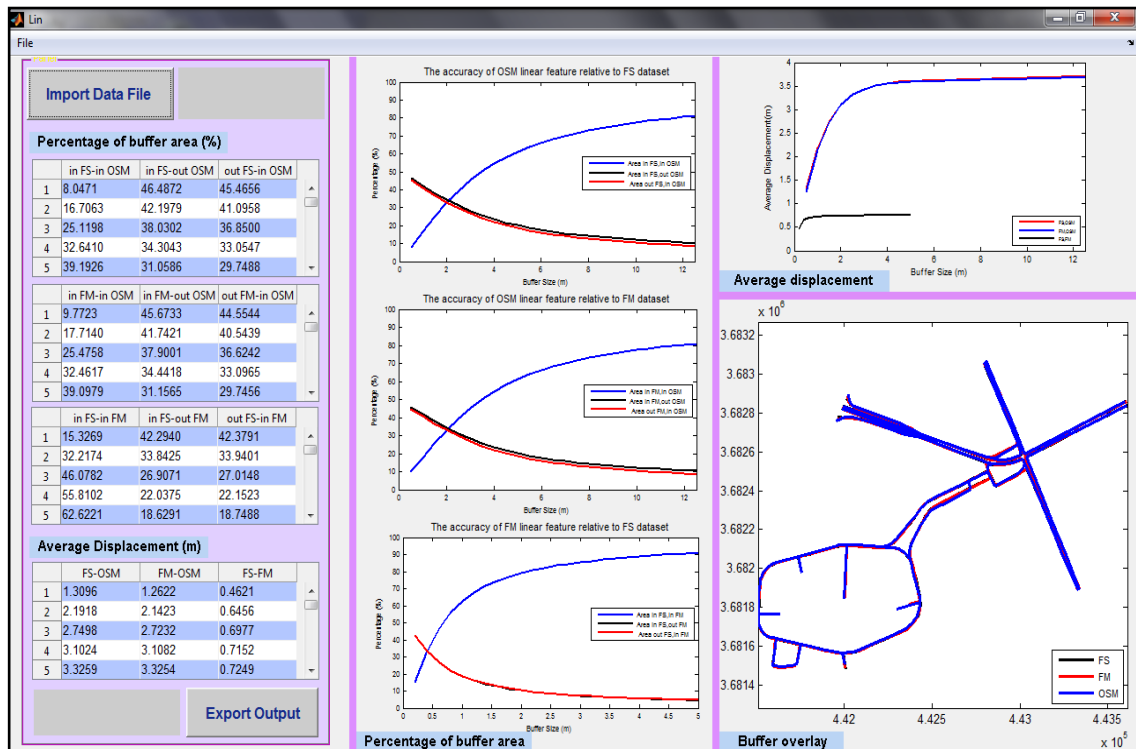


Figure 4.24 The results of the linear similarity measurement for the comparison of FS, GDS and OSM in Baghdad-Iraq

Table 4.16 The overlap percentage of buffering area between FS and GDS datasets in Baghdad-Iraq study area

Buffer size iteration (m)	The percentage of buffer area in FS-in GDS (%)	The percentage of buffer area in FS-out GDS (%)	The percentage of buffer area out FS-in GDS (%)
0.2	15	42	42
0.6	46	27	27
1.0	63	19	19
1.4	72	14	14
1.8	77	11	12
2.2	81	10	10
2.6	83	8	8
3.0	85	7	7
3.4	87	6	7
3.8	88	6	6
4.2	89	5	5
4.6	90	5	5
5.0	91	5	5

Table 4.17 The overlap percentage of buffering area between FS and OSM datasets in Baghdad-Iraq study area

Buffer size iteration (m)	The percentage of buffer area in FS-in OSM (%)	The percentage of buffer area in FS-out OSM (%)	The percentage of buffer area out FS-in OSM (%)
0.5	8	47	46
1.5	25	38	37
2.5	39	31	30
3.5	50	26	24
4.5	58	22	20
5.5	64	19	17
6.5	68	17	15
7.5	72	15	13
8.5	74	14	12
9.5	77	13	11
10.5	78	12	10
11.5	80	11	9
12.5	81	10	9

Table 4.18 The overlap percentage of buffering area between GDS and OSM datasets in Baghdad-Iraq study area

Buffer size iteration (m)	The percentage of buffer area in GDS-in OSM (%)	The percentage of buffer area in GDS-out OSM (%)	The percentage of buffer area out GDS-in OSM (%)
0.5	10	46	45
1.5	26	38	37
2.5	39	31	30
3.5	50	26	24
4.5	58	22	20
5.5	64	19	17
6.5	68	17	15
7.5	72	15	13
8.5	74	14	12
9.5	77	13	11
10.5	78	12	10
11.5	80	11	9
12.5	81	10	9

Table 4.19 The average displacement values for the comparisons of FS, GDS and OSM datasets in Baghdad-Iraq study area

Buffer size iteration (m)	The average displacement between FS and OSM datasets (m)	The average displacement between GDS and OSM datasets (m)	Buffer size iteration (m)	The average displacement between FS and GDS datasets (m)
0.5	1.310	1.262	0.2	0.462
1.5	2.750	2.723	0.6	0.698
2.5	3.326	3.325	1.0	0.725
3.5	3.510	3.505	1.4	0.735
4.5	3.588	3.576	1.8	0.738
5.5	3.612	3.596	2.2	0.741
6.5	3.626	3.610	2.6	0.744
7.5	3.639	3.622	3.0	0.747
8.5	3.652	3.635	3.4	0.750
9.5	3.665	3.647	3.8	0.753
10.5	3.677	3.660	4.2	0.757
11.5	3.690	3.672	4.6	0.760
12.5	3.702	3.683	5.0	0.763

4.5 Area shape similarity

Shape comparison among various geospatial objects can be considered one of the fundamental topics of the fields of both geographical investigations and multi-source geospatial data integration (Ali, 2002; Maceachren, 1985). For effective data integration processing, in addition to positional and linear similarity measurements, the representation, description and analysis of shape properties are also significant. In

general, there are two main properties for any planar shapes: perimeter and area, which can assist the basic descriptive analysis. In a discrete vector graphical shape, the perimeter of an object is the distance around the outside of the polygon, while the area of an object is simply the number of square units it takes to completely fill the enclosed object.

Many other shape descriptors have been studied and applied in practice. Some examples are: compactness, elongation, convexity and concavity (Stojmenović and Žunić, 2008; Esa et al., 2006; Ebdon, 1985; Austin, 1984). Of these, the compactness descriptor has been given the greatest attention due to its potential applicability to a wide range of geographical problems. The compactness index of an object (which is dimensionless) can be simply measured by dividing a shape perimeter squared into its area. A large variety of methods by which shape compactness can be applied have been presented in the literature; see for example, Montero and Bribiesca (2009), Bribiesca (2008), Bogaert et al. (2000), and Bribiesca (1997). There has already been a great deal of research into shape compactness measurements; for instance, Maceachren (1985) investigated the various methods for measuring compactness. He compared and evaluated the compactness indices, resulting in the identification of four compactness groups: perimeter-area measurement, single perimeters of related circle, direct comparison to a standard shape and dispersion of elements of a shape's area. All compactness indices were calculated for a sample of U.S. counties in order to identify the similarities and the differences among their shapes. Montero and Bribiesca (2009) carried out a number of investigations into the measurement of shape circularity and compactness. They compared different shapes under different conditions, such as shapes with holes or with noisy perimeters. The measure of compactness as an intrinsic property of an object is therefore invariant under geometric transformation such as rotation, translation and scaling.

Methods that use these dimensions to assess shape quality are not computationally complex, but do not tell much about shape, especially for irregular shapes (Ali, 2002). An alternative method for analyzing shapes involves moment invariant analysis. From the group of global scalar transform techniques, the moment methods can be considered the most popular. Moments were first used for mechanics purposes rather than shape descriptors. Hu (1962) was the first to set out the mathematical foundation for two

dimensional moment invariants and demonstrated their application to shape recognition. He proved that a proper combination of moments can provide translation, scale, and rotation invariant quantities. Moment and functions of moment have been used as a pattern in a number of image processing applications for recognition and classification. This technique was first applied to aircraft shapes and was shown to be quick and reliable (Dudani et al., 1977). In general, there are two procedures that can be followed for calculating moments, for areas of raster data or for boundaries of polygon vector data, as will be discussed in the following.

4.5.1 Area moments invariant

In this subsection, a brief review of Hu's invariant moment is presented. The two dimensional moments invariant are computed based on the information provided by the shape interior region. These traditional invariants need the coordinates of all the bidimensional object pixels in image space in order to be computed. The geometric moments of order $p+q$ with the basis set $(x^p y^q)$ can be defined as:

$$m_{pq} = \iint_N x^p y^q f(x, y) dx dy \quad (4-12)$$

For $p, q = 0, 1, 2, \dots$

m_{pq} is the two dimensional moment of the function $f(x, y)$. The order of the moment is $(p+q)$ where p and q are both natural numbers.

When the geometrical moment in equation (4.12) is referred to the object centroid (x_0, y_0) , it becomes the central moment, and it would be expressed by the equation:

$$\mu_{pq} = \iint_N (x - x_0)^p (y - y_0)^q f(x, y) dx dy \quad (4-13)$$

The coordinates of the centre of gravity of the shape can be calculated as follows:

$$x_0 = m_{10}/m_{00}, \text{ and } y_0 = m_{01}/m_{00}$$

The central moments μ_{pq} are invariant to translation and may be normalized to be also invariant to an area scaling change by the following formula. The quantities in equation (4.14) are called normalized central moments.

$$\partial_{pq} = \mu_{pq} / \mu_{00}^{\beta} \quad (4-14)$$

Where $\beta = ((p + q)/2) + 1$, for $p+q=2, 3, 4, \dots$

From these normalised central moments, Hu (1962) developed a set of seven compound spatial moments. They were invariant to translation, rotation and scale change. The set of seven moments can be used for a simple pattern recognition experiment to successfully identify various types of characters. The Hu invariance moment is constituted of order 2 and 3 normalized central moments as demonstrated in the following equation:

$$\begin{aligned} \Omega_1 &= \partial_{20} + \partial_{02} \\ \Omega_2 &= (\partial_{20} - \partial_{02})^2 + 4\partial_{11}^2 \\ \Omega_3 &= (\partial_{30} - 3\partial_{12})^2 + (\partial_{03} - 3\partial_{21})^2 \\ \Omega_4 &= (\partial_{30} + \partial_{12})^2 + (\partial_{03} + \partial_{21})^2 \\ \Omega_5 &= (\partial_{30} - 3\partial_{12})(\partial_{30} + \partial_{12})[(\partial_{30} + \partial_{12})^2 - 3(\partial_{21} + \partial_{03})^2] \\ &\quad + (3\partial_{21} - \partial_{03})(\partial_{21} + \partial_{03}) \times [3(\partial_{30} + \partial_{12})^2 - (\partial_{21} + \partial_{03})^2] \\ \Omega_6 &= (\partial_{20} - \partial_{02})[(\partial_{30} + \partial_{12})^2 - (\partial_{21} + \partial_{03})^2] \\ &\quad + 4\partial_{11}(\partial_{30} + \partial_{12})(\partial_{21} + \partial_{03}) \\ \Omega_7 &= (3\partial_{21} - \partial_{03})(\partial_{30} + \partial_{12})[(\partial_{30} + \partial_{12})^2 - 3(\partial_{21} + \partial_{03})^2] \\ &\quad + (3\partial_{12} - \partial_{30})(\partial_{21} + \partial_{03}) \times [3(\partial_{30} + \partial_{12})^2 - (\partial_{21} + \partial_{03})^2] \end{aligned} \quad (4-15)$$

Hu moments defined in equation (4.15) can be expressed as follows (Noh and Rhee, 2005):

Ω_1 : The sum of horizontal and vertical directed variance, more distributed towards horizontal and vertical axes, the values are enlarged.

Ω_2 : The covariance value of vertical and horizontal axes when the variance intensity of vertical axis and horizontal axis are similar.

Ω_3 : The result emphasizing the values inclined to left/right and upper/lower axes.

Ω_4 : The result emphasizing the values counterbalancing to left/right and upper/lower axes.

$\Omega_5, \Omega_6, \Omega_7$: The extraction of values invariant against size, rotation and location.

The values of these seven invariant moments (Ω_i), $1 \leq i \leq 7$ can be computed over the shape boundary associated with its interior part.

4.5.2 Line (improved) moment invariant

After Hu, several studies have revealed different methods to compute moments invariant. In 1993, Chen published a paper in which he introduced a convenient procedure to calculate the moment invariant emphasize the object boundary. These moments are called improved moments invariant and are a reformation of Hu's moments. They consist of a set of invariant functions devised in such a way as to be evaluated on the shape boundary only. The recognition of objects using moments of outlines is also possible and may lead to some simplification in computation when compared to the raster or area moments (Joo, 2005). In this case, the one dimensional moment of order $(p+q)$ over a general line is defined by the following equation:

$$m_{pq} = \int_C x^p y^q dl \quad (4-16)$$

For $p, q = 0, 1, 2, 3, \dots$

Where:

\int_C is a line integral along the curve C

$$dl = \sqrt{(dx)^2 + (dy)^2}$$

The central moments can be defined similarly as in Hu's model as follows:

$$\mu_{pq} = \int_C (x - x_0)^p (y - y_0)^q dl \quad (4-17)$$

Where:

$$x_0 = m_{10}/m_{00}, \text{ and } y_0 = m_{01}/m_{00}$$

The central moment is thus similar to that of area central moment, except that here there is only a single integral. The integral must be evaluated along the edge of the object. It is obvious that the modified central moments are invariant to translation. These new central moments can also be normalised so that they are invariant to change of scales. Chen (1993) used the same invariant functions given by equation (4.14), except that a new scaling factor α instead of β was introduced. Thus Chen's scale normalised central moments are given by:

$$\partial'_{pq} = \mu_{pq} / \mu_{00}^{\alpha} \quad (4-18)$$

Where:

$$\alpha = p + q + 1, \text{ for } p+q= 2, 3, 4, \dots\dots$$

The seven moments invariant values can then be calculated in a similar way to the area moments invariant situation by using the set of equations (4.15), except equation (4.18) should be used rather than equation (4.14). For the study here, this method has been adopted for the purpose of measuring shape similarity. Section 4.2 reported that the data used for this project consists of vector datasets. In any GI applications, the use of vector data can only supply information about polygon boundaries. For this reason, the improved moment invariant approach was chosen because this method is one of the more practical ways of calculating moments along shape outlines, as discussed earlier.

4.5.3 Distances measurements

After calculating the ordered sets of moments, shape quality assessment can be achieved through the moments' vector space model. This model can be used to measure the difference between the space distances of the moments of compared datasets. For instance, for two tested datasets, there are seven moments that can be calculated for each dataset and which can be denoted as (moments Ω_i^M) and (moments Ω_i^N), for $i = 1, 2, \dots, 7$. To calculate the space distance between Ω_i^M and Ω_i^N , there are different methods that can be applied such as Euclidean, Murkowski and Quadratic. In this research, Euclidean distance is adopted to calculate the space distance between the two sets of moments, as was recommended by Ali (2002). This can be shown as follows:

$$d_s(M, N) = \sqrt{\sum_{i=1}^7 (\Omega_i^M - \Omega_i^N)^2} \quad (4-19)$$

Where:

M and N : are the two compared datasets

Ω_i^M and Ω_i^N : are the sets of 7-moments invariant for the reference and compared datasets

As mentioned in preceding sections, the necessary condition to integrate two polygons from two different datasets is the shape similarity between them. The method above was used in order to determine degree of shape similarity. Valid shape integration between any two spatial datasets is found if the Euclidean distance between them is comparatively small, otherwise they can be considered different.

4.5.4 Polygon shape similarity analysis tool

To achieve this task, the model improved moments invariant were computed using a similar procedure to that applied for points and lines data. It was developed by the author using Matlab and the documented code is illustrated in Appendix B. The structure of the design of the shape similarity interface is similar to those of positional and linear similarity measurements. It basically consists of three parts, as exemplified in Figure 4.25. These involve the numerical analysis part for the quantitative values of the seven moments invariant of compared datasets, which is represented on the left section. The middle part of the interface includes a bar chart for the comparisons of the differences among seven moments of tested datasets. The differences have been calculated using the technique described in subsection 4.5.3. The third part of the interface involves the overlay drawing for all compared datasets. This can be used for visual interpretation of the possibility of formal and informal geospatial data integration.

The FS, OS, GDS and OSM polygon datasets were utilized to generate the necessary information for the subsequent vector (improved) moment invariant computations. Three data properties (E, N and ID) of a tested polygon should be extracted to files in order to calculate moments invariant. The data files can be stored in any appropriate format such as .txt or .xlsx for later use at the polygon shape similarity measurement interface. Note that the data files should saved into individual files for each of the datasets. After preparing the tested datasets, they can be imported into the developed interface by a specific button at the top of the interface. Once the files of compared datasets have been chosen, the shape similarity analysis processing will begin. The process of data analysis initially includes applying equation 4.17 in order to calculate the central moment. After that, the normalized central moment is computed using equation 4.18. The associated 7-moments can then be calculated through the set of equations (4.15). The last step is determining the Euclidean distance between the

compared samples in order to decide whether there is a similarity between them or not, as in equation 4.19. The output results can be saved in any format by clicking a special button developed for this purpose at the end of the interface. The shape similarity measurement tool, which has been developed here, was tested on various data sources and different study areas, as will be described in the following paragraphs.

4.5.4.1 Computing shape similarity of authoritative and VGI datasets

The goal of the initial test is to evaluate the possibility of polygon shape integration of FS, OS and OSM datasets in urban areas such as the Cramlington1-UK site. Figure 4.25 illustrates the results produced by the application of the seven moments invariant technique to assess the shape similarity for the compared datasets. Most of the features that are used for this experiment are common buildings which have relatively large areas, such as shops, restaurants, public library and police station. The left part of Figure 4.25 showed the mean values of the seven moments invariant for the polygons of each tested datasets: OS, FS and OSM respectively. In order to identify the similarity between any two datasets, the variations between their moments invariant list should be considered. The values showed that there is a significant similarity between most of the moment values of OS and FS datasets. Although the moment values of OS and FS data are not exactly alike, the obvious convergence between them indicates that a similarity between them does exist. On the other hand, the same tables revealed that there is a numerical separation between the moment values for the comparisons of OS and FS data versus OSM information. This indicates that there is a significant dissimilarity among those datasets.

On examining the bar chart in the middle part of the same figure, it can be observed that the respective Euclidean distances in 'moment's space' for OS/FS comparison is less than the distance between the mean values of the OS/OSM measurement. This is also true for the comparison of the space distance between the mean values of the 7-moments of OS/FS and FS/OSM. These measures reinforce the results obtained from the numerical analysis of the moment values. The right part of Figure 4.25 shows that the visual representation of OS data is very close to the FS datasets, while there is a considerable mismatching between both the reference and the formal datasets against informal data.

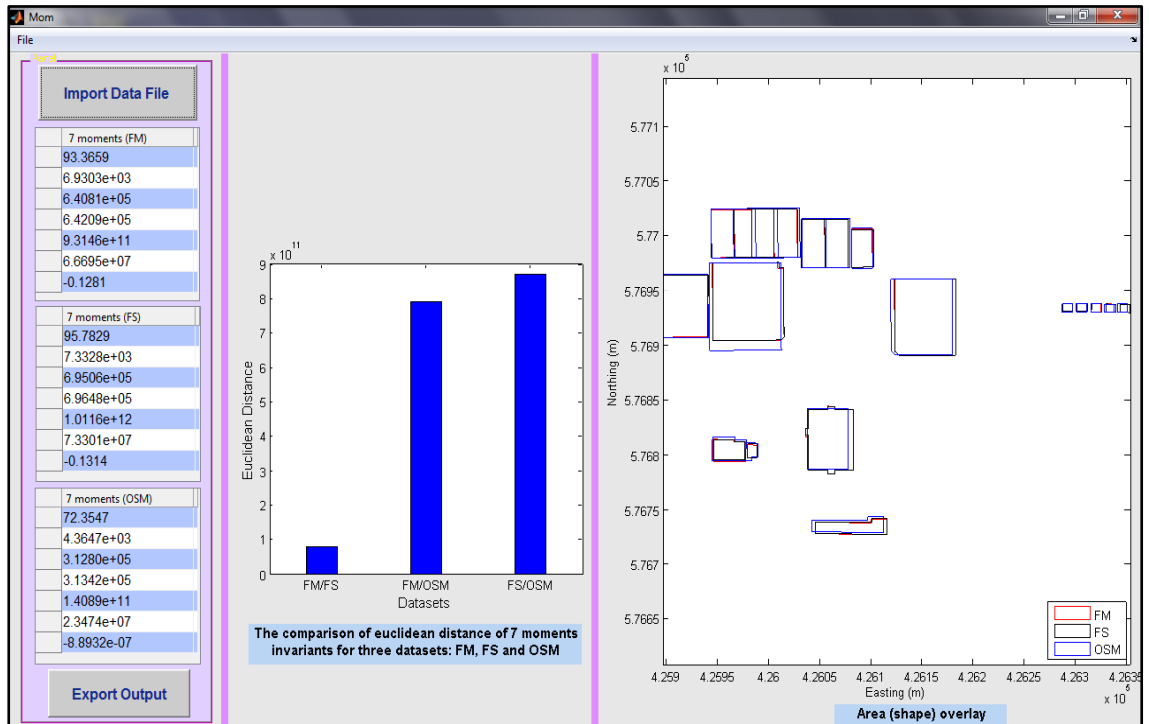


Figure 4.25 The interface of the output results of area shape similarity measurement for three datasets (FS, OS and OSM) in Cramlington1-UK

This part of the experiment was performed on three subsets of building (houses) features in an urban area. Again the same three data sources FS, OS and OSM were used, but in a different study area, Cramlington2-UK. All measures and analysis are displayed in Figure 4.26. In order to control the shape quality measurement of the compared datasets, it is helpful to look at the values of the mean of the 7-moments invariant of each of the datasets, shown on the left part of Figure 4.26. Data from this figure can be compared with the data in Figure 4.25 (previous test) which shows similarities in several ways. It is apparent from this analysis that the mean values of the moments for the OS and FS data demonstrate some similarities, whereas they are completely different to the OSM values. A distance bar chart between the mean values of the computed moments is also included in Figure 4.26. The bar chart of the distances shows that in the case of the OS/FS comparison, the space distance value is less than those in the other two situations. These findings were also confirmed by the overlay drawing which reflects clear discrepancies between both reference and formal data compared to OSM information and small differences between the reference and the formal datasets overlaying.

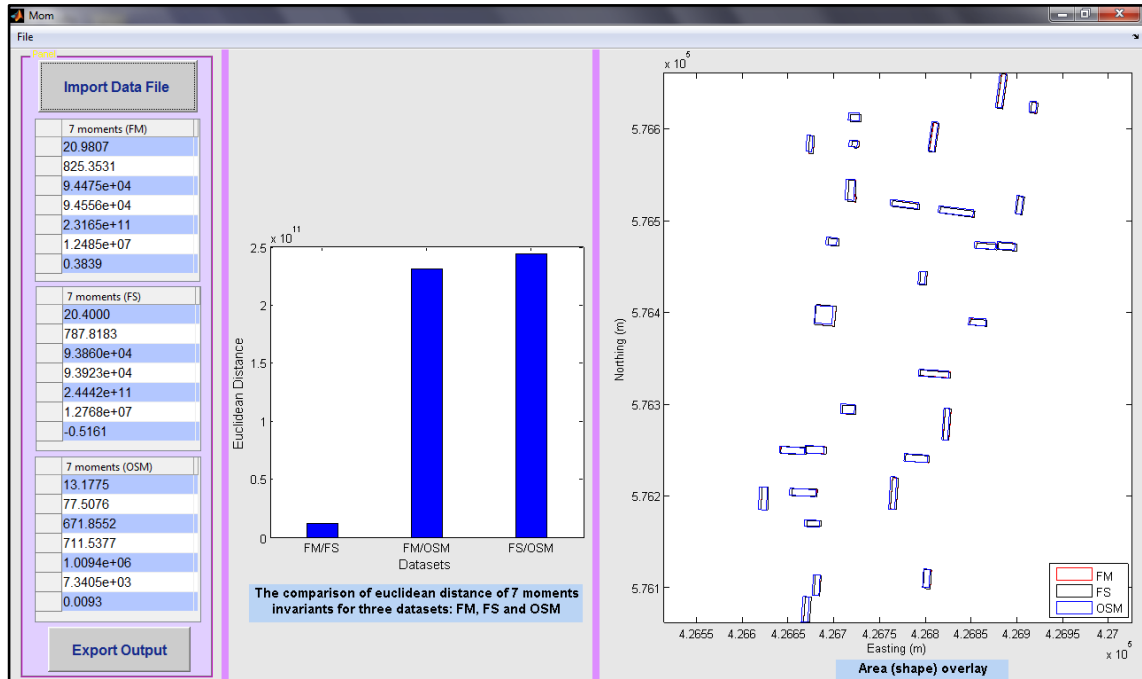


Figure 4.26 The interface of the output results of area shape similarity measurement for three datasets (FS, OS and OSM) in Cramlington2-UK

As mentioned in section 4.2, this study also involved testing of the ability of geospatial data integration in a rural area in order to include natural 'soft' features in the research flowline. The Clara Vale-UK site was selected as a testing area for the measuring of shape similarity among three datasets: FS, OS and OSM. Figure 4.27 contains the numerical and graphical results for this site. Although the figure shows that there is some separation between the reference distance and formal datasets in Clara Vale, it is clear that the difference between the values of FS and OS datasets is less than that between OSM and the other datasets. Comparison of the space distances values bar chart revealed that the mean of the 7-moments of the FS and OS datasets was also more homogenous (with a smaller difference) than the values between the FS and OS against OSM data. From the same figure, it can be seen that there are some discrepancies between the overlaid data from the reference and the formal data sources. However, these differences are relatively small compared to those between the OSM data and both FS and OS datasets.

In general, the analysis found that the accuracy of the polygons of OS formal data is very close to the reference FS datasets. However, the shape accuracy of OSM data does not match the reference or formal datasets. This is especially true for areas of 'hard' detail. This is probably because in the urban areas more regular shaped polygons were used, such as buildings and houses, while irregular complex shapes, such as woods and

golf courses, were characteristic of the rural area. The less similarity of the rural dataset may be explained by the shortcoming of the spatial information coverage resulting in fuzzy areas, such as the extent of woodland boundaries.

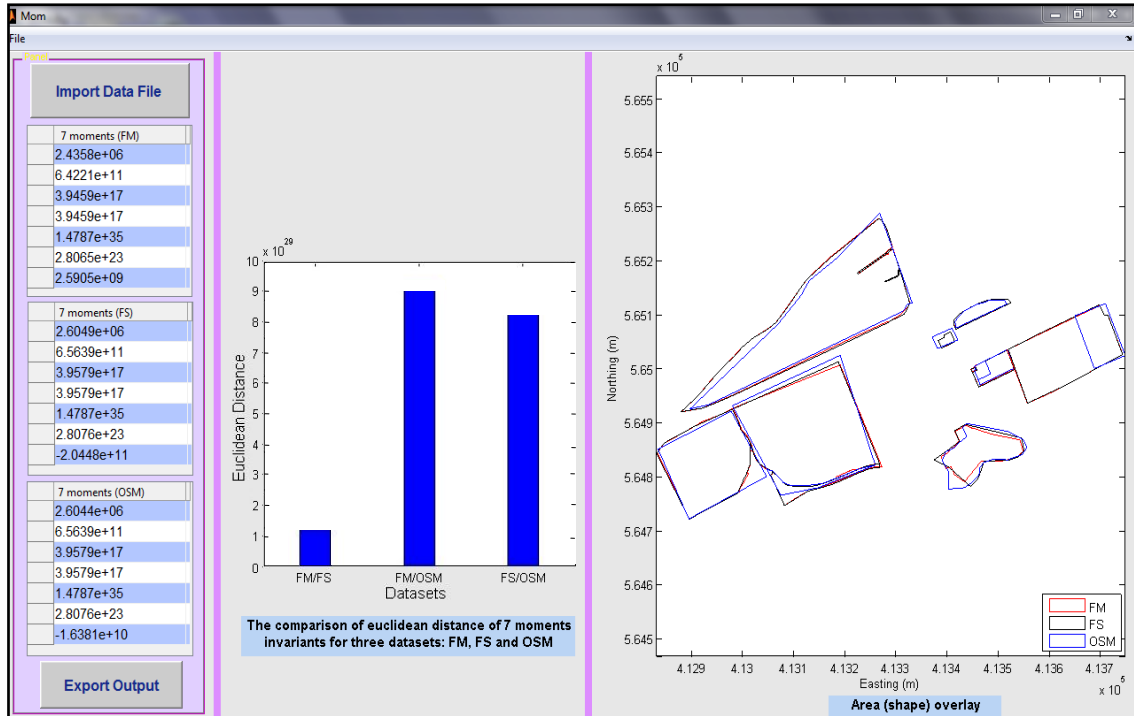


Figure 4.27 The interface of the output results of area shape similarity measurement for three datasets (FS, OS and OSM) in Clara Vale-UK

In the case of the Baghdad-Iraq site, most of the tested data were car parks and public buildings such as university or college buildings. The tested features were intended to be large area features. Figure 4.28 contains tables, graphs and diagrams showing the mean of the seven moments invariant for each dataset, the space distance between the mean of the seven moments invariant between each of the two datasets and the overlay datasets drawing. The conclusion with reference to the shape data in this study area is similar to that obtained from the shape similarity comparison in urban-UK study areas. The outlines of polygon features in the OSM datasets is less convergent with the reference and formal datasets than the polygon features in the reference and formal datasets are. This finding, while preliminary, suggests that it would not be an easy task to integrate the shape polygon features extracted from OSM datasets into the benchmark FS or authoritative GDS datasets, whereas it would be possible for the features sampled from the FS and GDS datasets. It is therefore likely that such connections exist between the conclusion of the shape similarity analysis and those obtained for both the positional and the linear similarity measurements.

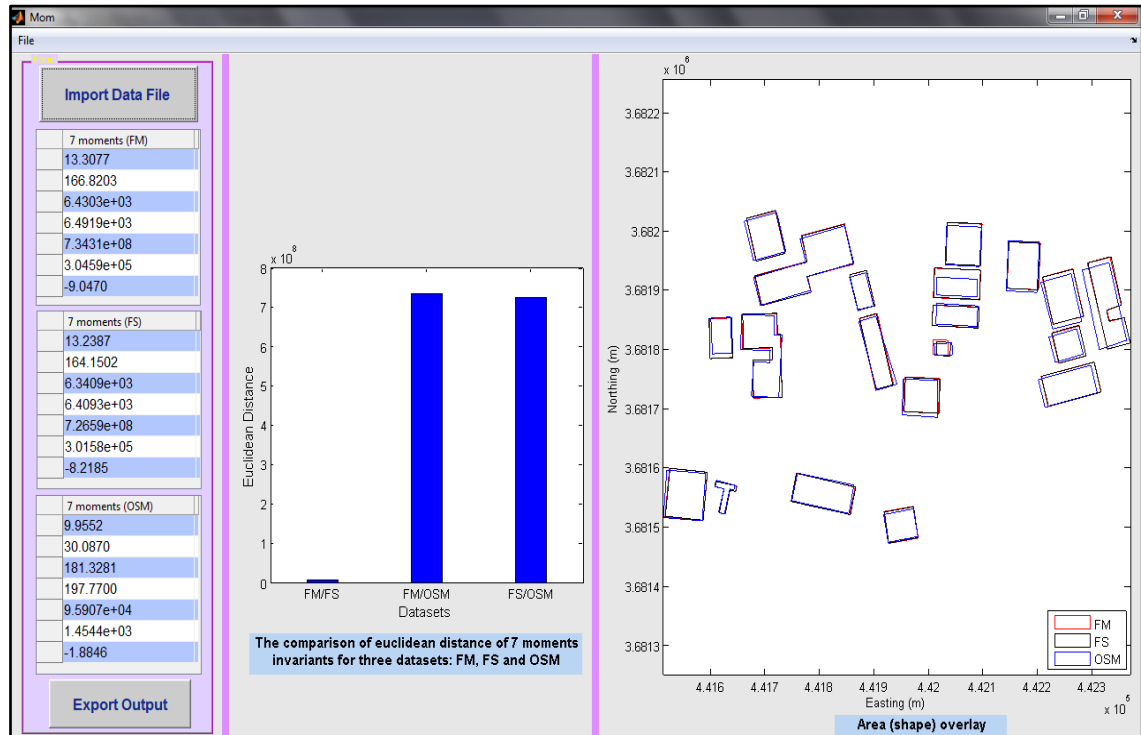


Figure 4.28 The interface of the output results of area shape similarity measurement for three datasets (FS, GDS and OSM) in Baghdad-Iraq

4.6 Chapter summary

This chapter has initially introduced the study areas. Three sites have been selected and tested; the one site in Cramlington and the other in Clara Vale are located in Northumberland, UK. The third area was chosen outside UK in the middle of Baghdad, Iraq. Different kinds of features and spatial data sources are presented in the case studies, which meant that the areas were ideal study sites for this research. Natural 'soft' and man-made 'hard' features for different data sources such as FS, OS, GDS and OSM were analysed and compared. These spatial data variations supported the study and were related to possible outcomes. The sampling scheme for field survey data acquisition, using different instruments and techniques, was also described.

Additionally, the methodology for assessing the possibility of geometrical integration from multi-sources of spatial datasets has been developed. The growth of VGI on the web has introduced new technologies and challenges for GI communities and organizations. The problems concerning integration between VGI and authoritative datasets to develop SDI, for example, are considered in this project as being the most vital issues regarding VGI datasets. Several mathematical bases of geometrical similarity measurement methods have been presented and discussed, and the advantages

and the limitations of each one highlighted. The National Standard for Spatial Data Accuracy (NSSDA) has been developed for such positional accuracy assessment. Positional accuracy also included circular observations assessment using a directional statistics technique. Shape accuracy has been examined in a study of the similarity of the curvature of linear features or the boundary of polygons. In research on line accuracy measurements, the double buffering method and an assessment of buffer overlay has been adopted. For area shape similarity measurement, shape metrics involving moments invariant such as Hu's (1962) invariant moments, and Chen's (1993) improved moment invariant have been applied.

The chapter also investigated how developing interfaces code could be implemented in order to follow the geometrical similarity measurement methodology outlined within the previous sections. Primarily, the structures of the similarity tools that would supply the most powerful and flexible programs were considered. All were encoded and developed within the Matlab environment. Each code consists of three main sections : input data, calculations and analysis, and outcome. The interface was designed so that the outcomes would be presented in three parts: numerical analysis, graphical representation and visualisation of overlay data.

This project found that in general there is a significant geometrical similarity between the formal data (OS and GDS) and the reference field survey datasets. This is particularly so for the hard details (urban) sites. However, the positional and shape accuracy of OSM data did not match the formal or the reference datasets. It is therefore likely that such shortcomings exist in the OSM data when investigated from a geometrical similarity perspective. The examinations of the differences among the reference, formal and OSM datasets were initially to determine the possibility of their data integration. For the large-scale data used in this project, the discrepancies in the comparisons indicated that such geometrical integration can be achieved for the reference and the formal datasets, but it would be difficult to match the reference and the OSM data or the formal and the OSM data. This is the major outcome of this part of the research.

Further investigations and analyses on the factors that may affect the geometrical quality of VGI datasets, and the reasons for the discrepancies between the formal and informal datasets, are described in the next chapter.

Chapter 5 Factors Affecting Geometrical Integration of OSM Information with Official Datasets

5.1 Introduction

Chapter 4 mentioned different approaches and techniques that can be used to assess the geometrical similarity matching between formal and informal datasets. This included the development of tools for the purpose of evaluating positional, linear and area shape similarity among reference field survey (FS) data, authoritative data, such as Ordnance Survey (OS) / UK data and General Directorate for Survey (GDS) / Iraq data, as well as Volunteered Geographic Information (VGI) data, such as OpenStreetMap (OSM) information, with the intention of assessing their possible integration. However, the problem of heterogeneity between formal and informal datasets emerged.

The main focus of this chapter will be on experimental analysis of factors that may affect OSM geometrical data quality and consequently affect successful integration with formal datasets. Three main factors have been selected for study in this project: data source, feature type and individuals. Factorial design studies were undertaken in order to develop and implement an experiment to perform such analysis. By examining various factors, factorial design can identify the factor that has the most effect on OSM geometrical data quality, in addition to determining the interaction between factors. This experiment was based on two different spatial data sources: the FS bench mark dataset and OSM information. The properties of FS datasets were introduced in Chapter 4, while OSM information was explored in detail in Chapter 3.

The chapter begins by introducing the factorial design approach. This essentially involves a full detailed description showing how factorial design can quantitatively estimate the effect of factors which can have influence on spatial data quality. The mathematical equations of factorial design, which involve examining combinations of factors, have also been included and described. In order to perform a successful experiment by applying the factorial design technique, it is necessary to follow certain systematic steps which are also described in this chapter. Subsequently, the preparation of datasets and the reasons for the importance of factors selected in this study are also discussed. The analysis of the experimental results with their discussion can be found before the conclusion section of the chapter.

5.2 Factorial design

5.2.1 An overview

The statistical significance of experimental design was originally investigated by Fisher in the 1920s and early part of the 1930s. Fisher found that there were difficulties in obtaining effective analysis of the data that were generated from an agricultural experiment system. Thus he introduced the principles of experimental design which include the concepts of factorial design and analysis of variance (Fisher and Mackenzie, 1923), as cited in Box (1980). Although Fisher's analysis can be considered as the pioneering work in the development of experimental design, there are many other researchers who contributed to developing and expanding the use of experimental design; see for example, Taguchi (1991), Kackar (1985) and Box and Wilson (1951).

Factorial design is the most efficient test to study the influence of different factors on the response variable. They are widely applied in experiments where there is a necessity to study the effect of several factors on the response variable. One of the most important parts of factorial design is the specific factors themselves which will be denoted here with index j . In most cases, each of these factors has two levels. The levels of experimental factors may be either numerical (quantitative) variables such as pressure, speed or time, or categorical (qualitative) such as 'low' and 'high' of the levels of the factor. The full combination of various factors and levels is known as a full factorial experiment. For example, with j factors each has two levels, the observations of the design would be represented as follows (Ryan, 2007):

$$\underbrace{2 \times 2 \times \dots \times 2}_{j} = 2^j \quad (5-1)$$

This is basically called 2^j factorial design. As the example in above equation has only two levels, the smallest number of experimental runs would be obtained. For instance, if there are two factors with two levels, the design will be 2^2 which generate 4 runs. In general, factorial design can be used to examine the effect of each factor on the response variable separately, which is usually called the main effect. It is defined as the effect of the factor alone averaged across the levels of other factors. Factorial design also allows looking at the joint (or interaction) between different factors that may affect the

response variable. The interaction is the differences among the variations between the means of different levels of one factor over different levels of the other factor. In this research, a 2^3 factorial design has been adopted, as three factors each with two levels that may affect the geometrical quality of OSM data have been selected. Further explanations and the practical results of the main effect of each factor and interaction among factors can be seen in section 5.6.

5.2.2 Estimation of the effects in the 2^3 factorial design

As discussed in the previous section, factorial design is characterised by a combination of a certain number of levels across factors of interest. For instance, if an experiment involves the effect of three factors, each with two levels; thus it would be useful to consider a two-levels-three-factors factorial design (2^3 factorial design) to run and analyse this class of experiments. Assume, for example, that there are three factors known as *A*, *B* and *C* and the two levels were denoted arbitrarily as 'low -' and 'high +' for each factor. As a result, eight treatments would be obtained from the combination of all factors and levels (Montgomery, 2001). These treatment combinations can be represented geometrically as a cube, as shown in the following Figure:

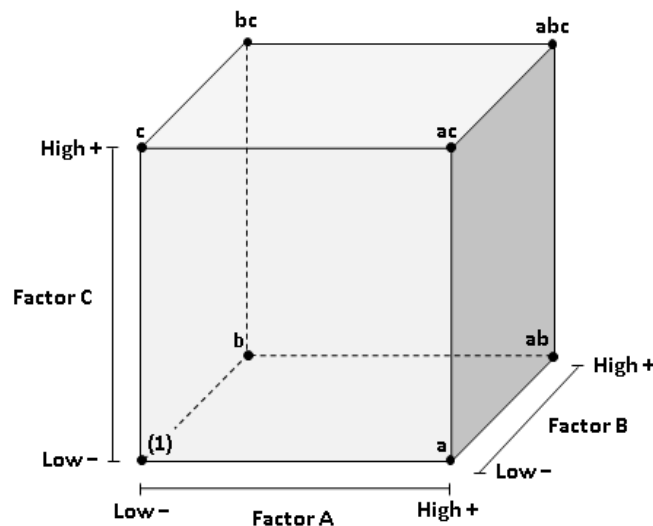


Figure 5.1 Graphical views of 2^3 factorial design (Montgomery, 2001)

It can be seen from Figure 5.1 that the effect factors have been denoted with a capital letter (*A*, *B*, *C*), whereas the eight treatment combinations have been represented by lowercase letters ((1), *a*, *b*, *c*, *ab*, *ac*, *bc*, *abc*). It can also be observed from the figure that the lowercase letters represent the high level of corresponding factors in treatment

combinations. In contrast, the lack of lowercase letters means low levels of factors. For example, a includes the treatment combination of factor A at high level, factor B at low level and factor C at low level. Another example can be represented by the treatment combinations at b which involves the processing of factor A at low level, factor B at high level and factor C at low level. On the other hand, the mixing of more than one lowercase letter represents the high level of equivalent factors. For instance, the treatment combinations at ab include high levels for both factors A and B and low level for factor C . While the processing at abc represents the treatment of the three factors A , B and C at high levels. From the figure, it can also be seen that (1) combination denotes the low levels of all factors A , B and C . The eight treatments can be scheduled in a standard order which is generally called a design matrix, as illustrated in Table 5.1.

Table 5.1 Design matrix of 2^3 factorial design (Montgomery, 2001)

Labels of treatment combination	Factor A	Factor B	Factor C
(1)	–	–	–
a	+	–	–
b	–	+	–
ab	+	+	–
c	–	–	+
ac	+	–	+
bc	–	+	+
abc	+	+	+

Each corner of the cube (which represents eight treatment combinations, as shown in the table above) will have a number associated with it. These numbers can be determined based on the values of the levels of each factor. Consequently, the values of these eight treatment combinations can be used to calculate the effect value of each factor, as will be described in the following paragraph.

The estimation of the effect of each factor can be obtained by taking the average of the four treatments of high levels of specific factor minus the average of the four treatments of the low levels of that factor. For example, if the average value of the high and low levels of factor A are y_{A+} and y_{A-} respectively, the effect of factor A can be represented as $A = y_{A+} - y_{A-}$. Similarly, it can calculate the effect of the other two factors B and C by considering the differences between the average values of their high and low levels.

In general the calculations of the main effect of each factor has been illustrated by Montgomery (2001) as follows:

$$A = y_{A^+} - y_{A^-}$$

$$A = \frac{1}{4n} [a + ab + ac + abc - (1) - b - c - bc] \quad (5-2)$$

$$B = y_{B^+} - y_{B^-}$$

$$B = \frac{1}{4n} [b + ab + bc + abc - (1) - a - c - ac] \quad (5-3)$$

$$C = y_{C^+} - y_{C^-}$$

$$C = \frac{1}{4n} [c + ac + bc + abc - (1) - a - b - ab] \quad (5-4)$$

The effects of the interaction between any two factors can be computed as the difference between the mean of the effects of one factor with respect to the two levels of the other factor. For example, the effect of the interaction between the two factors A and B can be measured as the difference of the average of A effect when B treatment is at high levels and the average of A effect when B treatments at low levels. Numerically these have been represented by Montgomery (2001) as follows:

$$AB = \frac{1}{4n} [abc - bc + ab - b - ac + c - a + (1)] \quad (5-5)$$

$$AC = \frac{1}{4n} [(1) - a + b - ab - c + ac - bc + abc] \quad (5-6)$$

$$BC = \frac{1}{4n} [(1) + a - b - ab - c - ac + bc + abc] \quad (5-7)$$

Montgomery (2001) also reported that the interaction among the all factors together can be estimated by taking the average difference between AB interactions for the two different levels of factor C , as follows:

$$ABC = \frac{1}{4n} [abc - bc - ac + c - ab + b + a - (1)] \quad (5-8)$$

Where n is the number of designs (experiments) replicated

5.3 Stages of factorial design

In general, the studying of the process of such a system is required first in order to design an experiment. Any system or model usually consists of different variables as input and one or more variables as output. Thus to design a successful experiment, it is necessary to understand what should be studied in advance and how the data has been collected and analysed. In order to control this issue, a specific procedure should be followed to outline and analyse the settings of an experiment. The first point, of course, would be to define the problem and objectives. This is usually followed by a consideration of how to select the factors that may affect the process, the levels of each of them and the response variables. Also, it is important to analyse the final results of the experiment and make recommendations. A brief overview of each of these points will be described in this section. Further details of these steps can be found in Montgomery (2009).

