

A Geographic Knowledge Discovery Approach to Property Valuation

Katerina Christopoulou

*Thesis submitted for the Degree of
Doctor of Philosophy (PhD)
University College London*

February, 2009

Abstract

This thesis involves an investigation of how knowledge discovery can be applied in the area Geographic Information Science. In particular, its application in the area of property valuation in order to reveal how different spatial entities and their interactions affect the price of the properties is explored. This approach is entirely data driven and does not require previous knowledge of the area applied.

To demonstrate this process, a prototype system has been designed and implemented. It employs association rule mining and associative classification algorithms to uncover any existing inter-relationships and perform the valuation. Various algorithms that perform the above tasks have been proposed in the literature. The algorithm developed in this work is based on the Apriori algorithm. It has been however, extended with an implementation of a ‘Best Rule’ classification scheme based on the Classification Based on Associations (CBA) algorithm.

For the modelling of geographic relationships a graph-theoretic approach has been employed. Graphs have been widely used as modelling tools within the geography domain, primarily for the investigation of network-type systems. In the current context, the graph reflects topological and metric relationships between the spatial entities depicting general spatial arrangements. An efficient graph search algorithm has been developed, based on the Dijkstra shortest path algorithm that enables the investigation of relationships between spatial entities beyond first degree connectivity.

A case study with data from three central London boroughs has been performed to validate the methodology and algorithms, and demonstrate its effectiveness for computer aided property valuation. In addition, through the case study, the influence of location in the value of properties in those boroughs has been examined. The

results are encouraging as they demonstrate the effectiveness of the proposed methodology and algorithms, provided that the data is appropriately pre processed and is of high quality.

Acknowledgements

I am indebted to a number of people who helped me during the period of this work. Although impossible to mention everyone, I would like to express my gratitude to the following.

First, I would like to thank my supervisor Dr Muki Haklay for his direction, encouragement and support throughout the research and the writing process.

My fellow PhD students, Claire Ellul, Jose Paulo de Almeida, Samson Ayugi and Mauren Abreu de Souza for always being there to help me and support me. Thanks also go to my friend and old colleague Anna Pothou for all the encouragement during this stressful period.

I would also like to thank my sponsors in Greece, the Greek State Scholarships Foundation (IKY) and the Hellenic Geographic Military Service (ΓΥΣ) for their financial support. Their support made the completion of this project possible.

Special thanks belong to my husband Dr. Charalambos Makatsoris for his continuous encouragement, support, guidance and especially patience during the whole period. In addition, I would like to thank baby Elle for her patience during the viva preparation period.

Finally, I would like to thank my family in Greece for their support both financial and emotional. Their constant encouragement provided, is greatly appreciated.

Contents

Abstract.....	2
Acknowledgements.....	4
Contents	5
List of figures	8
List of tables.....	11
List of tables.....	11
Introduction	12
1.1 Background	12
1.2 Aims and Objectives	14
1.3 Research Questions	16
1.4 Approach	18
1.5 Main Contributions	20
1.6 Structure of the Thesis	21
2 Knowledge Discovery in Geographic Information Science.....	23
2.1 Knowledge Discovery Process.....	23
2.1.1 Data mining	28
2.1.2 Data Mining Tasks	30
2.1.3 Data Mining Techniques	31
2.2 Knowledge Discovery and Spatial Data	50
2.2.1 GIS and Spatial Database Systems	51
2.2.2 Special characteristics of Spatial Data.....	53
2.2.3 Modelling Spatial Dependencies	56
2.3 Spatial Data Mining	61
2.3.1 Tasks	61
2.3.2 Techniques	62
2.3.3 Existing Systems	78
2.4 Summary	79

3	Property Valuation	82
3.1	Property Market	82
3.2	Property Value	83
3.2.1	Property Value Determinants	86
3.3	Property Valuation	90
3.3.1	Issues in Property Valuation.....	91
3.4	Property Valuation Methods and Techniques	94
3.4.1	Traditional valuation methods.....	94
3.4.2	Advanced techniques.....	97
3.5	Location theory in property valuation.....	103
3.6	Location in Property Valuation Research - Techniques	109
3.7	GIS in Property Valuation.....	116
3.8	Location aware property valuation in a knowledge discovery setting.....	122
3.9	Summary	124
4	Design of a Property Valuation System	126
4.1	Research Opportunities	126
4.2	A new approach.....	128
4.3	The modelling and Knowledge Discovery Algorithm.....	130
4.3.1	Graph-theoretic approach for modelling location.....	130
4.3.2	Graph traversal algorithm.....	135
4.3.3	Data mining algorithm.....	139
4.4	A Procedure for the design and implementation of the system	141
4.4.1	Analysis and Design Methodology	141
4.4.2	Conceptual Architecture and Non-functional Requirements of the System.....	142
4.4.3	System Design.....	144
4.4.4	Data Requirements	153
4.4.5	Database Design.....	154
4.5	Summary	157
5	Implementation	158
5.1	Software Platforms.....	158
5.2	Data Sources.....	160
5.3	Data Acquisition.....	162
5.4	Database Implementation.....	165
5.4.1	Description of the Initial Datasets	166
5.4.2	Level of Representation.....	169
5.4.3	Data Preparation.....	170
5.4.4	Population of the database.....	180
5.5	Study Area.....	183
5.6	Summary	188