5.3.1 Identifying the problem statement

To select an appropriate design for the experiment, the problem statement of the experiment should be specified accurately. A better understanding of the phenomenon under consideration may be achieved by the availability of a clear problem statement. It is also important to keep in mind the notion that a clear recognition of the problem leads to the optimum final solution of the experiment. Therefore, at the beginning of any experiment it is necessary to set up a list of the details of a problem or questions that are to be addressed by the experiment. However, the development of clear objectives and statements for some complex experiments is not a simple matter. For this reason, some authors, such as Montgomery (2001), recommended preparing a specialist team to approach designing this kind of experiment.

In this research, the problem of geometrical mismatching between OSM information and reference datasets has arisen. The results that confirm this issue were discussed in more detail in the previous chapter (chapter 4). Hence, experimentation is necessary to determine to what extent error values are influenced by individual factors and whether or not there is any significant interaction between factors. Thus factorial design has been chosen to investigate the effects of factors on error values in comparing FS and OSM

data. The selection of such factors that may affect the physical quality of OSM datasets will be described in next subsection.

5.3.2 Selection of factors and levels

It is important for constructive experimental design to include all factors that may affect the process of the study. The factors should be varied, but not overlapping. The factors should also be chosen as independent variables to ensure that the levels of one factor are not used for the other factors. Although selecting factors for designing an experiment appears to be a fairly straightforward task, more attention should be paid to how to measure or control the desired values of the factors. In this research, three different main factors have been selected for the purpose of studying their effect on OSM data quality: *data source, feature type and individuals*. The main reasons behind these selections are presented in section 5.4.

In order to begin the process of such experimental design, it is also necessary to address the appropriate number of levels of each factor. The selection of factor levels is fundamentally based on the factor type. For example, the choosing of levels for quantitative factors can be critical. If an experiment has one factor Y with two levels Y_1 and Y_2 denoting low and high level respectively, then the question of how to choose the levels will arise. For the experiment performed here, two quantitative levels for each factor were selected. The high levels (+) of the three factors represents *GPS, hard detail and same individual* respectively. The low levels (–) of the three factors represents *other source, soft detail and different individuals* respectively.

5.3.3 Choosing the response variable

The selection of the response variable required a certain degree of attention in this research. The variable should supply useful information to achieve the objectives of such an experimental process. It is also important to note that the response variable should be measurable and relevant to the experimental variability. Usually, the mean or median of a measured characteristic will be considered as the response variable. In this study, the linear error 'Euclidean distance' (defined in chapter 4) between the positional coordinates of FS and OSM datasets was chosen as the response variable.

5.3.4 Selecting experimental design and performing the experiment

There are several considerations that should be taken into account when choosing an appropriate experimental design. These include the number of the sample size, the order of the experimental trials and the amount of experimental replication. Replication refers to running the entire experimental design more than once. For each replicate, eight treatment combinations (i.e., corner of the design box) for the response variable will be calculated. Therefore, the mean value of the two replicates for the response variable of each treatment combination can be taken. This can make data analysis easier and more accurate. Also, there is a high possibility of obtaining some odd results, if there is only a single observation at each corner of the design box. Furthermore, replication provides an effective way to check if there is an outlier, and/or dispersion (consistency, variability) of the response that may affect the results of that setting.

The experimental design adopted in this project is presented in section 5.6. The experimental design can be performed effectively by means of several software packages such as Minitab. This software can assist the experimenter by selecting or suggesting a suitable design. This requires entering information about number and types of factors, levels and replications. However, when designing and implementing an experiment, it is vital to keep in mind the main objectives of that experiment such as identifying the most significant factor that may cause the variability of the response variable.

5.3.5 Experimental results analysis

As the interpretation of experimental results is not a straightforward task, it is often preferable to adopt statistical analysis, such as a *hypothesis test*, to obtain conclusions. Testing a hypothesis means trying to test scientific questions which are typically generated by researchers (Field and Hole, 2003). In general, the hypothesis testing involves several stages. The first step is to state the research question as two competing hypotheses or statements, the null hypothesis and alternative hypothesis. The null hypothesis is frequently denoted by H_0 . It is assumed to be true unless evidence is obtained to prove the converse. The alternative hypothesis is usually known as H_1 . It is assumed to be correct when it is difficult to find evidence supporting the null hypothesis. An example of this is the growth of VGI data, throughout the world, in

which a GI consumer may assume a null hypothesis such as VGI data shows a higher rate of growth in urban areas than in rural areas; the alternative hypothesis would be there is no difference between the growth rates of VGI datasets in both urban and rural areas.

Statistical testing is also necessary to reject or confirm the prediction hypothesis. It is, of course, fairly difficult to decide which hypothesis is correct, so the experimental work should conclude with probabilities. In practice, there are many statistical procedures that can be applied to hypothesis testing. The inferential statistics group is generally used to achieve this kind of analysis (Field and Hole, 2003). In this study, the analysis of variance (ANOVA), which is the principal statistical method for the analysis of data in experimental design, has been adopted.

In order to test the hypothesis assumptions, the laws of probabilities are commonly based on ANOVA statistical approach. Specifically, the test can be performed by comparing the calculated values of the statistical test to a certain critical value. The significance level of the critical value is usually applied at $\alpha = 0.05$. This corresponds to a 95% confidence level. This means that the null hypothesis should be rejected, if the statistical test revealed a probability value of less than 5%. In addition to the requirements above, a sample of observations should be available to perform an effective hypothesis testing.

For the test carried out in this project, the null hypothesis H_0 and alternative hypothesis H_1 are as follows:

H_0 = the factors have no effect on the geometrical quality of the OSM dataset.

H_1 = the factors affect the geometrical quality of the OSM dataset.

5.4 Why these factors are significant for this experiment

In order to better understand the suitability of OSM data for geospatial data integration, especially with formal sources, there are two questions that may arise and ought to be answered. One question that needs to be asked, however, is whether there are any factors that influence the physical quality of OSM information. Although many factors may affect the OSM dataset, the literature such as Ramm et al. (2011), Al-Bakri and Fairbairn (2010), and Haklay et al. (2010) has placed particular emphasis on the three that were defined in subsection 5.3.2. This section considers these factors by describing

the literature related to these issues. Another of the most interesting questions is related to deciding which one of these factors has the most effect on OSM datasets. The answer to this question will discuss in more detail in section 5.6.

The first factor (OSM data source) has gained most attention within the field of VGI data. Many data sources have been used to create OSM information including GPS devices, satellite images, aerial photographs or simply from local knowledge (Ramm et al., 2011). Although some GPS receivers may be able to perform very accurate measurements, most of the OSM data measurements have been achieved using low accuracy GPS such as Garmin, Holux and GPSlim236. In addition to the standard GPS devices, there is a host of other devices with built in GPS chips, such as mobile phones or car-navigation systems. The expected accuracy may vary, depending on the device used. The GPS accuracy also becomes more problematic when surveying buildings data for the OSM project. This is simply because the shadow of a building may prevent the positioning signals from reaching a GPS receiver. Instead of moving with GPS to collect OSM data, OSM also has two different sources for such imagery: the aerial or satellite imagery from Yahoo and NASA Landsat images. In general, producing data using this technique can be considered easier than the GPS technique. However, features produced from images may not always be correct. Their accuracy is fundamentally based on the familiarity of data producers with a given area. For example, some of the OSM Baghdad/Iraq data has been created by users living in the UK. This makes it more difficult to produce precise features such as tracing a street as a one-way or a dual carriageway. Thus, the local population usually has better knowledge regarding an area of interest. In addition to the previous OSM data sources, some features are averaged by eye from many GPS positions and tracks which are recorded by several individuals over time. As OSM data sources vary and show marked differences, it is worthwhile including them as one factor of the experiment performed in this work.

Next, the number of VGI data contributors has an impact on the data itself. In this context, Haklay et al. (2010) concluded that features created by large numbers of volunteers are likely to be more accurate than the entries produced by individuals. They mentioned that some errors may be introduced when VGI data has been produced by only one volunteer. For instance, they can forget to survey some objects or specify wrong locations for other features. Thus, increasing the number of contributors may

decrease such kinds of errors. 'Linus Law' was adopted by Haklay et al. to perform this investigation. Their results proposed that a positive relation exists between data quality and the number of contributors to OSM data mapping. Therefore, it is important to consider the number of data producers as a factor that can affect the quality of VGI data in general and OSM information in particular.

In addition, heterogeneities of VGI data may occur in the distribution of quality parameters, especially spatial accuracy, attribute accuracy and completeness between urban and rural areas. This may occur due to the direct distinction between feature types such as 'hard' and 'soft' features which are the most common features in urban and rural areas respectively. This view is supported by Al-Bakri and Fairbairn (2010) who concluded that there is an obvious difference between the quality of OSM data in the contrasting areas. They suggested that the lower accuracy of the rural dataset may be explained by the inconsistency of the data coverage resulting in fuzzy, interpolated or inferred boundaries such as the extent of woodland areas. Thus feature type is another significant element that may affect the quality of VGI data and plays a key role in the geospatial data integration process.

From the discussion above, it can be concluded that it is becoming increasingly difficult to ignore these factors without deciding the main effect of each of them on OSM data quality and the interaction effects between them, if any. The next sections describe the design, synthesis, characterization and evaluation of the factorial design experiment to test the effect of these factors on the geometrical quality of OSM datasets.

5.5 Preparing tested datasets

After selecting the factors, levels and response variable as explained in last two sections, the data should be collected to carry out an experiment. In order to conduct the experiment effectively, it was necessary to collect data that corresponded to the requirements of the selection of these factors, levels and response variable.

Four test sites were selected, each displaying different types of features, in order to test the flowline designed for this experiment. These sites were the city centre of Newcastle, Gosforth, Cramlington and Clara Vale, as shown in Figures 5.2, 5.3, 5.4 and 5.5 respectively which are represented as OS datasets. The same study areas are also presented in Appendix A (Figures A-1, A-2, A-4 and A-5), but from OSM data sources. Entities were selected to match each of the treatment combinations, as shown in Table

5.2. For example, as the first treatment in the table below shows low levels of all factors, thus the data samples of this treatment would be expected to be features that were produced by sources other than GPS, soft details features, and developed by different individuals. Each of the treatment combinations in Table 5.2 came from different study areas. For instance, the data of treatments one and two and five which are labelled as ((1), *a* and *c*) were extracted from the study area of the Gosforth site in Figure 5.3, while the data of the Newcastle city centre site in Figure 5.2 were used to cover the data of treatments three and four (*b* and *ab*). Similarly, the data from the Clara Vale and Cramlington study areas were applied into treatments six, seven and eight (*ac*) and (*bc* and *abc*) respectively. The planimetric coordinate values for the tested samples were extracted from FS and OSM datasets in order to calculate the experimental response variable values.

This practical testing used some of the areas already sampled – Cramlington and Clara Vale – supplemented by sites in Newcastle and Gosforth. The resulting range of points tested for geometrical accuracy therefore covered both 'hard' and 'soft' detail; both GPS sourced and non-GPS sourced (i.e. satellite imagery tracing, local knowledge) data; and data points forming part of an extensive dataset captured by one individual and datasets comprising multiple contributors. Thus each corner of the 3D cube representing the three-factor factorial design, and the combination of high and low levels for each factor, was tested. The testing was undertaken in the manner described in the previous chapter, involving high accuracy reference data capture (equivalent to the FS datasets mentioned already) and comparison with coordinates derived from the OSM data. The OSM data was extracted following the same procedure as that described in the last paragraph of subsection 3.5.1.2.

In the case of hard features, prominent features such as intersections of roads and pavements, corners of buildings, car parks and fences were selected and coordinates extracted. However, as it is difficult to obtain the corresponding points for comparing soft details' features, techniques based on the centroid of each feature were developed. They were simply based on creating radials with different angles from the centroid of each irregular feature. Subsequently, the corresponding points at a specific angle could be extracted for the compared datasets. Theoretically, these points should have the same planimetric coordinate values. However, since the comparison was performed between two different datasets, field survey and OSM datasets, there is no doubt that some

differences will have occurred. In total, 800 tested points were selected in order to carry out the experiment. As the whole experiment in this thesis was replicated twice, each replication involved 400 points which included 50 points for each experimental treatment, as illustrated in Table 5.2.



Figure 5.2 Newcastle city centre site

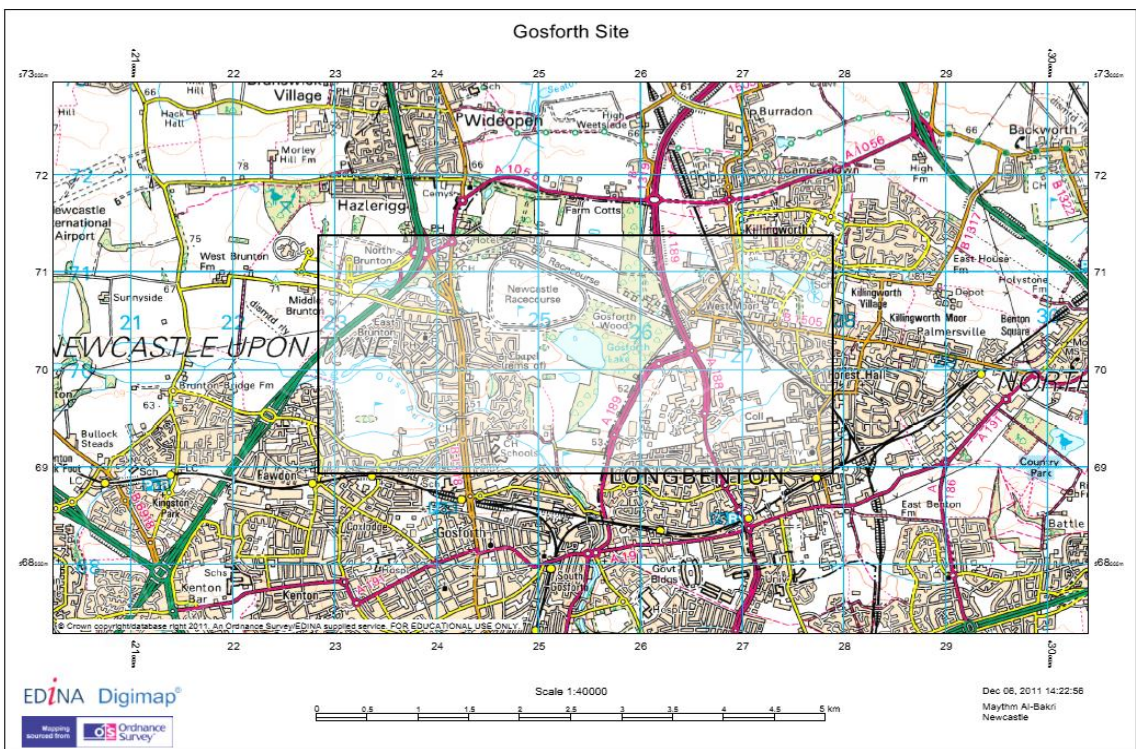


Figure 5.3 Gosforth site

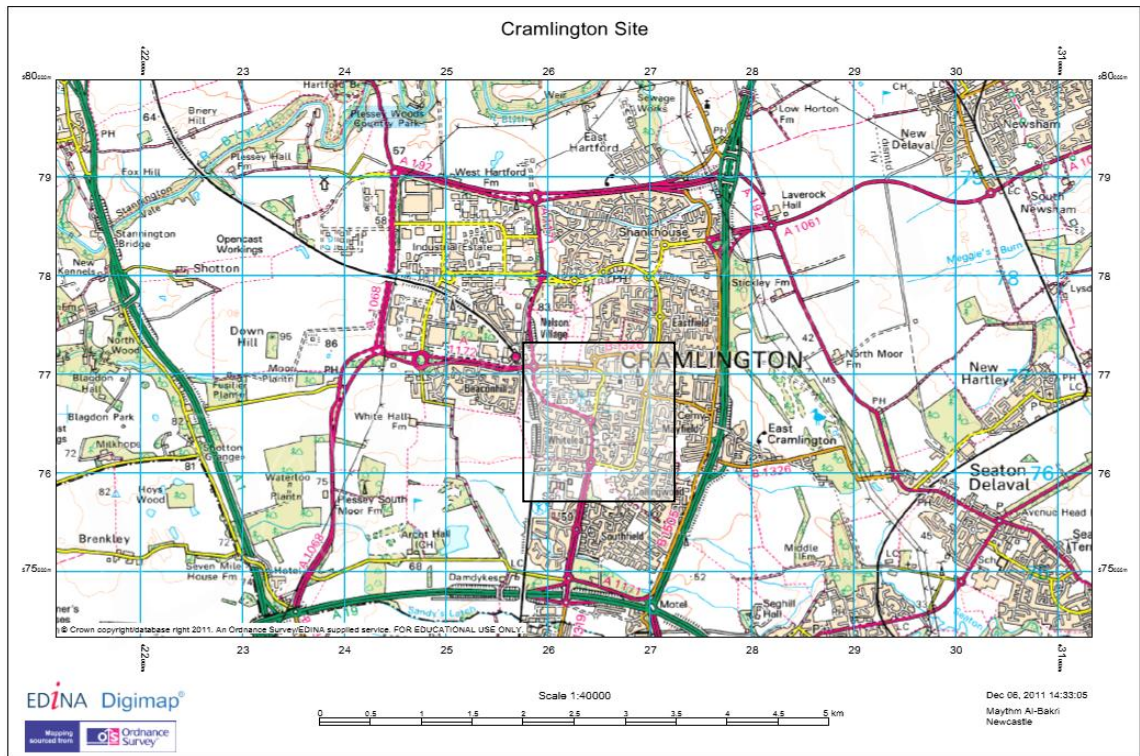


Figure 5.4 Cramlington site

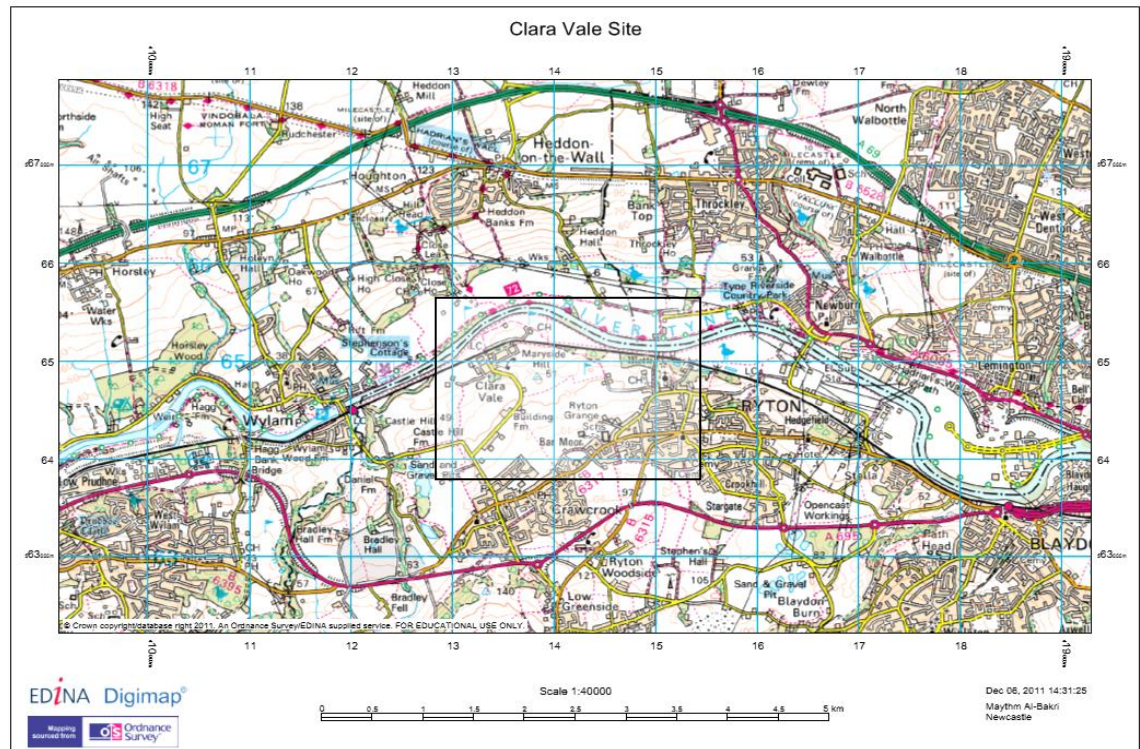


Figure 5.5 Clara Vale site

5.6 Implementing the experiment: discussion and analysis

5.6.1 Experimental settings

A factorial experiment was carried out to investigate the factors thought to influence the physical quality of OSM data. The three factors, *data source* (*A*), *feature type* (*B*) and *individuals* (*C*), were considered. Each of them was set at low and high level (as discussed in section 5.3.2). The design matrix obtained from the case studies of the experiment is shown in Table 5.2. From the statistics illustrated in table below, the mean value of the response variable for each treatment is mostly greater than the median value. In order to minimise the effect of outliers, the median value of Euclidean distance in each combination was selected as the response variable.

Table 5.2 Design matrix and statistics of the developed experiment

Labels of treatment combination	Data Source	Feature Type	Individuals	No. of samples	Euclidean distance (m)		
					Mean	Standard Deviation	Median
(1)	–	–	–	50	14.180	7.330	14.360
a	+	–	–	50	8.890	7.500	6.720
b	–	+	–	50	11.054	5.395	8.316
ab	+	+	–	50	8.955	5.977	7.686
c	–	–	+	50	13.551	6.786	10.934
ac	+	–	+	50	7.140	2.208	6.976
bc	–	+	+	50	8.817	5.715	8.839
abc	+	+	+	50	3.713	1.856	3.557
(1)	–	–	–	50	13.350	7.540	16.030
a	+	–	–	50	7.655	6.835	5.983
b	–	+	–	50	10.774	5.744	8.827
ab	+	+	–	50	8.520	6.056	6.581
c	–	–	+	50	13.741	6.558	12.084
ac	+	–	+	50	9.330	3.842	8.202
bc	–	+	+	50	9.074	5.432	9.646
abc	+	+	+	50	4.403	2.262	4.060

The design discussed here consists of three factors that imply eight runs. It is usually possible to visualise the treatment combinations' data over experiment space by cube plot (Montgomery, 2009), as explained in subsection 5.2.2. In this experiment, the

ranges of the response variable are graphically presented as a box in Figure 5.6. The numbers at the corners of the design cube refer to the average value for each of the treatment combinations. For instance, the number at the corner of the high level of factor *B* and the low levels of factors *A* and *C* is 8.572. This number is produced by taking the average of the response variable of the factors' levels that were in agreement with the previous statements. Hence, cube plot can easily show the worst and the best factor levels' combinations to determine the required response. For example, the graph displays that the ranges of the *response variable* are much larger when all factors are at their low levels (*soft details, other source and different individuals*). This indicates that these levels of the factors may lead to more variability in OSM datasets. On the other hand, it can also be found that when the factors are at higher levels (*hard details, GPS and same individual*), the ranges of the *response variable* are smaller than other combinations. This indicates that those factor levels can reduce the amount of errors in OSM information.

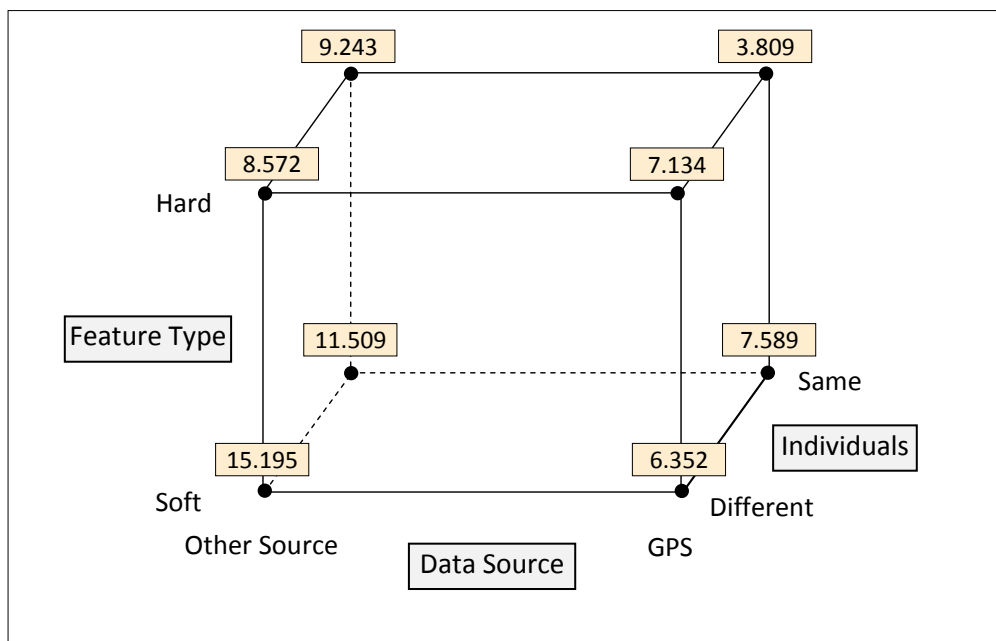


Figure 5.6 Cube plot for response variable 'Euclidean distance'

5.6.2 The main and interactions effects

The main effects of the three factors *A*, *B* and *C* can be presented graphically as shown in Figure 5.7. The graph shows that there has been a clear fall in the values of all factors from a high to low value which indicates that the main effects of all three factors are negative values. The magnitude and direction of each factor effect is based on the mean

response of the lower and higher levels of such factors (Frigon and Mathews, 1997). For example, the main effect of factor *data source* has been obtained by taking the average values of the response variable at two levels: high and low. In order to obtain the first point of the main effect diagram, the average of the low levels of the eight values of the response variable should be determined. From Table 5.2 the mean values of the response variable of low level (–) of data source factor is 11.129. It is clear from Figure 5.7 that this value coincides with the first value of the *data source* main effect plot. Similarly, the other point of the graph can be calculated by taking the average of the response variable for the factor data source at high level (+). After such calculations from Table 5.2, the mean value of 6.221 can be obtained. This value is totally in agreement with the value of the other point of the data source main effect graph.

To distinguish between the different possibilities of the main effect of the factors, the length and slope of the lines' plots were implemented. Lines with little or no slope have a smaller influence on the response variable than lines with longer and steeper gradients (Frigon and Mathews, 1997). It can be noted from Figure 5.7 that the three factors affect the response variable as they display a steep gradient. However, it is obvious that the *data source* is the most significant factor that can affect the response variable. This is because it is this factor that shows the longest and steepest line compared to other factors.

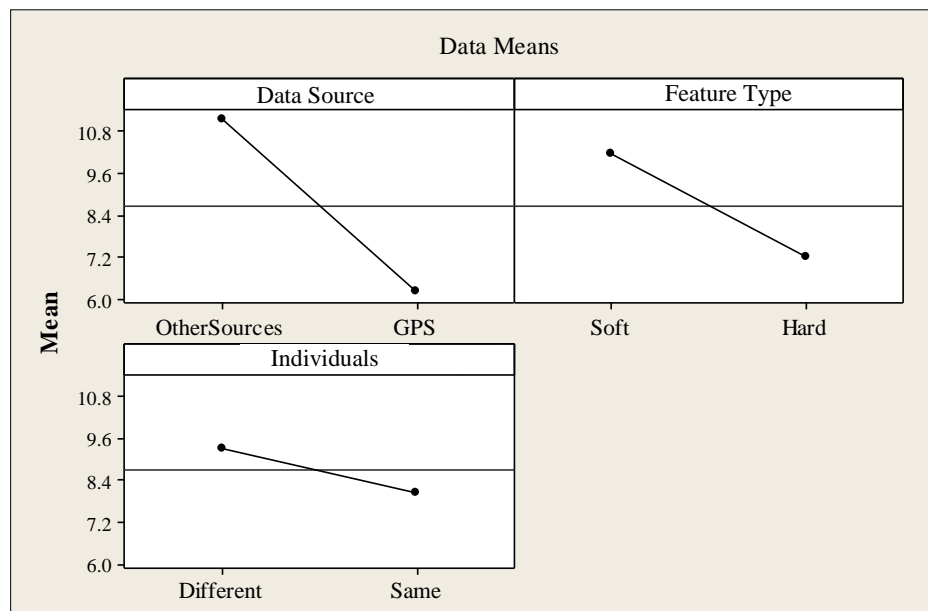


Figure 5.7 Main effects plot for Euclidean distance

Furthermore, to examine the main effects of the factors, it is important to investigate whether there are any significant interactions between them. A possible interpretation of

interaction testing outputs can be achieved by representing the results graphically (Montgomery, 2001). For instance, to investigate the interaction between *data source* and *feature type* factors, the mean of the response variable at different factor levels can be represented as two lines, as shown in Figure 5.8. The first point of the red line was obtained by taking the average of the response variable when the feature type was at low level (–) and data source at high level (+); the second point of the same line was produced from the average of the response variable when the feature type was at high level (+) and data source at high level (+). The same interpretation can be made with regard to the black line. The two lines represent the effect of one factor at the higher and lower level of the other factor.

If the two lines of interactions are parallel, this indicates that there is no interaction between factors. However, a lack of parallelism indicates an interaction between factors (Montgomery, 2001). The *Data source-Feature type*, *Data source-Individuals* and *Feature Type-Individuals* interactions are plotted in Figure 5.8. It is apparent from this diagram that the two lines of the (*Data Source-Feature Type*) interaction are not parallel. The vertical distances between the two lines at each level of *Feature Type* are different. This indicates an interaction between these factors. The data graphed in Figure 5.8 also show that the lines of other interactions being tested are approximately parallel. This leads to the conclusion that there are no interactions between other factors such as (*Data Source-Individuals*) and (*Feature Type-Individuals*).

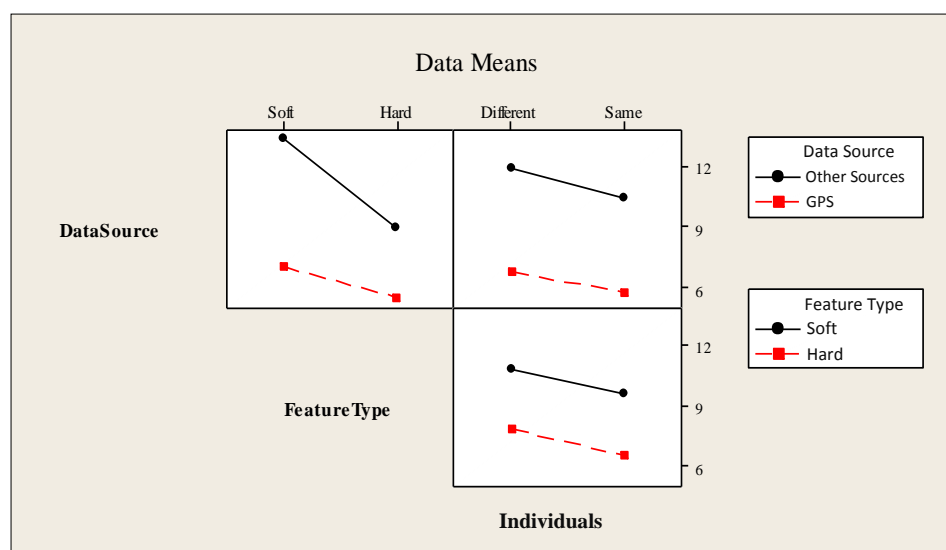


Figure 5.8 Interaction effects plot for Euclidean distance

5.6.3 Testing significant effects

5.6.3.1 Numerical analysis

In any experimental design, the magnitude and direction of the factors' effects should be examined to decide which variable is the most important. Although this was carried out graphically, as in the previous section, it is usually preferable to achieve some numerical calculations for testing and analysing the effects. In addition to the graphical representations, Minitab can analyse 2^k factorial design numerically, as shown in Table 5.3. The upper part of the table illustrates the estimations of the effects of the factors and the coefficients of regression for each experimental treatment. Examining the values of the effects (first column in the table) obviously shows that the Data Source (factor A) is the largest, followed by Feature Type (factor B) and the interaction Data Source×Feature Type× Individuals (A×B×C).

The regression coefficient values can be considered as additional indications of a quantitative value of the effects. The stronger effect of the observed factor should have a higher coefficient value. It can be observed from Table 5.3, for example, that the Data Source factor has the highest value of the regression coefficient magnitudes and the highest value of effect as well. The sign of the effects and the regression coefficients provides an idea regarding the suitability of the factor effect. A positive sign suggests that the response variable will increase by increasing the factor value, while a negative sign indicates that increasing the factor value will decrease the response variable. This panel of the table also reports the *t-statistics* for each of the individual factor effects. In this work, the *p-value*=0.05 and the effects can be considered significant if their *p-value* is less than this threshold. It can be shown from the same part of the table that all the main factors have the largest effects and, in addition, there are significant interactions between A and B factors and among three factors. This is in agreement with the preliminary investigation of the data which depended on graphics.

The lower portion of the table summarises the output of the experimental analysis of variance. It particularly focuses on the terms that were used in the model. For instance, the first row under source is the *main effects* which includes three main effects: *Data Source (A)*, *Feature Type (B)* and *Individuals (C)*. Each one of them only provided a single degree of freedom, thus 3 degrees of freedom can be obtained from all of them,

as shown in the column labelled DF. The row entitled 2-Way interactions includes the terms AB, AC and BC interactions, while 3-Way interactions refers to the interaction effect of all factors together and is denoted by ABC. The next three columns are headed as Seq SS, Adj SS and Adj MS respectively, which are the abbreviations of the sequential sum of squares, adjusted sum of squares and adjusted mean square. These are the primarily calculations that should be achieved in order to apply the *F-test*. For further details regarding such models and calculations, there are numerous sources have attempted to explain them; see for example, Montgomery (2009) and Ryan (2007).

The results in the lower part of the table may be used to confirm the significant of effects based on *p-value*. The column denoted by *F* refers to the results of the *F-test*. The F-test was applied to test the effect of groups. This involved the main effects group, 2-Way Interactions group and 3-Way Interactions. Clearly, there is real statistical significance for the main effects and 3-Way Interactions, as they have *zero p-value*. This is followed by the 2-Way Interactions. The *p-value* of this interaction is less than 0.05, thus it can be considered as a significant interaction at the level of 5%. This was in agreement with the t-test results of the individual factor effect. The main factors (*A, B and C*) have large effects, and there is interaction between the three factors. Furthermore, there may be some interaction between two factors such as the interaction between A and B. Thus the analysis of this part has confirmed the previous analysis.

Returning to the hypothesis/question posed at the beginning of this part of the study, it is now possible to state that the null hypothesis can be rejected and the alternative hypothesis accepted. The null hypothesis stated that the factors have no effect on the geometrical quality of the OSM dataset, whereas the alternative hypothesis supposed that the factors affect the geometrical quality of the OSM dataset. The conclusion to support the alternative hypothesis has been borne out from the analysis and statistical tests that were performed above.

Table 5.3 Minitab analysis for the experiment achieved in this project.

Estimated effects for Euclidean distance					
Term	Effect	Coef	SD Coef	T	P
Constant		8.675	0.183	47.49	0.000
Data Source (A)	-4.909	-2.454	0.183	-13.44	0.000
Feature Type (B)	-2.972	-1.486	0.183	-8.13	0.000
Individuals (C)	-1.276	-0.638	0.183	-3.49	0.008
Data Source×Feature Type (A×B)	1.473	0.736	0.183	4.03	0.004
Data Source×Individuals (A×C)	0.232	0.116	0.183	0.63	0.543
Feature Type×Individuals (B×C)	-0.051	-0.026	0.183	-0.14	0.892
Data Source×Feature Type×Individuals (A×B×C)	-2.230	-1.115	0.183	-6.10	0.000

Analysis of variance for Euclidean distance

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Main Effects	3	138.231	138.231	46.0771	86.29	0.000
2-Way Interactions	3	8.903	8.903	2.9677	5.56	0.023
3-Way Interactions	1	19.889	19.889	19.8894	37.25	0.000
Residual Error	8	4.272	4.272	0.5340		
Pure Error	8	4.272	4.272	0.5340		
Total	15	171.296				

5.6.3.2 Graphical analysis

The analysis of experimental results to determine which factor is the most important can be achieved graphically by using a Pareto chart as a tool for such interpretation (Ryan, 2007). The Pareto plot represents a graphical configuration for the significant effects and interactions. The graph can identify and compare the magnitudes of different effects at the same time, as shown in Figure 5.9. In this plot, the standardised value of each effect is represented as a bar. The effect estimations are ordered from the biggest value to the smallest value. The importance of the effects can be determined on the graph by comparing their values with the threshold value, which is usually plotted as a red line. The values that extend beyond the vertical red line are potentially significant values. As presented in Figure 5.9, the effects of the data source factor were reported as being significantly greater than other factors. Although the effects of the other two factors do exceed beyond the red line, the value of the effect of the first factor is bigger than the

others. It can be also seen from the graph that the interactions between the factors *A* and *B*, and among the factors and *A*, *B* and *C* are also statistically significant, as their effects values are bigger than the value of the red line.

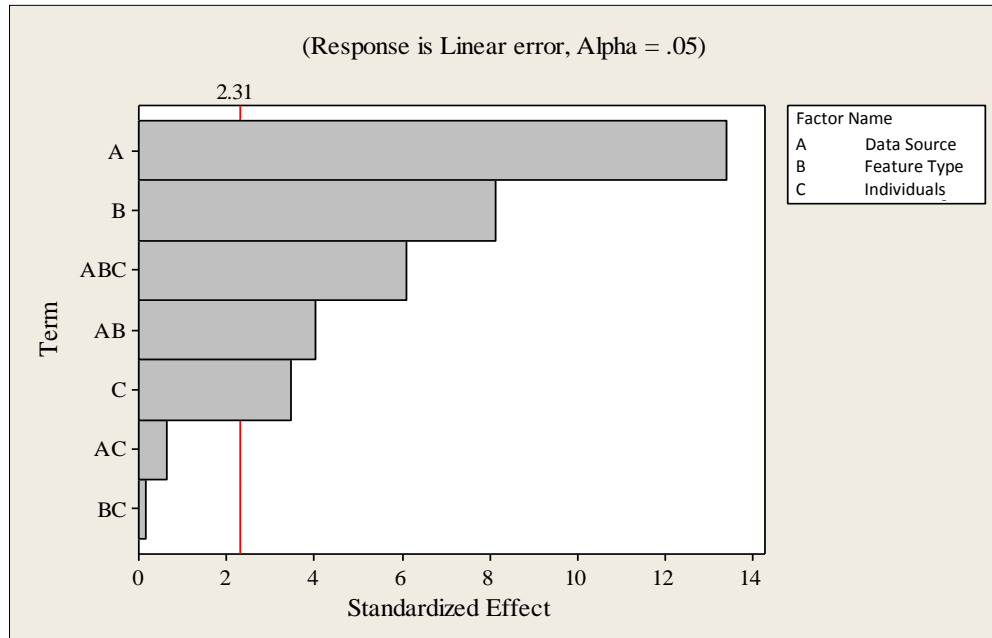


Figure 5.9 Pareto chart of the standardized effects

Another useful technical summary graph mentioned by Ryan (2007) is the normal probability plot (NNP). It can be used to confirm the significant analysis of the Pareto chart. The NNP is based on Lenth's (1989) approach to determine the significant effects of the data. It compares the standardised effect of the factors with cumulative probability or percent. If there are no effects the points will fall along a straight line, whereas the significant effects will appear as points falling further than the ends of the straight line.

Figure 5.10 presents the NNP for the effects of the factors of the experiment described above at 5% significant level. The graph displays that there has been a marked rise in the red points that fall off the straight line. This indicates that there are significant effects of these points. The active effects of the points represent the main affects of factors *A*, *B* and *C*, and interaction effects *AB* and *ABC*. In contrast, there are two black points that fall on or close to the straight line which represent the inactive effects of *AC* and *BC* interactions. It is apparent from this plot that the analysis of the results is exactly identical to that produced by the Pareto chart.

In addition, this plot can provide the direction of each effect. If the point appears on the right side of the NNP line then the effect is positive, while the effect can be considered as a negative if the point falls on the left side of the straight line of the plot. For instance, the interactions AB and AC have positive effects, as both of them are on the right side of the straight line. This means that the response increases when the low level changes into high level. On the other hand, all other factors and interactions have negative effects, as they appeared on the left side of the NP line. This infers that the response decreases by changing to the low level instead of high level. This analysis supports the analysis of results and the directions of effects for both main and interaction graphs that were discussed earlier in this section.

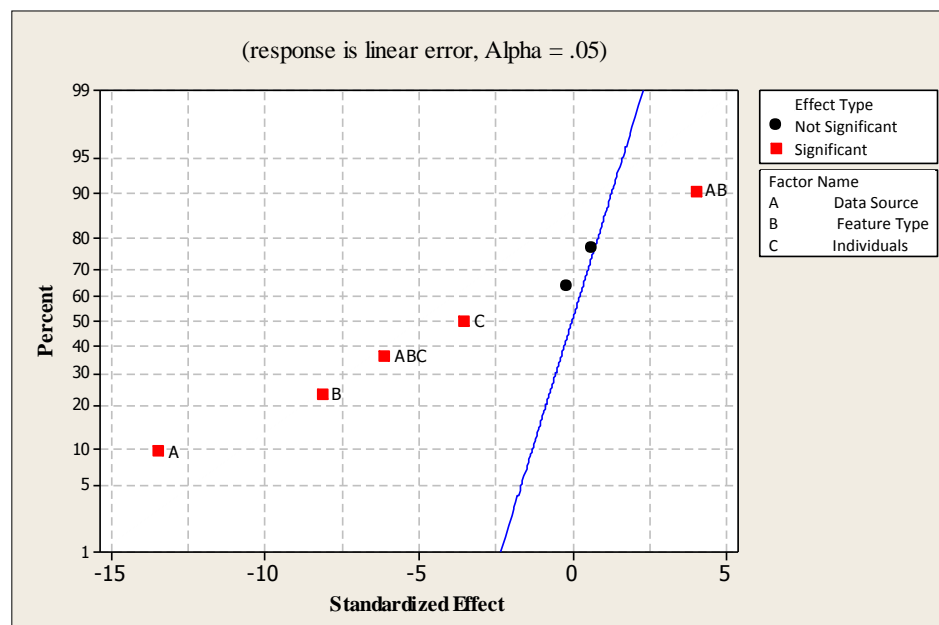


Figure 5.10 Normal probability plot of the main effects and interactions among factors

5.7 Chapter summary

This chapter described a technique that has been adopted to specify the effects of various factors on VGI geometrical data quality in general, and OSM mapping in particular. This was undertaken by setting up a factorial experiment to analyse the effects of three factors, *data source* (A), *feature type* (B) and *individuals* (C), which may affect the physical quality of OSM information. A factorial design approach was adopted to estimate the most significant effect factor. This initially involves selecting factors and levels and choosing an appropriate response variable. In this research, three factors were chosen, as mentioned above, with two levels for each factor. The high level

(+) of the three factors represents *GPS*, *hard detail* and *same individual* respectively. The low level (-) of the three factors represents *other source*, *soft detail* and *different individuals* respectively. The linear error 'Euclidean distance' between FS and OSM was selected as a response variable in this experiment. In order to obtain significant interpretation of the experimental results, it was necessary to depend on statistical analysis. In this study, the analysis of variance (ANOVA), which is the principal statistical method for the analysis of data in experimental design, was implemented.

By applying the methodology of the factorial analysis, the results indicated that the data source, among the three factors, has the most significant effect on OSM data quality, at 0.05 significant levels. The next important factor is the feature type, as it appears as zero p-value, which is less than the critical value of p-value = 0.05 of this experiment. The individual factor can also be considered to be a significant factor, as it has a p-value equal to 0.008. The results also showed that there are interactions between the factors *A* and *B*, and among the factors and *A*, *B* and *C*, as their effects values exceeded the value of the red line in Figure 5.9. This proved that the null hypothesis can be rejected and the alternative hypothesis can be accepted which they were stated in section 5.3.5. It is therefore likely that such effects exist for three factors on the geometrical data of OSM information.

Since semantic similarity is another fundamental notion that should be taken into consideration in GIS, in addition to geometrical similarity, to achieve beneficial interoperability among spatial data, Chapter Six provides a detailed discussion of the models and approaches that have been used for semantic similarity handling, while Chapter Seven will focus on developing a model for the assessment of the integration of feature classification of authoritative datasets such as OS and GDS, and VGI datasets such as OSM information.

Chapter 6 Semantic Similarity Models and Approaches

6.1 Introduction

In general, the term 'semantics' can be defined as the study of the meaning attached to words or symbols. The word itself was added to the English language relatively recently and Read (1948), in a detailed account, traced the history of the term. Today semantics is regarded as an important branch of linguistics as it can be used to study the relationship between meanings and words. It is necessary to have an understanding of semantics for the study of the acquisition of language (i.e. how humans can produce and use words in the context of social communication for the purpose of clarifying and interpreting a variety of activities). Semantics is not only a concern of linguistics; the subject can also be viewed from a philosophical perspective. Many philosophers have suggested that the study of 'ordinary language' can solve many philosophical problems; for instance, investigating the definitions and differences between the concepts of the words 'right' and 'wrong'. In the context of spatial data, semantics is always related to geometric and geographic aspects of the features; for example, on maps, polygons may represent buildings, while lines may correspond to motorways or roads. Therefore, graphic and geometric forms can suggest the spatial data's meaning or semantics. The semantics of spatial data also depend upon the database definition for a specific feature. For instance, an object represented by a polygon may correspond to a playground in one database, while the same geometrical form can represent a green field in another data base. Furthermore, similar semantic heterogeneity may occur when using the same descriptive term for different features in different datasets.

New trends in technology and science have led to an increase in the amount of geographic information available on the Internet, as mentioned in chapters 2 and 3. This has produced increasing expectations with regard to the semantic heterogeneity issue. Thus, semantic similarity plays an increasing role as a measure of overlap for different processing such as integration of classification schemes for different sources' datasets. The most important problem regarding heterogeneity in spatial datasets may occur in testing semantic and structural feature classification trees' similarity of data from different sources. It is common to find the same pictorial concept for different names in two datasets. At the same time, it is possible that a feature may belong to a sub-class in

one dataset and a super-class in another. As a consequence of semantic data inconsistency, the integration of multiple spatial datasets may remain one of the main challenges facing spatial data users. Hence, in addition to the issue of geometrical similarity (as described in Chapter 4), semantic similarity is another fundamental issue that should be taken into consideration in GIS for the purpose of achieving beneficial interoperability in spatial datasets. In this chapter, the models and approaches for semantic similarity handling will be discussed. The next chapter will focus on assessing the possibility of integrating the feature classification of authoritative datasets such as Ordnance Survey (OS) and General Directorate for Survey (GDS), as well as VGI data sets such as OpenStreetMap (OSM) information.

In the next section, a brief overview of the analysis of traditional approaches to semantic similarity assessment is presented. This involves an explanation of why the confusion matrices approach is not suitable for the purpose of measuring the similarity of authoritative and VGI classifications. Other models for semantic similarity measurements are considered in section 6.3. The advantages and disadvantages of each model are specified and described in this section. Section 6.4 presents in detail the definition and concepts of formal semantic databases (e.g. WordNet). An overview of the structure of WordNet hierarchy begins by discussing the nature of the nodes of the lexical network and the relationships that connect these nodes. The differences between semantic similarity and relatedness are also included in this section. Subsequently, further examination of semantic similarity and relatedness methods in WordNet::Similarity software is considered in section 6.5. The performance of WordNet::Similarity software is discussed in section 6.6. Each of the semantic similarity models is evaluated within the WordNet::Similarity software package and ultimately one is selected as the optimum model, before the chapter summary remarks are stated in section 6.7.

6.2 Traditional models for semantic similarity analysis

6.2.1 Standard confusion matrices

Chapter 2 indicated that accuracy assessment is an essential task in understanding and determining the suitability of feature classifications accuracy for different processes such as the process of geospatial data integration. In order to be able to evaluate and

analyse individual feature classifications' accuracy, the classifications of spatial data should be quantitatively assessed by comparing the tested area on the map against the reference area which should be of a higher degree of accuracy. The most common and popular method for assessing the accuracy of feature classifications is the confusion or error matrix. It is primarily used for remote sensing applications such as assessing the accuracy of classified images. This involves the comparison of the classes and locations of ground reference data with corresponding features on the classified image (Lillesand and Kiefer, 2000). The confusion or error matrix is a square array composed of a number of rows and columns. The reference data (which is considered correct data) is usually represented in the columns, while the rows represent the classified image data (the tested data which is used for assessment). For instance, in the more complex case of differential classifications between several spatial datasets there are more than two categories, such as 'Class A', 'Class B' and 'Class C'. The error matrix that can arise as a consequence is shown in Figure 6.1.

		Reference data			
		A	B	C	K_{i+}
Tested data	A	K_{AA}	K_{AB}	K_{AC}	K_{A+}
	B	K_{BA}	K_{BB}	K_{BC}	K_{B+}
	C	K_{CA}	K_{CB}	K_{CC}	K_{C+}
	K_{+j}	K_{+A}	K_{+B}	K_{+C}	K

Figure 6.1 An example of confusion matrix (Congalton and Green, 2009b)

Congalton and Green (2009b) illustrated that the elements of the matrix consist of the number of cells that fall in each class relative to the comparison value. The diagonal represents an effective discriminating system which should show a high proportion of correct classified elements. However, incorrect classified elements would be shown on both diagonal sides. The determination of the accuracy level can be achieved by calculating separate metrics: overall accuracy, producer's accuracy and user's accuracy. The overall accuracy is determined by dividing the total number of correctly classified

cells (diagonal cells) by the total number of cells tested, as shown in Equation (6-3). However, the overall accuracy cannot give comprehensive accuracy estimation for each of the different classes. If a high proportion of cells is recorded for a single class, this will create a bias in the overall accuracy measurement. Thus it is important to assess the accuracy of individual elements; for example, evaluating the producer's accuracy. This is simply the number of correctly classified elements of a specific category divided by the sum of the cells in that category as indicated by reference data, as illustrated in Equation (6-4). It can also be called the measure of error of omission.

Another way of representing individual class accuracy is called user's accuracy. This is calculated by dividing the number of correctly classified samples by the sum of that category as indicated by tested data, as in Equation (6-5). The user's accuracy can be considered as a measure of error of commission. Based on the example in Figure 6.1, the mathematical representations of the error matrix can be summarized as follows:

$$K_{i+} = \sum_{j=A}^C K_{ij} \quad (6-1)$$

$$K_{+j} = \sum_{i=A}^C K_{ij} \quad (6-2)$$

$$\text{overall accuracy} = \frac{\sum_{i=A}^C K_{ii}}{K} \quad (6-3)$$

$$\text{producer's accuracy} = \frac{K_{jj}}{K_{+j}} \quad (6-4)$$

$$\text{user's accuracy} = \frac{K_{ii}}{K_{i+}} \quad (6-5)$$

Where:

A, C : The first and last elements of confusion matrix of the example in Figure 6.1.

K_{ij} : The observation in row i and column j of confusion matrix

K_{ii}, K_{jj} : The observations of the major diagonal of confusion matrix

K_{i+} : The total number of tested samples classified into category i

K_{+j} : The total number of reference samples classified into category j

K : The total number of confusion matrix samples

The kappa coefficient is further common statistical measure that can be used to evaluate the accuracy of feature classifications (Congalton, 1991). It takes into account the rows and columns of the confusion matrix to estimate the total accuracy of classifications. It depends on the two classes to measure the accuracy of the classifier, while overall accuracy is based only on the diagonal values of the confusion matrix. Thus it can be considered a better measure of feature classification accuracy than the overall accuracy. The following equation can be used to calculate the kappa coefficient of the confusion matrix.

$$k = \frac{K \sum_{i=A}^C K_{ii} - \sum_{i=A}^C (K_{i+} \cdot K_{+j})}{K^2 - \sum_{i=A}^C (K_{i+} \cdot K_{+j})} \quad (6-6)$$

The value of the kappa coefficient is between 0.0 and 1.0. The maximum value of the kappa coefficient (1.0) indicates full agreement between classifications. While the value of the kappa coefficient will be zero if there is no agreement in classifications.

In this thesis, in addition to the evaluation of geometrical accuracy, accuracy measurement also covers the semantic data associated with the measured coordinated dataset. Testing using standard confusion matrices was attempted, but problems arose due to inconsistent classification schemes. This was essentially because the land use classification scheme in OSM was very different to OS and GDS datasets. Thus, a further study of the classifications of each dataset, authoritative and VGI data, was undertaken and a measure of 'closeness' of attribute information was attempted. The study considered the semantic similarity of individual corresponding features and XML schema elements in order to additionally assess the possibilities of data integration from VGI and official sources. Such a method was adopted by combining the information content approach and the XML schema structural component, a process which will be discussed in more detail in the next chapter.

6.3 Alternative similarity measurement models

Surveys such as that conducted by Schwering (2008) have shown that the existing semantic similarity assessment approaches may be classified into geometric, network, transformational, feature and alignment models. The resultant similarity measures different depending on the model chosen, as there are differences in how each model

defines the relationship of concepts and the metric properties of those relationships. In general the notion of semantic similarity models can be found in two semantic similarity measure forms: the commonalities and differences measure or the semantic distance measure (Schwering, 2008). By using the commonalities and differences notion, the higher the similarity between two concepts, the more commonalities and less difference there will be between them. While in the semantic distance measure, all concepts must be structured as a hierarchical tree and the semantic similarity can be measured by determining the shortest path length between the two concepts. Therefore, selecting the appropriate similarity model for such processing would be based on the task itself because each measure has different assumptions and properties as will be explained in the next paragraphs.

6.3.1 Geometric models

Geometric models can be considered as one of the most commonly used approaches in analysing semantic similarity. In these models, the semantic similarity can be measured by modelling the objects as points in a multidimensional space. This approach describes the similarity measure between the objects by comparing the spatial distance between the respective points in multidimensional space. The main assumption of these models is that similar concepts should be close to each other in multidimensional spaces. Thus, this model can be considered as one example of the second notion of the semantic similarity measure models (semantic distance models). The distances between entities in multidimensional spaces can be typically measured by Minkowski metrics, as follows:

$$d(c_1, c_2) = \left[\sum_{k=1}^n |X_{c_1,k} - X_{c_2,k}|^r \right]^{1/r} \quad (6-7)$$

Where:

n : The number of dimensions

$X_{c_1,k}$ and $X_{c_2,k}$: The value of entities c_1 and c_2 along the dimension k

r : The value of $r = 1$ when the equation 6.7 represents the 'city block' distance between two points. While the value of $r = 2$ when equation 6.7 yields 'Euclidean distance' (Suppes et al., 1989)

As Schwering (2008) argued, the geometric similarity models satisfy three metric assumptions. Firstly, the minimality: the two concepts are similar, if the spatial distance between them is zero. It means the same as the assumption that states that the maximum similarity can be obtained between the concept and itself. Secondly, the symmetry: this statement suggests that the semantic similarity and the distance between two concepts are identical from both sides. For instance, the semantic similarity will be the same when it is measured from concept A to B, for example, or from concept B to A. Finally, triangle inequality: this assumption states that the distance between two concepts is usually smaller than or equal to the distance between both concepts through a third one.

It can be seen that these models are based on the view of symmetric and transitive similarity assumptions. Thus for some applications it is difficult to determine the similarity with respect to the multi-dimensional space view of the geometric models. Therefore, other models can be adopted to calculate semantic similarity between concepts, as will be illustrated in the following subsections.

6.3.2 *Semantic network models*

Network models are different to geometrical models in that they are based on semantic nets rather than multidimensional space for measuring similarity. The main idea in this area is to organize the concepts as a taxonomical hierarchy. The hierarchical semantic network is usually composed of a number of label nodes linked by a series of edges. The nodes are terms, words or properties. The edges may represent a range of relationships that connect the nodes with each other (Lee et al., 1993). Figure 6.2 illustrates an example of semantic network hierarchy which shows the distance between nodes is not necessarily identical. For instance, it can be seen from the figure that node A is connected to node a_1 via one edge r_4 , while node A is connected to node C through two edges r_1 and r_3 . There are various kinds of relations that connect between nodes in the same semantic hierarchy. For example, the relations between superclass and subclass, is-a relation and part-of relation, which are usually represented as a directed edge between the concepts in the semantic hierarchy. The semantic relationships will be discussed in more detail in section (6.4).

Although geometric and network models are based on different techniques to measure semantic similarity, as mentioned above, both of them use the notion of distance

measurement between concepts. The distance notion is different when used by geometrical or structural (network) models. In geometrical models, the distance can be measured as a spatial distance, while the distance is usually measured on the hierarch or graph in network models. In general, the structural models can be categorised as metric and non-metric models. According to Schwering (2008), the network similarity measure is metric when the distance between nodes is considered without taking into account the direction of the edges. On the other hand, the similarity measure is non-metric, if the direction of the arcs between words has been taken into account. In general, most network models use the edge counting approach to calculate similarity. The idea underlying this technique is that the more similar the concepts, the shorter the path length will be between them in semantic hierarchy.

The network models have the shortcoming of measuring similarity with respect to the nodes' density. In practice, a closer distance between nodes would achieve higher density nodes and higher similarity. This is true for the concepts in the middle and high levels of the hierarchical structure. However, for the concepts in the lower sections of the hierarchy the distance is increased and the similarity will decrease. For example, it appears from the upper part of Figure 6.2 that the distance between the nodes A and B is only two edges (r_1 and r_2), whereas at the lowest hierarchical level of the same figure it can be observed that the distance between the nodes a_1 and b_1 is four edges (r_4 , r_1 , r_2 and r_7). Although this can be considered as one limitation of these models, the approaches have been widely followed in a number of domains such as measuring semantic and relatedness similarity, as will be presented in subsections 6.5.4 to 6.5.7.

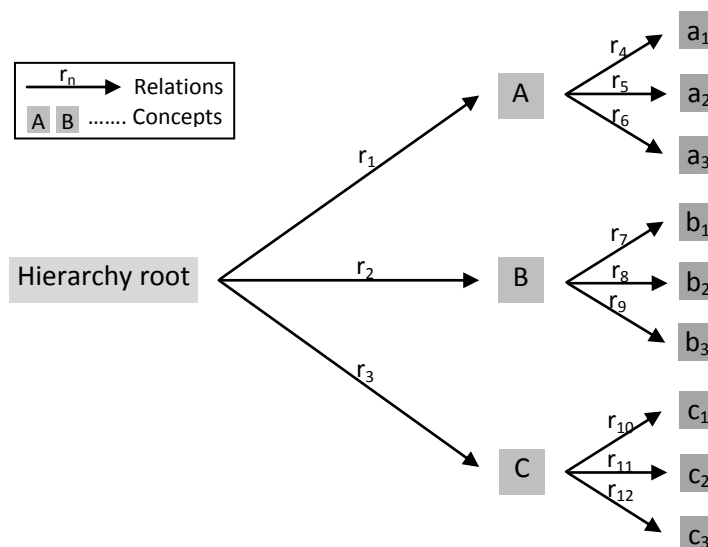


Figure 6.2 A hierarchical network structure

6.3.3 Transformational models

The main idea behind measuring similarity using these models is based on the transformational distance concept. Although these models use the distance measured as an indication of similarity, the distance concept is different to the distance notion that has been used in geometrical and network models. The distance of transformational models is calculated as the number of transformations that may be applied in order to obtain the required entity. This is equivalent to a spatial distance and a distance on a hierarchy for geometrical and network models respectively.

This approach has been suggested by Hahn et al. to provide a framework for measuring the complexity of transformation (Hahn et al., 2003; Hahn, 2001). They depended on Representational Distortion theory for measuring a similarity. Thus, the similarity magnitude of two entities can be expressed by the number of operations that are needed to transform (distort) one representation into another identical concept. When the number of transformation processing is increased, the similarity would be expected to decrease. For instance, the series *ABAB* needs two operations to turn into *AAAA*, while it requires three transformations to become *ACCC*. Therefore, *ABAB* can be considered to be more similar to *AAAA* than *ACCC*.

6.3.4 Feature models

Feature methods are based on the relationships or the properties of the terms in the taxonomy to calculate semantic similarity. By applying these models, the similarity increases with regard to common features and decreases as far as different features are concerned. Although the previous approaches (geometrical, network and transformational) utilize the properties of objects for the purpose of measuring similarity, the properties are different to those used in feature models. The feature models usually use the qualitative information of objects to measure the semantic similarity between concepts rather than distances, as in the other models.

In (1977) Tversky proposed a similarity measure which can be considered as one of the most common feature models. In Tversky model, the similarity process can be defined by a set-theoretic similarity measure. Thus the similarity between two features sets, c_1 and c_2 , for instance, can be determined as a function of common and different features. Using this approach, semantic similarity can be measured in three steps: determining the

number of features which are common for both objects c_1 and c_2 , which is the intersection of common features $C_1 \cap C_2$; determining the number of features that relate to object c_1 but not c_2 $C_1 - C_2$; and determining the number of features that belong to c_2 but not c_1 $C_2 - C_1$. Figure 6.3 illustrates a graphical diagram for measuring similarity by applying Tversky operations. Formally defined, the similarity between c_1 and c_2 is:

$$S(c_1, c_2) = \frac{l(C_1 \cap C_2)}{l(C_1 \cap C_2) + \varphi \times l(C_1 - C_2) + \omega \times l(C_2 - C_1)} \quad (6-8)$$

Where:

φ and ω : The weights for the common and distinctive sets of features. As Tversky (1977) reported the values of φ and ω should be greater than or equal to zero. According to Gregson (1975), Eisler and Ekman (1959), and Bush and Mosteller (1951), all cited in Tversky (1977), the weight values can be varied and different. Therefore, the framework in equation 6-8 can provide a wide variety of similarity models which are fundamentally based on weight values.

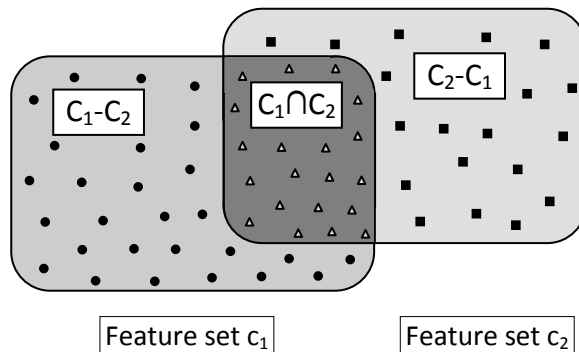


Figure 6.3 The relations between two sets of features (Tversky, 1977)

The main restrictions and limitations of feature models were outlined by Schwering (2008), Schwering (2005), and Tversky (1977). The feature models detect the whole features' similarity and no partial matches can be assessed. By using feature models, there is no ability to measure the structural relationships between two objects. Also, in these models, the concepts that are shared between two features must not be the same as the elements for any other features.

6.3.5 *Alignment models*

The similarity measurement using alignment models was derived from the idea of the structural alignment framework which was originally adopted by Gentner and Markman (1997). These models indicate that the similarity measure should rely on how features align with, or correspond to other elements, not only measuring the similarity of different and common features as in previous models. In general, the measuring of alignment similarity can be categorized into alignable and non-alignable relationships. The number of the levels of objects can be used in order to compare the ability to align two datasets. For example, Figure 6.4 illustrates two datasets each consisting of two levels of objects, and therefore alignable: they would be non-alignable with variable hierarchical levels.

The alignment relationships may also be affected by the characteristics of the compared objects themselves. For instance, in the examples shown in Figure 6.4, the building object is in the first level of dataset 1, whereas the highway is at the top of dataset 2. Since the two datasets are not similar at the same levels, the relation is non-alignable. On the other hand, the existence of corresponding objects in compared datasets may also affect alignment processing. For example, from Figure 6.4, it can be seen that the car park object in the left dataset does not match any object in the other configuration, therefore a non-alignable relation would be addressed for this situation.

For the notion of GIS processing, such as the integration of classification for multiple sources datasets, the alignment of the structure of the feature classification trees may play an increasing role as a measure of similarity between the classifications' hierarchies. For instance, for two names in two different datasets, it is common to find the same concept for both classifications trees. It is also possible to find the same features as a super-class in one hierarchy and a sub-class in another hierarchy. The assessing of the possible integration of structural feature classification of spatial data from official sources such as (OS) and (GDS) and VGI sources such as OpenStreetMap information will be illustrated in more detail in the next chapter.

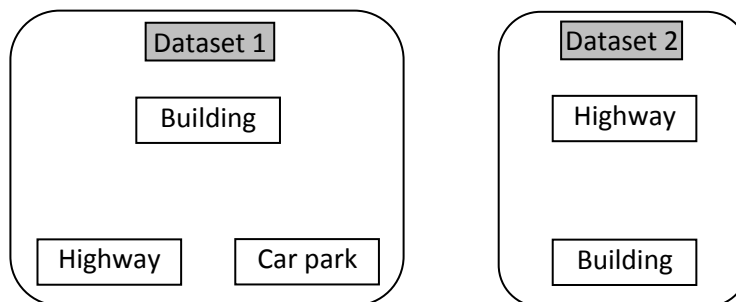


Figure 6.4 The geometric configuration of alignment differences

6.3.6 Information content based models

Rather than using properties of objects in order to identify their semantic similarity as in previous models, these models use the information content of the least common subsumes of the two concepts to measure their similarity. The information content matches the generality and specificity of the specific concept. The general idea of measuring similarity between two concepts using information based models is based on the extent to which they share information in common. This can be achieved by estimating the common information of the super-class of the two classes / concepts. The information content of a super-ordinate class can be obtained by determining the probabilities of occurrence of concepts in the corpus. Information based approaches have been widely used for such similarity measurement applications; see, for example Lin (1998), Jiang and Conrath (1997) and Resnik (1995b). These researchers have suggested more sophisticated methods for measuring similarity based on the same construct of information content models and combining them with other techniques. Most of these applications and models have been applied to measure semantic similarity in WordNet::Similarity software which will be described in more detail in sections 6.5.1 to 6.5.3.

6.4 Formal semantic lexical databases and their descriptions

6.4.1 WordNet database

WordNet is a lexical on-line database for English Language developed at Princeton University. WordNet was designed as a graph or network of information and each node of the network represents a specific concept. The node consists of a set of synonyms, or synsets, which represent the same meaning of the word or concept. For example, the concept of a car may be represented by a set of words such as 'car', 'auto', 'automobile'

and 'motorcar'. Thus synsets are a collection of words which represent the fundamental building block of WordNet. They are organised into a tree like hierarchical structure, which is constructed in the design of WordNet. For instance, in version 2.0 there are nine separate noun hierarchies that include 80,000 concepts, and 554 verb hierarchies that are made up of 13,500 concepts (Pedersen et al., 2004). The WordNet synset can be in the form of four types of Parts of Speech (POS) noun, verb, adjective, and adverb. The synsets are also structured into senses which correspond to different definitions or descriptions of the same term or concept. Figure 6.5 illustrates a fragment of the WordNet Is-A hierarchy.

Several relations have connected the WordNet synsets in lower and higher hierarchy. The relationship between the current synset and other synsets is defined through different types of explicit semantic relationships. The most common relationships are: hyponymy (is a kind of), hypernymy (is a generalization of), meronymy (is part/substance/member of) and holonym (i.e. part-of relationships). Thus, for example, synset K is connected to synset L through *is a kind of* relation when K is a hyponymy of synset L, and L is a hypernymy of K. For example, the synset containing 'laptop' is a hyponymy of the synset containing 'computer' and 'computer' is a hypernymy of 'laptop'. Also, it can be said that synset K is related to synset L during *is part of* relation when K is a meronym of L, and L is a holonym of K. For example, a synset containing a 'wheel' is a meronym of the synset containing a 'car' and a 'car' is a holonym of a 'wheel'. These relations can structure the synsets of words or concepts into large trees or hierarchies. For each hierarchy there is a root node which represents the more general concept or ancestor for other nodes in a tree.

The free availability and abundance of information on the WordNet on-line lexical system have made it an important lexical information resource for several applications. For instance, Resnik (1995a) presented an automatic method for revealing noun sense disambiguation within related noun sets using WordNet senses. In another major study, Leacock et al. (1998) described a statistical classifier that can be used to disambiguate parts of speech (a noun, a verb an adjective). They based their method fundamentally on WordNet relations in order to automatically identify the required sense from general text quantity. A further application of WordNet information was suggested by Fellbaum et al. (2001). They proposed a framework for extracting semantic distinctions and

syntactic clustering from the WordNet database. The application of WordNet is also found in other areas such as measuring semantic similarity and relatedness between compared words, as will be explained in the following subsections.

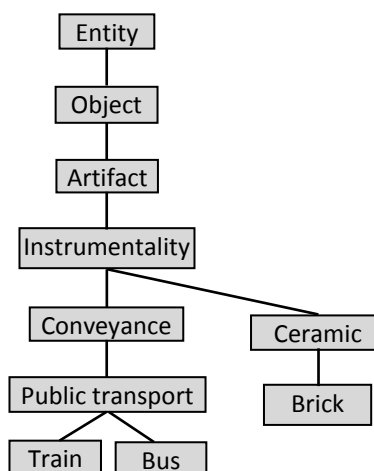


Figure 6.5. A fragment of the WordNet is-a hierarchy (Giannis et al., 2005)

6.4.2 *Semantic similarity and relatedness*

The importance of using semantic or relatedness similarity functions for data integration applications has been emphasized by many researchers; see for example, Guarino et al. (1999) and Lee et al. (1993). Semantic similarity measurements are also important in other topics such as word sense disambiguation (WSD), information retrieval and extraction, and automatic correction of word errors in a text. However, there is misinterpretation between definitions of 'semantic similarity' and 'relatedness' in the literature. It is important to note that the semantic similarity and relatedness measure are not identical (Pedersen et al., 2004). The relatedness similarity is more general than semantic similarity, and semantic similarity can be considered as a special case of relatedness similarity. If two concepts have similar meanings or they are synonyms, then the two terms can be semantically similar. For instance, measuring semantic similarity can show that a road is more similar to a highway than it is to a building; whereas, two objects are usually assumed to be related to each other when there is a link or relationship between them. For example, 'leaves' and 'tree' are related to each other because leaves are part of a tree. Also relatedness may exist for any kind of functional association or common relationship such as 'pen' and 'paper'.

Since the WordNet database contains a huge amount of information on the English Language and organizes this information into hierarchies, it is suitable for semantic

similarity and relatedness measures. WordNet::Similarity software can be used to measure semantic similarity and relatedness. It is a free software package and is based on a lexical database, WordNet, for measuring similarity and relatedness between a pair of words. A utility programme, *similarity.pl*, can provide a web interface for WordNet::Similarity to run a measure of similarity. Also, WordNet::Similarity can be installed on a desktop machine within the *Perl programme* and can call upon its methods for measuring semantic and relatedness similarity. The software allows the user to measure the similarity or relatedness of two concepts when sent without any specifications or sent by specifications associated with a concept such as *concept#pos#sense*. For example *path#n#2* means the second sense of *path* in the WordNet database.

Pedersen et al. (2004) showed that WordNet::Similarity provides six measures of semantic similarity and three measures of relatedness. Three of the six measures of similarity are based on information content of the least common subsume (LCS). These measures are: *res* (Resnik, 1995b), *lin* (Lin, 1998), and *jcn* (Jiang and Conrath, 1997). Three similarity measures are based on path length methods: *lch* (Leacock and Chodorow, 1998), *wup* (Wu and Palmer, 1994), and *path*. The relatedness methods include *hso* (Hirst and Onge, 1998), *lesk* (Banerjee and Pedersen, 2003), and *vector* (Patwardhan et al., 2003). The similarity and relatedness methods are described in more detail in the following section.

6.5 Methods for using semantic and structural models

This section describes some of the common methods for measuring semantic and relatedness similarity between terms, most of which are used by the WordNet::Similarity software package.

6.5.1 Resnik (*res*)

Resnik (1995b) proposed a method to compute semantic similarity based on the models of information content which were discussed in section 6.3.6. It was the first approach to bring together ontology and corpus. The method estimated the information content of terms by using statistical information from large groups. This semantic similarity method was introduced by following the standard argumentation of information theory.

The quantity of information that the two concepts share in common can be considered as the base for measuring semantic similarity. It can be defined as follows:

$$sim(c_1, c_2) = IC(lcs(c_1, c_2)) \quad (6-9)$$

Where:

IC : The information content of a concept

$lcs(c_1, c_2)$: The lowest common subsume of the two concepts (c_1, c_2)

If there is more than one subsume of the concepts c_1 and c_2 , then the semantic similarity between two concepts would be the maximum of the information content of concepts that subsume both concepts c_1 and c_2 . Hence the semantic similarity value in the Resnik method would be determined based on the assumption of the information theory, which was originally presented by Shannon (1948). It proposed that the information content of a concept (c) can be quantified as negative the log likelihood ($-\log p(c)$); where $p(c)$ is the probability of the occurrence of a concept c in a corpus. This quantitative characterization of information provides a new way to measure semantic similarity. Formally, defined as:

$$sim(c_1, c_2) = \max_{c \in S(c_1, c_2)} [-\log p(c)] \quad (6-10)$$

Where:

$S(c_1, c_2)$: The set of the common ancestors of concepts c_1 and c_2

$P(c)$: The probability of encountering an instance of a synset in some specific corpus

6.5.2 *Lin (lin)*

In (1998) Lin developed another method for measuring semantic similarity that was based on information content. This method depends on a lot of assumptions for measuring the semantic similarity of two concepts: the more two concepts have in common, the more similarity there is between them. The less similar two concepts, the less they will have in common. If the two concepts are identical, the maximum similarity score will be attained. For measuring semantic similarity between two concepts, Lin uses the sum of the information content of the two concepts c_1 and c_2 , in addition to the information content of the shared parents of these concepts. The information content

of the least common subsume has been scaled by this summation. This idea can be reflected by the following equation:

$$sim(c_1, c_2) = \frac{2 \times \log p(c^\circ)}{\log p(c_1) + \log p(c_2)} \quad (6-11)$$

Where:

c_1 and c_2 : The two concepts for which measuring semantic similarity is required

c° : The most specific ancestor that subsumes the two concepts c_1 and c_2

The zero value of information content should be given more attention. The above formula would provide undefined similarity when the denominator is zero. This can simply occur when the information content of the two concepts is zero. On the other hand, when the information content of the least common subsumes is zero, the similarity score will be zero.

6.5.3 Jiang and Conrath (*jcn*)

Jiang and Conrath (1997) introduced a measure of semantic similarity between words or concepts that relied on the combination of information content with edge counting. It incorporates the information content of the least common subsume of the two concepts along with the information content of the two concepts themselves. Also, it combines this information content with a lexical taxonomy structure. Thus, the model will be influenced by the information content of the two specific concepts and their subsumes. The distance between the two terms has been defined by Jiang and Conrath as follows:

$$dist_{jc}(c_1, c_2) = IC(c_1) + IC(c_2) - 2 \times IC(lcs(c_1, c_2)) \quad (6-12)$$

Where:

IC : The information content of the concept

$lcs(c_1, c_2)$: The lowest common subsume of the two concepts (c_1, c_2)

6.5.4 Leacock and Chodorow (*lch*)

The Leacock and Chodorow (1998) method is based on path length or the distance between words to measure semantic similarity between them. The measure takes into account the number of links between two concepts and the depth of the taxonomy,

rather than the information content that has been used in *res*, *lin* and *jcn* metrics. The similarity between two terms can be determined by finding the shortest path length between them divided by the double of the maximum depth of the taxonomy. The semantic similarity can be defined by using the following formula:

$$\text{sim}(c_1, c_2) = -\log\left(\frac{\text{shortestpath}(c_1, c_2)}{2D}\right) \quad (6-13)$$

Where:

c_1 and c_2 : The two concepts

$\text{shortestpath}(c_1, c_2)$: The shortest path length between the two concepts c_1 and c_2

D : The overall depth of the taxonomy

6.5.5 Wu and Palmer (*wup*)

Another measure of semantic similarity based on distances and depths was proposed by Wu and Palmer (1994). In particular, it relies on measuring path lengths between words in the WordNet taxonomy. The similarity of two concepts is defined by how closely they appear in the hierarchy: thus it is more a measure of their structural relations. The method considers the depth of the lowest common subsume (LCS) from the root node of the hierarchy, as well as the distance between both concepts and their LCS. To determine the semantic similarity of two terms, the depth of the LCS from the root node is calculated and is scaled by the summation of the distances of individual concepts. The semantic similarity measurement can be defined in Figure 6.6 and equation 6-14.

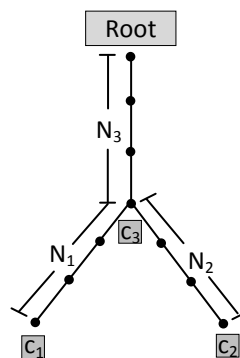


Figure 6.6 The concept of similarity measure (Wu and Palmer, 1994)

$$sim(c_1, c_2) = \frac{2 \times N_3}{N_1 + N_2 + 2 \times N_3} \quad (6-14)$$

Where:

N_1 : The number of nodes on the path from c_1 to c_3

N_2 : The number of nodes on the path from c_2 to c_3

N_3 : The number of nodes on the path from c_3 to the root of the hierarchy.

For example, in Figure 6.6 the depth of the nodes c_1 and c_2 to c_3 are both 4, and the distance between the root node and their LCS (c_3) is 4. Thus by using the adapted Wu and Palmer's formulation, the semantic similarity score between c_1 and c_2 is $(2 \times 4) / (4 + 4 + (2 \times 4)) = 0.50$.

6.5.6 Path

This method can be considered one of the simplest method for measuring semantic similarity in the WordNet::Similarity software package. It also uses the distance and path length for measuring semantic similarity between synsets. The semantic similarity can be calculated by determining the inverse of the shortest path length between the two concepts in the hierarchy (Pedersen et al., 2004). It can be defined as follows:

$$sim(c_1, c_2) = \frac{1}{shortestdist(c_1, c_2)} \quad (6-15)$$

For example, in Figure 6.6 the distance between c_1 and c_2 is 6, therefore the semantic similarity score by using this metric is $1/6$.

6.5.7 Hirst and Onge (hso)

The Hirst and Onge (1998) method used a lexical chain in WordNet to determine the relatedness between words. Different to the measures of semantic similarity that only considered path length (*lch*, *wup* and *path*), the Hirst & St-Onge method uses all the semantic relations that are defined in WordNet. These relationships can be classified as upward, downward and horizontal. This method also takes into account the links relations between the concepts in WordNet: extra-strong, strong and medium-strong. The extra-strong relation would occur when the two terms are exactly the same. The words would be related by strong relation when they are synonyms. While the medium-strong relation occurs between words which are usually connected by a short path and

does not have many direction changes. The following equation can be used to define the strength of the relatedness between two terms.

$$rel(c_1, c_2) = C - path\ length - k \times d \quad (6-16)$$

Where:

C and k : The constants

d : The number of changes in path direction

The values of C and k are 8 and 1 respectively. These values were used for the Hirst & St-Onge experiment. To show how this method can calculate the relatedness between two concepts, the example in Figure 6.6 will be considered. The length of the path linking between c_1 and c_2 is 6. There are upward and downward relations, thus there is one direction change. By applying equation 6.17 the relatedness between c_1 and c_2 is $8-6-1 \times 1 = 1$.

6.5.8 Banerjee and Pedersen (*lesk*)

Banerjee and Pedersen (2003) illustrated a measure of semantic relatedness that was inspired by Lesk (1986). The Lesk (1986) method is based on counting the overlapping concepts between the dictionary definitions of the two concepts in order to define their relatedness. This approach was suggested before the development of WordNet, and it is essentially designed to be used with traditional dictionaries. Thus Banerjee and Pedersen (2003) improved the Lesk method by using it with a huge online source of information, the WordNet database. Their measure is mainly based on the incorporation of the WordNet glosses information. This measure can assign the relatedness value of two concepts by measuring the overlapping terms of the two concepts' glosses.

6.5.9 Vector

Patwardhan (2003) adapted the approach of Schutze (1998), which relies on context vectors, to introduce another measure of semantic relatedness. The method suggested that each term in WordNet should be represented as a gloss vector. The gloss vector can be defined as a context vector created by including the gloss of WordNet as a context. The method of Patwardhan can measure the relatedness of two concepts c_1 and c_2 by comparing the equivalent gloss vectors of c_1 and c_2 . Thus the relatedness between two

concepts c_1 and c_2 can be determined as the cosine of the angle between the normalized gloss vectors of the two concepts. It can be calculated as follows:

$$rel_{vector}(c_1, c_2) = \cos(\text{angle}(\vec{v}_1, \vec{v}_2)) \quad (6-17)$$

Where:

\vec{v}_1 and \vec{v}_2 : The gloss vectors of the two concepts c_1 and c_2

Angle : The angle between two vectors

From the above description, it can be noted that a range of possible similarity and relatedness measures can be offered using WordNet::Similarity software. These measures are based on a number of different metrics and models. Thus one of the intentions of this study is to examine the ability of these methods of comparing feature classifications of discrete datasets. This can help in determining the possible integration of map legend categories or database attribute codings for different sources' datasets. The next section will cover and report the results of WordNet::Similarity performance. This includes an explanation concerning the specific method that was used to test and record the matching process between two concepts.

6.6 The achievement of WordNet::Similarity software

It would be reasonable to evaluate the performance of WordNet similarity measurements of semantic similarity by comparing them with human judgement results. The correlation between human judgement and WordNet similarity calculations can be examined by setting up an experiment to rate the similarity of a set of word pairs. It can be used with the same sample of 30 name pairs that were selected in an experiment when only human subjects were involved. The data of human ratings are from the publication of previous results of Miller and Charles (1991).

The correlation has been calculated using Pearson's correlation function (as in equation 6-18). Suppose that there are two variables X and Y , with means μ_x and μ_y respectively, and standard deviation sd_x and sd_y respectively. The correlation is computed as:

$$r = \frac{\sum_{i=1}^n (X_i - \mu_x)(Y_i - \mu_y)}{(n - 1) sd_x sd_y} \quad (6-18)$$

The Pearson correlation is (+1) in the case of a perfect positive (increasing) linear relationship, (-1) in the case of a perfect decreasing (negative) linear relationship, and some value between -1 and 1 in all other cases, indicating the degree of linear dependence between the variables. As the value approaches zero, there is less of a relationship. The closer the coefficient is to either -1 or 1, the stronger the correlation between the variables.

It seems from Figure 6.7 that all methods have a correlation score between (0.35) and (0.89). The information content of the least common subsumer (LCS) similarity methods have a minimum score of (0.47) by jcn (Jiang and Conrath, 1997) and maximum score of (0.81) according to the method used by Philip (1995b). Path length methods have a minimum score of (0.75) in the path method and maximum score of (0.78) in the method proposed by Leacock and Chodorow (1998). The relatedness methods have a minimum score of (0.35) which is given by the lesk method and a maximum correlation of (0.89) by the vector method.

The results of the experiment confirm that the relatedness method proposed by Patwardhan et al. (2003) (vector method) performs quite well and it has a correlation score of (0.89). This means that the similarity scores of this method are close enough to human judgement results. Path length methods seem to perform very well, especially the lch method, where the correlation of (0.78) was obtained. Regarding least common subsumer (LCS) similarity approaches, the res method performed better than other methods in this family, with a correlation of (0.81). In the information content similarity approaches, the lin and res methods were very close to each other, both performing better than other methods in this family. For semantic similarity in this study, Lin's (1998) approach has been adopted because it emphasises the meaning relationship, and the range of similarity scores is between 0 and 1, thus a normalizing process is not required (unlike the Resnik method).

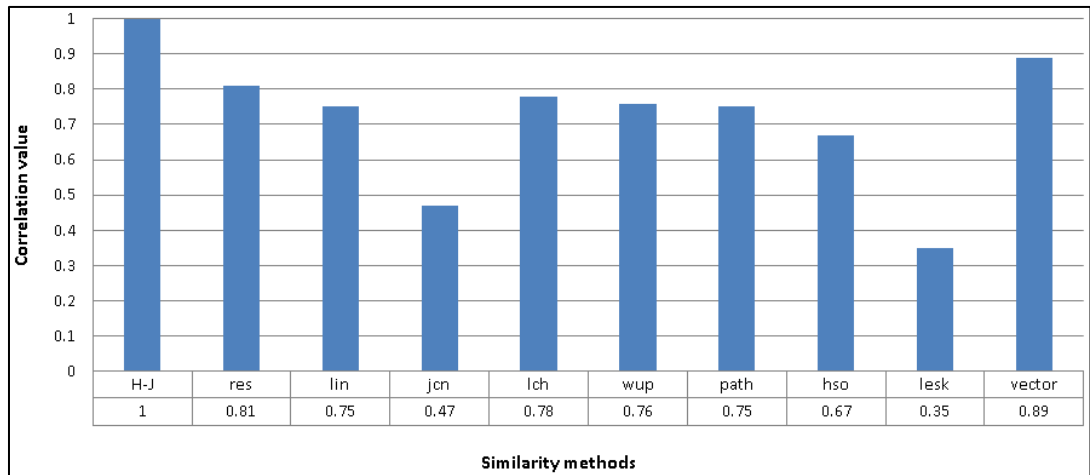


Figure 6.7 The correlation between human judgement results and WordNet::Similarity methods

6.7 Chapter summary

Different approaches to semantic similarity assessment are presented and compared in this chapter. Firstly, standard confusion matrices metrics were discussed. For this project, testing using standard confusion matrices was attempted, but problems resulted due to inconsistent feature classifications between formal data such as OS and GDS and VGI data such as OSM information. Thus, more advanced approaches to the computation of semantic similarity were described and adopted. The semantic similarity measurement models are usually used to compare the concepts (terms) or objects by taking into account the feature description or the semantic relationships. These include: geometric models, semantic network models, transformational models, feature models, alignment models and information content based models.

There is a great difference between these models in assessing semantic similarity. For example, geometric models measure similarity between objects by comparing the spatial distance between the respective points in the multidimensional space. The similarity of the network and transformational models is also founded on distance measurement. However, the distance concept is different to the one that is used in geometrical models. For network models the distance is usually measured on the hierarchy or graph, while the distance of transformational models is calculated as the number of transformations. Whereas to define the similarity of two objects using feature models, the distinctive or the common features are employed. Alignment models indicate that the similarity measure can rely on how features align with, or correspond

to other elements. Finally, Information content models are geared to use the information content of the least common subsumes of the two concepts to measure their similarity. From the comparison above, it can be noted that each model has special properties and each one should be used for specific similarity measurements.

The concepts of these models have been used to develop many metrics for measuring semantic similarity and relatedness between objects. For example, the WordNet::Similarity methods were founded on some of these models such as network models and information content based models. This chapter also illustrated and described the specifications and the metrics of each WordNet::Similarity method. Then the chapter moved on to assess the performance of the WordNet::Similarity software methods. The correlation between human judgement and WordNet similarity calculations was examined by setting up an experiment to rate the similarity of a set of word pairs. It is clear from the results that were obtained that the *lin* and *res* methods (information content similarity family) are very close to each other and that they perform better than the *jcn* method. Thus in this study, Lin's (1998) approach has been adopted for measuring semantic similarity. This is due to the fact that this approach uses meaning relationship and the semantic similarity score is between 1 and 0, therefore there is no need for normalising the results, as in the Resnik method.

The next chapter explains the assessment of the possibility of integration of feature classifications from official data sources, such as (OS) and (GDS), and VGI datasets, such as the (OSM) project. This will initially include evaluating integration using a feature-by-feature technique, then the XML schema of compared features will be examined in order to see if it is possible to integrate the schemas of different spatial data sources (i.e. official and informal).

Chapter 7 Finding Semantic and Structural Similarity for Classes and Instances

7.1 Introduction

The notion of semantic similarity measurement with regard to spatial data refers to measuring the heterogeneous meaning of the same real features on the Earth that are obtained from different spatial data sources. Semantic similarity can be utilised to overcome the difficulties of sharing and integrating multi-source data by measuring the degree of the semantic compatibility of geographic features (Riedemann et al., 1999). As mentioned in Chapter 2, the area of geospatial data integration processing has become more important recently due to the growth of network connections services such as the Internet and the increasing interest in reusing spatial data from different sources. However, the semantic inconsistency among integrated datasets may make the integration processing more difficult and complex.