6	Case Study	190
6.1	KD Process.....	190
6.2	Design of Experiments.....	194
6.3	Test cases	203
6.4	Summary	227
7	Conclusions.....	229
7.1	Thesis Review	229
7.2	Research Questions Revisited.....	236
7.3	Research Outcome	237
7.4	Future work recommendations.....	239
7.5	Conclusion	241
	References	242
	Appendix A- I2I Landuse	269
	Appendix B - Output .txt file.....	270
	Appendix C- Procedures	272
	Appendix D - Results	274
	Appendix E – CD Contents	286

List of figures

Figure 1-1: Document Structure Overview	21
Figure 2-1: Knowledge Discovery Process	25
Figure 2-2: Example.....	38
Figure 2-3: GIS elements	51
Figure 2-4: Metric Spatial Relationships.....	60
Figure 2-5: Example of generalised predicates	68
Figure 2-6: Example decision tree.....	68
Figure 2-7 : Example of co-location patterns	69
Figure 2-8: Classification of Outlier Detection Methods	73
Figure 3-1: Property valuation approaches.....	98
Figure 3-2: von Thunen model.....	106
Figure 3-3: Alonso's Model	107
Figure 3-4: Assignment process for modelled capital values	119
Figure 3-5: Methodology Overview.....	121
Figure 4-1: Proposed Methodology.....	129
Figure 4-2: Structure of graph that models spatial relationships.....	133
Figure 4-3: Example.....	136
Figure 4-4: Data Structure	137
Figure 4-5: Data mining process overview	139
Figure 4-6: Conceptual System design.....	143
Figure 4-7: Use Cases Diagram	145
Figure 4-8: Package Diagram.....	147
Figure 4-9: StartUp & DBAccess Class Diagram	149
Figure 4-10: JDM Wrapper class diagram	150
Figure 4-11: CBA Manage class diagram	151
Figure 4-12: Sequence diagram.....	152

Figure 4-13: Conceptual Database Schema.....	155
Figure 4-14: Logical Schema	156
Figure 5-1: PROVIDER Website.....	163
Figure 5-2: OS Mastermap dataset.....	164
Figure 5-3: I2I Dataset (Landuse Classification)	165
Figure 5-4: Initial Datasets Relationships	167
Figure 5-5: Address Point	168
Figure 5-6: Points of Interest.....	168
Figure 5-7: Data Integration methodology	171
Figure 5-8: Generalisation.....	172
Figure 5-9: Implemented PL/SQL procedures	173
Figure 5-10: Spatial Metadata & Indexing SQL Statements	174
Figure 5-11: Landuse Taxonomy	177
Figure 5-12: Property & Road Taxonomy	178
Figure 5-13: Database schema population	181
Figure 5-14: 1 st & 2 nd Order Spatial Relationship Graph	182
Figure 5-15: Study Area.....	183
Figure 5-16: Ethnic Composition (Hammersmith & Fulham)	184
Figure 5-17: Accommodation Type (Hammersmith & Fulham).....	185
Figure 5-18: Ethnic Composition (Kensington and Chelsea).....	186
Figure 5-19: Accommodation Type (Kensington & Chelsea).....	186
Figure 5-20: Ethnic Composition (Westminster)	187
Figure 5-21: Accommodation Type (Westminster).....	188
Figure 6-1: Knowledge Discovery process	192
Figure 6-2: Study Parameters	195
Figure 6-3: Experiment Guide.....	197
Figure 6-4: Hammersmith&Fulham_Flat_P11_DescL1_All_15Class.....	200
Figure 6-5: Sample Geographic Distribution	202
Figure 6-6: Geographic distribution for the detached properties sample	205
Figure 6-7: Geographic distribution for the semi-detached properties sample.....	206