Models and techniques for semantic similarity measurements were discussed and compared in Chapter 6. Furthermore, the performance of WordNet::Similarity software was determined by ascertaining the effectiveness of its methods in order to use it for semantic similarity measurements in this research. This chapter focuses on developing models for assessing the semantic similarity for possible integration of feature classifications from official spatial data sources such as Ordnance Survey (OS) and General Directorate for Survey (GDS), and VGI data sources such as OpenStreetMap (OSM). The models initially compared the semantic similarity of corresponding features of tested datasets using a feature-by-feature approach, as demonstrated in subsection 7.2.1. This chapter also describes the application of an XML schema similarity measurements model. This involves the measuring of three aspects - semantic similarity, structural similarity and data type similarity - between the elements of compared schemas, as described in section 7.3. Finally, concluding remarks are presented in the last section of this chapter.

7.2 Semantic similarity approaches

The semantic similarity measure of spatial concepts presented here is based on two sets of analyses. The first examines the integration possibility of the corresponding feature classifications of different datasets. The second set of analyses involves developing a

model for assessing the ability of XML schema matching of feature classifications from formal and informal spatial data sources, as described in the following sections.

7.2.1 Feature based approach

This set of experiments focuses on analysing the semantic similarity scores of equivalent feature pairs of authoritative datasets, such as OS and GDS data, and VGI datasets, such as OSM information. This includes applying specific statistical tests in order to decide the ability of integrating the corresponding feature classifications.

7.2.1.1 Testing the semantic similarity suitability for feature classification matching purposes

Based on the semantic similarity measurement results of feature classifications for three study areas - Cramlington-UK, Clara Vale-UK and Baghdad-Iraq - the research evaluates the possibility of feature-by-feature semantic matching processing for each area by applying a one-sample t-test procedure. Campbell and Swinscow (2009) reported that this statistical test can assist to investigate the comparison of a population mean of tested data to a desired target. For instance, the semantic similarity scores of corresponding features in a variety of study areas show what appear to be low accepted marks for a successful semantic integration approach (i.e. 0.5) (Al-Bakri and Fairbairn, 2012). The one-sample t-test was applied to evaluate the success of the matching process.

In order to statistically examine the comparison of the population mean of a tested dataset with the value of the desired target, null and alternative hypotheses should be stated. For this study the null hypothesis (H_0) assumed that the population mean of semantic similarity scores is greater or equal to 0.5, while the alternative hypothesis (H_1) supposed that the population mean is less than 0.5.

According to Campbell and Swinscow (2009) the *p-value* can be used to accept or reject the null hypothesis. After calculating the value of the one-sample t-test, the corresponding *p-value* can be easily obtained from specialist statistical tables. In order to assess such data by applying the one-sample t-test, there are two possible conclusions that can be drawn: if the *p-value* is less than the significance level, which is usually 0.05, the null hypothesis should be rejected at the predetermined confidence level of 95%. In

contrast, if the *p-value* is greater than 0.05, then there is not enough evidence to reject the null hypothesis. Thus it should be accepted at 95% confidence level.

For the project realised here, the one-sample t-test is applied at significance level $\alpha = 0.05$ (95% confidence level). The results are shown in Figure 7.1; the *p-values* are less than 0.05 for all study areas. This indicates that the null hypothesis should be rejected at the significance level of 5%. This means that the population mean of semantic similarity scores of the compared feature classifications is not equal or greater than (0.5). The 95% confidence interval values were also used to predict the population mean values. As can be seen from the figure below, the confidence interval values are between 0.1765 and 0.2537, 0.2140 and 0.3682, and 0.4094 and 0.4577 for Cramlington-UK, Clara Vale-UK and Baghdad-Iraq respectively. Thus the results of this analysis show that the population means of semantic similarity scores to be 0.5 is exaggerating. This is because the confidence intervals of all study areas do not include the comparison value of 0.5. Hence the findings of this part of study are consistent with those of *p-values* to reject the null hypothesis.

One-Sample T: Semantic similarity score							
Test of mu = 0.5 vs not = 0.5							
Variable	N	Mean	StDev	SE Mean	95% CI	T	P
Cramlington,UK	172	0.2151	0.2564	0.0196	(0.1765, 0.2537)	-14.57	0.000
Clara Vale,UK	63	0.2911	0.3062	0.0386	(0.2140, 0.3682)	-5.41	0.000
Baghdad,Iraq	225	0.4336	0.1838	0.0123	(0.4094, 0.4577)	-5.42	0.000

Figure 7.1 One-sample t-test outcomes for the feature classifications semantic similarity scores of the comparison of FM and OSM datasets for three study areas: Cramlington-UK, Clara Vale-UK and Baghdad-Iraq.

In addition to the numerical results of the one-sample t-test, the data can be presented graphically in Minitab. For example, a box plot can display a summary of the distribution of the sampled data. It generally consists of an upper quartile or 75th percentile (the upper line of the box), lower quartile or 25th percentile (the lower line of the box), median or 50th percentile (the middle line of the box), upper whisker (extended from the maximum data point with 1.5 of box height), lower whisker (extended from the minimum data point of the box with 1.5 of box height) and outliers (the data that lies beyond upper and/or lower whiskers) which are usually represented as black dots. The

box plot graph associated with the one-sample t-test can also show the 95% confidence interval for the population mean and the null hypothesis reference value.

For the current study, the graphical analysis of the one-sample t-test for the first study area (urban area-UK) is presented in Figure 7.2. The plot shows that the 50th percentile or median value (the green line) is 0.09 and it is near to the lower quartile (Q_1). This means that the semantic matching performance is very poor for this part of the study. The scores in this test are in the range of zero and 0.7633. The lowest 25% of the semantic similarity score is zero; the middle scores of the test range between zero and 0.4026; the top 25% of the test score are between 0.4026 and 0.7633. The blue line on the same figure represents the lower and upper values of the 95% confidence interval (0.1765, 0.2537); while the red dot indicates the hypothesized population mean which is stated at the null hypothesis as 0.5. The plot in figure 7.2 indicates that the confidence interval does not include the reference mean, as the red dot lies outside the range of the blue line. It is therefore suggested that the population mean of the semantic similarity score for this part of the study differs from the hypothesized value. Thus, there is an agreement between the graphical analysis and the numerical analysis of the one-sample t-test by rejecting the null hypothesis and accepting the alternative hypothesis.

Figure 7.3 illustrates the graphical analysis of the one-sample t-test for the possibility of semantic matching of different datasets in a rural area, UK. It can be observed from the graph that the median value (the green line) is 0.1869 and is closer to the bottom of the box than the top of the box. This indicates that the possibility of semantic integration regarding this part of the project is difficult. Although the median value shows that the semantic similarity scores of Clara Vale-UK are better than Cramlington-UK, the median of semantic similarity of Clara Vale-UK is still relatively low. The scores in this test range from zero to one. The lowest 25% of the marks is between zero and 0.0493, the middle 50% of the scores range from 0.0493 to 0.4547, and the upper 25% of the marks is between 0.4547 and 1. By comparing the range for Clara Vale with the range of Cramlington, it can be seen that the Clara Vale data gives better similarity scores, although still less than 0.5.

Figure 7.3 also displays the values of 95% confidence interval which is between 0.2140 and 0.3682. It is represented by the blue line in the figure. What is interesting in this

figure is that the hypothesized population mean (red dot) is out of the range of the confidence interval (blue line). Hence, there is evidence to reject the null hypothesis and accept the alternative hypothesis. Data from this plot is in agreement with the numerical data in Figure 7.1.

The last part of this graphical analysis is related to the Baghdad-Iraq site and is reported in Figure 7.4. The box plot displays that the median value (the green line) of the semantic similarity scores of this study area is 0.5591. It is equal to the upper quartile or 75th percentile value. This indicates that the data quality of this test is comparatively better than the quality of the previous two datasets (UK study areas) as their median values are near the lower end of the box. The data of this analysis ranges from zero (the smallest outlier) to 0.8650 (the largest outlier). For the top 25% of the data, it includes the group of all the scores beyond the top of the box plot (upper whisker and outliers) which is between 0.5591 and 0.8650. The middle part of the box represents 50% of the tested data and it ranges from 0.3780 to 0.5591, while the lowest 25% involves everything less than the lower quartile value (lower whisker and outliers) which in this test is between zero and 0.3780.

The value of 95% confidence interval is also reported in Figure 7.4 as between 0.4094 and 0.4577. This means that there is 95% confidence that the true population mean of the semantic similarity scores is between 0.4094 and 0.4577. It can be noted that 0.5 is not inside that confidence interval, and the null hypothesis should therefore be rejected. Although the results of the last test (Baghdad-Iraq) are better than the two UK study areas, the analysis indicated failing to accept the null hypothesis as the hypothesized population mean is out of the range of confidence interval.

Therefore, a similar picture of difficult feature classifications integration in the other two study areas is revealed for the Baghdad study area. The mismatches between the feature classifications of FM and VGI suggest that using OSM data to initiate or revise categorisations in the formal datasets would be extremely difficult.

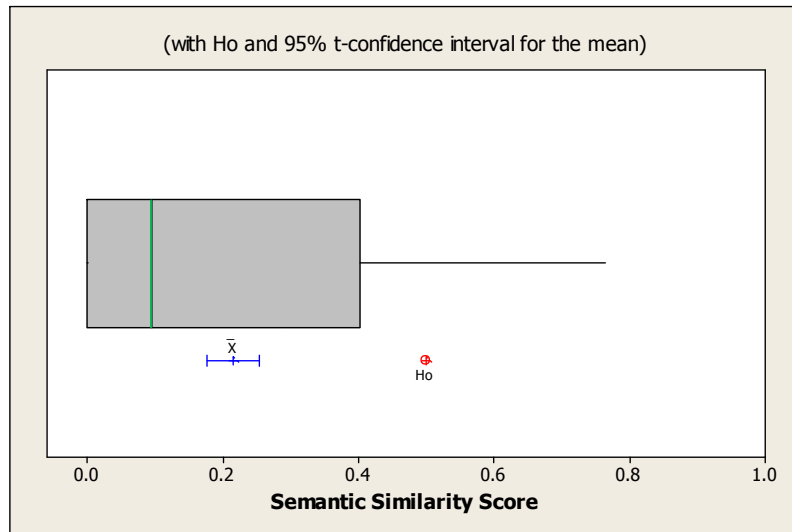


Figure 7.2 Box plot of the one-sample t-test results for the semantic similarity scores of feature classifications in Cramlington-UK site.

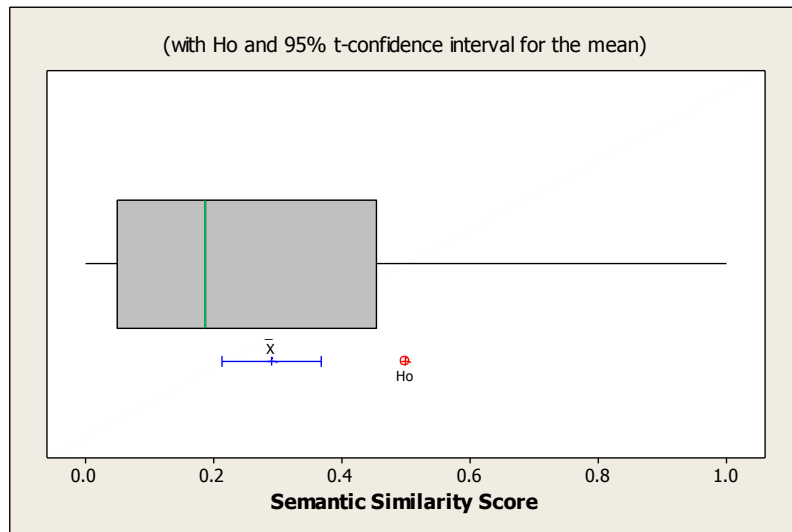


Figure 7.3 Box plot of the one-sample t-test results for the semantic similarity scores of feature classifications in Clara Vale-UK site

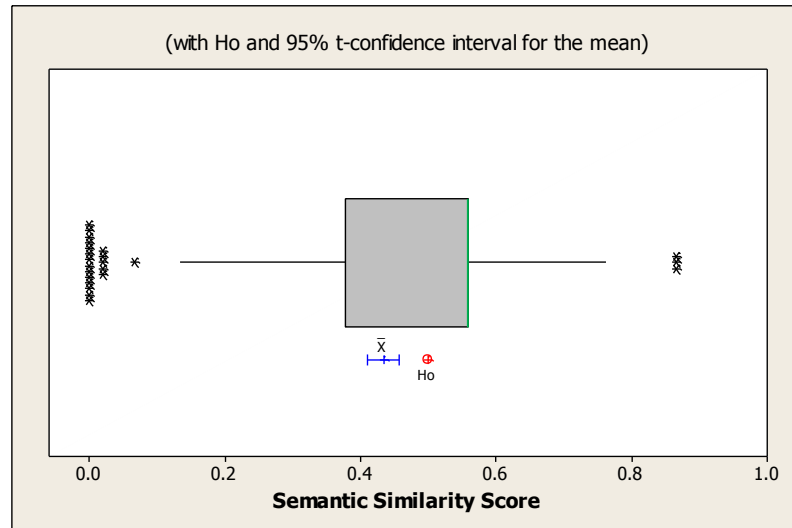


Figure 7.4 Box plot of the one-sample t-test results for the semantic similarity scores of feature classifications in Baghdad-Iraq site

7.2.2 Schema similarity approach

The second set of experiments transferred the tests from individual features classifications similarity measurements into hierarchical classification schema similarity measurements. The schema was defined by Saleem et al. (2008) as a rooted label tree which consists of a set of nodes that is connected by a unique path or edge. The nodes are usually constructed inside schema as a hierarchy of several levels of ancestors, parents and children. Generally, the relations between nodes of any source and target schemas may be classified into three main types: single correspondences, multiple correspondences and missing correspondences. In the following subsection, the schema node matching relationships will be described along with an example for each pattern.

7.2.2.1 A categorization of schema relationships

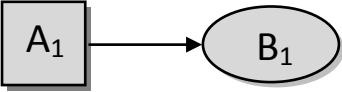
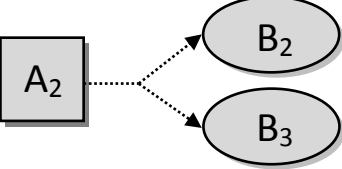
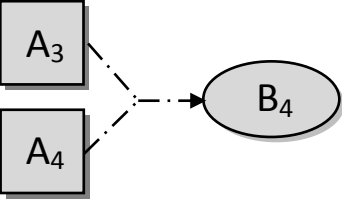
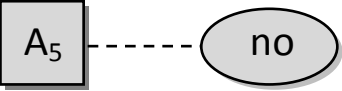
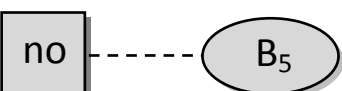
Recent evidence suggests that there are several categories of schema nodes relationships (Howard et al., 2010). Table 7.1 presents a common classification of the relationships between related schemas. The table distinguishes several correspondences based on how many members are in the relationships. For instance, the schema situation related to 'One-To-One' correspondence can be seen in the upper part of the same table. This is an optimum relation as each participating entity can only have one exclusive relationship. In other words, node that is required by the source schema can be found in the target schema in the same format; e.g. elements A_1 and B_1 in table 7.1. It is assumed that one

element (A_1) from the first group directly matches one element (B_1) from other group and vice versa.

The table also displays another relation, 'One-To-Many', that can be used to express the schema relationships. This class is different to the previous one in a number of respects. For example, in this instance there is one class in the source schema equivalent to two or more classes in the target schema. A classic example of this point is displayed in the second part of table 7.1. It can be observed that the element A_2 from the first group matches two objects B_2 and B_3 in the corresponding group. Compared to the 'One-To-One' instance, this relation can be considered as a one way or non-reverse relation. The example in the table shows that there is a difference in the construction of the relation if matching from one or the other direction. In fact this can provide another category of node instances which is commonly known as a 'Many-To-One' relationship. It is the reverse of the 'One-To-Many' relation. A multiple elements in the source schema can be similar to one category in the target schema. The table below illustrates a clear example of this point in which the two classes A_3 and A_4 are directly matched to one class B_4 .

Howard et al. (2010) also reported that missing instances between source and target schemas may occur. In general, missing correspondences can be divided into two categories: 'Target Lacks Data' when classes in the source schema have no correspondences in the target schema, and 'Source Lacks Data' when data in the target schema have no correspondence in the source schema. The evidence of these situations can be clearly seen in the last part of table 7.1. For this thesis, the tests were undertaken in order to analyse the relationships between the nodes of feature classification schemas of FM data such as OS and GDS datasets and VGI data such as OSM information, as will be described in section 7.3.

Table 7.1 A summary of schema node relationships

Representation	Interpretation
	One-To-One
	One-To-Many
	Many-To-One
	Target Lacks Data
	Source Lacks Data

7.3 Evaluating the similarity between different classes' features

7.3.1 Pre-processing phase

This part of the experimental testing was performed in two stages: firstly, a straightforward analysis 'tokenization' was undertaken in order to deal with the classes that consist of multiple words. Then, the feature classifications were modelled as a rooted labelled graph called a schema tree. This has been included to show the structural and semantic differences between category classes.

7.3.1.1 Pre-Processing of classes' names

From a linguistic perspective, a feature classification name may be either a single word, which can be called an 'atomic name', or composed of several or multiple words. Before any further GI processing such as assessing the possibility of schema integration, each compound name needs to be segmented into identifiable, individual words. This is usually carried out in order to prepare the semantic data for processing using WordNet::Similarity software. The sentences or multiple names are split up into token lists, as operation known as tokenization. Generally, there are two types of token that

can be obtained from this operation. One is related to the organisation of items such as numbers, punctuation marks, quotation marks and dates, while the other corresponds to units, such as words, which can be linguistically analysed (Grefenstette and Tapanaines, 1994).

Tokenization can be considered as one of the most important steps that may affect the final results of a variety of matching tasks. For instance, Grefenstette and Tapanaines (1994) showed that the tokenization method can influence the outcomes of complex tasks such as name entity recognition. Although there are several positive aspects of tokenization processes, such as isolating words from a text, there are some considerations that should be taken into account when the tokenization procedure is carried out. For example, the full stops that may be included in names or phrases cannot always be considered as punctuation. They may serve as abbreviation marks such as in 'G.I. ' or 'U.T.M'. They may also represent an ordinal number as in many languages such as German. For example, the number '12th' is typically written in German with a trailing period as '12.'. Thus it is necessary to distinguish between marks representing abbreviations and a mark indicating a full stop at the end of a sentence. Furthermore, Frunza (2008) reported that some languages such as French and Romanian use dots to present large numbers such as millions or thousands (e.g. 1.500,65), while the English language applies commas (e.g. 1,500.65). Hence, distinguishing between several types of punctuation marks can be very difficult in the tokenization process and an ambiguity may occur between different languages.

The analysis of the current study provides a description concerning the tokenization approach that has been applied for pre-processing semantic data of compared datasets. This involves the category names of VGI and authoritative datasets for three study areas: Cramlington-UK, Clara Vale-UK and Baghdad-Iraq. The method developed in this research uses the 'MorphAdorner' tokenizer (MorphAdorner, 2009). This is a java command-line program which provides methods for text handling such as tokenization. This software has the ability to distinguish between words or symbols based on the whitespace between them in the texts or sentences. The output is an unambiguous sequence of basic tokens such as names, punctuation marks and numbers.

The feature classifications data that are used in this project have different patterns. Some of them are single words, thus the tokenization process is not required. On the other hand, there is a need to tokenize others as they are composed of multiple expressions. Figure 7.5 presents the results obtained from the preliminary analysis of the data rates that require tokenization. It can be observed from the bar graph that 22% of OSM information and 47% of OS data in Cramlington-UK are compound classes and need tokenization. The composite of the semantic data in Clara Vale-UK is practically similar to those of the Cramlington-UK area. The rates reflect that 21% and 46% of OSM and OS datasets in Clara Vale should be tokenized. The same figure shows that all the OSM classes of the Baghdad-Iraq study area are single words and there is no requirement for the tokenization process. On the other hand, GDS data has 43% of the categories that need to be tokenized or split up. In general, therefore, it seems that most of the OSM classes are single words and they can be considered as simple categories compared to the formal datasets in which approximately half of the classes are complex terms. Even though the formal data tokenization rates are not too high (each less than 50%), the pre-processing analysis supports the fact that specific tokenization needs to be applied to both formal and informal classes' datasets before the semantic similarity approach can be undertaken.

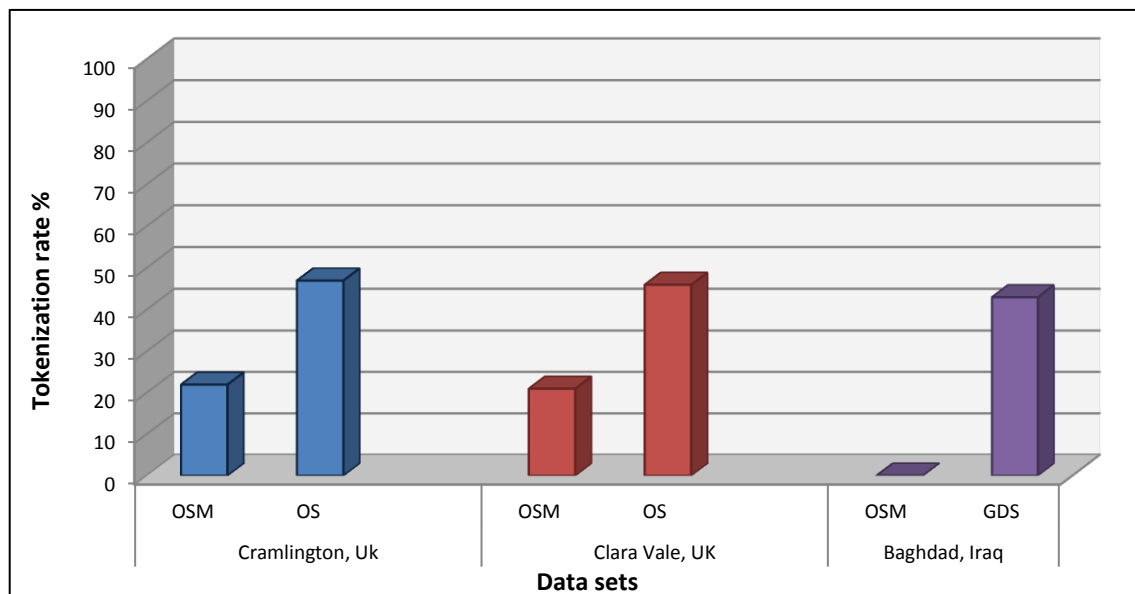


Figure 7.5 The comparison of tokenization rates of formal and informal datasets in three study areas: Cramlington-UK, Clara Vale-UK and Baghdad-Iraq

7.3.1.2 *Encoding feature classifications as an XML schema*

As a second step of pre-processing the names of classes, each feature classification list will be coded as an XML schema. Extensible Markup Language (XML) is emerging as the standard for the formatting and exchange of data over the web. In general, there are two types of XML data: XML document and XML schema. An XML document can be defined as a set of rules that is usually used for encoding a document in a standard format. An example of an XML document based on ship-order data is illustrated in Figure 7.6. From this figure, it can be seen that the root element of an XML document is the 'shiporder' element and there are three child elements included within the root element. This involves 'orderperson', 'shipto' and 'item' elements. It can be also seen from the same figure that the item element is represented two times and it consists of 'title', 'note' optional appearance, 'quantity' and 'price' elements (W3Schools, 2012). The XML schema provides and describes the data structure of an associated XML document which is defined by a set of elements and type declarations.

The data model that represents XML schema can be considered one of the significant means of measuring XML schema similarity. Several authors have suggested various approaches for modelling XML schema, as described in Erhard and Philip (2001) and Melnik et al. (2002). However, a recent study by Thang and Nam (2010) reported that the most significant approach is representing schema as a labelled graph. The hierarchical nature of the XML schema tree can be defined in many relationships such as parent-child, order relationships, or ancestor-descendant. The XML schema typically comprises a set of schema components such as element declarations, attribute declarations, simple type definitions and complex type definitions.

Figure 7.7 presents an example of an XML schema file which corresponds to the XML document of figure 7.6. It can be observed from these figures that the XML schema must define the elements that may appear in the XML document. It also must identify the number and order of child elements. Furthermore, it has the ability to define the data types of the elements and decide whether the elements contain text or not. In this research, type definitions are important. The simple data type has a simple content which can be defined as string, decimal, integer, Boolean, etc. Complex data types relate to an XML schema element that contains one or a sequence of child elements (Formica, 2008). For instance, in the XML schema example of figure 7.7, the element

'shiporder' was declared as a complex data type because it already has three direct children: 'orderperson', 'shipto' and 'item'. Each of them is declared as a different node through directed labelled edges.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<shiporder orderid="889923"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:noNamespaceSchemaLocation="shiporder.xsd">
  <orderperson>John Smith</orderperson>
  <shipto>
    <name>Ola Nordmann</name>
    <address>Langgt 23</address>
    <city>4000 Stavanger</city>
    <country>Norway</country>
  </shipto>
  <item>
    <title>Empire Burlesque</title>
    <note>Special Edition</note>
    <quantity>1</quantity>
    <price>10.90</price>
  </item>
  <item>
    <title>Hide your heart</title>
    <quantity>1</quantity>
    <price>9.90</price>
  </item>
</shiporder>
```

Figure 7.6 An example of XML document (W3Schools, 2012)

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema">
<xs:element name="shiporder">
  <xs:complexType>
    <xs:sequence>
      <xs:element name="orderperson" type="xs:string"/>
      <xs:element name="shipto">
        <xs:complexType>
          <xs:sequence>
            <xs:element name="name" type="xs:string"/>
            <xs:element name="address" type="xs:string"/>
            <xs:element name="city" type="xs:string"/>
            <xs:element name="country" type="xs:string"/>
          </xs:sequence>
        </xs:complexType>
      </xs:element>
      <xs:element name="item" maxOccurs="unbounded">
        <xs:complexType>
          <xs:sequence>
            <xs:element name="title" type="xs:string"/>
            <xs:element name="note" type="xs:string" minOccurs="0"/>
            <xs:element name="quantity" type="xs:positiveInteger"/>
            <xs:element name="price" type="xs:decimal"/>
          </xs:sequence>
        </xs:complexType>
      </xs:element>
    </xs:sequence>
    <xs:attribute name="orderid" type="xs:string" use="required"/>
  </xs:complexType>
</xs:element>
</xs:schema>
```

Figure 7.7 An example of the respective XML schema for the document of figure 7.6 (W3Schools, 2012)

As examples of the graphic representation of XML schema models, the feature classifications have been represented as an XML rooted tree for the topographic layer of the OS MasterMap and OSM dataset in Cramlington and Clara Vale in UK. In addition, the feature classifications of GDS and OSM of the Baghdad-Iraq site have also been coded as an XML rooted tree. Both the FM and VGI datasets cover the same studied regions. Each schema element or attribute is translated into a node. The XML schema editor (within Liquid XML Studio) was used to code the feature classifications into XML schema graphs for all study areas, as illustrated in Appendix C. The schematic graphs for the UK and Iraq datasets are shown in Figures 7.8 to 7.14 respectively.

From the data in figures 7.8 and 7.9, it is apparent that the structures of the feature classifications of the two datasets in Cramlington-UK are not exactly the same. For instance, most of the OSM data has four levels and a regular hierarchy, while the OS data has a more varied branching structure. It can be seen that there are some features that have the same meaning in both schemas such as 'path', but are positioned in different schema locations: 'path' is located in the fourth level of the OSM classes' tree while it is in the third level of the OS data. It is also the case that for some schema elements, it is difficult to find corresponding features of some of the OSM data, such as 'school' or 'library', in the corresponding OS datasets.

The feature classification schemas in figures 7.10 and 7.11 represent the OSM and OS data in Clara Vale-UK. The comparison indicates that, as for the other case studies, some features in both datasets are the same, but differences may occur in the classifications, categories and concepts definitions. For example, it can be noticed from the figures that there are some concepts such as 'track' that have the same meaning and the same level in both schemas. Furthermore, there are some concepts in the OS data that correspond to different levels of the OSM schema. For instance, 'rail' is placed in the third level of the OS schema, but 'rail' is in the fourth level of the OSM schema. For the term 'path' in the OS tree graph there is no exact equivalent term in the OSM schema, but there is a synonym concept, 'footway'. Furthermore, for some features in the OSM, such as 'playground' or 'garden', there is no correspondence at all in the OS data and vice versa.

For the Baghdad-Iraq datasets, the schemas were modelled as shown in figures 7.12 to 7.14. Figure 7.12 illustrates the schema development of OSM information in the English language, while figure 7.13 presents the XML schema tree of the GDS feature classifications in the Arabic language. As the XML schema matching process is fundamentally based on measuring the semantic similarity of corresponding labels, as will be discussed in the next subsection, it is impossible to integrate two schemas that have been created in different languages. Thus all the tag names of the Arabic schema have been translated into the English language using the Oxford dictionary (2010), a well-trusted source of translation. The translated schema can be seen in Figure 7.14. Similar observations of relationship and matching of the UK study areas can be made for the comparison of the GDS and OSM schemas. The semantic and structural similarity matching derived from a study of XML schema can play an increasing role as a measure of the possibility of classification schemas' integration for different data sources.



Figure 7.8 XML Schema for feature classifications of OSM information in Cramlington-UK

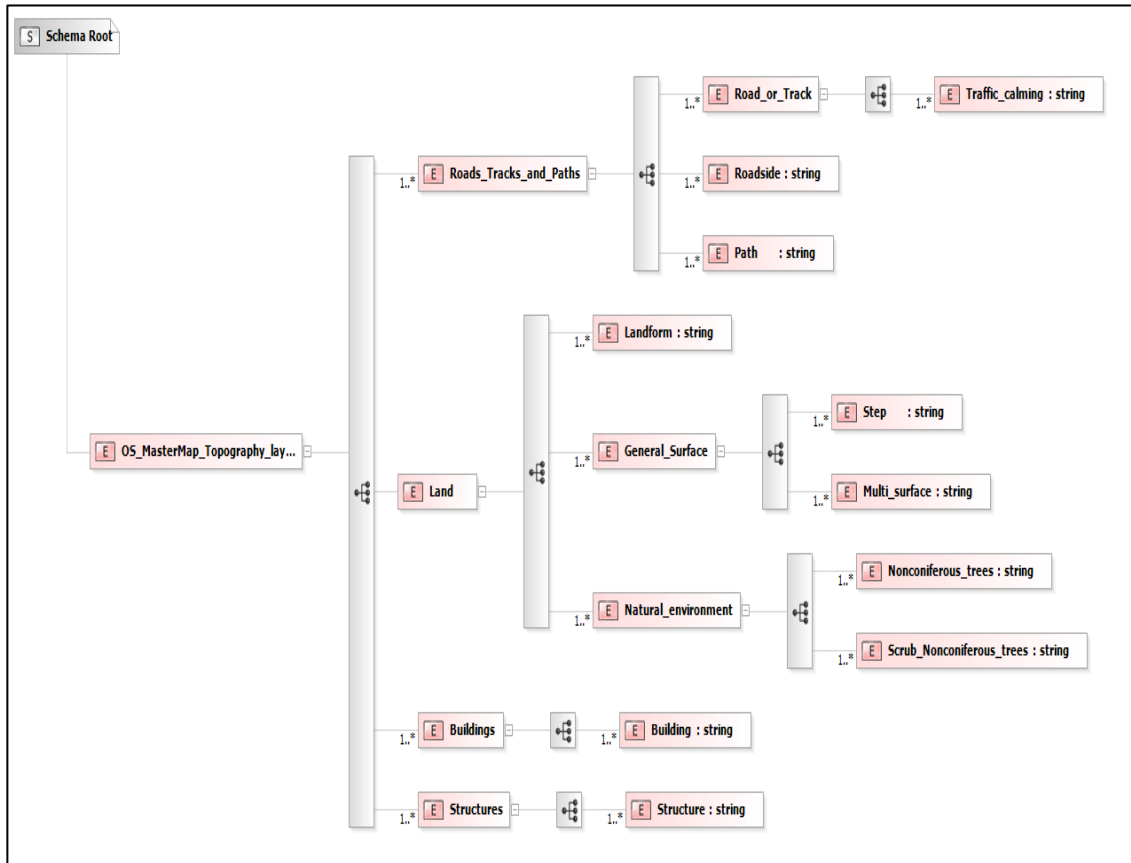


Figure 7.9 XML Schema for feature classifications of OS datasets in Cramlington-UK

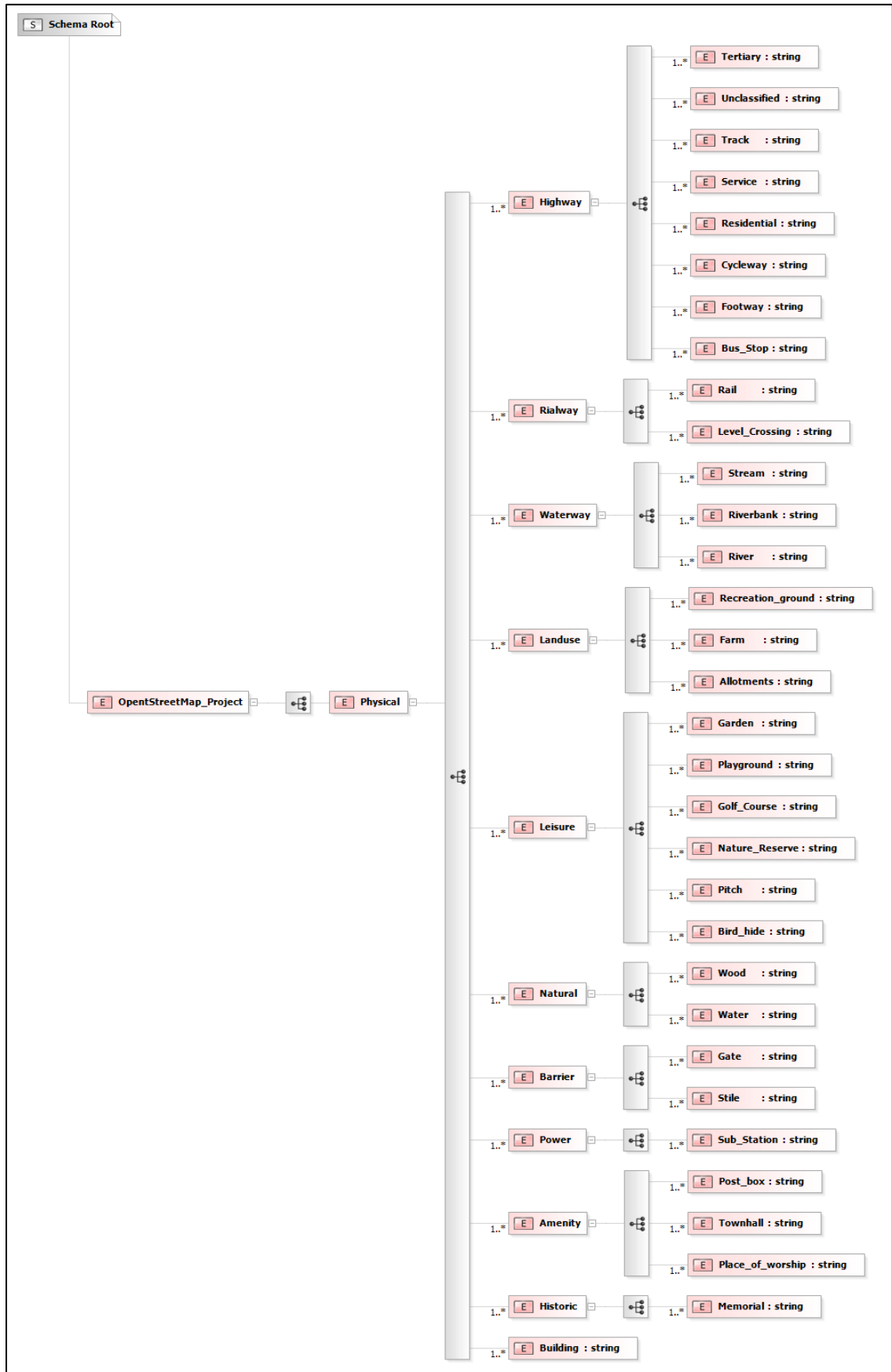


Figure 7.10 XML Schema for feature classifications of OSM information in Clara Vale-UK

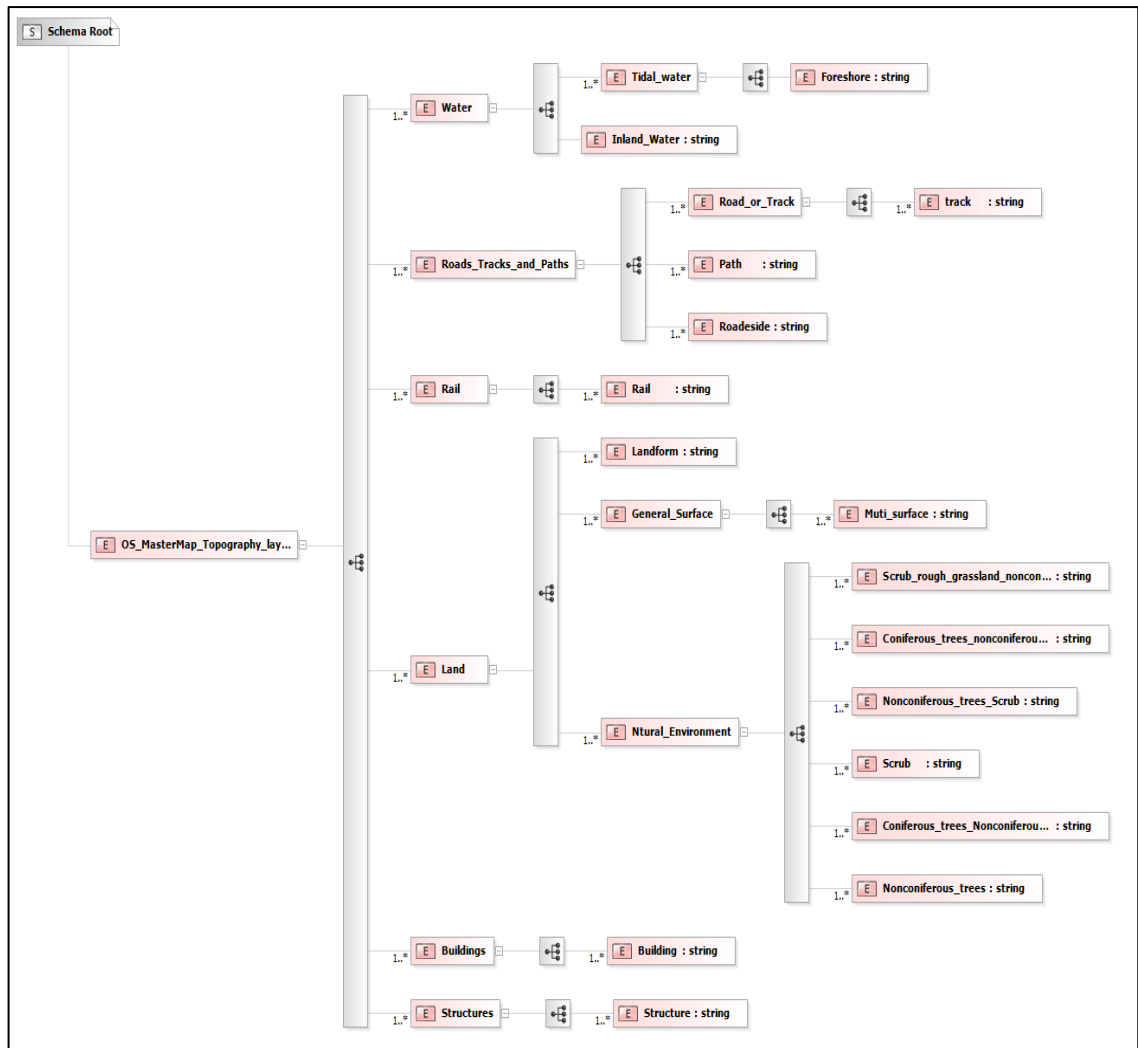


Figure 7.11 XML Schema for feature classifications of OS datasets in Clara Vale- UK

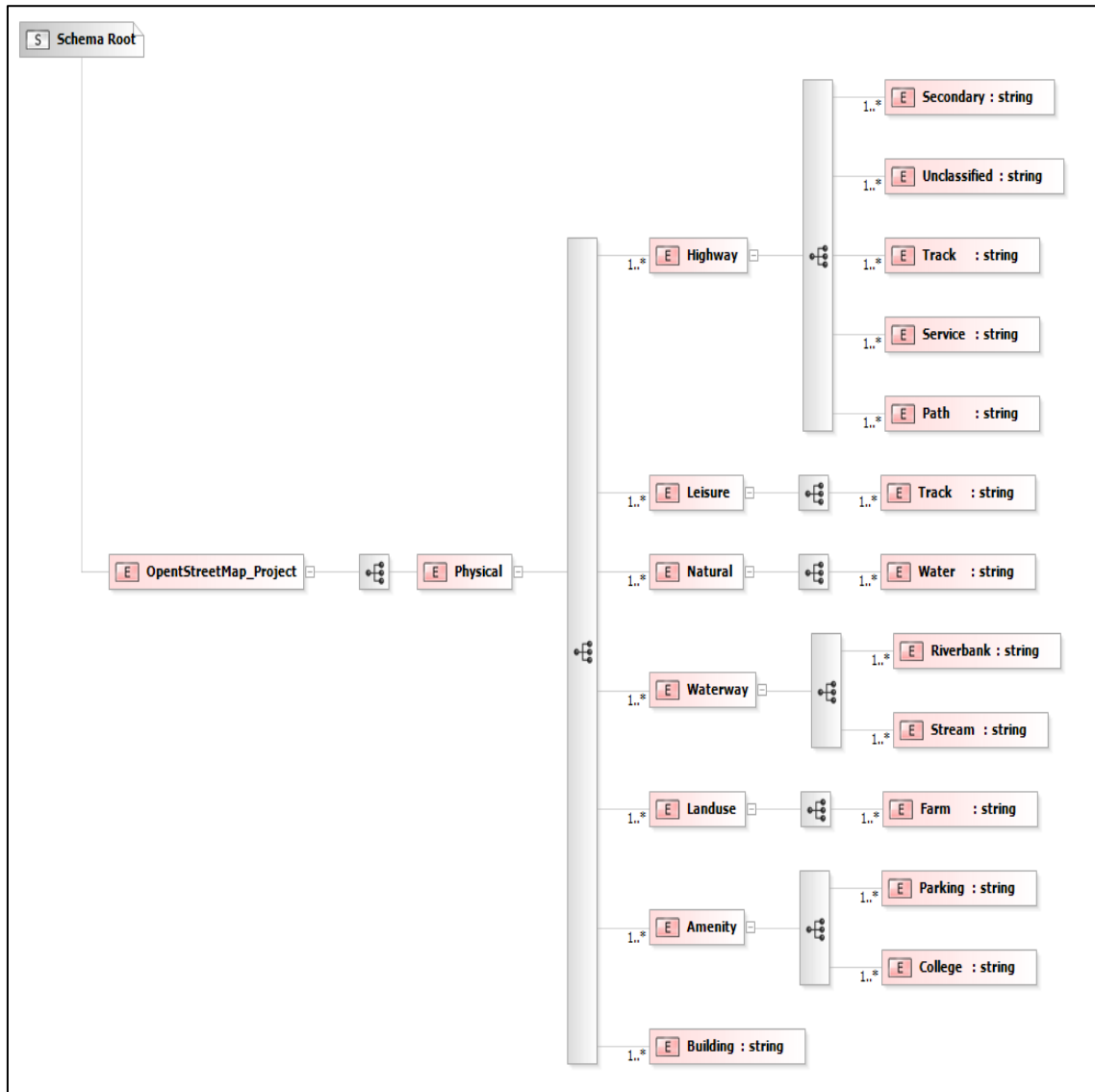


Figure 7.12 XML Schema for feature classifications of OSM information in Baghdad-Iraq

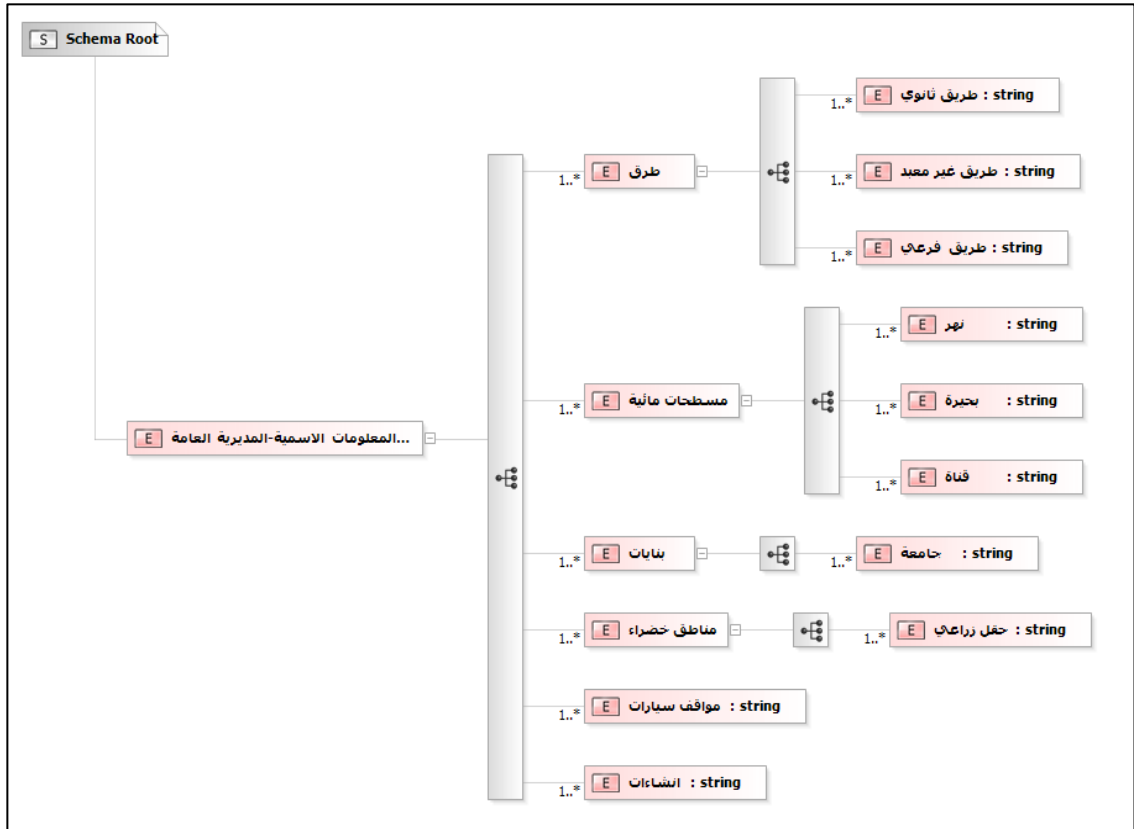


Figure 7.13 XML Schema (in Arabic) for feature classifications of GDS datasets in Baghdad-Iraq

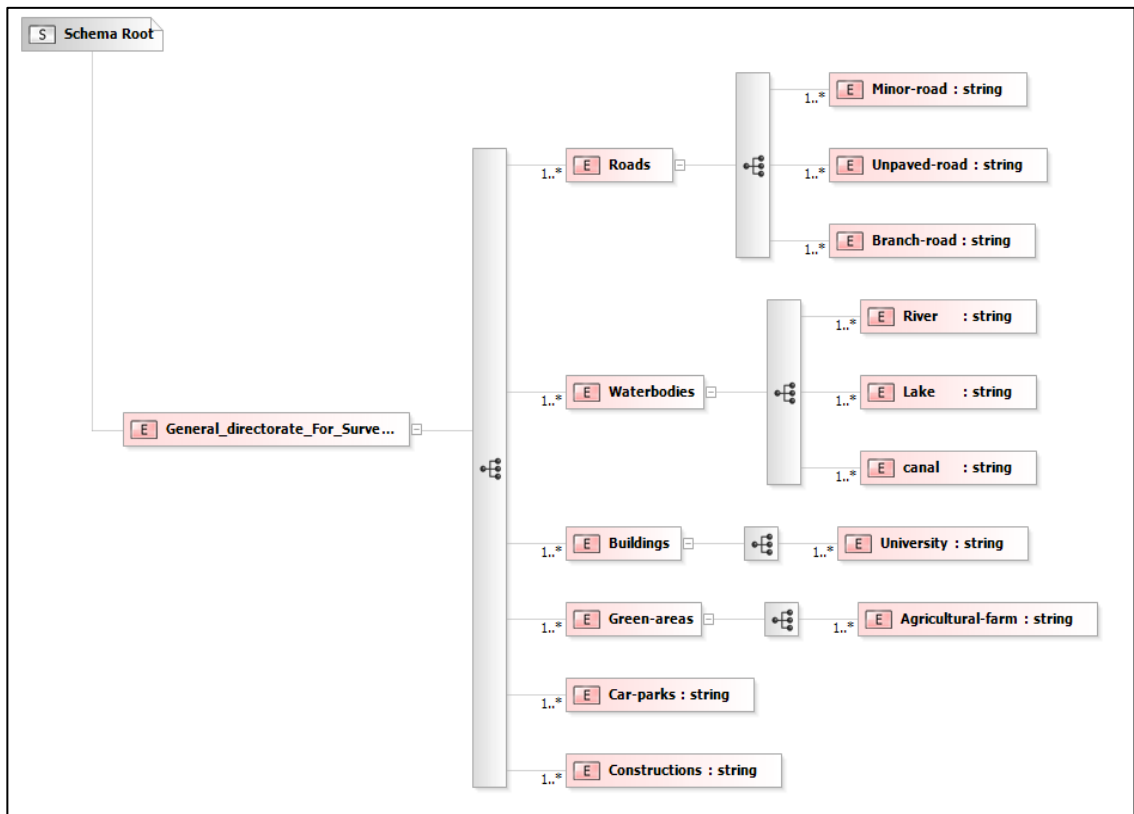


Figure 7.14 XML Schema (in English) for feature classifications of GDS datasets in Baghdad-Iraq

7.3.2 Node similarity measurement phase

In this stage of the research, many similarity evaluation approaches are used together to assess the integration of schema feature classifications from official and informal spatial data sources. Firstly, the semantic similarity is measured between the classes of compared schema nodes. The structural similarity between the compared schema trees is also taken into account. In addition to these measurements, data type similarity is determined and considered. In order to obtain the overall similarity score, the individual similarities are combined together.

7.3.2.1 Label name similarity

The tag names in different schemas may be semantically similar or different; consequently, it is important to include the name similarity measure to compute the degree of similarity between the nodes of two schemas. Prior to commencing this processing, each composed class should be normalized into a token list, 'a list of words', in order to make it comparable, as was discussed in section 7.3.1.1. After breaking the classes into a list of tokens, the similarity between two names can be computed by determining the similarity of those two token lists. This can be carried out by calculating the average of the best similarity measure for each source token with a target token as follows (Tansalarak and Claypool, 2007):

$$N_{sim}(N_1, N_2) = \frac{\sum_{l_1 \in L_1} [\max_{l_2 \in L_2} sim(l_1, l_2)] + \sum_{l_2 \in L_2} [\max_{l_1 \in L_1} sim(l_2, l_1)]}{|L_1| + |L_2|} \quad (7-1)$$

Where:

$Sim(l_1, l_2)$: is the semantic similarity measure between each pair of tokens.

$|L_1|$ and $|L_2|$: is the size of token sets of two words (N_1, N_2) where the semantic similarity calculation between them should be performed.

This operation can assist in determining how linguistically close the names of the two nodes are to each other. For the current work, the similarity between the category words was computed by means of the WordNet::Similarity software package. The output of this processing is a mark ranging between 0 and 1. The score reflects the name matching strength of the compared words. The low values correspond to a difference of compared

names, while high values indicate similar names (i.e. 1 corresponds to completely identical names).

As most terms in the WordNet::Similarity database have many definitions, the best match has been adopted here. This involves recording the maximum value of the semantic scores from a matching of every two tested features. For instance, the term 'track' has ten definitions in WordNet::Similarity ontology and the term 'path' has four definitions. In this case, when submitting the two terms to the WordNet::Similarity software without any pre-defined senses, definition number 1 of 'track' will match optimally with definition number 4 of 'path'. The optimum matching means the semantic similarity score is one, and this value has been recorded for this example.

For the classes of schemas in Figures 7.8 to 7.14, matrices of semantic similarity scores were created. A valid semantic relationship was assumed when the semantic similarity score of the compared terms was higher than 0.5. The semantic correspondences of schema node classifications in the UK and Iraq study areas were examined. The results obtained from these analyses are illustrated in tables 7.2 to 7.4 respectively. The data of one-to-one relations from the tables below reported that only 2%, 0% and 5% of the features in compared FM and OSM schemas have an optimum single relation in the Cramlington-UK, Clara Vale-UK and Baghdad-Iraq sites respectively. It is somewhat surprising that the rates of the one-to-one relations are very low in all three cases of this study. In fact there is no one-to-one relation in the Clara Vale-UK site. This indicates that most recorded semantic similarity scores for a unique relation between the source and the target classification schemas are less than 0.5. The values of one-to-one relation suggest that a weak link may exist between single classes in each of the compared datasets.

As can be seen from the same tables, the multiple node relations reported confused relations which are indicated by the significant rate values. For example, the rate values of the relation one-to-many are 46%, 43% and 26% for the three study areas respectively. For the UK case studies, approximately half of the nodes of one schema match two or more from other schema, whereas about a quarter of the nodes in the source schema are similar to two or more nodes in the target schema in the Iraqi site. Also, it can be observed that nearly half of the nodes in all the study areas have many-

to-one relations (it is necessary to note that some classes may appear in both one-to-many and many-to-one relations). In addition, there are considerable numbers of feature classes in one schema which have no correspondence in other schema. This was reflected in the results for the categories of missing correspondences. The findings of these analyses are rather disappointing. The results confirm that the rates of the linguistic similarity between the feature classifications nodes of the formal sources such as OS or GDS datasets and VGI information such as OSM are very low. Thus these results were not very encouraging with regard to integrating their classification schemas.

Table 7.2 Results of schema relationships in Cramlington-UK

Node relations	Rate (%)
Single correspondences (One class in Source-To-One class in Target)	2
Single correspondences (One class in Source-To-Many classes in Target)	46
Single correspondences (Many classes in Source-To-One class in Target)	52
Missing correspondence (Source Lacks Data)	46
Missing correspondence (Target Lacks Data)	35

Table 7.3 Results of schema relationships in Clara Vale-UK

Node relations	Rate (%)
Single correspondences (One class in Source-To-One class in Target)	0
Single correspondences (One class in Source-To-Many classes in Target)	43
Single correspondences (Many classes in Source-To-One class in Target)	57
Missing correspondence (Source Lacks Data)	43
Missing correspondence (Target Lacks Data)	39

Table 7.4 Results of schema relationships in Baghdad-Iraq

Node relations	Rate (%)
Single correspondences (One class in Source-To-One class in Target)	5
Single correspondences (One class in Source-To-Many classes in Target)	26
Single correspondences (Many classes in Source-To-One class in Target)	47
Missing correspondence (Source Lacks Data)	47
Missing correspondence (Target Lacks Data)	29

7.3.2.2 *Structural similarity*

As discussed earlier, element names play a fundamental role in XML schema similarity measurement. However, identical concepts or similar words of any given two schema graphs may be structured differently. Hence, Cao et al. (2010) argued that structural relationships measurement is another essential element that should be considered for effective schema matching. It is necessary here to clarify exactly what is meant by 'structural similarity'. Throughout this thesis, the term 'structural similarity' is used to refer to computing the similarity between two labels as a function of the length of the distance in schema hierarchy. It also indicates the position or the depth of nodes in a schema graph.

There have been many approaches to measuring the structural similarity between two labels in the same schema tree, as has been shown in Chapter 6 and will be discussed in this chapter. For example, Rada et al (1989) suggested a method for measuring structural similarity between two nodes in the same schema. Their approach defined 'distance' and is a technique to measure the average minimum path length over all pairs' combinations of nodes for two different subsets of nodes. However, the present research is aiming to measure the similarity between two elements in two different schemas. Thus, a different structural similarity approach has been followed and it was essentially based on the semantic similarity value of the context of the node of two different schemas (Amarintrarak et al., 2009). Nodes are set in context by assessing their relationship with children, parents and siblings. In order to apply this approach, a series of requirements are needed. The number of child elements that are semantically similar should be determined. The number and the position of ancestor and siblings nodes in both source and target schemas should be known. Formally, structural similarity can be expressed as following:

$$StrSim(n_1, n_2) = (1/p) \times (sls/sib) \quad (7-2)$$

Where:

p : is the shortest distance between the parent and the child node n_1 which is semantically similar to the parent of node n_2 . Amarintrarak et al. (2009) proposed that the shortest distance length can be considered acceptable if it is

only for less than 10 levels. However, if the number of levels is more than 10, they argued that the similarity will be very small and the p value will be set to 10.

s/s : is the number of n_1 and n_2 siblings which have semantic similarity.

sib : is equal to the greater number between n_1 and n_2 siblings.

Figures 7.15 to 7.17 illustrate sub trees from the schemas in Figures 7.8 to 7.14. Consider, for example, structural similarity measurements are required for the following scenarios. In Figure 7.15, the comparison could be between the 'path' of OS data and the 'path' of OSM information. From Figure 7.16 the measurement could be between the 'inland water' of OS data and the 'stream' of OSM information. From Figure 7.17 the assessment could be between the 'unpaved road' of GDS data and the 'track' of OSM information.

For the first instance, the structural similarity was 0.02. This was computed by applying equation 7.2. It is essentially based on the semantic similarity value of the parents of the compared nodes which was 0.44. Following the above technique the p value should be 10. This is because the semantic similarity value of the two parents is low. The number of siblings or sib value is 10, as there are 10 siblings for the 'path' node in the OSM data. The s/s value is 2, as only two siblings of 'path' nodes in each schema have semantic similarity (exceeding the 0.5 threshold). It can be observed that, although the names of the two nodes are exactly the same ('path') and their semantic similarity is one, their structural similarity is extremely low.

The structural similarity of the second case was 0. A possible explanation for this might be that there is no semantic similarity between the siblings of 'inland water' of OS data and the siblings of 'stream' of OSM information. Thus the s/s value is 0. Although the parents of the compared nodes have semantic similarity (0.68) and the sib value is 2, the overall structural similarity will be zero when applying equation 7.2.

On the other hand, when equation 7.2 was used for computing the structural similarity of the last situation, the structural similarity result was 0.25. Compared to the structural similarity values of the other two cases, the value of this test is comparatively higher, although still low overall. This is probably because the p value is 1. The two parents

(roads and highway) of the compared elements (unpaved road and track) are semantically similar (0.85). Following the approach above, the p value should be the shortest path between the parent and the compared node, and in this case it is 1. Although the number of siblings or sib value is 4, only one sibling is semantically similar with the other group of siblings. In other words the s/s value is 1. This may be the reason for the low structural similarity of this case.

This combination of findings provides some support for the conceptual premise that the structural similarity varies for each individual feature pairs in compared schemas and its value is based on the different elements. The present study was designed to determine the structural similarity between the feature classification schemas of FM data such as OS and GDS datasets, and VGI data such as OSM information. The results revealed that the structural similarity rates are very low. This is mainly due to the fact that semantic similarity between the compared datasets is already low.

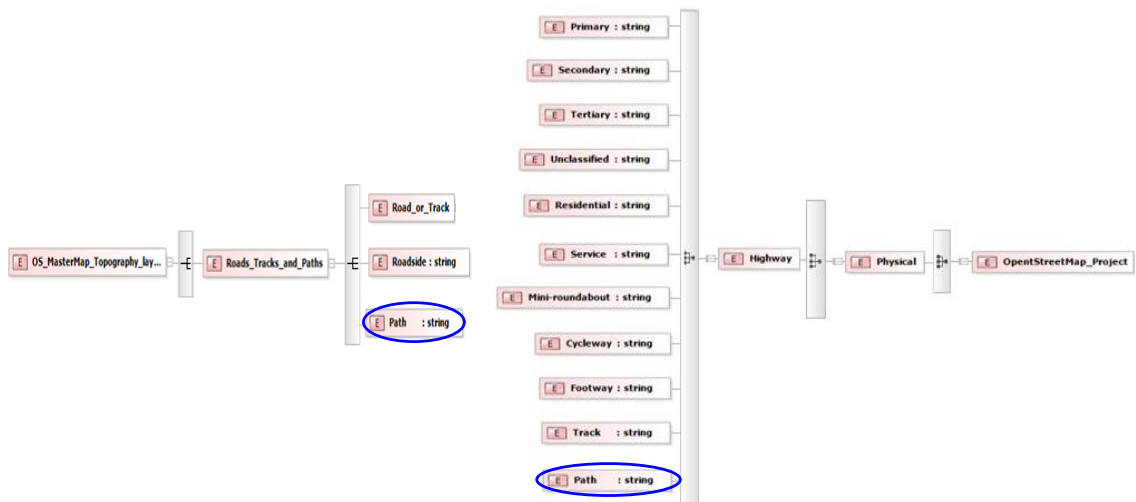


Figure 7.15 An example of structural similarity for part of OS and OSM schemas in Cramlington-UK



Figure 7.16 An example of structural similarity for part of OS and OSM schemas in Clara Vale-UK

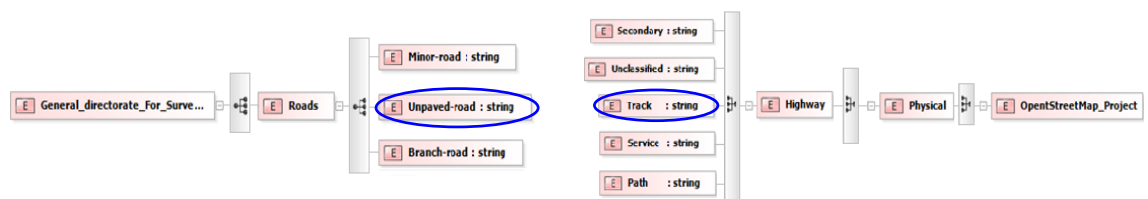


Figure 7.17 An example of structural similarity for part of GDS and OSM schemas in Baghdad-Iraq

7.3.2.3 Data type similarity

The name and structural similarity can be considered as most important when comparing XML schemas. However, they are insufficient by themselves for overall similarity measurement. For some instances, the semantic and structural similarities between schema nodes exist, but they have different data type. This similarity may be known as a false positive similarity, as 'data type' may contribute in determining similarity between different schemas. Algergawy et al. (2009) reported that in order to reduce some of these false positive similarities, it is necessary to include other schema information measurements such as measuring data type similarity. As discussed in section 7.3.1.2, the elements of any XML schema may be composed of different data types and formats. In general, the XML schema elements can be either a complex type which is a parent, or a derived node which is a child of the parent with no children itself. In any XML schema graph, only the atomic types or leaf nodes have data type. XML schema supports 44 built-in data types, as can be seen in Figure 7.18.

The creation of a data type similarity table is one of the more practical ways of determining data type similarity. This is initially based on the approach proposed by Hong-Minh and Smith (2007). In order to measure two data types' compatibility, they suggested a function that can be applied on the data type tree of Figure 7.18. Formally defined, the compatibility (c) between two data types d_1 and d_2 can be calculated as follows:

$$c(d_1, d_2) = e^{-\beta l} \times \frac{e^{\alpha h} - e^{-\alpha h}}{e^{\alpha h} + e^{-\alpha h}}, \quad \text{where } d_1 \neq d_2 \quad (7-3)$$

$$c(d_1, d_2) = 1, \quad \text{where } d_1 = d_2 \quad (7-4)$$

Where:

d_1 and d_2 : are the data types of node one (n_1) and node two (n_2) respectively.

l : is the shortest path length between d_1 and d_2 .

h : is the depth of the subsuming node of d_1 and d_2 .

α and β : are tuning parameters (determined experimentally by Hong-Minh and Smith (2007) to both equal 0.3057).

Table 7.5 illustrates a portion of data type similarity results. The evidence shows that the data type similarity value lies in the range of zero to one. It can be observed from

the table that the data type similarity is set to one for the elements that have the same data type such as the T_1 : string and T_2 : string. Furthermore, data type similarity values can vary for the different categories of data types, as can be seen from other examples in the same table. For the datasets of the current study, the nature of most of the data type comparison was String:String. Therefore, the data type similarity between them yields a value of 1.0. However, there were some similarity values that were set to 0.0 which reflect Simple:String and Complex data types.

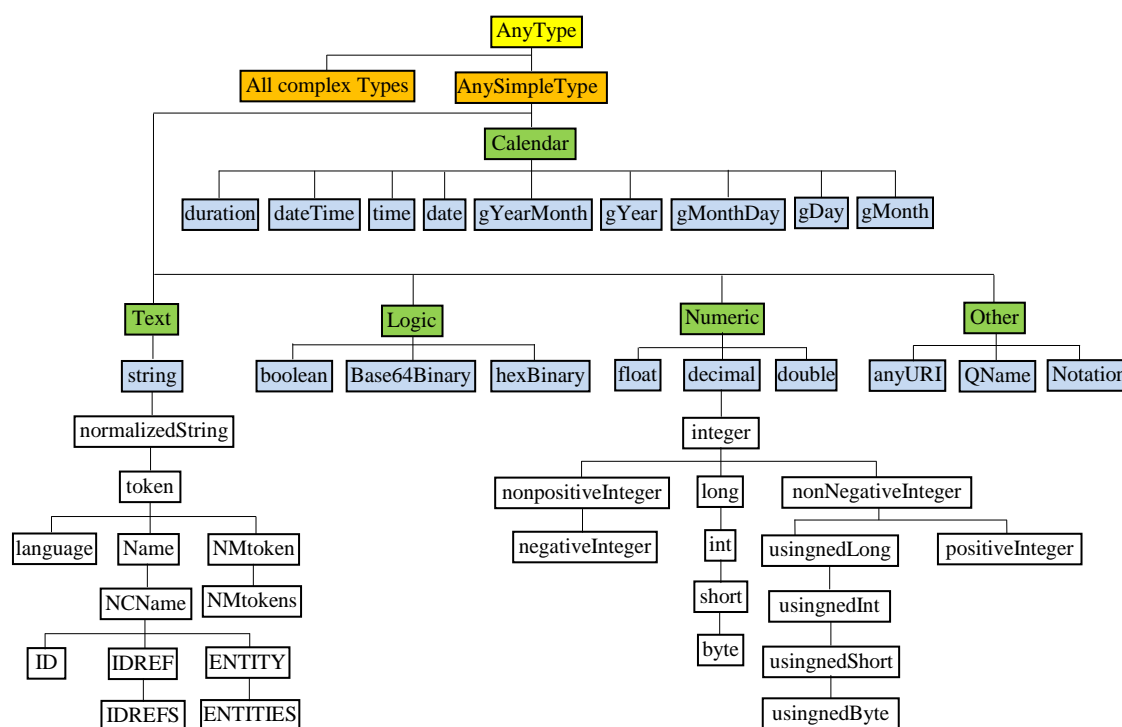


Figure 7.18 The hierarchy of XML schema data types (Hong-Minh and Smith, 2007)

Table 7.5 Portion of data type similarity

Type 1	Type 2	Similarity
String	String	1
String	Token	0.4
Token	Name	0.7
Float	Integer	0.2
Double	Decimal	0.3
.....

7.3.2.4 Similarities combination

The previous subsections have described how each similarity measurement represents a specific schema element. For instance, the label name similarity was applied only for the semantic element; the structural similarity was used to measure the paths' length between the compared nodes; and data type similarity measurement only assessed the data type element. The relative metrics of each similarity measure have the ability to examine the relationship between different components of spatial data schemas. However, using these individual components for evaluating the possibility of schema matching is insufficient for most comparisons. Thus, it is necessary to combine all these similarity measurements in order to assess the ability of schema integration.

The combination of similarity elements can be performed by different methods. For instance, consider three elements a, b and c. The similarity combination can be obtained by taking the average of them; by considering the maximum measure of the three elements; by additive combination (e.g. $(1 - (1 - \text{similarity}_a)(1 - \text{similarity}_b)(1 - \text{similarity}_c))$); or by a weighted method which is recommended by Kim et al. (2008). It can be expressed as follows:

$$Sim(A, B) = (W_N * N_{sim}) + (W_S * S_{sim}) + (W_T * T_{sim}) \quad (7-5)$$

Where:

N_{sim} , S_{sim} and T_{sim} : are label name similarity, structural similarity and data type similarity respectively.

W_N , W_S and W_T : are similarity weights, $W_N + W_S + W_T = 1$

The weight has been utilised in order to determine the importance of each individual similarity. Kim et al. (2008) pointed out that this value can be acquired from human semantic mapping data or it can be obtained from the domain experts. For example, Amarintrarak et al (2009) assigned the highest weight value to the name similarity ($W_N=0.5$). They considered the name similarity to be a vital element for measuring the overall similarity. The structural similarity is weighted as ($W_S=0.35$). It is the next most important element and is used to decide the correctness of the element position in a schema graph. The weight value of data type similarity is set out as ($W_T=0.15$). This measure might be less important than the other two similarities.

In the present study, the combining of semantic, structural and data type similarities was undertaken and the weighted approach was adopted. The analysis of combined similarity for OS, GDS and OSM schema classification features for the UK and Iraqi sites are presented in tables 7.6 to 7.8 respectively. These findings are rather disappointing, as they reported very low rates for the node relationships and very high rates for the missing correspondences. It is apparent from the tables that the rates of One-To-One relations, for example, are less than 20% for all study areas, while they are more than 65% for all the missing correspondences (source lacks data). One of the more significant findings to emerge from this part of study is that it is difficult to integrate the feature classification schemas of FM datasets such as OS or GDS data with VGI sources such as OSM datasets.

Table 7.6 Results of combined similarity between OS and OSM classification in Cramlington-UK

Similarity relation	Rate (%)
Single correspondences (One -To-One)	2
Single correspondences (One-To-Many)	2
Single correspondences (Many -To-One)	13
Missing correspondence (Source Lacks Data)	85
Missing correspondence (Target Lacks Data)	65

Table 7.7 Results of combined similarity between OS and OSM classification in Clara Vale-UK

Similarity relation	Rate (%)
Single correspondences (One -To-One)	5
Single correspondences (One-To-Many)	10
Single correspondences (Many -To-One)	19
Missing correspondence (Source Lacks Data)	74
Missing correspondence (Target Lacks Data)	58

Table 7.8 Results of combined similarity between GDS and OSM classification in Baghdad-Iraq

Similarity relation	Rate (%)
Single correspondences (One -To-One)	16
Single correspondences (One-To-Many)	16
Single correspondences (Many -To-One)	16
Missing correspondence (Source Lacks Data)	68
Missing correspondence (Target Lacks Data)	50

7.4 Chapter summary

The lack of semantic and structural similarity of multi-source spatial datasets is one of the most important issues of heterogeneity in spatial datasets. This can particularly affect all forms of geospatial data integration processing. The differences among the nodes of feature classification schemas, and also the relationships between their meanings and the structures, can be considered as the main concerns of successful geospatial data matching. This chapter has illustrated in detail the approach used in this project, leading to an evaluation of the possibility of integrating feature classifications from national mapping agencies such as OS and GDS, and VGI sources such as OSM data. The various processes included feature by feature comparison and similarity measurement among feature classifications schema trees. Most of the pre-processing was successfully performed within the MorphAdorner tokenizer and XML schema editor (within Liquid XML Studio). The semantic similarity measurement was carried out using the Lin method in the WordNet::Similarity software.

The initial stage of analysis involved measuring semantic similarity between corresponding features of tested datasets. Specific statistical tests were applied to assist in determining population means of semantic similarity scores. The subsequent results of statistical tests led to a decision regarding the ability of feature classifications matching. In a one-sample t-test, the population mean of semantic similarity scores for each study area were tested against the hypothesised semantic similarity value of 0.5. The results revealed that the population means of all tested areas were less than the hypothesized value. Thus, it can be concluded that the integration of feature classifications from official and informal sources would be a difficult task.

This chapter also set out with the aim of assessing the possibility of semantic and structural integration of feature classification schemas of authoritative and VGI datasets. The approach included tokenization in order to break up composed classes into a single word. The capabilities provided by the MorphAdorner tokenizer in handling sentences or multiple words helped to ease the acquisition of single words. The results showed that most of the FM classifications, such as OS and GDS data, are composite and need tokenization. In contrast, most of OSM categories are single words. Therefore tokenisation may only be necessary for a few of them. Subsequently, the feature classifications for all compared data were coded as an XML schema via Liquid XML Studio. Figures 7.8 and 7.14 showed that the structure of feature classification schemas for FM and OSM is different for all comparisons. The differences may occur in the names of the labels, the structures of schema sub-trees, or even in data type formats.

Semantic similarities were measured between the corresponding schema nodes using WordNet::Similarity software. The Lin method was used to evaluate the semantic similarity between compared data. Tables 7.2 to 7.4 illustrated the evaluation results for schema nodes comparison. The results revealed that the rates of One-To-One relations are very low for all datasets. Furthermore, the One-To-Many and Many-To-One instances demonstrated confused relationships. On the other hand, the missing correspondence rates were relatively high. Most of the recorded similarity scores were less than 0.5 which can be considered to be disappointing values for the process of successful schema matching.

The structural similarity measurement between the schema trees was based on semantic similarity scores of the ancestors and siblings of the compared nodes. The approach also depended on the number of siblings of the node. The findings indicated that the rates of structural similarity of the tested schemas were extremely low. The next step was to compute the data type similarity between corresponding nodes in different schemas. This was basically determined by creating a table of data type similarity. For this project, the data types of the compared schema elements were String:String or String:Complex data types, therefore the outputs were 1 or 0.

Once these three similarities had been computed, a weighed combination technique was adopted in order to obtain the overall similarity. What is surprising is that in all three

cases of this study the rates of the similarity combination of node relationship were very low, while the missing correspondence rates were very high. These results were not very encouraging for integrating feature classifications from official sources such as OS and GDS datasets and informal data such as OSM information.

Chapter 8 Conclusions and Recommendations for Further Work

8.1 Thesis overview

This research has considered the development and design of a set of analytical steps to see if it is possible to integrate spatial datasets from official and volunteer geographic information (VGI) data sources. The evaluation of spatial data quality can be utilised to generate a priori knowledge for the assessment of the possibility for geospatial data integration processing. The geometrical (positional and shape) and semantic similarity requirements for geospatial data integration purposes were evaluated. Different geospatial data integration assessment methodologies were developed and examined in this thesis. The National Standard for Spatial Data Accuracy (NSSDA), directional discrepancies analysis procedures, and traditional positional accuracy assessment methods, were demonstrated to check the suitability of formal and informal positional dataset integration. For shape similarity measurement (linear and area shape), the buffer overlay and moments invariant models were adopted. Subsequently, the potential effects of various factors on informal geometrical data quality were investigated by developing and implementing an experiment by applying factorial design methodology. The semantic similarity measurement was analysed by means of developing models for feature by feature comparison and XML schema tree classifications combining. This specifically included semantic similarity, structural and data type similarity assessment. The implementation of the developed methodology was based on different sites and different data sources in the UK and Iraq.

The need for assessing the possibility of a multi-source geospatial data integration process was described in Chapter 1. An overview of the effect of spatial data quality on geospatial data integration tasks, concepts and issues of geospatial data integration with its applications was included in Chapter 2 in order to provide a comprehensive background of considerations in the geospatial data integration process. An illustration of the characteristics of formal and informal data sources, and the main differences between the two concepts was examined in Chapter 3. The methodology of positional and shape similarity measurement with the developing of practical interfaces to achieve the calculation and present the similarity outputs was explained in Chapter 4. Chapter 5 examined the factorial design approach for discovering the proportion of the effect of

different factors on the geometrical quality of VGI datasets. In Chapters 6 and 7, the models and approaches for measuring similarity measurement with the outcomes of assessing the integration feature classifications from official and VGI datasets were discussed and demonstrated.

Conclusions regarding the ability of integration of formal and VGI datasets, and the potential effect of different factors on VGI data quality, based on the methodologies and models of geometrical (positional and shape) and semantic similarity measurements, in addition to the experimental design procedure, are drawn and discussed in this chapter. Subsequently, revisiting the objectives of the research, as presented in Chapter 1, is addressed in this chapter in order to see if they have been achieved. This is followed by suggestions for future work in order to improve this research area.

8.2 Evaluation of the thesis objectives

This thesis has developed tools and models capable of assessing the integration of formal and VGI datasets. The research flowline started with developing tools for assessing geometrical similarity measurements (positional, linear and area shape) in addition to assessing the various factors' effects on VGI geometrical data quality. It also included the assessment of semantic similarity measurements, as explained in Chapters 4, 5 and 7. In Chapter 1 of this thesis several objectives were defined and the following paragraphs are concerned with determining whether these objectives were achieved.

Generally speaking spatial data users can save cost, time and effort by employing various forms of geospatial data integration processing. However, one of the fundamental issues that may face successful geospatial data integration is spatial data quality. By increasing the availability of free spatial datasets such VGI data, the geospatial data integration process has become more complex as this data is created and shared on the web by non-professional people. Therefore, a general overview of data quality issues was discussed in order to assist in understanding the impacts of these issues on the geospatial data integration process. The descriptions illustrated in this thesis also included the different applications and concepts of geospatial data integration processing, in addition to presenting some existing attempts at formal and VGI data integration research, to build a strong background on what is going on in the research area.

The thesis examined the nature of formal and VGI datasets from the perspective of geospatial data integration. The role of producing formal spatial datasets (e.g. topographic mapping) for different national mapping agencies such as the OS in the UK and GDS in Iraq was highlighted. A detailed discussion of contemporary data sources, such as user generated content (UGC) in general and VGI as a special case of it, was addressed. Several examples of VGI data were illustrated and OSM was employed as an informal spatial dataset as it provides the most specific spatial data (coordinates and feature classifications) compared to other VGI data sources. Limitations of VGI data regarding the quality of spatial data were highlighted. In contrast to official datasets, the main limitations were the production of spatial data by untrained contributors and the absence of full metadata both of which may affect the quality of the VGI data.

A computer interface programme was developed to assess geometrical or physical similarity, such as positional, linear and area shape; measurements among these compared datasets for evaluating of the possibility of formal and VGI data integration. The interfaces successfully read the input data from import files, calculated the geometrical similarity and displayed the outputs as three parts: numerical, graphical and visual representation for each geometrical similarity element. The geometrical similarity methodology mentioned in the previous section was adopted to assist various geometrical similarity measurements.

A factorial design approach was used to design a practical experiment for assessing the potential effects of various factors on VGI data quality and to check if there is any interaction effect between selected factors. Three factors were selected and studied with two levels of each factor, as described in subsection 5.3.2, based on the importance of factors and establishment of priorities as explained in section 5.4. Such experimental implementation can assist in recognising the factor which has the most influence on the geometrical quality of VGI data and consequently affecting the integration with official datasets. Accordingly, the VGI community can be informed to take care over this factor in order to improve the overall VGI data quality.

From the descriptions presented in Chapter 2, one of the most important issues of inconsistency in spatial datasets is the lack of semantic similarity between compared datasets. Consequently, this can affect the integration process of multi-sources of

geospatial data integration. Since the aim of this research is to assess the possibility of geospatial data integration from different data sources, models for assessing the semantic similarity measurements between formal data (FM) and VGI datasets for geospatial data integration purposes were proposed and developed. The models looked first at how semantic similarity, feature-by-feature, of two datasets can be compared and measured (subsection 7.2.1). Then the developed models investigated the assessment of the integration of feature classification schemas of FM and VGI datasets (subsection 7.2.2). The developed models therefore addressed the semantic similarity measurements issues in the thesis.

Different study areas and feature types were successfully used to test the research methodology. For each part of the research, various results and conclusions upon the importance of the findings were obtained, analysed and discussed. Therefore, the objectives of the research set out in section 1.2 were addressed and consequently the main aim of this research was also achieved.

8.3 Major conclusions of the thesis

There are several outcomes and findings that have emerged from each analysis and experiment in this research which may be mainly divided into: the assessment of geometrical integration of formal and VGI datasets, the factors affecting the geometrical quality of VGI datasets and assessment of the semantic similarity of formal and VGI datasets for integration purposes.

8.3.1 The assessment of geometrical integration of formal and VGI datasets

This thesis proposed developing tools for this purpose based on the methodology described in Chapter 4. These tools were implemented using the Matlab environment. The tests included different spatial data sources: reference field survey data (FS), formal data (FM) such as data from the Ordnance Survey (OS) UK and General Directorate for Survey (GDS) Iraq, and VGI sources such as OpenStreetMap (OSM) information. The findings of this part of the research will be discussed as three groups of positional, linear and area shape similarity as follows:

The evaluation of positional similarity processing found that the RMSE and NSSDA accuracy of the comparisons of FS/OS, FS/OSM and OS/OSM in the urban area

(Cramlington1-UK) site were 0.492 and 0.846m, 5.429m and 9.143m, and 5.331m and 8.989m respectively. Similar observations can be made for the other rural site in UK (Cramlington2) which reflected 0.342m and 0.590m, 4.500m and 7.714m and 4.564m and 7.796m for the values of the RMSE and NSSDA accuracy for the comparisons of FS/OS, FS/OSM and OS/OSM datasets respectively. The RMSE and NSSDA values comparing reference dataset (FS) with OS and OSM data and comparing OS with OSM datasets in a rural area (Clara Vale-UK) were 1.843m and 3.189m, 11.650m and 20.161m and 10.887m and 18.832m respectively. For the study area outside the UK where a different national mapping agency data source has been examined, the GDS Iraqi data comparisons results analysis revealed that the RMSE and NSSDA values are relatively not too far from the urban areas in the UK comparisons. They were 1.246m and 2.149m for FS/GDS datasets, 5.903m and 10.190m for FS/OSM datasets, and 5.806m and 10.012m for GDS/OSM. In addition to the above analysis, the descriptive statistics, such as mean, standard deviation, median, maximum, minimum and interquartile range of the components' errors between the planimetric coordinates (E, N) of tested points are also included in the positional similarity assessment methodology. They reveal that the E and N components of errors for FS and formal data are very close to each other. However, there is a significant difference in the E and N elements when comparing FS/OSM and OS/OSM.

The directions of errors were also included and analysed in this part of research. The results showed that the discrepancies between FS and formal datasets are slightly more concentrated around the mean direction comparisons with informal datasets, indicating less variability in official data. In general, the positional similarity assessment findings values indicated that the results of the comparison of the reference dataset with the tested datasets were very high for OSM data, but less so with the formal datasets. Although the RMSEs of OSM positional data in urban areas are lower than the RMSEs of OSM data in the rural area, there is still such a wide range of discrepancy between the compared datasets that the use of OSM information for geospatial data integration purposes is difficult.

As mentioned above, the geometrical similarity assessment also included the assessment of the linear similarity measurements among FS, FM and OSM datasets for integration processing. The outcomes of comparing OSM information, FS and OS linear datasets in

the UK study area showed that the overlap percentage was approximately 96% when the buffer size was 4.25m for the comparison of FS/OS, while for the comparisons of FS/OSM and OS/OSM the overlap percentages were 84% and 84% respectively when the buffer size was even larger at 12.5m (see graph on p.103 showing magnitude and stability of such comparisons). In addition, the results of other linear similarity measurement analysis (average displacement) showed that the average displacement value of the tested linear features was 0.25m for the 1m buffer size at which stable results are evident for FS/OS datasets. The comparisons also indicated that the average displacement values were approximately 3.40m where the buffer size was 12.5m for FS/OSM and OS/OSM respectively.

The same observations were made for overlap percentage and average displacement values of comparison of FS, GDS and OSM information in the Iraq study area. The findings showed that the overlap percentage value between FS and GDS datasets was 91% when the buffer size was 5m, while it was 81% when the buffer size was larger at 12.5m for FS/OSM and GDS/OSM comparisons. The findings also revealed that the average displacement value was 0.75m for buffer size 2.5m for FS/GDS comparison, whereas for FS/OSM and GDS/OSM, it was 3.70m when buffer size was 12.5m. This study has found that generally both overlap percentages and average displacement values for both the UK and Iraq sites indicated convergence between FS and FM datasets, while there is a significant linear difference between the comparison of OSM data with each of the FS and FM datasets. Therefore, for the current study, the same conclusion can be drawn from linear features analysis as was noted from positional similarity measurements. Thus, integrating FS and FM datasets is possible and integrating OSM with FS or FM datasets is problematic.

In addition to the previous two geometrical elements (positional and linear), this study determined the polygon shape similarity among tested datasets using moments invariant. The findings of comparing FS, OS polygon data and OSM information in the two UK urban sites indicated that there is a convergence between the 7-moments of FS and OS; however, the study revealed significant separation between the 7-moments values when comparing OSM with the FS and OS datasets. Consequently, the space distance between the moments of FS and OS was relatively smaller than those between OSM and each of the FS and OS datasets. On the other hand, the rural-UK area test showed

some separation between the values of the moments of FS and OS datasets; however, comparing both FS and OS data with OSM information showed a significant difference between the moments' values of the compared datasets. For the Iraq urban area, the analysis showed similar shape similarity outcomes to those obtained from urban areas in the UK. In general, therefore, the evidence from this study suggests a similar conclusion to that obtained from the comparisons of positions and lines features. The differences between the 7-moments invariant favour the possible integration of FS and OS datasets; however, from the findings it seems successful integration between OSM data and each of the FS and FM datasets is difficult.

8.3.2 The evaluation of various factors affect geometrical quality of VGI datasets

The current research considered the different factors that may affect the geometrical quality of VGI data in general and OSM information specifically. As explained in Chapter 5, a factorial design procedure was followed in order to design the experiment of this analysis. Three factors were chosen to undertake the analysis: *data source (A)*, *feature type (B)* and *individuals (C)*. Each factor was set out with two levels: The high level (+) included *GPS*, *hard detail* and *same individual* respectively, while the low level (-) involved *other source*, *soft detail* and *different individuals*. The response variable was selected to be the Euclidean distance between the bench mark field survey data and OSM datasets.

The numerical results of this experiment showed that the factor 'data source' has the most significant effect (compared to the other two factors) on the geometrical quality of OSM information within the critical value of p-value = 0.05. The 'feature type' factor showed a p-value equal to zero which is less than the critical value of p-value = 0.05. This indicated that this factor can be considered as the next factor that can affect the geometrical quality of OSM datasets. The 'individual' factor also showed a p-value less than the critical value of p-value = 0.05. It achieved a p-value equal to 0.008; although this value is greater than the p-values of the other two factors, it also reflected the significance of this factor as well. These numerical analysis findings were also confirmed by the graphical analysis.

In addition to the analysis of each of the factors separately, the experimental analysis also included the analysis of the interaction effects among the three factors. The results

of the numerical and graphical analysis showed that there is an interaction effect among the three factors (A, B, and C) that were chosen for the current study. In general the findings confirmed the alternative and rejected the null hypothesis that was set out in subsection 5.3.5. This proved that there are significant effects of the three factors of this experiment with different levels of effects. The data source was addressed as the most significant factor that may affect the geometrical data quality of OSM information. Therefore, it would be advantageous for the OSM community to be careful about the data sources that are used to produce OSM datasets. This may improve the geometrical data quality of OSM and subsequently could improve the integration with official datasets. Whilst it is acknowledged that the testing of three factors in this case study has concluded that care with data sources must be taken, further case studies may reveal differing results. The wide variety of examples of OSM data projects, presented in Chapter 3, may well be affected in differing ways by variability (or lack of variability) in the factors tested here. In addition, the determination of which are the most important factors may vary from location to location. The factorial design testing undertaken here can be usefully applied to understand such variation.

8.3.3 The assessment of semantic similarity of formal and VGI datasets for integration purposes

The research steps also involved the assessment of the possibility of integration of feature classifications from official data sources such as OS and GDS, and VGI data such as OSM information. As explained in section 8.2, this included developing two models for assessing the semantic similarity, one for feature-by-feature comparison and the other for XML schema integration assessment. The WordNet::Similarity software was used to measure the semantic similarity between the compared datasets. The Lin method was used to evaluate the semantic similarity between compared data, as discussed in section 6.6. The first similarity assessment model was based on the one-sample t-test to determine the population mean of semantic similarity scores of corresponding features. Then this population mean was compared with the hypothesised semantic similarity score 0.5. The results of this analysis indicated that the population mean of the semantic similarity scores in all study areas were less than the hypothesised value 0.5. This can lead to the conclusion that there is a mismatching between the feature classifications of formal and OSM datasets which cautions against the integration of these datasets.