Figure 6-8: Geographic distribution for the terraced properties sample.....	206
Figure 6-9: Geographic distribution for the flats and maisonettes sample	207
Figure 6-10: Actual and study volumes of sales comparison per property type.....	208
Figure 6-11: Actual and study volumes of sales comparison per year	210
Figure 6-12: Westminster_Flat_P11_DescL1_3Class	211
Figure 6-13: Westminster_Terrace_P11_DescL1_3Class	212
Figure 6-14: Westminster_SemiDetached_P11_DescL1_3Class	212
Figure 6-15: Westminster_Detached_P11_DescL1_3Class.....	213
Figure 6-16: Kensington&Chelsea_Flat_P11_DescL1_3Class	214
Figure 6-17: Kensington&Chelsea_Terrace_P11_DescL1_3Class.....	214
Figure 6-18: Kensington&Chelsea_SemiDetached_P11_DescL1_3Class.....	215
Figure 6-19: Kensington&Chelsea_Detached_P11_DescL1_3Class.....	215
Figure 6-20: Hammersmith&Fulham_Flat_P11_DescL1_3Class.....	216
Figure 6-21: Hammersmith&Fulham_Terrace_P11_DescL1_3Class.....	216
Figure 6-22: Hammersmith&Fulham_SemiDetached_P11_DescL1_3Class.....	217
Figure 6-23: Hammersmith&Fulham_Detached_P11_DescL1_3Class.....	218
Figure 6-24: 3D value model	219
Figure 6-25: HydePark_Flat_P11_DescL1_3Class.....	223
Figure 6-26: HydePark_Flat_P11_DescL1_3Class Association Rule Extract	224
Figure 6-27: Ward_Flat_P11_DescL1,2,3_All	226

List of tables

Table 2-1: Academic communities and knowledge discovery	27
Table 2-2: Data mining tasks and techniques	32
Table 2-3: Associative Classification Algorithms	44
Table 3-1: RICS Valuation Framework	86
Table 3-2: Value Determinants	87
Table 3-3: Examples of feedforward and recurrent networks	99
Table 3-4: Example model of a neuron	100
Table 3-5: Location aware Hedonic Studies	112
Table 3-6: Location aware ANN studies	115
Table 3-7: Spatial Statistics Studies	116
Table 3-8: GIS use in Property Related Research	117
Table 5-1: Potential Data Providers	161
Table 5-2: Oracle Spatial Operators Used	175
Table 5-3: 2001 Census Variables	179
Table 6-1: Unique vs. All transactions experiments	198
Table 6-2: Comparative results (Landuse Description Level	218
Table 6-3: Comparative Results (Landuse Description Level 2)	221
Table 6-4: Comparative Results (Landuse Description Level3)	221
Table 6-5: Borough_Level Classification Tests	225
Table 7-1: Comparison of approaches	232

1 Introduction

1.1 Background

When considering the important economic value of the land, the process of property valuation stands out as a significant element in land management. Property Valuation involves the estimation of the market value of a property. It is a non-trivial process since it involves the consideration of a variety of underlying factors of the market and the way they affect the value of the property at a given time. Such factors may include governmental policies, geographical factors or even factors such as fashion; season etc. Property valuation also depends on the purpose (e.g. sale, taxation, financing) and the type of the property (residential or commercial), for which it is exercised.

It is widely recognised that there are five main standard valuation methods (Lawrance *et al.*, 1971), of which the Comparative Method is considered as the most reliable but also heavily dependant on the quality of the selected comparables. Recently, a number of techniques for determining the value of property by trying to mimic the thought process of the actors of the market have been developed (Pagourtzi *et al.*, 2003).

The successful application of a valuation method is heavily dependant on the quality and the variety of the data. Among the factors (legal, physical, economic) that influence the value, location is considered to be of outmost importance. Location in terms of proximities to infrastructure or amenities, neighbourhood quality, environmental quality and topology plays an important role in the formulation of the value, therefore can generate variations in price among similar properties. Despite the recognised importance of location in the ways it affects the value of a property (Kauko, 2003) it is currently under-represented in existing valuation models (Wyatt

& Ralphs, 2003). This is mainly due to the modelling difficulties that relate to the wide variety of spatial factors and their interactions that may or may not affect the property in question at given instances of time (Deddis *et al.*, 2002). As a result, in the majority of the cases the incorporation of location is based on a valuer's knowledge and experience (Wyatt & Ralphs, 2003). Examples of research projects that face these challenges include: the multi-level hedonic modelling with location (Orford, 1999); Artificial Neural Networks (Jenkins *et al.*, 1998) and Accessibility Index (Wyatt, 1995). However, there is still a need for new more efficient and accurate location based valuation models (Deddis *et al.*, 2002).