The other model was initially included to assess the semantic similarity between the nodes of compared schemas. The semantic similarity was measured for many relations, such as One-To-One, One-To-Many, Many-To-One, and missing correspondence relations, between the tested schemas with various results. The findings revealed low semantic similarity scores for the One-To-One relations and confused relations for the One-To-Many and Many-To-One cases. The results of the semantic similarity scores for 'missing correspondence' relation achieved high rates. This makes the opportunities of successful integration between the compared datasets more difficult or complex. In addition to semantic similarity, the structural similarity between compared schemas was also considered. The results of this test showed very low structural similarity between tested schemas. Data type similarity was also taken into account in the similarity measurement model. The 'data type' output of compared schemas was one or zero (matching data type or not). Afterwards, a weighted combination method was undertaken to calculate the overall similarity scores. The most interesting findings were that the rates of missing correspondence relation were very high, while the rates of similarity combinations of all other node relations were very low. In general, these findings are rather disappointing for achieving successful integration of XML schema of feature classifications from official data sources such as OS and GDS with VGI data sources such as OSM data. Results shown in this study reveal that attempts to integrate the particular classification datasets tested here would result in confusion, for example, informal data being tagged in such a way that formal data could not be updated using it. However, the rich set of OSM semantics, such as those evident in OSM land use classes (Taginfo, 2012), can reflect the ability of OSM contributors to define most of the natural and man-made feature types. Subsequently, more discussions and suggestions for using the richness of OSM semantic data will be illustrated in sections 8.5 and 8.6.

8.4 Data handling of VGI for integration with formal datasets

From the discussion of the main findings of this project, as described in the previous section, major issues emerge relating to the handling of inconsistency among VGI data sources such as OSM information, field survey data, and formal data such as that from OS and GDS sources. Suggestions regarding how to improve VGI data quality and how to minimise heterogeneity for such geospatial data integration processing will be made in this section. It is advised that the role of VGI data collection techniques should be carefully considered. Chapter 5 of this thesis illustrated how there are several data

sources that are commonly used to collect and create VGI datasets, such as local knowledge, satellite and aerial imagery, as well as GPS devices. The factorial design experiment which was developed in the same chapter showed that these data sources have the most significant effect on VGI geometrical data quality. At the same time, the results of the experiment revealed that using GPS as a source of VGI data acquisition can reduce the value of the response variable 'Euclidean distance' between the reference (FS) data and OSM datasets. However, the error values are still at the level of several metres which can be considered relatively high for integration purposes, especially with formal datasets. Therefore, it is proposed to raise awareness among VGI communities, through VGI websites or specific workshops, of the importance of utilising, as far as possible, reliable and accurate GPS instruments rather than alternative, lower quality, VGI data sources.

Chapter 4 of this thesis demonstrated an evaluation of the integration possibilities based on geometrical data quality of formal and VGI datasets for different sites and features. The results of this analysis indicated that it is more effective in urban areas, where 'hard' detail (e.g. buildings, car parks and kerb lines) predominates, than in rural areas, where 'soft' detail (e.g. vegetation boundaries and terrain features) is the more common type of object. The geometrical data quality differences in rural areas may occur due to the incompleteness of the data coverage resulting in fuzzy, interpolated or inferred boundaries, such as the extent of woodland areas. Therefore, it is suggested that there should be an increase in focusing on how to create VGI data in rural areas or country sites. This may involve making advice available, through VGI websites, to VGI contributors in these areas, concerning the procedures for collecting spatial data in 'fuzzy' areas, with an emphasis on surveying all the details of the features' boundaries, as well as endeavouring to survey as many features as possible (not only a framework of man-made objects) in such rural areas.

The findings of the assessment of feature classifications similarity from formal and VGI data sources for integration processing suggested difficulties for successful data integration, as shown in Chapter 7. Avoiding the invention of user-specific tags for VGI features classifications and trying to use something similar to the existing classification schemas (e.g. formal data legends) may help to establish a greater degree of accuracy on this matter and improve the possibility of the integration of feature classifications from these spatial data sources. The adoption of similar names for parents' classes (e.g.

'highway') and their appropriate sub-classes (e.g. 'primary route', 'secondary route' etc) in the XML schema trees may contribute to achieving better similarities of this kind between compared XML schemas. Using the online descriptive and photographic examples of feature classifications that are available on VGI websites (e.g. OSM) more rigorously is also suggested in order to have an understanding of the correct classifications of features before suggesting new or superfluous feature classes. Improvement of integration may also be enhanced by modification of the generic ontologies in WordNet::Similarity. Enhancement in all of these fields may assist in an improvement in the role of VGI datasets in the flowline of official spatial data handling.

8.5 Utilising the distinct nature of VGI

This research has shown that there are significant incompatibilities preventing the integration of VGI with formal datasets. However, some of these can be addressed in a positive manner with regard to certain aspects such as the richness of datasets that have been offered by VGI data sources (Ballatore and Bertolotto, 2011). For example, the richness of feature type definition in OSM could enhance the value of integration. The growing availability of the subclasses of OSM features may enable a wealth of new opportunities to enhance and update the quality of feature classification of formal or governmental data. For instance, by comparing the children levels (i.e. fourth level) of XML schema trees of OSM datasets with those of formal datasets, as shown in Chapter Seven, it can be seen that the schema of OSM feature classifications display more distribution and classes especially at the latest or end levels. The dynamic nature in terms of the frequency of updating and gathering detailed features of OSM data have established these aspects of feature types of OSM project as being particularly useful. This can be considered as one positive aspect of OSM or informal datasets which could enable these kinds of spatial datasets to be beneficial or advantageous.

Although the current study demonstrated that a divergence exists when evaluating the integration of the geometrical elements of VGI and formal spatial data sources, Zielstra and Hochmair (2011) showed that the integration of pedestrian routes' accessibility to transit stations (bus and metro stations) of VGI data, such as the OSM project, into the data produced from Tele Atlas and/or NAVTEQ can be useful in US and German cities. The reason for including OSM data in this integration process as a worthy source of pedestrian routes has been explained by the authors as being a potential rich and

valuable source of pedestrians' data that can be supplied from OSM project. This merit of OSM data has been proved by the findings of the numerical analysis of Zielstra and Hochmair (2011). They showed that the information about pedestrian segments of OSM data can increase the usage of the transit facilities when the commercial data is not available, as in the case studies in Germany and some US cities such as Chicago and San Francisco. It seems possible that these results are due to the effective efforts of OSM communities in Germany and some US cities to develop comprehensive OSM pedestrian path networks.

The positive aspects of VGI data that were mentioned in Chapter 3, such as the easy access, free use and the rapid growth of these kinds of datasets, makes it possible to envisage further practical applications of them. For example, the combining of the richness of the OSM database into the addresses of places around the world, by following mapping mashup technologies, may produce a compatible and helpful database. This could be used for tourist or navigation purposes rather than using traditional formal and expensive tourist maps. Another useful VGI application was suggested by Neis et al. (2010). They explained that the integration of up-to-date OSM data into the UN Spatial Data Infrastructure for Transportation would make it possible to manage disaster relief by creating crisis maps. The obvious example of this situation was when an earthquake hit Haiti in 2010 and how the use of OSM helped in the rescue of people, as described in section 2.4.

The availability of other VGI data sources beyond OSM can also be useful and valuable for such applications. The other VGI data sources, especially those which do not provide pure spatial datasets, such as Flickr, may be used for such applications that do not need to consider the spatial data quality elements in the processing flowline. For instance, Schade et al. (2011) proposed a workflow for using VGI data such as Flickr images for risk detection events. They initially described the positive characteristics of this kind of dataset and why it can be applied for this application. The massive growth of images through this website, the multiplicity of options for uploading images and the ability of users to geotag images can be considered the main positive aspects of this service (more details concerning the properties of the Flickr website were illustrated in subsection 3.5.2). As an example of their application, Flickr images were used to detect the risk of flood in the UK for the period from 2007 to 2009. The analysis of the approach taken was basically based on the fundamental information that can be obtained

from the Flickr website, such as the location of the picture, the time and date a picture was taken and the time and the data the pictures were uploaded onto the website.

Another example of a benefit gained from VGI data was illustrated by Spinsanti and Ostermann (2010). They developed a methodology to test the framework of assessing the contribution of VGI data to detecting the spread of fire events, and compared this with the official information sources that can be obtained from the European forest fire information system. For their project, the authors used picture data from Flickr and text data from Tweeter. They argued that using VGI data for communicating during disasters is an effective method of crisis management which may reduce the amount of risks and losses.

The above discussion has shown that although this research project concluded that it is difficult to integrate VGI data with formal spatial datasets, there is a wide range of applications that VGI data can serve and assist with; especially those which do not need high quality spatial datasets. Updating transit maps, enhancing pedestrian navigation systems and addressing disaster scenarios are all examples of small-scale topographic data which can be more easily integrated with formal datasets. Large-scale data including some of the datasets tested in this research show more limited promise due to geometric and semantic mismatching. Therefore, the conclusion that formal data providers would not be interested in integrating their datasets with, for example, OSM data is correct. VGI users, on the other hand, are certainly keen to incorporate official datasets into their own, as demonstrated in Chapter 2, and the results obtained in this research would give confidence to such procedures. The issue of data quality is taken seriously throughout the OSM community, for example. The section on Quality Assurance in the OSM blog shows that a number of tools are available and in development for detection, reporting and monitoring of data quality (Wiki-OpenStreetMap, 2012).

8.6 Recommendations for future work (research)

This research has thrown up many suggestions regarding the issues of VGI's ability to integrate with other datasets which require further investigation and they are stated as follows:

- In this research, geometrical and semantic similarity measurements were undertaken and processed to assess the integration of formal and VGI datasets.

Other parameters such as temporal accuracy, completeness, lineage and further processing (e.g. generalisation) may be incorporated into the assessment of the integration model.

- Additional VGI data sources beyond the field surveyed information that was used in this project (OSM), such as images (e.g. Flickr), textual descriptions and user updating (e.g. TomTom MapShare) could also contribute to 'crowdsourced' datasets and their accuracy requires close examination.
- Besides the practical testing of assessing the ability of integrating geospatial datasets that was accomplished here, considerably more work in order to establish more comparative studies might also be required as a future step. For instance, evolving and implementing a framework to investigate and compare the possibility of integrating OSM data with official datasets in developed, developing and undeveloped countries would allow for further integration to be assessed. In particular, it would be interesting to study and contrast the OSM data properties (e.g. geometrical and semantic data) in different geographical areas (e.g. Africa against Europe). Examining the performance of such integration framework can provide insight on the suitability of using OSM data for such data integration processing around varied parts of the world.
- The semantic similarity assessment analysis that was achieved in this project is the initial or preparation step for assessing the integration of feature classifications from official and VGI data sources. Future research should therefore concentrate on the investigation of developing a robust system to integrate the subclasses of OSM datasets under the corresponding classes of formal datasets if they were impossible to categorise at a finer level these subclasses.
- As the main conclusions of this thesis found that the geospatial data integration processing between VGI data such as OSM and formal data sources may be problematic, developing a spatial data quality intelligent security system, in order to assess the quality of VGI data before uploading such data into VGI websites, is needed. This system should have the ability to check the spatial data quality elements based on input threshold values. If the quality of data uploaded into VGI platforms matches the required threshold value, the system will allow such updating; otherwise the uploading process should be stopped.

- Another area that may need more exploration is the assessment of the factors that may affect OSM quality. The experiment of factorial design in Chapter 5 was implemented using three factors with two levels for each factor. For further experimental investigations it would better to examine other factors in other study areas especially those described in subsection 3.5.1.3. This may provide different analysis which may help in the development of the role of VGI data in the flowline of official datasets handling.
- So far in this research, the integration assessment included the 2D datasets. It would be interesting to assess the possibility of the integration of a 3D city model, for instance, produced from terrestrial laser scanning data or from point clouds derived from VGI data sources such as Flickr. The software capabilities of Microsoft Photosynth and Autodesk's 123D Catch packages allow for point cloud creation and subsequent comparison opportunities for such purposes.
- In order to enhance the specifications of VGI data quality, and as a result improving its integration with official datasets, it is suggested that further research could be undertaken in areas of improving VGI accuracy such as peer-to-peer reliability assessment by using network methods to determine the information credibility. People in the same field could correct and maintain the error of spatial datasets between each other.
- More broadly, research is also needed to control the contribution data of the OSM project. It is recommended that the Flanagin and Metzger (2008) suggestion could be followed to increase the author transparency of VGI data. Tools, such as WikiScanner, which is used in Wikipedia to reveal the identity (via an IP address) of the contributors, could be utilised in order to refine VGI data for geospatial data integration processing.

These nine recommendations can be addressed in future research and, as the area of VGI is becoming ever more interesting, it is felt that there is significant scope for continuing this work.

References

- Al-Bakri, M. and Fairbairn, D. (2010) 'Assessing the accuracy of crowdsourced data and its integration with official spatial data sets', *The Ninth International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*. University of Leicester / UK, pp. 317-320.
- Al-Bakri, M. and Fairbairn, D. (2011) 'User generated content and formal data sources for integrating geospatial data ', *25th International Cartographic Conference* Paris, France, pp. 1-8.
- Al-Bakri, M. and Fairbairn, D. (2012) 'Assessing similarity matching for possible integration of feature classifications of geospatial data from official and informal sources', *International Journal of Geographical Information Science*, 26, (8), pp. 1437-1456.
- Algergawy, A., Nayak, R. and Saake, G. (2009) 'XML schema element similarity measures: a schema matching context', in *On the Move to Meaningful Internet Systems: OTM 2009*. Vilamoura, Portugal, pp. 1246-1253.
- Ali, A. B. H. (2002) 'Moment representation of polygons for the assessment of their shape quality', *J Geograph Syst*, 4, pp. 209–232.
- Amarintrarak, N., Saikeaw, K. R., Tongshima, S. and Wiwatwattana, N. (2009) 'SAXM : semi-automatic XML schema mapping', *The 24th International Technical Conference on Circuits/Systems, Computers and Communications*. Jeju Island, Korea, pp. 44 - 47.
- ANSI. (1998) *National Committee on Information Technology Standards 320 (Spatial Data Transfer Standard)*. Washington D.C.: U.S. Department of Commerce: <http://mcmcweb.er.usgs.gov/sdts/standard.html#view>
- ANZLIC (2001) *ANZLIC Metadata Guidelines: Core Metadata Elements for Geographic Data in Australia and New Zealand, version 2*. Available at: http://www.ga.gov.au/image_cache/GA9364.pdf (Accessed: 19-10-2011).
- Aradóttir, Á. L., Robertson, A. and Moore, E. (1997) 'Circular statistical analysis of birch colonization and the directional growth response of birch and black cottonwood in south Iceland', *Agricultural and Forest Meteorology*, 84, (1-2), pp. 179-186.
- Arnold, B. and SenGupta, A. (2006) 'Recent advances in the analyses of directional data in ecological and environmental sciences', *Environmental and Ecological Statistics*, 13, (3), pp. 253-256.

- ASPRS. (1989) 'Accuracy standards for large scale maps', *Photogrammetric Engineering and Remote Sensing*, 56, pp. 1038-1040.
- Auer, M. and Zipf, A. (2009) 'How do free and open geodata and open standards fit together? from scepticism versus high potential to real applications.', *The First Open Source GIS UK Conference*. Nottingham, UK, pp. 1-6.
- Austin, R. F. (1984) 'Measuring and representing two dimensional shapes', in *Spatial Statistics and Models*. Dordrecht: Reidel D, pp. 293-312.
- Balasubramaniam, N. (2009) 'User-Generated Content', in Michahelles, F.(ed), *Business Aspects of the Internet of Things*. Seminar of advanced topics, FS 2009, Zurich, Switzerland, pp 28-33.
- Ballatore, A. and Bertolotto, M. (2011) 'Semantically enriching VGI in support of implicit feedback analysis', in Tanaka, K., Fröhlich, P. and Kim, K.-S.(eds) *Web and Wireless Geographical Information Systems, volume 6574 of Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, pp. 78-93
- Banerjee, S. and Pedersen, T. (2003) 'Extended gloss overlaps as a measure of semantic relatedness', *Proceedings of the 18th international joint conference on artificial intelligence*. Acapulco, Mexico, pp. 805-810.
- Barrett, H. C. (2012) *Web 2.0 tools for lifelong & life wide learning*. Available at: <http://electronicportfolios.com/web2/class/index.html> (Accessed: 27-03-2012).
- Bartoschek, T. and Keßler, C. (2013) 'VGI in education—from K-12 to graduate studies', in Sui, D., Elwood, S. and Goodchild, M.(eds) *Crowdsourcing Geographic Knowledge. Volunteered Geographic Information (VGI) in Theory and Practice*. Springer, pp. 341-360.
- Bishr, M. and Kuhn, W. (2007) 'Geospatial information bottom-up: A matter of trust and semantics', *The European Information Society - Leading the Way with Geoinformation*. Springer-Verlag Berlin Heidelberg, pp. 365-387.
- Blakemore, M. (1984) 'Generalisation and error in spatial data bases', *Cartographica*, 21, (2&3), pp. 131-139.
- Bogaert, J., Rousseau, R., Hecke, P. V. and Impens, I. (2000) 'Alternative area-perimeter ratios for measurement of 2D shape compactness of habitats', *Appl. Math. Comput.*, 111, (1), pp. 71-85.
- Bohme, R. (1993) *Inventory of World Topographic Mapping*. Oxford: Elsevier Science Publishers Ltd., Pergamon Press.

- Boin, A. T. and Hunter, G. J. (2006) 'Do spatial data consumers really understand data quality information?', *7th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*. pp. 215-224.
- Bowers, J. A., Morton, I. D. and Mould, G. I. (2000) 'Directional statistics of the wind and waves', *Applied Ocean Research*, 22, (1), pp. 13-30.
- Box, G. E. P. and Wilson, K. B. (1951) 'On the experimental attainment of optimum conditions', *Journal of the Royal Statistical Society*, 13, (1), pp. 1-45.
- Box, J. F. (1980) 'R.A.Fisher and the design of experiments, 1922-1926', *The American Statistician*, 34, (1), pp. 1-7.
- Bribiesca, E. (1997) 'Measuring 2-D shape compactness using the contact perimeter', *Computers & Mathematics with Applications*, 33, (11), pp. 1-9.
- Bribiesca, E. (2008) 'An easy measure of compactness for 2D and 3D shapes', *Pattern Recognition*, 41, (2), pp. 543-554.
- Brimicombe, A. (2003) *GIS, Environmental Modelling and Engineering*. London: Taylor and Francis.
- Burrough, P. A. and McDonnell, R. A. (1998) *Principles of Geographical Information Systems*. New York: Oxford University Press Inc.
- Bush, R. R. and Mosteller, F. A. (1951) 'A model for stimulus generalization and discrimination', *Psychological Review*, 58, pp. 413-423.
- Butenuth, M., Gösseln, G. v., Tiedge, M., Heipke, C., Lipeck, U. and Sester, M. (2007) 'Integration of heterogeneous geospatial data in a federated database', *ISPRS Journal of Photogrammetry and Remote Sensing*, 62, (5), pp. 328-346.
- Cai, G. (2002) 'A GIS approach to the spatial assessment of telecommunications infrastructure', *Networks and Spatial Economics*, 2, (1), pp. 35-63.
- Campbell, M. J. and Swinscow, T. D. (2009) *Statistics at Square One*. 11th ed West Sussex, UK: John Wiley & Sons Ltd.
- Cao, H., Qi, Y., Candan, K. S. and Sapino, M. L. (2010) 'XML data integration: schema extraction and mapping', in Li, C. and Ling, T. W.(eds) *Advanced Applications and Structures in XML Processing: Label Streams, Semantics Utilization and Data Query Technologies*. USA: IGI Global pp. 308-312.
- Casado, M. L. (2006) 'Some basic mathematical constraints for the geometric conflation problem', *the 7th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*. Lisbon, Portugal, pp. 264-274.

- Castelein, W., Grus, L., Crompvoets, J. and Bregt, A. (2010) 'A characterization of volunteered geographic information', *13th AGILE International Conference on Geographic Information Science*. Guimarães, Portugal, pp. 1-10.
- Chen, C.-C. (1993) 'Improved moment invariants for shape discrimination', *Pattern Recognition*, 26, (5), pp. 683-686.
- Chen, C.-C., Knoblock, C. A. and Shahabi, C. (2008) 'Automatically and accurately conflating raster maps with orthoimagery', *Geoinformatica*, 12, (3), pp. 377-410.
- Chilani, C. D. (2010) *Adjustment Computations: Spatial Data Analysis*. 5th ed Hoboken, New Jersey: John Wiley & Sons, Inc.
- Chilton, S. (2009) 'Crowdsourcing is radically changing the geodata landscape: case study of OpenStreetMap ', *24th International Cartographic Conference* Santiago, Chile, pp. 1-7.
- Cho, G. (2005) *Geographic Information Science: Mastering the Legal Issues*. England: John Wiley & Sons, Ltd.
- Choudhury, S., Chakrabarti, D. and Choidhury, S. (2009) *An Introduction to Geographic Information Technology*. New Delhi, India: I.K. International Publishing House Pvt.Ltd.
- Chrisman, N. R. (1982) 'A theory of cartographic error and its measurement in digital data bases', *Auto-Carto 5 - Proceedings*. Crystal City, Virginia, pp. 159-168.
- Ciepluch, B., Mooney, P., Jacob, R. and Winstanley, A. C. (2009) 'Using OpenStreetMap to deliver location-based environmental information in Ireland', *SIGSPATIAL Special*, 1, (3), pp. 17-22.
- Coleman, D., Georgiadou, Y. and Labonte, J. (2009) 'Volunteered geographic information: the nature and motivation of producers', *International Journal of Spatial Data Infrastructures Research*, 4, (1), pp. 332 - 358.
- Coleman, D. J. (2010) 'Volunteered geographic information in spatial data infrastructure: an early look at opportunities and constraints', *GSDI 12 world conference* Leuven University Press, pp. 131-147
- Congalton, R. G. (1991) 'A review of assessing the accuracy of classifications of remotely sensed data', *Remote Sensing of Environment*, 37, pp. 35-46.
- Congalton, R. G. and Green, K. (2009a) 'Positional accuracy', in *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices*. USA: CRC Press, pp. 19-54.

- Congalton, R. G. and Green, K. (2009b) 'Thematic accuracy ', in *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices*. USA: CRC Press, pp. 55-59.
- Corcoran, J., Chhetri, P. and Stimson, R. (2009) 'Using circular statistics to explore the geography of the journey to work', *Papers in Regional Science*, 88, (1), pp. 119-132.
- Cormode, G. and Krishnamurthy, B. (2008) 'Key differences between Web1.0 and Web2.0', *First Monday*, 13, (6), pp. 1-30.
- Craglia, M. (2007) 'Volunteered geographic information and spatial data infrastructures: when do parallel lines converge?', *VGI specialist meeting*. Santa Barbara, USA, pp. 1-3, Available at : http://www.ncgia.ucsb.edu/projects/vgi/docs/position/Craglia_paper.pdf.
- Cuartero, A., Armesto, J., Rodríguez, P. G. and Arias, P. (2010) 'Error analysis of terrestrial laser scanning data by means of spherical statistics and 3D graphs', *Sensors*, 10, (11), pp. 10128-10145.
- Delavar, M. R. and Devillers, R. (2010) 'Spatial data quality: From process to decisions', *Transactions in GIS*, 14, (4), pp. 379-386.
- Deparday, V. (2010) *Enhancing volunteered geographical information (VGI) visualization with open source web-based software*. thesis. University of Waterloo, Canada. Available at: http://uwspace.uwaterloo.ca/bitstream/10012/5709/1/Deparday_Vivien.pdf.
- Devillers, R. and Jeansoulin, R. (2006) *Fundamentals of Spatial Data Quality*. Great Britain: Antony Rowe Ltd, Chippenham, Wiltshire.
- Dey, S. and Ghosh, P. (2008) 'GRDM--A digital field-mapping tool for management and analysis of field geological data', *Computers & Geosciences*, 34, (5), pp. 464-478.
- Donker, F. W. and Loenen, B. v. (2006) 'Transparency of accessibility to government-owned geoinformation', *12th EC-GI & GIS Workshop*, Innsbruck, Austria, pp. 1-12.
- Drecki, I. (2007) 'Geographic information uncertainty: The concept and representational challenges', *Proceedings of the 23rd International Cartographic Conference*. Moscow, Russia, pp. 1-13.
- Dudani, S. A., Breeding, K. J. and McGhee, R. B. (1977) 'Aircraft identification by moment invariants', *Computers, IEEE Transactions on*, C-26, (1), pp. 39-46.

- Ebdon, D. (1985) *Statistics in Geography: A practical Approach* 2nd ed United Kingdom: Wiley-Blackwell.
- Edward, D. and Simpson, J. (2002) 'Integration and access of multi-source vector data', *Symposium of Geospatial Theory, Processing and application*. Ottawa, Canada, pp. 1-8.
- Eisler, H. and Ekman, G. (1959) 'A mechanism of subjective similarity', *Acta Psychologica*, 16, pp. 1-10.
- Elwood, S. (2008) 'Volunteered geographic information: future research directions motivated by critical, participatory, and feminist GIS', *Geojournal*, 72, (3), pp. 173–183.
- Elwood, S. (2009) 'Geographic information science: new geovisualization technologies – emerging questions and linkages with GIScience research', *Progress in Human Geography*, 33, (2), pp. 256–263.
- eMarketer (2009) *User-generated content draws fans*. Available at: <http://www.emarketer.com/Article.aspx?R=1006895> (Accessed: 17-04-2012).
- Erhard, R. and Philip, A. B. (2001) 'A survey of approaches to automatic schema matching', *The VLDB Journal*, 10, (4), pp. 334-350.
- Esa, R., Mikko, S. and Janne, H. (2006) 'A new convexity measure based on a probabilistic interpretation of images', *IEEE Trans. Pattern Anal. Mach. Intell.*, 28, (9), pp. 1501-1512.
- Exel, M. v., Dias, E. and Fruijtier, S. (2010) 'The impact of crowdsourcing on spatial data quality indicators', *GIScience 2010*. Zurich, Switzerland pp. 1-4.
- Fellbaum, C., Palmer, M., Trang Dang, H., Delfs, L. and Wolff, S. (2001) 'Manual and automatic semantic annotation with WordNet', *Proceedings of the NAACL Workshop on WordNet and Other Lexical Resources: Applications, Customizations*. Carnegie Mellon University, Pittsburg, pp. 1-8.
- FGDC. (1994) *Content Standards for Digital Geospatial Metadata*. Washington D.C., USA:
- FGDC (1998a) *Content Standards for Digital Geospatial Metadata, Version 2*. Available at: <http://www.fgdc.gov/library/newsletters/98news/summer98/> (Accessed: 19-10-2011).
- FGDC. (1998b) *Geospatial positioning accuracy standards. part 3: National Standard for Spatial Data Accuracy. FGDC-STD-007.3-1998* Washington, DC: Federal Geographic Data Committee

- Field, A. and Hole, G. (2003) *How to Design and Report Experiments*. London, UK: SAGE Publications Ltd.
- Finn, M. P., Utery, E. L., Starbuck, M., Weaver, B. and Jaromack, G. M. (2004) 'Integration of the national maps', *XXth ISPRS Congress*. Istanbul, Turkey, pp. 1-3.
- Fisher, N. I. (1993) *Statistical Analysis of Circular Data*. New York: Cambridge University Press.
- Fisher, P. F., Comber., A. and Wadsworth, R. (2006) 'Approaches to uncertainty in spatial data', in Devillers, R. and Jeansoulin, R.(eds) *Fundamentals of Spatial Data Quality*. London: ISTE.
- Fisher, R. A. and Mackenzie, W. A. (1923) 'Studies in crop variation. II. The manurial response of different potato varieties', *Journal of Agricultural Science*, 13, pp. 311-320
- Flanagin, A. J. and Metzger, M. J. (2008) 'The credibility of volunteered geographic information', *Geojournal*, 72, pp. 137–148.
- Flickr (2012) Available at: <http://www.flickr.com/map/> (Accessed: 16-04-2012).
- Fonseca, F., Egenhofer, M., Davis, C. and Câmara, G. (2002) 'Semantic granularity in ontology-driven geographic information systems', *Annals of Mathematics and Artificial Intelligence*, 36, (1-2), pp. 121-151.
- Foody, G. and Atkinson, P. (2002) *Uncertainty in remote sensing and GIS*. New York: Wiley.
- Formica, A. (2008) 'Similarity of XML-schema elements: A structural and information content approach', *Computer Journal*, 51, pp. 240-254.
- Frigon, N. L. and Mathews, D. (1997) *Practical Guide to Experimental Design*. Canada: John Wiley & Sons, Inc.
- Friis-Christensen, A., Schade, S. and Peedell, S. (2005) 'Approaches to solve schema heterogeneity at the European level', *Proceedings of the 11th EC-GIS Workshop*. pp. 1-10.
- Frunza, O. (2008) 'A Trainable Tokenizer, solution for multilingual texts and compound expression tokenization', *Language Resources and Evaluation Conference, LREC*. Marrakech, Morocco, pp. 581-584.
- Gentner, D. and Markman, A. B. (1997) 'Structure mapping in analogy and similarity', *American Psychologist*, 52, (1), pp. 45-56.

- George, C. and Skerri, J. (2007) 'Web 2.0 and User-Generated Content: legal challenges in the new frontier', *Journal of Information, Law and Technology*, 12, (2), pp. 1-22.
- Giannis, V., Epimenidis, V., Paraskevi, R., Euripides, G. M. P. and Evangelos, E. M. (2005) 'Semantic similarity methods in wordNet and their application to information retrieval on the web', *Proceedings of the 7th annual ACM international workshop on Web information and data management*. Bremen, Germany, ACM, pp. 10-16.
- Girres, J.-F. and Touya, G. (2010) 'Quality assessment of the French OpenStreetMap dataset', *Transactions in GIS*, 14, (4), pp. 435-459.
- Givens, J. M. (1999) *Positional accuracy handbook using the National Standard for Spatial Data Accuracy to measure and report geographic data quality*. Available at: http://www.mnplan.state.mn.us/pdf/1999/lmic/nssda_o.pdf (Accessed: 16-01-2012).
- Goodchild, M. (2010) 'The role of volunteered geographic information in a postmodern GIS world', *ArcUser*, Spring 2010, p.20-21.
- Goodchild, M. and Glennon, A. (2010) 'Crowdsourcing geographic information for disaster response: a research frontier', *International Journal of Digital Earth*, 3, (3), pp. 231-241.
- Goodchild, M. F. (2007a) 'Citizens as sensors: the world of volunteered geography', *Geojournal*, 69, (4), pp. 211–221.
- Goodchild, M. F. (2007b) 'Citizens as voluntary sensors: spatial data infrastructure in the world of web 2.0', *International Journal of Spatial data Infrastructures Research*, 2, pp. 24-32.
- Goodchild, M. F. (2008a) 'Imprecision and Spatial Uncertainty', in Shekhar, S. and Xiong, H.(eds) *Encyclopedia of GIS* New York: SpringerScience+Business Media, LLC
- Goodchild, M. F. (2008b) 'Spatial Accuracy 2.0', *Proceeding of the 8th international symposium on spatial accuracy assessment in natural resources and environmental sciences*. Shanghai, pp. 1-9.
- Goodchild, M. F. and Hunter, G. J. (1997) 'A simple positional accuracy measure for linear features', *International Journal of Geographical Information Science*, 11, (3), pp. 299 - 306.
- Goodchild, M. F. and Li, L. (2012) 'Assuring the quality of volunteered geographic information', *Spatial Statistics*, 1, pp. 110-120.

- GoogleMaps (2012) *Google maps/earth additional terms of service* Available at: http://maps.google.com/help/terms_maps.html (Accessed: 27-03-2012).
- Greenwalt, C. and Shultz, M. (1962) *Principles of error theory and cartographic applications*. St Louis MO: Aeronautical Chart and Information Center, US Airforce. Available at: <http://www.dtic.mil/dtic/tr/fulltext/u2/276978.pdf>
- Grefenstette, G. and Tapanaines, P. (1994) 'What is a word, what is a sentence? problems of tokenization', *International Conference on Computational Lexicography*. Budapest, pp. 79-87.
- Gregson, R. A. M. (1975) *Psychometrics of Similarity*. New York Academic Press.
- Griethe, H. and Schumann, H. (2005) 'Visualizing uncertainty for improved decision making', *Proceedings of the 4th International Conference on Business Informatics Research BIR 2005*. University of Skovde, Skovde, Sweden, pp. 1-11.
- Gruber, G., Menard, C. and Schachinger, B. (2008) 'Evaluation of the geometric accuracy of automatically recorded 3D – City Models compared to GIS-data', in Bernard, L., Friis-Christensen, A. and Pundt, H.(eds) *Lecture Notes in Geoinformation and Cartography (The European Information Society)* Springer Berlin Heidelberg, pp. 67-78.
- Guarino, N., Masolo, C. and Vetere, G. (1999) 'OntoSeek: content-based access to the Web', *Intelligent Systems and their Applications, IEEE*, 14, (3), pp. 70-80.
- Hagenauer, J. and Helbich, M. (2012) 'Mining urban land-use patterns from volunteered geographic information by means of genetic algorithms and artificial neural networks', *International Journal of Geographical Information Science*, 26, (6), pp. 963-982.
- Hahn, U. (2001) 'Similarity: A transformational approach', *23 Annual Meeting of the Cognitive Science Society*. Lawrence Erlbaum Associates, Inc, pp. 393-398.
- Hahn, U., Chater, N. and Richardson, L. B. (2003) 'Similarity as transformation', *Cognition*, 87, pp. 1–32.
- Haklay, M. (2010) 'How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets', *Environment and Planning B: Planning and Design*, 37, (4), pp. 682 -703.
- Haklay, M., Basiouka, S., Antoniou, V. and Ather, A. (2010) 'How many volunteers does it take to map an area well? the validity of linus law to volunteered geographic information', *Cartographic Journal, The*, 47, pp. 315-322.

- Haklay, M., Singleton, A. and Parker, C. (2008) 'Web mapping 2.0: The neogeography of the GeoWeb', *Geography Compass*, 2, (6), pp. 2011-2039.
- Haklay, M. M. and Weber, P. (2008) 'OpenStreetMap: user-generated street maps', *IEEE Pervasive computing*, 7, (4), pp. 12-18.
- Hanbury, A. (2003) 'Circular statistics applied to colour images', *Proceedings of the 8th Computer Vision Winter Workshop*. pp. 55-60.
- Harding, J. (2006) 'Vector data quality: a data provider's perspective', in Devillers, R. and Jeansoulin, R.(eds) *Fundamentals of Spatial Data Quality*. London: ISTE, pp. 141-158.
- Harris, D. (2008) *Web 2.0 evolution into the intelligent Web 3.0*. London, UK: Emereo Pty Ltd.
- Hatzopoulos, J. N. (2008) *Topographic Mapping: Covering the Wider Field of Geospatial Information Science & Technology (GIS & T)*. Florida, USA: Universal Publishers.
- Hirst, G. and Onge, D. S. (1998) 'Lexical chains as representation of context for the detection and correction malapropisms', in Fellbaum, C.(ed), *WordNet: an electronic lexical database*. Cambridge: MIT press, pp. 305-332.
- Ho, S. and Rajabifard, A. (2010) 'Learning from the crowd: the role of volunteered geographic information in realising a spatially enabled society', *GSDI 12 World Conference: Realising Spatially Enabled Societies*. Singapore, pp. 1-23.
- Holland, D. and Murray, K. (2000) 'A digital national framework for topographic data in Great Britain', *International Archives of Photogrammetry and Remote Sensing*. Amsterdam, pp. 303-308.
- Hong-Minh, T. and Smith, D. (2007) 'Hierarchical approach for datatype matching in XML schemas', *24th British National Conference on Databases*. UK, University of Glasgow, pp. 120-129.
- Howard, M., Payne, S. and Sunderland, R. (2010) *Technical guidance for the INSPIRE schema transformation network service*.
- Hu, M.-K. (1962) 'Visual pattern recognition by moment invariants', *IRE Transactions on Information*, 8, pp. 179-187.
- Hunter, G. J. (1999) 'New tools for handling spatial data quality: moving from academic concepts to practical reality', *URISA Journal*, Vol. 11, (No. 2), pp. 25-34.
- ISO/TC211 (2003) *Geographic Information / Geometrics*. Available at: <http://www.isotc211.org/> (Accessed: 20-10-2011).

- Jakobsson, A. (2002) 'Data quality and quality management - examples of quality evaluation procedures and quality management in European national mapping agencies', in Shi, W., Fisher, P. F. and Goodchild, M. F.(eds) *Spatial Data Quality*. London: Taylor & Francis, pp. 216–229.
- Jammalamadaka, S. R. and SenGupta, A. (2001) *Topics in Circular Statistics*. Singapore: World Scientific Press.
- Jensen, J., Saalfeld, A., Broome, F., Cowen, D., Price, K., Ramsey, D., Lapine, L. and User, E. L. (2005) 'Spatial data acquisition and integration', in McMaster, R. B. and User, E. L.(eds) *A Research Agenda for Geographic Information Science*. London: CRC Press, pp. 17-57.
- Jiang, J. J. and Conrath, D. W. (1997) 'Semantic similarity based on corpus statistics and lexical taxonomy', *International Conference Research on Computational Linguistics (ROCLING X)*. Taiwan, pp. 19-33.
- Joo, J. M. (2005) 'Improved moment invariants know how, why and when', *Revista de Investigacion de Fisica* 8, (2), pp. 82-90.
- Kacker, R. N. (1985) 'Off-line quality control, parameter design, and the Taguchi method', *Journal of Quality Technology*, 17, pp. 176-188.
- Kim, J., Peng, Y., Kulvatunyou, S., Ivezic, N. and Jones, A. (2008) 'A layered approach to semantic similarity analysis of XML schemas', *The 2008 IEEE International Conference on Information Reuse and Integration*. Las Vegas, pp. 274 - 279
- Korte, G. B. (2001) *The GIS Book, How to Implement, Manage, and Assess the Value of Geographic Information Systems*. 5th ed Albany, New York: Onword Press.
- Koukoletsos, T., Haklay, M. and Ellul, C. (2012) 'Assessing data completeness of VGI through an automated matching procedure for linear data', *Transactions in GIS*, 16, (4), pp. 477-498.
- Krumm, J., Davies, N. and Narayanaswami, C. (2008) 'User-generated content', *Pervasive Computing, IEEE*, 7, (4), pp. 10-11.
- Kumi-Boateng, B. and Yakubu, I. (2010) 'Assessing the quality of spatial data', *European Journal of Scientific Research*, 43, (4), pp. 507-515.
- Leacock, C. and Chodorow, M. (1998) 'Combining local context with WordNet similarity for word sense identification', in Fellbaum, C.(ed), *WordNet: A Lexical Reference System and its Application*. Cambridge: MIT Press, pp. 265-283.

- Leacock, C., Miller, G., A. and Chodorow, M. (1998) 'Using corpus statistics and WordNet relations for sense identification', *Comput. Linguist.*, 24, (1), pp. 147-165.
- Lee, J. H., Kim, M. H. and Lee, Y. J. (1993) 'Information retrieval based on conceptual distance in is-a hierarchies', *Journal of Documentation*, 49, (2), pp. 188-207.
- Leica (2012a) *Leica flexline TS02/06/09*. Available at: http://www.leica-geosystems.com/en/Construction-Surveying-TPS-Leica-FlexLine-TS020609_72053.htm (Accessed: 10-05-2012).
- Leica (2012b) *Leica GPS1200 series: technical data*. Available at: http://www.leica-geosystemssolutionscenters.com/Site/Instrument%20PDF's/GPS%20Systems/SmartRover%20&%20GPS1200/GPS1200_TechnicalData_en.pdf (Accessed: 13-04-2012).
- Lenth, R. V. (1989) 'Quick and easy analysis of unreplicated factorials', *Technometrics*, 31, (4), pp. 469-473.
- Lesk, M. (1986) 'Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone', *In Proceedings of the Special Interest Group for Design of Communications Conference*. Toronto, Ontario, pp. 24-26.
- Leyk, S., Boesch, R. and Weibel, R. (2005) 'A conceptual framework for uncertainty investigation in map-based land cover change modelling', *Transactions in GIS*, 9, (3), pp. 291-322.
- Lillesand, T. M. and Kiefer, R. W. (2000) *Remote Sensing and Image Interpretation*. New York: John Wiley and Sons.
- Lin, D. (1998) 'An information-theoretic definition of similarity', *Proceedings of the Fifteenth International Conference on Machine Learning*. Morgan Kaufmann Publishers Inc., pp. 296-304.
- Lo, C. P. and Yeung, A. K. W. (2007) *Concepts and techniques of geographic information systems*. 2nd ed Toronto: Pearson Education Canada, Inc.
- Ludwig, I., Voss, A. and Krause-Traudes, M. (2011) 'A Comparison of the street networks of navteq and OSM in Germany', in Geertman, S., Reinhardt, W., Toppen, F., Cartwright, W., Gartner, G., Meng, L. and Peterson, M. P.(eds) *Advancing Geoinformation Science for a Changing World*. Vol. 1 Springer Berlin Heidelberg, pp. 65-84.

- Lv, W., Liao, W., Wu, D. and Xie, J. (2008) 'A new road network model and its application in a traffic information system', *Fourth International Conference on Autonomic and Autonomous Systems*. Gosier, Guadeloupe, pp. 160-164.
- Maceachren, A. (1985) 'Compactness of geographic shape: comparison and evaluation of measures', *Human Geography*, 67, (1), pp. 53-67.
- Map-Kibera (2012) Available at: http://mapkibera.org/wiki/index.php?title=Main_Page (Accessed: 15-09-2012).
- Mardia, K. V. and Jupp, P. E. (2000) *Directional Statistics*. Chichester: John Wiley & Sons.
- Marsden, L. E. (1960) 'How the national map accuracy standards were developed', *Surveying and Mapping*, 20, (4), pp. 427-439.
- MathWorks. (2012) *MATLAB® Creating Graphical User Interfaces*. Natick, USA: The MathWorks, Inc.
- McDougall, K. (2009) 'Volunteered geographic information for building SDI', *2009 Surveying and Spatial Sciences Institute Biennial International Conference (SSC2009)*. Adelaide, Australia, Surveying and Spatial Sciences Institute, pp. 645-653.
- McDougall, K. (2012) 'An assessment of the contribution of volunteered geographic information during recent natural disasters', in Rajabifard, A. and Coleman, D.(eds) *Spatially Enabling Government, Industry and Citizens: Research and Development Perspectives*. Needham, MA, United States: GSDI Association Press, pp. 201-214.
- Melinik, S., Garcia-Molina, H. and Rahm, E. (2002) 'Similarity flooding: a versatile graph matching algorithm and its application to schema matching', *Proceedings of the 18th International Conference on Data Engineering*. San Jose, CA , USA pp. 117-128.
- Miliard, M. (2008) 'Wikipediots: who are these devoted, even obsessive contributors to Wikipedia', *City Weekly*. Available at: <http://www.cityweekly.net/utah/article-5129-feature%20wikipediots-%09who%20are-these-devoted-even-obsessive-contributors-to-%09wikipedia.html>.
- Miller, G. and Charles, W. (1991) 'Contextual correlates of semantic similarity', *Language and Cognitive Processes*, 6, (1), pp. 1-28.
- Moellering, H. (1997) 'An introduction to world database transfer standards', in Moellering, H. and Hogan, R.(eds) *Spatial Database Standards 2: Characteristics for assessing and Full Descriptions of the National and International Standards in the World*. Oxford: Elsevier Science.

- Mohammadi, H., Rajabifard, A., Binns, A. and Williamson, I. P. (2008) 'Geo-web service tool for spatial data integrability', *11th AGILE 2008 Conference on GI Science*. Girona, Spain, pp. 1-17.
- Mohammadi, H., Rajabifard, A. and Williamson, I. P. (2009) 'Enabling spatial data sharing through multi-source spatial data integration', *GSDI 11 World Conference*. Rotterdam Netherlands, pp. 1-19.
- Mohammadi, H., Rajabifard, A. and Williamson, I. P. (2010) 'Development of an interoperable tool to facilitate spatial data integration in the context of SDI', *International Journal of Geographical Information Science*, 24, (4), pp. 487-505.
- Montero, R. S. and Bribiesca, E. (2009) 'State of the art of compactness and circularity measures', *International Mathematical Forum*, 4, (27), pp. 1305 - 1335.
- Montgomery, D. C. (2001) *Design and Analysis of Experiments*. 5th ed New York: John Wiley & Sons, Inc.
- Montgomery, D. C. (2009) *Design and Analysis of Experiments*. 7th ed Hoboken: John Wiley & Sons, Inc.
- Mooney, P. and Corcoran, P. (2012) 'The annotation process in OpenStreetMap', *Transactions in GIS*, 16, (4), pp. 561-579.
- Mooney, P., Corcoran, P. and Ciepluch, B. (2012) 'The potential for using volunteered geographic information in pervasive health computing applications', *Journal of Ambient Intelligence and Humanized Computing*, special issue on Pervasive Health Computing, pp. 1-15.
- Mooney, P., Corcoran, P. and Winstanley, A. C. (2010) 'Towards quality metrics for OpenStreetMap', *GIS: Proceedings of the ACM International Symposium on Advances in Geographic Information Systems*. San Jose, CA, pp. 514-517.
- MorphAdorner (2009) Available at: <http://morphadorner.northwestern.edu/morphadorner/wordtokenizer/example/> (Accessed: 10-06-2012).
- Moussa, W. and Fritsch, D. (2010) 'A Simple approach to link 3D photorealistic models with contents of bibliographic repositories', *EuroMed 2010* Lemessos, Cyprus, pp. 482-491.
- Mustière, S. and Devogele, T. (2008) 'Matching networks with different levels of detail', *Geoinformatica*, 12, (4), pp. 435-453.
- Neis, P., Singler, P. and Zipf, A. (2010) 'Collaborative mapping and emergency routing for disaster logistics -Case studies from the Haiti earthquake and the UN portal for Afrika', *Geoinformatik 2010*. Kiel, Germany, pp. 1-6.

- Neis, P., Zielstra, D. and Zipf, A. (2011) 'The street network evolution of crowdsourced maps: OpenStreetMap in Germany 2007–2011', *Future Internet*, 4, (1), pp. 1-21.
- Neis, P. and Zipf, A. (2012) 'Analyzing the contributor activity of a volunteered geographic information project -The case of OpenStreetMap', *ISPRS International Journal of Geo-Information*, 1, (2), pp. 146-165.
- Newsam, S. (2010) 'Crowdsourcing what is where: Community-contributed photos as volunteered geographic information', *Multimedia, IEEE*, 17, (4), pp. 36-45.
- Noh, J. S. and Rhee, K. H. (2005) 'Palmprint Identification algorithm using Hu invariant moments and otsu binarization', *Proceedings of the Fourth Annual ACIS International Conference on Computer and Information Science*. Jeju Island, South Korea, pp. 94-99.
- Nordin, N. L. B. M. (2008) *Interface developing for hata model using Matlab*. thesis. Universiti Teknologi Malaysia.
- O'Reilly, T. (2005) *What is Web 2.0: design patterns and business models for the next generation of software*. Available at: <http://oreilly.com/web2/archive/what-is-web-20.html?page=1> (Accessed: 13-02-2012).
- Ochoa, X. and Duval, E. (2008) 'Quantitative analysis of user-generated content on the web', *Proceedings of the First International Workshop on Understanding Web Evolution* Beijing, China, pp. 19-26.
- Omran, E. E. and van Etten, J. (2007) 'Spatial-data sharing: applying social-network analysis to study individual and collective behaviour', *International Journal of Geographical Information Science*, 21, (6), pp. 699 - 714.
- Oort, P. v. (2006) *Spatial data quality: from description to application*. thesis. Wageningen University.
- OpenStreetMap (2011) *Beginners' guide*. Available at: http://wiki.openstreetmap.org/wiki/Beginners'_guide (Accessed: 08-11-2011).
- OpenStreetMap (2012a) *Component overview*. Available at: http://wiki.openstreetmap.org/wiki/Component_overview (Accessed: 21-01-2012).
- OpenStreetMap (2012b) *Database*. Available at: <http://wiki.openstreetmap.org/wiki/Database> (Accessed: 17-04-2012).
- OpenStreetMap (2012c) *Downloading data*. Available at: http://wiki.openstreetmap.org/wiki/Downloading_data (Accessed: 05-09-2012).

- OpenStreetMap (2012d) *Map features*. Available at: http://wiki.openstreetmap.org/wiki/Map_features (Accessed: 28-08-2012).
- OrdnanceSurvey (2003) *German experience with PAI: DEW example*. Available at: http://www.ordnancesurvey.co.uk/oswebsite/pai/pdfs/german_experience_DEW.pdf (Accessed: 02-10-2011).
- OrdnanceSurvey. (2009) *OS MasterMap topography layer: User guide and technical specification*.
- OrdnanceSurvey (2012) Available at: <http://www.ordnancesurvey.co.uk/oswebsite/products/os-mastermap/index.html> (Accessed: 19-05-2012).
- OSM-stats (2012) *OpenStreetMap stats*. Available at: http://www.openstreetmap.org/stats/data_stats.html (Accessed: 21-12-2012).
- OSM-wiki (2012) *Stats*. Available at: <http://wiki.openstreetmap.org/wiki/Stats> (Accessed: 10-04-1012).
- Oxford. (2010) *Oxford Essential Arabic Dictionary* Ner York: Oxford University Press Inc.
- Pang, A. (2001) 'Visualizing uncertainty in geo-spatial data', *Proceedings of the Workshop on the Intersections between Geospatial Information and Information Technology*. Arlington, VA, USA, pp. 1-14.
- Parry, R., B and Perkins, C., R. (1987) *World Mapping Today*. UK: Butterworth & Co. (Publishers) Ltd.
- Patwardhan, S. (2003) *Incorporating dictionary and corpus information into a context vector measure of semantic relatedness*. thesis. University of Minnesota.
- Patwardhan, S., Banerjee, S. and Pedersen, T. (2003) 'Using measures of semantic relatedness for word sense disambiguation', *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics (CICLING-03)*. Mexico city, pp. 241-257.
- Pauly, A. and Schneider, M. (2010) 'VASA: An algebra for vague spatial data in databases', *Information Systems*, 35, (1), pp. 111-138.
- Pedersen, T., Patwardhan, S. and Michelizzi, J. (2004) 'WordNet::Similarity - Measuring the Relatedness of Concepts ', *Proceedings of Fifth Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-2004)*. Boston, pp. 38-41.

- Perkal, J. (1956) 'On epsilon length', *Bulletin de l'Academie Polonaise des Sciences*, 4, pp. 399–403.
- Perkal, J. (1965) *On the length of empirical curves*. Available at: <http://www-personal.umich.edu/~copyright/image/micmg/perkal1/perkal1.pdf> (Accessed: 07-07-2011).
- Perkins, C. and Dodge, M. (2008) 'The potential of user-generated cartography: a case study of the OpenStreetMap project and Mapchester mapping party', *North West Geography*, 8, pp. 19-32.
- Piwowar, J. M. and LeDrew, E. F. (1990) 'Integrating spatial data: A user's perspective', *Photogrammetric Engineering and Remote Sensing*, 56, pp. 1497-1502.
- Polo, M.-E. and Felicísimo, A. M. (2010) 'Full positional accuracy analysis of spatial data by means of circular statistics', *Transactions in GIS*, 14, (4), pp. 421-434.
- Programmableweb (2012) *Top mashup tags*. Available at: <http://www.programmableweb.com/mashups> (Accessed: 13-04-2012).
- Rada, R., Mili, H., Bicknell, E. and Blettner, M. (1989) 'Development and application of a metric on semantic nets', *IEEE Transactions on Systems, Man, and Cybernetics*, 19, (1), pp. 17-30.
- Rajabifard, A., Feeney, M. E. and Williamson, I. (2003) 'Spatial data infrastructure: concept, nature and SDI hierarchy', in Williamson, I. P., Rajabifard, A. and Feeney, M. E. (eds) *Developing Spatial Data Infrastructure: from Concept to Reality*. London, UK: Taylor and Francis, pp. 17-40.
- Rak, A., Coleman, D. and Nichols, S. (2012) 'Legal liability concerns surrounding volunteered geographic information applicable to Canada', *International Workshop on Geospatial Data Quality*. Quebec City, Canada, pp. 125-141.
- Ramm, F., Topf, J. and Chilton, S. (2011) *OpenStreetMap - using and enhancing the free map of the world*. Cambridge, England: UIT Cambridge Ltd.
- Read, A. W. (1948) 'An account of the word 'semantics'', *Word* IV. pp. 78-97.
- Resnik, P. (1995a) 'Disambiguating noun groupings with respect to WordNet senses', *Proceedings of the Third Workshop on Very Large Corpora*. Cambridge, Massachusetts, pp. 54-68.
- Resnik, P. (1995b) 'Using information content to evaluate semantic similarity in a taxonomy', *Proceedings of the 14th international joint conference on Artificial intelligence - Volume 1*. Montreal, Quebec, Canada, Morgan Kaufmann Publishers Inc., pp. 448-453.

- Rhind, D. W., Green, N. P. A., Mounsey, H. M. and Wiggins, J. C. (1984) 'The Integration of geographical data', *Proceedings of Australa Carto Perth, Australian Cartographic Association*. pp. 273-293.
- Ridley, H. M., Atkinson, P. M., Aplin, P., Muller, J.-P. and Dowman, I. (1997) 'Evaluating the potential of the forthcoming commercial U.S. high-resolution satellite sensor imagery at the Ordnance Survey', *Photogrammetric Engineering & Remote Sensing*, 63, (8), pp. 997-1005.
- Riedemann, C., Pundt, H., Harvey, F., Kuhn, W. and Bishr, Y. (1999) 'Semantic interoperability: A central issue for sharing geographic information', *The Annals of Regional Science*, 33, (2), pp. 213-232
- Rifat, M. R., Moutushy, S., Ahmed, S. I. and Ferdous, H. S. (2011) 'Location based information system using OpenStreetMap', *IEEE Student Conference on Research and Development (SCOReD)*. Putrajaya, Selangor, Malaysia, pp. 397-402.
- Rinner, C., Kebler, C. and Andrulis, S. (2008) 'The use of Web 2.0 concepts to support deliberation in spatial decision-making', *Computers, Environment and Urban Systems*, 32, (5), pp. 386-395.
- Ryan, T. P. (2007) *Modern Experimental Design*. Hoboken, New Jersey Wiley & Sons, Inc.
- Saalfeld, A. (1988) 'Conflation: automated map compilation', *International Journal of GIS*, 2, (3), pp. 217-228.
- Saleem, K., Bellahsene, Z. and Hunt, E. (2008) 'PORSCH: Performance oriented schema mediation', *Journal Information Systems* 33, (7-8), pp. 637-657.
- Samadzadegan, F. (2004) 'Data integration related to sensors, data and models', *XXth ISPRS Congress*. Istanbul, Turkey, pp. 1-6.
- Schade, S., Díaz, L., Ostermann, F., Spinsanti, L., Luraschi, G., Cox, S., Nuñez, M. and Longueville, B. D. (2011) 'Citizen-based sensing of crisis events: sensor web enablement for volunteered geographic information', *Applied Geomatics*, Online First™, 19 July 2011. DOI 10.1007/s12518-011-0056-y, pp. 1-16.
- Schutze, H. (1998) 'Automatic word sense discrimination', *Computational Linguistics*, 24, (1), pp. 97-123.
- Schwering, A. (2005) 'Hybrid model for semantic similarity measurement', in Meersman, R. and Tari, Z.(eds) *On the Move to Meaningful Internet Systems 2005: CoopIS, DOA, and ODBASE*. Vol. 3761 Springer Berlin / Heidelberg, pp. 1449-1465.

- Schwering, A. (2008) 'Approaches to semantic similarity measurement for geo-spatial data: a survey', *Transactions in GIS*, 12, (1), pp. 5-29.
- Seeger, C. J. (2008) 'The role of facilitated volunteered geographic information in the landscape planning and site design process', *GeoJournal*, 72, (3), pp. 199-213.
- Servigne, S., Lesage, N. and Libourel, T. (2006) 'Quality components, standards, and metadata', in Devillers, R. and Jeansoulin, R.(eds) *Fundamentals of Spatial Data Quality*. London, UK: ISTE Ltd, pp. 179-208.
- Shannon, C. (1948) 'A mathematical theory of communication', *Bell System Technical Journal*, 27, pp. 379-423 & 623-656.
- Shi, W., Fisher, P. F. and Goodchild, M. F. (2002) *Spatial Data Quality*. London: Taylor & Francies.
- Spinsanti, L. and Ostermann, F. O. (2010) 'Validation and relevance assessment of volunteered geographic information in the case of forest fires', *2nd International Workshop on Validation of Geo-Information Products for Crisis Management*. Ispra, Italy, pp. 1-9.
- Stein, A. (2010) 'Fuzzy methods in image mining ', in Jeansoulin, R., Papini, O., Prade, H. and Schockaert, S.(eds) *Methods for Handling Imperfect Spatial Information*. Chennai, India: Scientific Publishing Services Pvt. Ltd., pp. 243-268.
- Stojmenović, M. and Žunić, J. (2008) 'Measuring elongation from shape boundary', *J. Math. Imaging Vis.*, 30, (1), pp. 73-85.
- Sui, D. (2008) 'The wikification of GIS and its consequences: Or Angelina Jolie's new tattoo and the future of GIS', *Computers, Environment and Urban Systems*, 32, (1), pp. 1-5.
- Suppes, P., Krantz, D. M., Luce, R. D. and Tversky, A. (1989) *Foundations of Measurement: Geometrical, Threshold, and Probabilistic Representations*. San Diego: CA, Academic Press.
- Taginfo (2012) *Landuse, for describing the primary use of areas of land*. Available at: <http://taginfo.openstreetmap.org/keys/landuse#values> (Accessed: 1-10-2012).
- Taguchi, G. (1991) *Introduction to Quality Engineering*. New York: White Plains.
- Tahir, H. H. and Pareja, T. F. (2010) 'MATLAB package and science subjects for undergraduate studies', *International Journal for Cross-Disciplinary Subjects in Education*, 1, (1), pp. 38-42.
- Tansalarak, N. and Claypool, K. T. (2007) 'QMatch - using paths to match XML schemas', *Data & Knowledge Engineering*, 60, (2), pp. 260-282.

- Thang, H. Q. and Nam, V. S. (2010) 'XML schema automatic matching solution', *International Journal of Electrical, Computer, and Systems Engineering*, 4, (1), pp. 68-74.
- Thellufsen, C., Rajabifard, A., Enemark, S. and Williamson, I. (2009) 'Awareness as a foundation for developing effective spatial data infrastructures', *Land Use Policy*, 26, (2), pp. 254-261.
- Topcon (2012a) *Electronic total station GTS-220 series*. Available at: <http://www.topcon.com.sg/survey/cs230.html> (Accessed: 07-05-2012).
- Topcon (2012b) *GR-3 GPS: Basic specifications*. Available at: http://www.topografia-global.com/catalogo/gps/doblef/to_gr3.pdf (Accessed: 26-03-2012).
- Turner, A. J. (2006) *Introduction to Neogeography*. Short Cuts Series: O'Reilly Media, Inc
- Tveite, H. and Langaas, S. (1999) 'An accuracy assessment method for geographical line data sets based on buffering', *International Journal of Geographical Information Science*, 13, (1), pp. 27 - 47.
- Tversky, A. (1977) 'Features of similarity', *Psychological Review*, 84, (4), pp. 327-352.
- Uitermark, H. T. and Dutch, C. (1996) 'The integration of geographic databases: realising geodata interoperability through the hypermap metaphor and a mediator architecture', *Proceedings of the Second Joint European Conference and Exhibition on Geographical Information* Barcelona, Spain, IOS Press, pp. 92-95.
- Ulubay, A. and Altan, M. O. (2002) 'A different approach to the spatial data integration', *The Symposium on Geospatial Theory, Processing and Applications*. Ottawa, Canada, pp. 1-6.
- USBB. (1947) *United States National Map Accuracy Standards*. Washington, D.C., U.S. : Bureau of the Budget
- Usery, E., L, Finn, M., P. and Starbuck, M. (2005) 'Integrating data layers to support the national map of the united states', *International Cartographic Conference*. A Corua, Spain, pp. 1-9.
- Veljanovski, T., Kanjir, U., Pehani, P., Oštir, K. and Kovačič, P. (2012) 'Object-based image analysis of VHR satellite imagery for population estimation in informal settlement Kibera-Nairobi, Kenya', in Escalante-Ramírez, B.(ed), *Remote Sensing-Applications*. Rijeka, Croatia: In Tech, pp. 407-434.

- Veregin, H. (1999) 'Data quality parameters', in Longley, P., Goodchild, M., Maguire, D. and Rhind, D.(eds) *Geographical Information Systems Principles and Technical Issues*. USA: John Wiley & Sons, Inc, pp. 177-189.
- Vickery, G. and Wunsch-Vincent, S. (2007) *Participative Web and User-Created Content: Web 2.0, wikis and social networking*. USA: OECD.
- W3Schools (2012) *An XSD Example*. Available at: http://www.w3schools.com/schema/schema_example.asp (Accessed: 08-05-2012).
- Wald, L. (1999) 'Some terms of reference in data fusion', *Geoscience and Remote Sensing, IEEE Transactions on*, 37, (3), pp. 1190-1193.
- Weaver, B. (2004) *Implementing of the national map road database*. Nashville, Tennessee, The United States:
- Wiki-OpenStreetMap (2011) *State of the map 2012/call for venues/Tokyo*. Available at: http://wiki.openstreetmap.org/wiki/State_of_the_Map_2012/Call_for_venues/Tokyo (Accessed: 28-08-2012).
- Wiki-OpenStreetMap (2012) *Quality assurance*. Available at: http://wiki.openstreetmap.org/wiki/Quality_assurance (Accessed: 12-09-2012).
- Wiki-Project-Kenya (2012) Available at: http://wiki.openstreetmap.org/wiki/WikiProject_Kenya (Accessed: 10-09-2012).
- Wiki-Project-Thailand (2012) Available at: http://wiki.openstreetmap.org/wiki/WikiProject_Thailand (Accessed: 06-09-2012).
- Wikimapia-statistics (2012) *Users account, Places total*. Available at: http://wikimapia.org/stats/action_stats/?fstat=6&period=2&year=2009&month=6 (Accessed: 24-04-2012).
- Wolf, E. B., Matthews, G. D., McNinch, K. and Poore, B. S. (2011) *OpenStreetMap collaborative prototype, phase one*. Reston, Virginia: U.S. Geological Survey
- Wu, Z. and Palmer, M. (1994) 'Verb semantics and lexical selection', *32nd. Annual Meeting of the Association for Computational Linguistics*. Las Cruces, New Mexico, USA, pp. 133-138.
- Yahoo (2012) *Flickr, photo-sharing passions*. Available at: <http://advertising.yahoo.com/article/flickr.html> (Accessed: 21-04-2012).
- Zandbergen, P. (2008) 'Positional accuracy of spatial data: non-normal distributions and a critique of the national standard for spatial data accuracy', *Transactions in GIS*, 12, (1), pp. 103-130.

- Ziegler, P. and Dittrich, K. R. (2004) 'Three decades of data integration - All problems solved?', *In 18th IFIP World Computer Congress (WCC 2004)*. Toulouse, France, pp. 3-12.
- Zielstra, D. and Hochmair, H. (2011) 'Comparative study of pedestrian accessibility to transit stations using free and proprietary network data', *Transportation Research Record: Journal of the Transportation Research Board*, 2217, pp. 145-152.
- Zielstra, D. and Zipf, A. (2010) 'Quantitative studies on the data quality of OpenStreetMap in Germany', *Sixth International Conference on Geographic Information Science, GIScience 2010*. Zurich, Switzerland, pp. 1-7.
- Zook, M., Graham, M., Shelton, T. and Gorman, S. (2010) 'Volunteered geographic information and crowdsourcing disaster relief: a case study of the Haitian earthquake', *World Medical & Health Policy*, 2, (2), pp. 7-33.

Appendix A Maps of the Test Areas from OpenStreetMap Project



Figure A-1 Screenshot of OSM map for Cramlington 1 and 2-UK study area (image sampled on 05/09/2012, rendered at zoom level -14/18). It is available at: <http://www.openstreetmap.org/>



Figure A-2 Screenshot of OSM map for Clara Vale-UK study area (image sampled on 05/09/2012, rendered at zoom level -15/18). It is available at: <http://www.openstreetmap.org/>



Figure A-3 Screenshot of OSM map for Baghdad-Iraq study area (image sampled on 05/09/2012, rendered at zoom level -15/18). It is available at: <http://www.openstreetmap.org/>

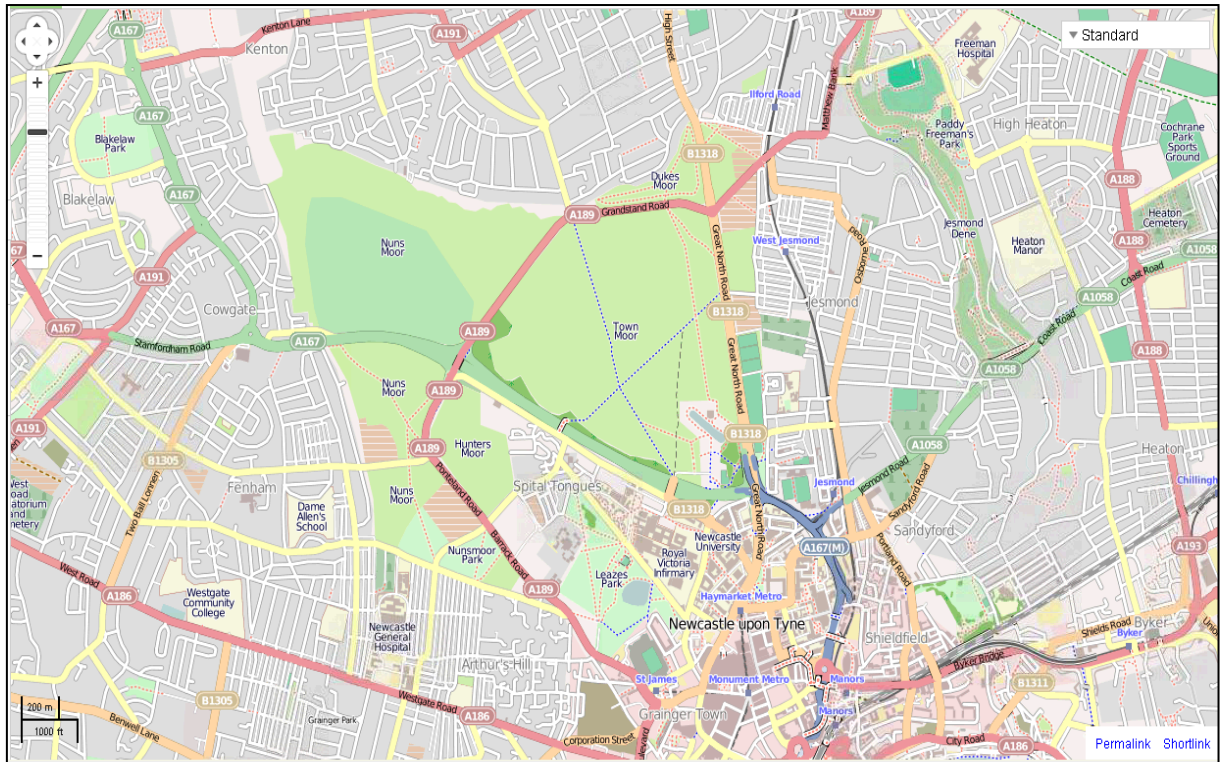


Figure A-4 Screenshot of OSM map for Newcastle city centre-UK study area (image sampled on 05/09/2012, rendered at zoom level -14/18). It is available at: <http://www.openstreetmap.org/>

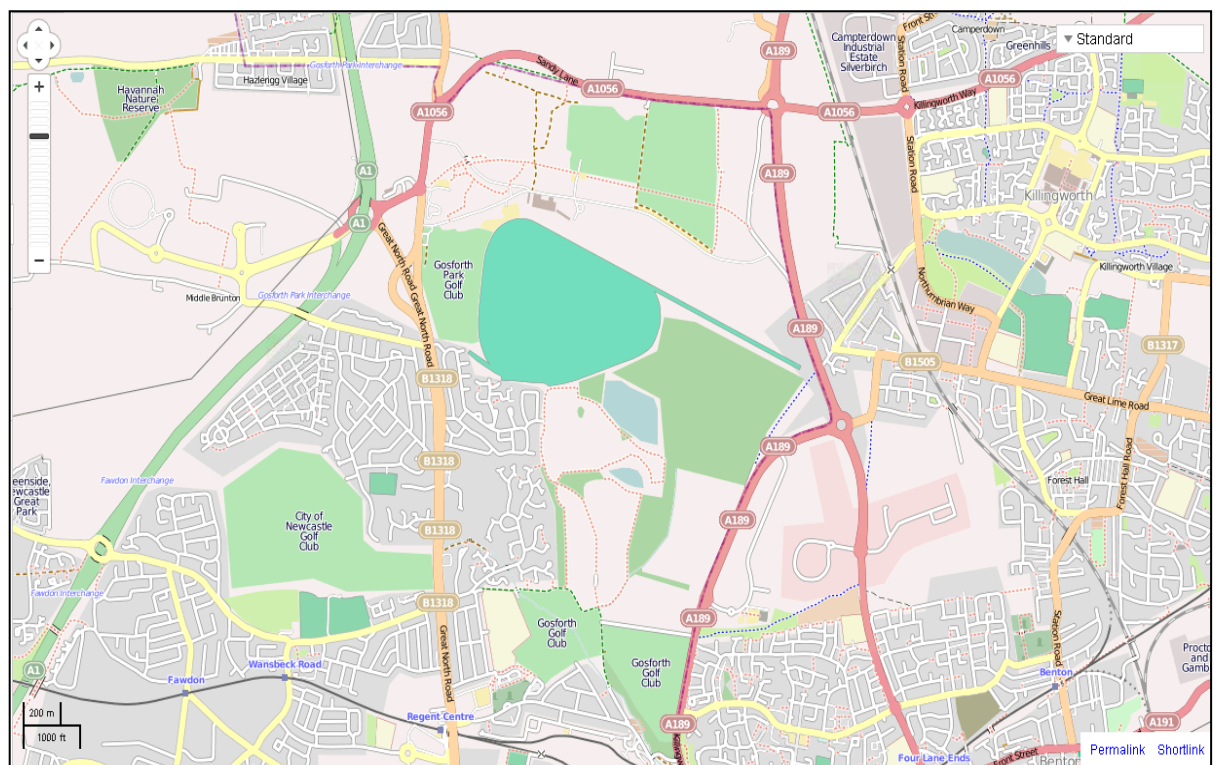


Figure A-5 Screenshot of OSM map for Gosforth-UK study area (image sampled on 05/09/2012, rendered at zoom level -14/18). It is available at: <http://www.openstreetmap.org/>

Appendix B The Programs of Geometrical Similarity Measurements Interfaces

B.1 The positional similarity measurement interface's program (Pos.m)

```
% #####Function for creating the interface#####

function varargout = Pos(varargin)
gui_Singleton = 1;
gui_State = struct('gui_Name',       mfilename, ...
                  'gui_Singleton',   gui_Singleton, ...
                  'gui_OpeningFcn', @Pos_OpeningFcn, ...
                  'gui_OutputFcn',  @Pos_OutputFcn, ...
                  'gui_LayoutFcn',   [], ...
                  'gui_Callback',    []);
if nargin && ischar(varargin{1})
    gui_State.gui_Callback = str2func(varargin{1});
end

if nargout
    [varargout{1:nargout}] = gui_mainfcn(gui_State, varargin{:});
else
    gui_mainfcn(gui_State, varargin{:});
end

function Pos_OpeningFcn(hObject, eventdata, handles, varargin)

handles.output = hObject;

#####Update handles structure#####
guidata(hObject, handles);

function varargout = Pos_OutputFcn(hObject, eventdata, handles)

varargout{1} = handles.output;

% #####Executes on button press in pushbutton1#####
function pushbutton1_Callback(hObject, eventdata, handles)

global pathname

[namefile,pathname]=uigetfile({'*.txt','Text Files (*.txt)'},'Choose Input
File');
[NO,E,N,ET,NT]=textread(strcat(pathname,namefile));
global Datx
global TT
global PRMSE
global cir

Datx=zeros(6,2);
[rz1, cz1]=size(ET);
PRMSE=zeros(rz1,1);
set(handles.Tab1,'data',PRMSE);

dir4=zeros(rz1,1);
for i=1:rz1
    DE(i,1)=ET(i,1)-E(i,1);
    DN(i,1)=NT(i,1)-N(i,1);
end

% #####Defining the symbols#####

MME=mean(DE);
```

```

MMN=mean (DN) ;
SE=std (DE) ;
SN=std (DN) ;
MDE=median (DE) ;
MDN=median (DN) ;
MAXE=max (DE) ;
MAXN=max (DN) ;
MINE=min (DE) ;
MINN=min (DN) ;
IRE=IQR (DE) ;
IRN=IQR (DN) ;

Datx=[MME MMN;SE SN;MDE MDN;MAXE MAXN;MINE MINN;IRE IRN];

Datx=roundn (Datx, -3);

% #####Statistical calculations#####

for i=1:rz1
    DES (i,1)=(DE (i,1))^2;
    DNS (i,1)=(DN (i,1))^2;
    PRMSE (i,1)=(DES (i,1)+DNS (i,1))^0.5;
end
a=sum (DES) ;
RMSEx=sqrt (a/rz1) ;
may1=sprintf ('%0.3f', RMSEx) ;
set (handles.TXT1, 'String', may1) ;
b=sum (DNS) ;
RMSEy=sqrt (b/rz1) ;
may2=sprintf ('%0.3f', RMSEy) ;
set (handles.TXT2, 'String', may2) ;

c=DE-RMSEx;
for i=1:rz1
    d (i,1)=(c (i,1))^2;
end
e=sum (d) ;
Sx=sqrt (e/ (rz1-1)) ;

cc=DN-RMSEy;
for i=1:rz1
    dd (i,1)=(cc (i,1))^2;
end
ee=sum (dd) ;
Sy=sqrt (ee/ (rz1-1)) ;

Sh=(Sx+Sy) /2;
avg=(sum (DES)+sum (DNS)) /rz1;
TRMSE=(avg)^0.5;
n=((1.96)^2*(Sh)^2)/((0.20)*TRMSE)^2;

if RMSEx==RMSEy
    NSSDAstatistic= 1.7308*TRMSE;

else NSSDAstatistic= 2.4477*0.5*(RMSEx+RMSEy);
end

TT=[RMSEx;RMSEy;TRMSE;NSSDAstatistic];
may3=sprintf ('%0.3f', NSSDAstatistic) ;

set (handles.TXT3, 'String', may3) ;

may4=sprintf ('%0.3f', TRMSE) ;

set (handles.TXT4, 'String', may4) ;

axes (handles.axes1) ;

```

```

cla

quiver(E,N,DE,DN);hold on
plot(E,N,'r.');
```

%Title('Resultant Shift in Points','fontsize',10);

```

xlabel('Easting (m)','fontsize',12);
ylabel('Northing (m)','fontsize',12);
axis tight
```

Points numbers on graph

```

for i=1:rz1
    n1=num2str(i);

    st1=strcat('p',n1);

text(E(i,1),N(i,1),st1);
end

for i=1:rz1
    dir1(i,1)=(DN(i,1)/DE(i,1));
end
dir2=atan(dir1);
dir3=dir2*180/pi;
for i=1:rz1
if DE(i,1)>0 & DN(i,1)>0;
    dir4(i,1)=90-dir3(i,1);
else
    if DE(i,1)>0 & DN(i,1)<0;
        dir4(i,1)=90+(abs(dir3(i,1)));
    else
        if DE(i,1)<0 & DN(i,1)<0;
            dir4(i,1)=270-dir3(i,1);
        else
            if DE(i,1)<0 & DN(i,1)>0;
                dir4(i,1)=270+(abs(dir3(i,1)));
            end
        end
    end
end
end
end
end

dir3;
dir4;
```

#####Circular Statistics#####

```

dir5=(dir4*pi/180);
s=sum(sin(dir5));
c=sum(cos(dir5));
r=sqrt((c)^2+(s)^2);
if s>0 & c>0
angrad=atan(s/c);
else
    if c<0
        angrad=(atan(s/c))+pi;
    else
        if s<0 & c>0
            angrad=(atan(s/c))+(2*pi);
        end
    end
end
end
angdeg=angrad*180/pi;
set(handles.Txt5, 'String', angdeg);
rb=r/rz1;

set(handles.Txt6, 'String', rb);
```



```

v=1-rb ;
set(handles.Txt7, 'String', v);

vs=(180/pi)*sqrt(2*v);

cir=[angdeg;rb;v];

#####Compass plots#####

axes(handles.axes4);
cla;
[x,y]=pol2cart(dir5,PRMSE);
compass(x,y);
camorbit(0,180);
camroll(90);

axes(handles.axes1);
PRMSE1=PRMSE;
[Tr,Tc]=size(PRMSE);

PRMSE=roundn(PRMSE,-3);

PRMSE1(:,2)=PRMSE;
PRMSE1(:,1)=1:Tr;

set(handles.Tab1,'data',PRMSE1);
hleg1=legend('Planimetric error','Reference dataset','fontsize','12');
set(hleg1,'location','southeast','location','northwest');
set(handles.Tab2,'data',Datx);

% #####Function for creating the interface#####
function FileMenu_Callback(hObject, eventdata, handles)

function OpenMenuItem_Callback(hObject, eventdata, handles)
file = uigetfile('*.fig');
if ~isequal(file, 0)
    open(file);
end

% -----
function PrintMenuItem_Callback(hObject, eventdata, handles)

printdlg(handles.figure1)

% -----
function CloseMenuItem_Callback(hObject, eventdata, handles)

selection = questdlg(['Close ' get(handles.figure1,'Name') '?'],...
                    ['Close ' get(handles.figure1,'Name') '...'],...
                    'Yes','No','Yes');
if strcmp(selection,'No')
    return;
end

delete(handles.figure1)

% --- Executes on selection change in popupmenu1.
function popupmenu1_Callback(hObject, eventdata, handles)

% --- Executes during object creation, after setting all properties.
function popupmenu1_CreateFcn(hObject, eventdata, handles)

if ispc && isequal(get(hObject,'BackgroundColor'),
get(0,'defaultUiControlBackgroundColor'))
    set(hObject,'BackgroundColor','white');
end

```

```

set(hObject, 'String', {'plot(rand(5))', 'plot(sin(1:0.01:25))',
'bar(1:.5:10)', 'plot(membrane)', 'surf(peaks)'});

% --- Executes on button press in togglebutton1.
function togglebutton1_Callback(hObject, eventdata, handles)

function TXT1_Callback(hObject, eventdata, handles)

% --- Executes during object creation, after setting all properties.
function TXT1_CreateFcn(hObject, eventdata, handles)

if ispc && isequal(get(hObject,'BackgroundColor'),
get(0,'defaultUiControlBackgroundColor'))
    set(hObject,'BackgroundColor','white');
end

% --- Executes when entered data in editable cell(s) in Tab1.
function Tab1_CellEditCallback(hObject, eventdata, handles)

data = get(hObject, 'data');

function TXT2_Callback(hObject, eventdata, handles)

function TXT2_CreateFcn(hObject, eventdata, handles)

if ispc && isequal(get(hObject,'BackgroundColor'),
get(0,'defaultUiControlBackgroundColor'))
    set(hObject,'BackgroundColor','white');
end

function TXT3_Callback(hObject, eventdata, handles)

function TXT3_CreateFcn(hObject, eventdata, handles)

if ispc && isequal(get(hObject,'BackgroundColor'),
get(0,'defaultUiControlBackgroundColor'))
    set(hObject,'BackgroundColor','white');
end

function TXT4_Callback(hObject, eventdata, handles)

function TXT4_CreateFcn(hObject, eventdata, handles)

if ispc && isequal(get(hObject,'BackgroundColor'),
get(0,'defaultUiControlBackgroundColor'))
    set(hObject,'BackgroundColor','white');
end

% --- Executes during object creation, after setting all properties.
function Tab1_CreateFcn(hObject, eventdata, handles)

function Txt5_Callback(hObject, eventdata, handles)

function Txt5_CreateFcn(hObject, eventdata, handles)
if ispc && isequal(get(hObject,'BackgroundColor'),
get(0,'defaultUiControlBackgroundColor'))
    set(hObject,'BackgroundColor','white');
end
function Txt6_Callback(hObject, eventdata, handles)

function Txt6_CreateFcn(hObject, eventdata, handles)
if ispc && isequal(get(hObject,'BackgroundColor'),
get(0,'defaultUiControlBackgroundColor'))
    set(hObject,'BackgroundColor','white');
end
end

```

```

function Txt7_Callback(hObject, eventdata, handles)

function Txt7_CreateFcn(hObject, eventdata, handles)

if ispc && isequal(get(hObject,'BackgroundColor'),
get(0,'defaultUiControlBackgroundColor'))
    set(hObject,'BackgroundColor','white');
end

function Txt8_Callback(hObject, eventdata, handles)

function Txt8_CreateFcn(hObject, eventdata, handles)

if ispc && isequal(get(hObject,'BackgroundColor'),
get(0,'defaultUiControlBackgroundColor'))
    set(hObject,'BackgroundColor','white');
end
function edit21_Callback(hObject, eventdata, handles)

function edit21_CreateFcn(hObject, eventdata, handles)

if ispc && isequal(get(hObject,'BackgroundColor'),
get(0,'defaultUiControlBackgroundColor'))
    set(hObject,'BackgroundColor','white');
end

function edit22_Callback(hObject, eventdata, handles)

function edit22_CreateFcn(hObject, eventdata, handles)

if ispc && isequal(get(hObject,'BackgroundColor'),
get(0,'defaultUiControlBackgroundColor'))
    set(hObject,'BackgroundColor','white');
end

function edit23_Callback(hObject, eventdata, handles)

function edit23_CreateFcn(hObject, eventdata, handles)

if ispc && isequal(get(hObject,'BackgroundColor'),
get(0,'defaultUiControlBackgroundColor'))
    set(hObject,'BackgroundColor','white');
end

function edit24_Callback(hObject, eventdata, handles)

function edit24_CreateFcn(hObject, eventdata, handles)

if ispc && isequal(get(hObject,'BackgroundColor'),
get(0,'defaultUiControlBackgroundColor'))
    set(hObject,'BackgroundColor','white');
end

function axes1_CreateFcn(hObject, eventdata, handles)

function axes1_ButtonDownFcn(hObject, eventdata, handles)

function pushbutton5_Callback(hObject, eventdata, handles)

R=uigetdir;
global Datx
N=strcat(R,'\','Datx.xlsx');
xlswrite(N,Datx);
global TT
N=strcat(R,'\','TT.xlsx');
xlswrite(N,TT);
global PRMSE

```

```
N=strcat(R, '\\', 'PRMSE.xlsx');
xlswrite(N, PRMSE);
global cir
N=strcat(R, '\\', 'cir.xlsx');
xlswrite(N, cir);
```

B.2 The linear similarity measurement interface's program (Lin.m)

```
% #####Function for creating the interface#####
function varargout = Lin(varargin)
gui_Singleton = 1;
gui_State = struct('gui_Name',       mfilename, ...
                  'gui_Singleton',   gui_Singleton, ...
                  'gui_OpeningFcn', @Lin_OpeningFcn, ...
                  'gui_OutputFcn',  @Lin_OutputFcn, ...
                  'gui_LayoutFcn',  [], ...
                  'gui_Callback',    []);
if nargin && ischar(varargin{1})
    gui_State.gui_Callback = str2func(varargin{1});
end

if nargout
    [varargout{1:nargout}] = gui_mainfcn(gui_State, varargin{:});
else
    gui_mainfcn(gui_State, varargin{:});
end

% --- Executes just before Lin is made visible.
function Lin_OpeningFcn(hObject, eventdata, handles, varargin)
handles.output = hObject;
guidata(hObject, handles);

function varargout = Lin_OutputFcn(hObject, eventdata, handles)
varargout{1} = handles.output;

% --- Executes on button press in pushbutton1.
function pushbutton1_Callback(hObject, eventdata, handles)
global pathname

[namefile,pathname]=uigetfile({'*.txt','Text Files (*.txt)'},'Choose Input
File');
[FSB1,OSMB1,INTB1,OSB2,OSMB2,INTB2,FSB3,OSB3,INTB3,fsx,fsy,osx,osy,osmx,osmy]=
textread(strcat(pathname,namefile));

global OLa
global OLsa
global OLfa
global ADa

% #####Reading the input data files#####
F1='fs.xlsx';
[Dat1]=xlsread(strcat(pathname,F1));

[r1,c1]=size(Dat1);

F2='fm.xlsx';
[Dat2]=xlsread(strcat(pathname,F2));
[r2,c2]=size(Dat2);

F3='osm.xlsx';
[Dat3]=xlsread(strcat(pathname,F3));

[r3,c3]=size(Dat3);
```

```

% ##### Plotting three percentage graphs #####
j=0;
for i=1:25
    j=j+1;
    sub1(i,1)=FSB1(i,1)-INTB1(i,1);
    sub2(i,1)=OSMB1(i,1)-INTB1(i,1);
    suma(i,1)=INTB1(i,1)+sub1(i,1)+sub2(i,1);
    OL1(i,1)=(INTB1(i,1)/suma(i,1))*100;
    OL2(i,1)=(sub1(i,1)/suma(i,1))*100;
    OL3(i,1)=(sub2(i,1)/suma(i,1))*100;

    subs1(i,1)=OSB2(i,1)-INTB2(i,1);
    subs2(i,1)=OSMB2(i,1)-INTB2(i,1);
    sumas(i,1)=INTB2(i,1)+subs1(i,1)+subs2(i,1);
    OLs1(i,1)=(INTB2(i,1)/sumas(i,1))*100;
    OLs2(i,1)=(subs1(i,1)/sumas(i,1))*100;
    OLs3(i,1)=(subs2(i,1)/sumas(i,1))*100;

    subf1(i,1)=FSB3(i,1)-INTB3(i,1);
    subf2(i,1)=OSB3(i,1)-INTB3(i,1);
    sumaf(i,1)=INTB3(i,1)+subf1(i,1)+subf2(i,1);
    OLf1(i,1)=(INTB3(i,1)/sumaf(i,1))*100;
    OLf2(i,1)=(subf1(i,1)/sumaf(i,1))*100;
    OLf3(i,1)=(subf2(i,1)/sumaf(i,1))*100;
end

OLa=OL1;
OLa(:,2)=OL2;
OLa(:,3)=OL3;

OLsa=OLs1;
OLsa(:,2)=OLs2;
OLsa(:,3)=OLs3;

OLfa=OLf1;
OLfa(:,2)=OLf2;
OLfa(:,3)=OLf3;
x=0.5:0.5:12.5;
axes(handles.axes1);
cla
plot(x,OL1,'b',x,OL2,'k',x,OL3,'r','linewidth',1.5)

set(get(gcf,'CurrentAxes'),'FontName','Arial','FontSize',6);
title('The accuracy of OSM linear feature relative to FS
dataset','fontsize',8);
%tt=title({'The accuracy of OSM linear feature';'relative to FS dataset'});
set(gca,'ylim',[0 100], 'ytick',[0:10:100]);
ylabel('Percentage (%)');
xlabel('Buffer Size (m)');
hleg1=legend('Area in FS,in OSM','Area in FS,out OSM',' Area out FS,in OSM');
set(hleg1,'Location','East','fontsize',5);
set(gca,'xlim',[0 12.5])
%
x=0.5:0.5:12.5;
axes(handles.axes2);
cla
plot(x,OLs1,'b',x,OLs2,'k',x,OLs3,'r','linewidth',1.5)

set(get(gcf,'CurrentAxes'),'FontName','Arial','FontSize',6);
title('The accuracy of OSM linear feature relative to FM
dataset','fontsize',8);
set(gca,'ylim',[0 100], 'ytick',[0:10:100]);
ylabel('Percentage (%)');
xlabel('Buffer Size (m)');
hleg1=legend('Area in FM,in OSM','Area in FM,out OSM',' Area out FM,in OSM');
set(hleg1,'Location','East','fontsize',5);
set(gca,'xlim',[0 12.5])