A common characteristic of these approaches is that good knowledge of the area under investigation is required. This knowledge is used for the determination of the key variables in these models. This implies that there is a bias in these models as a result of this knowledge, leading to biased valuations. An alternative is to use an entirely data-driven approach where a priori assumptions about the role of location are not necessary. Knowledge discovery approaches are data driven and are designed to determine unknown patterns and relationships in data that may exist.

Knowledge discovery in databases is the non-trivial process of discovering of valid, novel, potentially useful, and ultimately understandable patterns in data (Fayyad *et al.*, 1996A). A number of methodologies have been proposed for the knowledge discovery processes that are mainly variations of the general process: data preparation-data mining-interpretation of the extracted knowledge. In this study the methodology proposed by Fayyad *et al.* (1996B) will be adopted and involves five basic activities: selection, pre-processing, transformation, data mining and interpretation. These five activities are also relevant when analysis is focused on geographical information (Miller & Han, 2001).

As explained in detail in Section 2.1, the actions involved in each activity are the following. Selection, Pre-processing and Transformation are data preparatory stages, which lead to the core knowledge discovery process – data mining. Data mining is commonly broken into three more sub-stages. The first involves the search and identification of the generic pattern type. The second step includes the identification of the specific data mining technique that is relevant for the problem at hand. The third and final step is the application of the selected technique for pattern search. The

final stage of knowledge discovery is the interpretation/evaluation in which visualization techniques are being used and the discovered knowledge is either integrated into a knowledge-based database or used in a report. In many applications, these activities are not carried out in a sequential manner, but rather iteratively.

When we look at the specifics of data mining, the high-level aims have been identified by Fayyad *et al.* (1996B) as being the prediction and description of the datasets. These are accomplished through the selection and application of an appropriate data mining task. Data mining tasks include (Miller & Han, 2001): Segmentation which can further analysed to Clustering and Classification, Dependency Analysis, Deviation and Outlier Analysis, Trend Detection and finally Generalisation and Characterisation (see Section 2.1.2).

Although knowledge discovery is a quite well established area in conventional databases its application in spatial databases is a new but very promising area for research (Ester *et al.*, 2001). The complexity of geographical phenomena (Gahegan, 2001) along with the large size of spatial datasets not only justifies the application of knowledge discovery to spatial datasets but they also make it highly attractive. The term Geographic Knowledge Discovery is used to describe the application of a general knowledge discovery procedure to geographical data. Spatial data mining is used to describe the data mining step.

Adoption of such an approach into geographical problem solving presents a number of challenges. Among them are, the modification of existing or the development of new algorithms that can handle spatial data, the representation and storage of the extracted knowledge into spatial databases and further incorporation into the model, the role of visualisation in such a methodology and also the mining of disparate and different in format data (Koperski *et al.*, 1998A; Gahegan, 2001; Miller, 2004).

1.2 Aims and Objectives

Knowledge discovery is applied in complex problems to reveal previously unknown information or structures within data that describe complex systems. A problem that is suitable for application of knowledge discovery must satisfy two conditions.

Firstly, it must be a non-trivial problem. That is, a complex problem which is mainly described with models containing uncertain variables. Secondly, it must involve datasets that have high data volume and diversity. The way location affects the value of a property is such a problem and is the main subject of the present research.

Today the most common way to take into account the location of a property in a valuation effort is the experience of the valuer or by using small and similar areas that compare with the property in question. In addition to such traditional methods, automated property valuation systems that are currently in use, employ computer-based valuation models where location is not always the main component.

Valuation modelling may include data that involve the structure and general character of the property in question, locational characteristics, environmental characteristics, transactional data and so on. These potential datasets, apart from the fact that they vary in type, may also be found in disparate and heterogeneous data sources. Developing coherent data models and representations for such datasets to integrate them and use them in a spatial data mining application is another key element in research of this type.

The main aim of this project is to research the application of knowledge discovery in Geographical Information Science (GIScience) and in particular in understanding the effects of location in the value of a property using spatial data mining technology. This has been accomplished through the design and implementation of an integrated location-aware knowledge-based methodology and system, aiming at automated property valuation. The system uses spatial data constructs and integrated data mining algorithms implemented using a prototype architecture and does not address these issues as a collection of isolated functions. The approach is data-driven. Therefore, it does not only rely on specific theories that attempt to explain the role of location in the property value using a priori assumptions or fixed mathematical models but rather identifies relationships and knowledge that is hidden in existing and readily available datasets. This knowledge is extracted automatically in the form of rules that are used to classify properties, provide a better understanding of how location affects their value and ultimately determine that value in conjunction to its specific location.