```

```

%
x=0.2:0.2:5;
axes(handles.axes3);
cla

plot(x,OLf1,'b',x,OLf2,'k',x,OLf3,'r','linewidth',1.5)
set(get(gcf,'CurrentAxes'),'FontName','Arial','FontSize',6);
title('The accuracy of FM linear feature relative to FS
dataset','fontsize',8);
set(gca,'ylim',[0 100], 'ytick',[0:10:100]);
ylabel('Percentage (%)');
xlabel('Buffer Size (m)');
hleg1=legend('Area in FS,in FM','Area in FS,out FM',' Area out FS,in FM');
set(hleg1,'Location','East','fontsize',5);
set(gca,'xlim',[0 5])
%

% ##### Plotting average displacement graphs #####
j=0;
for i=1:25
    j=j+0.5;
    QN1(i,1)=OSMB1(i,1)-INTB1(i,1);
    AD1(i,1)=pi*j*(QN1(i,1)/FSB1(i,1));
end

x=0.5:0.5:12.5;
axes(handles.axes4);
cla
plot(x,AD1,'r','linewidth',1.5)
set(get(gcf,'CurrentAxes'),'FontName','Arial','FontSize',7);
set(gca,'xlim',[0 12.5]);
ylabel('Average Displacement (m)','FontSize',6);
xlabel('Buffer Size (m)');

hold on

j=0;
for i=1:25
    j=j+0.5;
    QN2(i,1)=OSMB2(i,1)-INTB2(i,1);
    AD2(i,1)=pi*j*(QN2(i,1)/OSB2(i,1));
end

x=0.5:0.5:12.5;
plot(x,AD2,'b','linewidth',1.5)
set(get(gcf,'CurrentAxes'),'FontName','Arial','FontSize',7);
set(gca,'xlim',[0 12.5]);
ylabel('Average Displacement (m)');
xlabel('Buffer Size (m)');

hold on

j=0;
for i=1:25
    j=j+0.2;
    QN3(i,1)=OSB3(i,1)-INTB3(i,1);
    AD3(i,1)=pi*j*(QN3(i,1)/FSB3(i,1));
end

x=0.2:0.2:5;
plot(x,AD3,'k','linewidth',1.5)
set(get(gcf,'CurrentAxes'),'FontName','Arial','FontSize',7);
set(gca,'xlim',[0 12.5]);
ylabel('Average Displacement (m)');
xlabel('Buffer Size (m)');
hleg1=legend('FS,OSM','FM,OSM','FS,FM');
set(hleg1,'Location','southeast','fontsize',4);
    
```

```

% ##### Plotting linear data overlay#####
axes(handles.axes5);
cla
i=0;
for i=1:2:c1
    X1=Dat1(:,i);
    X2=Dat1(:,i+1);

    X1=X1(X1~=0);
    X2=X2(X2~=0);
h1=line (X1,X2,'linewidth',3,'Color','k');
hold on
end

i=0;
for i=1:2:c2
    XX1=Dat2(:,i);
    XX2=Dat2(:,i+1);

    XX1=XX1(XX1~=0);
    XX2=XX2(XX2~=0);

h2=line(XX1,XX2,'linewidth',3,'Color','r');
hold on
end

i=0;
for i=1:2:c3
    XXX1=Dat3(:,i);
    XXX2=Dat3(:,i+1);

    XXX1=XXX1(XXX1~=0);
    XXX2=XXX2(XXX2~=0);

h3=line(XXX1,XXX2,'linewidth',3,'Color','b');
hold on

axis equal
end

hleg1=legend([h1 h2 h3],{'FS', 'FM','OSM'});
set(hleg1,'location','southeast','linewidth',1,'fontSize',8);

ADa=AD1;
ADa(:,2)=AD2;
ADa(:,3)=AD3;
set(handles.Tab1,'data',OLa);
set(handles.Tab2,'data',OLsa);
set(handles.Tab3,'data',OLfa);
set(handles.Tab4,'data',ADa);

% #####Function for creating the interface#####
function FileMenu_Callback(hObject, eventdata, handles)
% -----
function OpenMenuItem_Callback(hObject, eventdata, handles)
file = uigetfile('*.fig');
if ~isequal(file, 0)
    open(file);
end
% -----
function PrintMenuItem_Callback(hObject, eventdata, handles)
prindlg(handles.figure1)
% -----
function CloseMenuItem_Callback(hObject, eventdata, handles)
selection = questdlg(['Close ' get(handles.figure1,'Name') '?'],...
    ['Close ' get(handles.figure1,'Name') '...'],...
    'Yes','No','Yes');

```

```

if strcmp(selection, 'No')
    return;
end

delete(handles.figure1)
% --- Executes on selection change in popupmenu1.
function popupmenu1_Callback(hObject, eventdata, handles)
function popupmenu1_CreateFcn(hObject, eventdata, handles)
if ispc && isequal(get(hObject, 'BackgroundColor'),
get(0, 'defaultUiControlBackgroundColor'))
    set(hObject, 'BackgroundColor', 'white');
end
set(hObject, 'String', {'plot(rand(5))', 'plot(sin(1:0.01:25))',
'bar(1:.5:10)', 'plot(membrane)', 'surf(peaks)'});

% --- Executes on button press in togglebutton1.
function togglebutton1_Callback(hObject, eventdata, handles)

function TXT1_Callback(hObject, eventdata, handles)

% --- Executes during object creation, after setting all properties.
function TXT1_CreateFcn(hObject, eventdata, handles)
if ispc && isequal(get(hObject, 'BackgroundColor'),
get(0, 'defaultUiControlBackgroundColor'))
    set(hObject, 'BackgroundColor', 'white');
end

% --- Executes when entered data in editable cell(s) in Tab1.
function Tab1_CellEditCallback(hObject, eventdata, handles)
data = get(hObject, 'data');
function TXT2_Callback(hObject, eventdata, handles)

% --- Executes during object creation, after setting all properties.
function TXT2_CreateFcn(hObject, eventdata, handles)
if ispc && isequal(get(hObject, 'BackgroundColor'),
get(0, 'defaultUiControlBackgroundColor'))
    set(hObject, 'BackgroundColor', 'white');
end
function TXT3_Callback(hObject, eventdata, handles)

% --- Executes during object creation, after setting all properties.
function TXT3_CreateFcn(hObject, eventdata, handles)
if ispc && isequal(get(hObject, 'BackgroundColor'),
get(0, 'defaultUiControlBackgroundColor'))
    set(hObject, 'BackgroundColor', 'white');
end
function TXT4_Callback(hObject, eventdata, handles)

% --- Executes during object creation, after setting all properties.
function TXT4_CreateFcn(hObject, eventdata, handles)
if ispc && isequal(get(hObject, 'BackgroundColor'),
get(0, 'defaultUiControlBackgroundColor'))
    set(hObject, 'BackgroundColor', 'white');
end

% --- Executes during object creation, after setting all properties.
function Tab1_CreateFcn(hObject, eventdata, handles)
function Txt5_Callback(hObject, eventdata, handles)

% --- Executes during object creation, after setting all properties.
function Txt5_CreateFcn(hObject, eventdata, handles)
if ispc && isequal(get(hObject, 'BackgroundColor'),
get(0, 'defaultUiControlBackgroundColor'))
    set(hObject, 'BackgroundColor', 'white');
end
function Txt6_Callback(hObject, eventdata, handles)

```



```

% --- Executes during object creation, after setting all properties.
function Txt6_CreateFcn(hObject, eventdata, handles)
if ispc && isequal(get(hObject,'BackgroundColor'),
get(0,'defaultUiControlBackgroundColor'))
    set(hObject,'BackgroundColor','white');
end
function Txt7_Callback(hObject, eventdata, handles)

% --- Executes during object creation, after setting all properties.
function Txt7_CreateFcn(hObject, eventdata, handles)
if ispc && isequal(get(hObject,'BackgroundColor'),
get(0,'defaultUiControlBackgroundColor'))
    set(hObject,'BackgroundColor','white');
end
function Txt8_Callback(hObject, eventdata, handles)

% --- Executes during object creation, after setting all properties.
function Txt8_CreateFcn(hObject, eventdata, handles)
if ispc && isequal(get(hObject,'BackgroundColor'),
get(0,'defaultUiControlBackgroundColor'))
    set(hObject,'BackgroundColor','white');
end
function edit21_Callback(hObject, eventdata, handles)

% --- Executes during object creation, after setting all properties.
function edit21_CreateFcn(hObject, eventdata, handles)
if ispc && isequal(get(hObject,'BackgroundColor'),
get(0,'defaultUiControlBackgroundColor'))
    set(hObject,'BackgroundColor','white');
end

function edit22_Callback(hObject, eventdata, handles)

% --- Executes during object creation, after setting all properties.
function edit22_CreateFcn(hObject, eventdata, handles)
if ispc && isequal(get(hObject,'BackgroundColor'),
get(0,'defaultUiControlBackgroundColor'))
    set(hObject,'BackgroundColor','white');
end
function edit23_Callback(hObject, eventdata, handles)

% --- Executes during object creation, after setting all properties.
function edit23_CreateFcn(hObject, eventdata, handles)
if ispc && isequal(get(hObject,'BackgroundColor'),
get(0,'defaultUiControlBackgroundColor'))
    set(hObject,'BackgroundColor','white');
end
function edit24_Callback(hObject, eventdata, handles)

% --- Executes during object creation, after setting all properties.
function edit24_CreateFcn(hObject, eventdata, handles)
if ispc && isequal(get(hObject,'BackgroundColor'),
get(0,'defaultUiControlBackgroundColor'))
    set(hObject,'BackgroundColor','white');
end

% --- Executes during object creation, after setting all properties.
function axes1_CreateFcn(hObject, eventdata, handles)

% --- Executes on mouse press over axes background.
function axes1_ButtonDownFcn(hObject, eventdata, handles)

% --- Executes on button press in pushbutton5.
function pushbutton5_Callback(hObject, eventdata, handles)

R=uigetdir;
global OLa
N=strcat(R, '\', 'OLa.xlsx');

```

```

xlswrite(N,OLa);
global OLsa
N=strcat(R,'\','OLsa.xlsx');
xlswrite(N,OLsa);
global OLfa
N=strcat(R,'\','OLfa.xlsx');
xlswrite(N,OLfa);
global ADa
N=strcat(R,'\','ADa.xlsx');
xlswrite(N,ADa);

```

B.3 The area (polygon) similarity measurement interface's program (Mom.m)

```

% #####Function for creating the interface#####
function varargout = Mom(varargin)
gui_Singleton = 1;
gui_State = struct('gui_Name',       mfilename, ...
                  'gui_Singleton',   gui_Singleton, ...
                  'gui_OpeningFcn',  @Mom_OpeningFcn, ...
                  'gui_OutputFcn',   @Mom_OutputFcn, ...
                  'gui_LayoutFcn',   [] , ...
                  'gui_Callback',     []);
if nargin && ischar(varargin{1})
    gui_State.gui_Callback = str2func(varargin{1});
end

if nargout
    [varargout{1:nargout}] = gui_mainfcn(gui_State, varargin{:});
else
    gui_mainfcn(gui_State, varargin{:});
end

% --- Executes just before Mom is made visible.
function Mom_OpeningFcn(hObject, eventdata, handles, varargin)
handles.output = hObject;

guidata(hObject, handles);

% --- Outputs from this function are returned to the command line.
function varargout = Mom_OutputFcn(hObject, eventdata, handles)
varargout{1} = handles.output;

% --- Executes on button press in pushbutton1.
function pushbutton1_Callback(hObject, eventdata, handles)
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
gui_Singleton = 1;
gui_State = struct('gui_Name',       mfilename, ...
                  'gui_Singleton',   gui_Singleton, ...
                  'gui_OpeningFcn',  @Mom_OpeningFcn, ...
                  'gui_OutputFcn',   @Mom_OutputFcn, ...
                  'gui_LayoutFcn',   [] , ...
                  'gui_Callback',     []);

global pathname
axes(handles.axes7);
cla;
axes(handles.axes3);
cla;
[namefile,pathname]=uigetfile({'*.txt;*.xlsx;*.xls','Data Files
(*.txt,*.xlsx,*.xls)'],'Chose Data File');

global abcd1
momFM=0;
momFS=0;
momOSM=0;

```

```

set(handles.Tab1,'data',momFM);

set(handles.Tab2,'data',momFS);
set(handles.Tab3,'data',momOSM);

global momFS
global momFM
global momOSM
Filnum=abcd1;

[S1,S2]=size(namefile);
SAM3=namefile(S2-4:S2);

SAM4=namefile(1:2)
if SAM4=='OS'
SAM2=namefile(4:S2-6);
else
    SAM2=namefile(3:S2-6);
end
[DatAA]=xlsread(strcat(pathname,namefile));
Filnum=DatAA(2,3);
for Filnam0=1:Filnum
    Fill=num2str(Filnam0);

for pm=1:3 % ##### Repeating the code tree times #####

% ##### Reading data files #####
    pm
    if pm==1
        SAM1='FM';
    elseif pm==2
        SAM1='FS';
    else
        SAM1='OSM';
    end
    namefile=[ SAM1 SAM2 Fill SAM3 ];

    [Dat0]=xlsread(strcat(pathname,namefile));
    Cent=Dat0(1,3:4); %% Centroid
    Dat0=Dat0(:,1:2);

    [Rs1,Cs1]=size(Dat0);
    Dat1=Dat0;
    Dat1(Rs1+1,:)=Dat1(1,:);

ord1=[0 0;1 0;1 1; 2 0 ;0 2 ;2 1;1 2; 0 3 ; 3 0]; %%moment orders#####
syms x
j=0;
for i=1:Rs1
    i1=i+1;
    x01=Dat1(i,1);
    y01=Dat1(i,2);

    x02=Dat1(i1,1);
    y02=Dat1(i1,2);

    #####Line equation#####
    y=y01+((y02-y01)/(x02-x01))*(x-x01);

    dyx=diff(y);
    dx1=sqrt(1+dyx^2);
    dx=subs(dx1);

    j=j+1;
for pq=1:9

```

```

        p=ord1(pq,1);
        q=ord1(pq,2);
        p_q=p+q;

#####for order 0,0 #####
        if p_q==0
            y4=dx*(x02-x01);

            else

#####Central moment#####
            y1=(x-Cent(1,1))^p*((y-Cent(1,2))^q)*dx;

            y2=subs(y1);

##### Integration Calculations #####

            y3=int(y2,x01,x02);
            y4=subs(y3);

            end

##### Output arrangement #####
            outDat(j,1:2)=Dat1(i,1:2);
            outDat(j,3:4)=Dat1(i1,1:2);
            jz=4+pq;

#####Central moment (out)#####
            outDat(j,jz)=y4;

            end
        end

        for i=1:Rs1

##### Normalized moment#####
            n11(i,1)=(outDat(i,7))/((outDat(i,5))^3);
            n02(i,1)=(outDat(i,9))/((outDat(i,5))^3);
            n20(i,1)=(outDat(i,8))/((outDat(i,5))^3);
            n12(i,1)=(outDat(i,11))/((outDat(i,5))^4);
            n21(i,1)=(outDat(i,10))/((outDat(i,5))^4);
            n03(i,1)=(outDat(i,12))/((outDat(i,5))^4);
            n30(i,1)=(outDat(i,13))/((outDat(i,5))^4);

##### Hu 7 moments invariants #####
            phi1(i,1)=n20(i,1)+n02(i,1);
            phi2(i,1)=(n20(i,1)-n02(i,1))^2+(4*(n11(i,1))^2);
            phi3(i,1)=(n30(i,1)-(3*n12(i,1)))^2+(n03(i,1)-(3*n21(i,1)))^2;
            phi4(i,1)=(n30(i,1)+n12(i,1))^2+(n03(i,1)+n21(i,1))^2;

            a(i,1)=(n30(i,1)-3*n12(i,1))*(n30(i,1)+n12(i,1));
            b(i,1)=(n30(i,1)+n12(i,1))^2-3*(n21(i,1)+n03(i,1))^2;
            c(i,1)=(3*n21(i,1)-n03(i,1))*(n21(i,1)+n03(i,1));
            d(i,1)=3*(n30(i,1)+n12(i,1))^2-(n21(i,1)+n03(i,1))^2;
            phi5(i,1)=(a(i,1)*b(i,1)+(c(i,1)*d(i,1)));

            e(i,1)=(n20(i,1)-n02(i,1))*((n30(i,1)+n12(i,1))^2-
            (n21(i,1)+n03(i,1))^2);
            f(i,1)=4*n11(i,1)*(n30(i,1)+n12(i,1))*(n21(i,1)+n03(i,1));
            phi6(i,1)=e(i,1)+f(i,1);

            g(i,1)=(3*n21(i,1)-n03(i,1))*(n30(i,1)+n12(i,1));
            f(i,1)=(n30(i,1)+n12(i,1))^2-3*(n21(i,1)+n03(i,1))^2;
            h(i,1)=(3*n12(i,1)-n30(i,1))*(n21(i,1)+n03(i,1));
            m(i,1)=3*(n30(i,1)+n12(i,1))^2-(n21(i,1)+n03(i,1))^2;
            phi7=(g(i,1)*f(i,1)+(h(i,1)*m(i,1)));

        end
    
```

```

##### Moment sum for each object#####
mom(1,1)=sum(phi1);
mom(2,1)=sum(phi2);
mom(3,1)=sum(phi3);
mom(4,1)=sum(phi4);
mom(5,1)=sum(phi5);
mom(6,1)=sum(phi6);
mom(7,1)=sum(phi7);

#####Data for each object#####
if pm==1
    DatFM=Dat1;
    momFM1(1:7,Filnam0)=mom;
else
    if pm==2
        DatFS=Dat1;
        momFS1(1:7,Filnam0)=mom;
    else
        if pm==3
            DatOSM=Dat1;
            momOSM1(1:7,Filnam0)=mom;
        end
    end
end
end

#####Plotting area (shape) overlay#####
x1=DatFM(:,1);
y1=DatFM(:,2);
x2=DatFS(:,1);
y2=DatFS(:,2);
x3=DatOSM(:,1);
y3=DatOSM(:,2);
axes(handles.axes7);

xlabel('Easting (m)','fontsize',10);
ylabel('Northing (m)','fontsize',10);
patch(x1,y1,'w','edgecolor','r','LineWidth',1)
alpha (.5)
patch(x2,y2,'w','edgecolor','k','LineWidth',1)

patch(x3,y3,'w','edgecolor','b','LineWidth',1)

hleg1=legend('FM','FS','OSM');
set(hleg1,'Location','southeast');

axis equal
end
hold off

#####Euclidean distance calculations#####
momFM=mean(momFM1');
momFM=momFM';
momFS=mean(momFS1');
momFS=momFS';
momOSM=mean(momOSM1');
momOSM=momOSM';

    dist1=momFM-momFS;
    dist2=dist1.^2;

    dist4=momFM-momOSM;
    dist5=dist4.^2;

    dist7=momFS-momOSM;

```

```

    dist8=dist7.^2;

dist3=(sum(dist2))^(0.5);

dist6=(sum(dist5))^(0.5);

dist9=(sum(dist8))^(0.5);

#####Plotting bar graph for compared moments#####
axes(handles.axes3);
cla;
y=[dist3;dist6;dist9];
bar(y,0.3,'FaceColor','b','EdgeColor','k','LineWidth',1.1);
xlabel('Datasets','fontsize',11);
ylabel('Euclidean Distance','fontsize',11);
set(gca,'XTickLabel',{'FM/FS','FM/OSM','FS/OSM'},'fontsize',10);

% #####Function for creating the interface#####
set(handles.Tab1,'data',momFM);
set(handles.Tab2,'data',momFS);
set(handles.Tab3,'data',momOSM);

% -----
function FileMenu_Callback(hObject, eventdata, handles)

% -----
function OpenMenuItem_Callback(hObject, eventdata, handles)
file = uigetfile('*.fig');
if ~isequal(file, 0)
    open(file);
end

% -----
function PrintMenuItem_Callback(hObject, eventdata, handles)
printdlg(handles.figure1)

% -----
function CloseMenuItem_Callback(hObject, eventdata, handles)
selection = questdlg(['Close ' get(handles.figure1,'Name') '?'],...
    ['Close ' get(handles.figure1,'Name') '...'],...
    'Yes','No','Yes');
if strcmp(selection,'No')
    return;
end

delete(handles.figure1)

% --- Executes on selection change in popupmenu1.
function popupmenu1_Callback(hObject, eventdata, handles)

% --- Executes during object creation, after setting all properties.
function popupmenu1_CreateFcn(hObject, eventdata, handles)
if ispc && isequal(get(hObject,'BackgroundColor'),
get(0,'defaultUiControlBackgroundColor'))
    set(hObject,'BackgroundColor','white');
end

set(hObject, 'String', {'plot(rand(5))', 'plot(sin(1:0.01:25))',
'bar(1:.5:10)', 'plot(membrane)', 'surf(peaks)'});

% --- Executes on button press in togglebutton1.
function togglebutton1_Callback(hObject, eventdata, handles)

```

```

function TXT1_Callback(hObject, eventdata, handles)

% --- Executes during object creation, after setting all properties.
function TXT1_CreateFcn(hObject, eventdata, handles)
if ispc && isequal(get(hObject,'BackgroundColor'),
get(0,'defaultUiControlBackgroundColor'))
    set(hObject,'BackgroundColor','white');
end

% --- Executes when entered data in editable cell(s) in Tab1.
function Tab1_CellEditCallback(hObject, eventdata, handles)
data = get(hObject, 'data');

function TXT2_Callback(hObject, eventdata, handles)

% --- Executes during object creation, after setting all properties.
function TXT2_CreateFcn(hObject, eventdata, handles)
if ispc && isequal(get(hObject,'BackgroundColor'),
get(0,'defaultUiControlBackgroundColor'))
    set(hObject,'BackgroundColor','white');
end

function TXT3_Callback(hObject, eventdata, handles)

% --- Executes during object creation, after setting all properties.
function TXT3_CreateFcn(hObject, eventdata, handles)
if ispc && isequal(get(hObject,'BackgroundColor'),
get(0,'defaultUiControlBackgroundColor'))
    set(hObject,'BackgroundColor','white');
end

function TXT4_Callback(hObject, eventdata, handles)

% --- Executes during object creation, after setting all properties.
function TXT4_CreateFcn(hObject, eventdata, handles)
if ispc && isequal(get(hObject,'BackgroundColor'),
get(0,'defaultUiControlBackgroundColor'))
    set(hObject,'BackgroundColor','white');
end

% --- Executes during object creation, after setting all properties.
function Tab1_CreateFcn(hObject, eventdata, handles)

function Txt5_Callback(hObject, eventdata, handles)

% --- Executes during object creation, after setting all properties.
function Txt5_CreateFcn(hObject, eventdata, handles)
if ispc && isequal(get(hObject,'BackgroundColor'),
get(0,'defaultUiControlBackgroundColor'))
    set(hObject,'BackgroundColor','white');
end

function Txt6_Callback(hObject, eventdata, handles)
% --- Executes during object creation, after setting all properties.
function Txt6_CreateFcn(hObject, eventdata, handles)
if ispc && isequal(get(hObject,'BackgroundColor'),
get(0,'defaultUiControlBackgroundColor'))
    set(hObject,'BackgroundColor','white');
end

function Txt7_Callback(hObject, eventdata, handles)

% --- Executes during object creation, after setting all properties.
function Txt7_CreateFcn(hObject, eventdata, handles)

```

```

if ispc && isequal(get(hObject,'BackgroundColor'),
get(0,'defaultUicontrolBackgroundColor'))
    set(hObject,'BackgroundColor','white');
end

function Txt8_Callback(hObject, eventdata, handles)

% --- Executes during object creation, after setting all properties.
function Txt8_CreateFcn(hObject, eventdata, handles)
if ispc && isequal(get(hObject,'BackgroundColor'),
get(0,'defaultUicontrolBackgroundColor'))
    set(hObject,'BackgroundColor','white');
end

function edit21_Callback(hObject, eventdata, handles)

% --- Executes during object creation, after setting all properties.
function edit21_CreateFcn(hObject, eventdata, handles)
if ispc && isequal(get(hObject,'BackgroundColor'),
get(0,'defaultUicontrolBackgroundColor'))
    set(hObject,'BackgroundColor','white');
end

function edit22_Callback(hObject, eventdata, handles)

% --- Executes during object creation, after setting all properties.
function edit22_CreateFcn(hObject, eventdata, handles)
if ispc && isequal(get(hObject,'BackgroundColor'),
get(0,'defaultUicontrolBackgroundColor'))
    set(hObject,'BackgroundColor','white');
end

function edit23_Callback(hObject, eventdata, handles)

% --- Executes during object creation, after setting all properties.
function edit23_CreateFcn(hObject, eventdata, handles)
if ispc && isequal(get(hObject,'BackgroundColor'),
get(0,'defaultUicontrolBackgroundColor'))
    set(hObject,'BackgroundColor','white');
end

function edit24_Callback(hObject, eventdata, handles)

% --- Executes during object creation, after setting all properties.
function edit24_CreateFcn(hObject, eventdata, handles)
if ispc && isequal(get(hObject,'BackgroundColor'),
get(0,'defaultUicontrolBackgroundColor'))
    set(hObject,'BackgroundColor','white');
end

% --- Executes during object creation, after setting all properties.
function axes1_CreateFcn(hObject, eventdata, handles)

% --- Executes on mouse press over axes background.
function axes1_ButtonDownFcn(hObject, eventdata, handles)
% --- Executes on button press in pushbutton5.
function pushbutton5_Callback(hObject, eventdata, handles)

R=uigetdir;
global momFS
N=strcat(R,'\','momFS.xlsx');
xlswrite(N,momFS);
global momFM
N=strcat(R,'\','momFM.xlsx');
xlswrite(N,momFM);
global momOSM

```



```

N=strcat(R, '\\', 'momOSM.xlsx');
xlswrite(N, momOSM);

function abcd1_Callback(hObject, eventdata, handles)
global abcd1
abcd1 = str2double(get(hObject, 'String'));
if isnan(abcd1)
    set(hObject, 'String', 0);
    errordlg('Input must be a number', 'Error');
end

% --- Executes during object creation, after setting all properties.
function abcd1_CreateFcn(hObject, eventdata, handles)
if ispc && isequal(get(hObject, 'BackgroundColor'),
get(0, 'defaultUicontrolBackgroundColor'))
    set(hObject, 'BackgroundColor', 'white');
end

% --- Executes during object creation, after setting all properties.
function figure1_CreateFcn(hObject, eventdata, handles)

```

Appendix C XML Schema Codes

C.1 XML schema code for OSM data in Cramlington-UK

```

<?xml version="1.0" encoding="utf-8"?>
<xs:schema targetNamespace="http://tempuri.org/XMLSchema.xsd"
elementFormDefault="qualified"
xmlns="http://tempuri.org/XMLSchema.xsd"
xmlns:mstns="http://tempuri.org/XMLSchema.xsd"
xmlns:xs="http://www.w3.org/2001/XMLSchema">
<xs:element name="OpentStreetMap_Project">
  <xs:complexType>
    <xs:sequence>
      <xs:element name="Physical">
        <xs:complexType>
          <xs:sequence>
            <xs:element name="Highway" minOccurs="1" maxOccurs="unbounded">
              <xs:complexType>
                <xs:sequence>
                  <xs:element name="Primary" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
                  <xs:element name="Secondary" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
                  <xs:element name="Tertiary" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
                  <xs:element name="Unclassified" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
                  <xs:element name="Track" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
                  <xs:element name="Service" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
                  <xs:element name="Cycleway" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
                  <xs:element name="Mini-roundabout" type="xs:string"
minOccurs="1" maxOccurs="unbounded" />
                  <xs:element name="Footway" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
                  <xs:element name="Residential" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
                </xs:sequence>
              </xs:complexType>
            </xs:element>
            <xs:element name="Shop" minOccurs="1" maxOccurs="unbounded">
              <xs:complexType>
                <xs:sequence>
                  <xs:element name="Doityourself" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
                  <xs:element name="Appliance" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
                  <xs:element name="Car" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
                  <xs:element name="Supermarket" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
                </xs:sequence>
              </xs:complexType>
            </xs:element>
            <xs:element name="Leisure" minOccurs="1" maxOccurs="unbounded">
              <xs:complexType>

```

```

    <xs:sequence>
      <xs:element name="Sports-centre" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
    </xs:sequence>
  </xs:complexType>
</xs:element>
  <xs:element name="Landuse" minOccurs="1" maxOccurs="unbounded">
    <xs:complexType>
      <xs:sequence>
        <xs:element name="Retail" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
        <xs:element name="Residential" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
      </xs:sequence>
    </xs:complexType>
  </xs:element>
  <xs:element name="Amenity" minOccurs="1"
maxOccurs="unbounded">
    <xs:complexType>
      <xs:sequence>
        <xs:element name="Doctors" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
        <xs:element name="Post-box" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
        <xs:element name="Parking" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
        <xs:element name="fast-food" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
        <xs:element name="Library" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
        <xs:element name="Police" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
        <xs:element name="Post-office" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
        <xs:element name="Bicycle-parking" type="xs:string"
minOccurs="1" maxOccurs="unbounded" />
        <xs:element name="Pub" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
        <xs:element name="Atm" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
      </xs:sequence>
    </xs:complexType>
  </xs:element>
</xs:sequence>
</xs:complexType>
</xs:element>
</xs:schema>

```

C.2 XML schema code for OS data in Cramlington-UK

```

<?xml version="1.0" encoding="utf-8"?>
<xs:schema targetNamespace="http://tempuri.org/XMLSchema.xsd"
elementFormDefault="qualified"
xmlns="http://tempuri.org/XMLSchema.xsd"

```

```

xmlns:mstns="http://tempuri.org/XMLSchema.xsd"
xmlns:xs="http://www.w3.org/2001/XMLSchema">
<xs:element name="OS_MasterMap_Topography_layer">
  <xs:complexType>
    <xs:sequence>
      <xs:element name="Roads_Tracks_and_Paths" minOccurs="1"
maxOccurs="unbounded">
        <xs:complexType>
          <xs:sequence>
            <xs:element name="Road_or_Track" minOccurs="1"
maxOccurs="unbounded">
              <xs:complexType>
                <xs:sequence>
                  <xs:element name="Traffic_calming" type="xs:string"
minOccurs="1" maxOccurs="unbounded" />
                </xs:sequence>
              </xs:complexType>
            </xs:element>
            <xs:element name="Roadside" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
            <xs:element name="Path" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
          </xs:sequence>
        </xs:complexType>
      </xs:element>
      <xs:element name="Land">
        <xs:complexType>
          <xs:sequence>
            <xs:element name="Landform" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
            <xs:element name="General_Surface" minOccurs="1"
maxOccurs="unbounded">
              <xs:complexType>
                <xs:sequence>
                  <xs:element name="Step" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
                  <xs:element name="Multi_surface" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
                </xs:sequence>
              </xs:complexType>
            </xs:element>
            <xs:element name="Natural_environment" minOccurs="1"
maxOccurs="unbounded">
              <xs:complexType>
                <xs:sequence>
                  <xs:element name="Nonconiferous_trees" type="xs:string"
minOccurs="1" maxOccurs="unbounded" />
                  <xs:element name="Scrub_Nonconiferous_trees" type="xs:string"
minOccurs="1" maxOccurs="unbounded" />
                </xs:sequence>
              </xs:complexType>
            </xs:element>
            <xs:element name="Buildings" minOccurs="1"
maxOccurs="unbounded">
              <xs:complexType>
                <xs:sequence>

```

```

        <xs:element name="Building" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
    </xs:sequence>
</xs:complexType>
</xs:element>
    <xs:element name="Structures" minOccurs="1"
maxOccurs="unbounded">
    <xs:complexType>
    <xs:sequence>
        <xs:element name="Structure" type="xs:string"
minOccurs="1" maxOccurs="unbounded" />
    </xs:sequence>
    </xs:complexType>
    </xs:element>
    </xs:sequence>
    </xs:complexType>
</xs:element>
</xs:schema>

```

C.3 XML schema code for OSM data in Clara Vale-UK

```

<?xml version="1.0" encoding="utf-8"?>
<xs:schema targetNamespace="http://tempuri.org/XMLSchema.xsd"
elementFormDefault="qualified"
xmlns="http://tempuri.org/XMLSchema.xsd"
xmlns:mstns="http://tempuri.org/XMLSchema.xsd"
xmlns:xs="http://www.w3.org/2001/XMLSchema">
<xs:element name="OpentStreetMap_Project">
    <xs:complexType>
    <xs:sequence>
        <xs:element name="Physical">
    <xs:complexType>
    <xs:sequence>
        <xs:element name="Highway" minOccurs="1" maxOccurs="unbounded">
    <xs:complexType>
    <xs:sequence>
        <xs:element name="Tertiary" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
        <xs:element name="Unclassified" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
        <xs:element name="Track" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
        <xs:element name="Service" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
        <xs:element name="Residential" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
        <xs:element name="Cycleway" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
        <xs:element name="Footway" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
        <xs:element name="Bus_Stop" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
    </xs:sequence>
    </xs:complexType>
    </xs:element>
        <xs:element name="Rialway" minOccurs="1" maxOccurs="unbounded">
    <xs:complexType>
    <xs:sequence>

```

```

        <xs:element name="Rail" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
        <xs:element name="Level_Crossing" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
    </xs:sequence>
</xs:complexType>
</xs:element>
    <xs:element name="Waterway" minOccurs="1" maxOccurs="unbounded">
<xs:complexType>
<xs:sequence>
    <xs:element name="Stream" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
    <xs:element name="Riverbank" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
    <xs:element name="River" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
</xs:sequence>
</xs:complexType>
</xs:element>
    <xs:element name="Landuse" minOccurs="1" maxOccurs="unbounded">
<xs:complexType>
<xs:sequence>
    <xs:element name="Recreation_ground" type="xs:string"
minOccurs="1" maxOccurs="unbounded" />
    <xs:element name="Farm" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
    <xs:element name="Allotments" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
</xs:sequence>
</xs:complexType>
</xs:element>
    <xs:element name="Leisure" minOccurs="1" maxOccurs="unbounded">
<xs:complexType>
<xs:sequence>
    <xs:element name="Garden" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
    <xs:element name="Playground" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
    <xs:element name="Golf_Course" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
    <xs:element name="Nature_Reserve" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
    <xs:element name="Pitch" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
    <xs:element name="Bird_hide" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
</xs:sequence>
</xs:complexType>
</xs:element>
    <xs:element name="Natural" minOccurs="1" maxOccurs="unbounded">
<xs:complexType>
<xs:sequence>
    <xs:element name="Wood" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
    <xs:element name="Water" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
</xs:sequence>
</xs:complexType>
</xs:element>
    <xs:element name="Barrier" minOccurs="1" maxOccurs="unbounded">

```

```

    <xs:complexType>
    <xs:sequence>
        <xs:element name="Gate" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
        <xs:element name="Stile" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
    </xs:sequence>
</xs:complexType>
</xs:element>
    <xs:element name="Power" minOccurs="1" maxOccurs="unbounded">
<xs:complexType>
<xs:sequence>
    <xs:element name="Sub_Station" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
</xs:sequence>
</xs:complexType>
</xs:element>
    <xs:element name="Amenity" minOccurs="1"
maxOccurs="unbounded">
<xs:complexType>
<xs:sequence>
    <xs:element name="Post_box" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
    <xs:element name="Townhall" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
    <xs:element name="Place_of_worship" type="xs:string"
minOccurs="1" maxOccurs="unbounded" />
</xs:sequence>
</xs:complexType>
</xs:element>
    <xs:element name="Historic" minOccurs="1"
maxOccurs="unbounded">
<xs:complexType>
<xs:sequence>
    <xs:element name="Memorial" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
</xs:sequence>
</xs:complexType>
</xs:element>
    <xs:element name="Building" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
</xs:sequence>
</xs:complexType>
</xs:element>
</xs:sequence>
</xs:complexType>
</xs:element>
</xs:schema>

```

C.4 XML schema code for OS data in Clara Vale-UK

```

<?xml version="1.0" encoding="utf-8"?>
<xs:schema targetNamespace="http://tempuri.org/XMLSchema.xsd"
elementFormDefault="qualified"
xmlns="http://tempuri.org/XMLSchema.xsd"
xmlns:mstns="http://tempuri.org/XMLSchema.xsd"
xmlns:xs="http://www.w3.org/2001/XMLSchema">
<xs:element name="OS_MasterMap_Topography_layer">

```

```

<xs:complexType>
<xs:sequence>
  <xs:element name="Water" minOccurs="1" maxOccurs="unbounded">
<xs:complexType>
<xs:sequence>
  <xs:element name="Tidal_water" minOccurs="1"
maxOccurs="unbounded">
<xs:complexType>
<xs:sequence>
  <xs:element name="Foreshore" type="xs:string" />
</xs:sequence>
</xs:complexType>
</xs:element>
  <xs:element name="Inland_Water" type="xs:string" />
</xs:sequence>
</xs:complexType>
</xs:element>
  <xs:element name="Roads_Tracks_and_Paths" minOccurs="1"
maxOccurs="unbounded">
<xs:complexType>
<xs:sequence>
  <xs:element name="Road_or_Track" minOccurs="1"
maxOccurs="unbounded">
<xs:complexType>
<xs:sequence>
  <xs:element name="track" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
</xs:sequence>
</xs:complexType>
</xs:element>
  <xs:element name="Path" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
  <xs:element name="Roadside" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
</xs:sequence>
</xs:complexType>
</xs:element>
  <xs:element name="Rail" minOccurs="1" maxOccurs="unbounded">
<xs:complexType>
<xs:sequence>
  <xs:element name="Rail" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
</xs:sequence>
</xs:complexType>
</xs:element>
  <xs:element name="Land" minOccurs="1" maxOccurs="unbounded">
<xs:complexType>
<xs:sequence>
  <xs:element name="Landform" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
  <xs:element name="General_Surface" minOccurs="1"
maxOccurs="unbounded">
<xs:complexType>
<xs:sequence>
  <xs:element name="Muti_surface" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
</xs:sequence>
</xs:complexType>
</xs:element>

```



```

    <xs:element name="Ntural_Environment" minOccurs="1"
maxOccurs="unbounded">
    <xs:complexType>
    <xs:sequence>
    <xs:element name="Scrub_rough_grassland_nonconiferous_trees"
type="xs:string" minOccurs="1" maxOccurs="unbounded" />
    <xs:element name="Coniferous_trees_nonconiferous_trees_scrub"
type="xs:string" minOccurs="1" maxOccurs="unbounded" />
    <xs:element name="Nonconiferous_trees_Scrub" type="xs:string"
minOccurs="1" maxOccurs="unbounded" />
    <xs:element name="Scrub" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
    <xs:element name="Coniferous_trees_Nonconiferous_trees"
type="xs:string" minOccurs="1" maxOccurs="unbounded" />
    <xs:element name="Nonconiferous_trees" type="xs:string"
minOccurs="1" maxOccurs="unbounded" />
    </xs:sequence>
    </xs:complexType>
    </xs:element>
    </xs:sequence>
    </xs:complexType>
    </xs:element>
    <xs:element name="Buildings" minOccurs="1"
maxOccurs="unbounded">
    <xs:complexType>
    <xs:sequence>
    <xs:element name="Building" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
    </xs:sequence>
    </xs:complexType>
    </xs:element>
    <xs:element name="Structures" minOccurs="1"
maxOccurs="unbounded">
    <xs:complexType>
    <xs:sequence>
    <xs:element name="Structure" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
    </xs:sequence>
    </xs:complexType>
    </xs:element>
    </xs:sequence>
    </xs:complexType>
    </xs:element>
    </xs:schema>

```

C.5 XML schema code for OSM data in Baghdad-Iraq

```

<?xml version="1.0" encoding="utf-8"?>
<xs:schema targetNamespace="http://tempuri.org/XMLSchema.xsd"
elementFormDefault="qualified"
xmlns="http://tempuri.org/XMLSchema.xsd"
xmlns:mstns="http://tempuri.org/XMLSchema.xsd"
xmlns:xs="http://www.w3.org/2001/XMLSchema">
<xs:element name="OpentStreetMap_Project">
    <xs:complexType>
    <xs:sequence>
    <xs:element name="Physical">
    <xs:complexType>
    <xs:sequence>

```

```

        <xs:element name="Highway" minOccurs="1" maxOccurs="unbounded">
        <xs:complexType>
        <xs:sequence>
            <xs:element name="Secondary" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
            <xs:element name="Unclassified" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
            <xs:element name="Track" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
            <xs:element name="Service" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
            <xs:element name="Path" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
        </xs:sequence>
        </xs:complexType>
        </xs:element>
        <xs:element name="Leisure" minOccurs="1" maxOccurs="unbounded">
        <xs:complexType>
        <xs:sequence>
            <xs:element name="Track" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
        </xs:sequence>
        </xs:complexType>
        </xs:element>
        <xs:element name="Natural" minOccurs="1" maxOccurs="unbounded">
        <xs:complexType>
        <xs:sequence>
            <xs:element name="Water" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
        </xs:sequence>
        </xs:complexType>
        </xs:element>
        <xs:element name="Waterway" minOccurs="1"
maxOccurs="unbounded">
        <xs:complexType>
        <xs:sequence>
            <xs:element name="Riverbank" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
            <xs:element name="Stream" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
        </xs:sequence>
        </xs:complexType>
        </xs:element>
        <xs:element name="Landuse" minOccurs="1"
maxOccurs="unbounded">
        <xs:complexType>
        <xs:sequence>
            <xs:element name="Farm" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
        </xs:sequence>
        </xs:complexType>
        </xs:element>
        <xs:element name="Amenity" minOccurs="1"
maxOccurs="unbounded">
        <xs:complexType>
        <xs:sequence>
            <xs:element name="Parking" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
            <xs:element name="College" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />

```

```

</xs:sequence>
</xs:complexType>
</xs:element>
  <xs:element name="Building" type="xs:string" minOccurs="1"
maxOccurs="unbounded">
  </xs:element>
</xs:sequence>
</xs:complexType>
</xs:element>
</xs:sequence>
</xs:complexType>
</xs:element>
</xs:schema>

```

C.6 XML schema code for GDS data in Baghdad-Iraq (in Arabic)

```

<?xml version="1.0" encoding="utf-8"?>
<xs:schema targetNamespace="http://tempuri.org/XMLSchema.xsd"
elementFormDefault="qualified"
xmlns="http://tempuri.org/XMLSchema.xsd"
xmlns:mstns="http://tempuri.org/XMLSchema.xsd"
xmlns:xs="http://www.w3.org/2001/XMLSchema">
<xs:element name="لمساحة العامة المديرية-الاسمية المعلومات">
  <xs:complexType>
  <xs:sequence>
    <xs:element name="طرق" minOccurs="1" maxOccurs="unbounded">
  <xs:complexType>
  <xs:sequence>
    <xs:element name="ثانوي طريق" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
    <xs:element name="معبد غير طريق" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
    <xs:element name="فرعي طريق" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
  </xs:sequence>
  </xs:complexType>
  </xs:element>
    <xs:element name="مائية مسطحات" minOccurs="1"
maxOccurs="unbounded">
  <xs:complexType>
  <xs:sequence>
    <xs:element name="نهر" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
    <xs:element name="بحيرة" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
    <xs:element name="فناة" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
  </xs:sequence>
  </xs:complexType>
  </xs:element>
    <xs:element name="بنايات" minOccurs="1" maxOccurs="unbounded">
  <xs:complexType>
  <xs:sequence>
    <xs:element name="جامعة" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
  </xs:sequence>
  </xs:complexType>
  </xs:element>

```

```

        <xs:element name="حضراء مناطق" minOccurs="1"
maxOccurs="unbounded">
        <xs:complexType>
        <xs:sequence>
        <xs:element name="زراعي حقل" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
        </xs:sequence>
        </xs:complexType>
        </xs:element>
        <xs:element name="سيارات موافق" type="xs:string"
minOccurs="1" maxOccurs="unbounded" />
        <xs:element name="انشاءات" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
        </xs:sequence>
        </xs:complexType>
        </xs:element>
</xs:schema>

```

C.7 XML schema code for GDS data in Baghdad-Iraq (in English)

```

<?xml version="1.0" encoding="utf-8"?>
<xs:schema targetNamespace="http://tempuri.org/XMLSchema.xsd"
elementFormDefault="qualified"
xmlns="http://tempuri.org/XMLSchema.xsd"
xmlns:mstns="http://tempuri.org/XMLSchema.xsd"
xmlns:xs="http://www.w3.org/2001/XMLSchema">
<xs:element name="General_directorate_For_Surveying-data">
<xs:complexType>
<xs:sequence>
<xs:element name="Roads" minOccurs="1" maxOccurs="unbounded">
<xs:complexType>
<xs:sequence>
<xs:element name="Minor-road" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
<xs:element name="Unpaved-road" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
<xs:element name="Branch-road" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
</xs:sequence>
</xs:complexType>
</xs:element>
<xs:element name="Waterbodies" minOccurs="1"
maxOccurs="unbounded">
<xs:complexType>
<xs:sequence>
<xs:element name="River" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
<xs:element name="Lake" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
<xs:element name="canal" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
</xs:sequence>
</xs:complexType>
</xs:element>
<xs:element name="Buildings" minOccurs="1"
maxOccurs="unbounded">
<xs:complexType>
<xs:sequence>
<xs:element name="University" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />

```

```
</xs:sequence>
</xs:complexType>
</xs:element>
  <xs:element name="Green-area" minOccurs="1"
maxOccurs="unbounded">
  <xs:complexType>
  <xs:sequence>
    <xs:element name="Agricultural-farm" type="xs:string"
minOccurs="1" maxOccurs="unbounded" />
    </xs:sequence>
  </xs:complexType>
  </xs:element>
    <xs:element name="Car-park" type="xs:string" minOccurs="1"
maxOccurs="unbounded" />
    <xs:element name="Constructions" type="xs:string"
minOccurs="1" maxOccurs="unbounded" />

  </xs:sequence>
</xs:complexType>
</xs:element>
</xs:schema>
```