To meet this aim, the following objectives had been set:

- Review the areas of knowledge discovery and property valuation in general and specifically the areas of Geographic Knowledge Discovery and location aware property valuation modelling. Awareness of existing methodologies in both areas will assist in the better formulation of the proposed methodology. It will form the basis on which the whole design of the system will be based. The property valuation review will also assist in the whole data mining process in the form of background knowledge.
- Use standard data that is readily available and relatively cheap to access.
- Develop a model that takes location explicitly into account.
- Design and implement a prototype system. The prototype system will assist in the demonstration of the proposed methodology.
- Test the methodology on real data from three central London boroughs. This application will allow the evaluation of the proposed methodology.
- Analyse and discuss the methodology in the light of the test.

1.3 Research Questions

This research was structured in such way that answers four main questions. These were formulated after the literature review that initially carried out in the fields of Geographical Knowledge Discovery and property valuation. The research questions that set the framework for the remaining of this research are:

- What knowledge can be extracted from existing standard data sources? How could this be represented and stored into a spatial database?

Discovering patterns and relationships within existing, standard data sources can unveil information relating to the dynamics between spatial objects. Such information can then be used to determine how the value of properties is affected by such spatial relations. This approach can be a viable alternative to the traditional approaches employed for property valuation. The use of standardised datasets that can be readily available makes such an approach economically attractive as it will

not require the capturing of data that is difficult or expensive to source. An implication of this approach is the discovery of new knowledge that is inherent but not obvious or known in commonly used data sources, professionals are familiar with. From the technological perspective, the representation, management and storage of the data sets and the knowledge discovered in a spatial database is an interesting challenge to be addressed. A database designed for this purpose will also enable the usage and further analysis of this data.

- How can location be modelled and successfully incorporated to a knowledge-based valuation model?

Successful modelling involves the representation of a system in the form of a set of variables and their relationships. As location is a complex term that can be expressed in various ways (e.g. proximity, environmental indicators), its modelling is not a trivial process. It involves addressing questions from level of detail and type of representation to collection of data and integration.

- How can the spatial arrangement of landuses affect the property value based on real-world data?

In the literature, the need for further investigation of the ways location influences the property value is highlighted. Extensive research has been carried out involving the structural characteristics of properties resulting to the development of reconciliation procedures that quantify the structural and legal influences on property value. However, in spite of its importance, the influences of location to property value have not been adequately addressed. Research is therefore required to develop a better understanding of how spatial relations affect property value and lead to the development of reconciliation procedures that take into account location too.

- Could such a location-driven methodology produce meaningful results? Does such an approach add value to valuation process and how does this method compare to existing approaches?

Current research has mainly focused on the structural description of properties and the way it influences their value. The role of location in these models is only minor. This work is primarily concerned with the development of a property valuation

approach that is entirely driven by location and explores whether this can be used independently or in conjunction with other well established practices.

1.4 Approach

The design of the research methodology followed in this thesis involved two basic considerations. The first was the identification of meaningful research questions in the related area (see Section 1.3) and second to lead towards the meeting of these goals. There were six main steps in the method and a review that links them to the research questions follows.

Step 1: Review of existing knowledge discovery methods and property valuation practices

An extensive literature review in both areas has been carried out that led to the formulation of the above questions. Initially the area of knowledge discovery in general and in relation to geography in particular, has been reviewed and the possible research opportunities have been identified. One of these was the need for a real-world application. This, in relation to the data-driven nature of such methods introduced the need for defining the application area of this project at an early stage. A number of potential areas that could benefit from such an approach have been identified. For reasons that connect to the research potentials the application of knowledge discovery in the area of property valuation was chosen (see Section 1.2). Therefore, to complete the theoretical background of the research, the property valuation review was carried out. This extended from the basic aspects of property valuation to the more specific issue of involving location to such models.

Due to the interdisciplinary nature of this work that involved two quite wide research areas, this step required extensive work which is reflected in Chapters 2 and 3.

Step 2: Identification of possible data sources and acquisition of final datasets

A database specification has been developed to meet the requirements derived from the literature in conjunction with the identification of potential sources. Data involved were divided to location and property specific. The acquisition of property

related information was a very time consuming task. Another consideration involved the degree of standardisation in the selected datasets. The location model was constructed based on data provided by Ordnance Survey, GeoInformation Group and the Office for National Statistics. Structural information was inferred from data provided by the GeoInformation Group.

Step 3: Design of the general system and detailed specification of the data mining algorithm

This step involved the general design of the database and the system, its basic parts and also the detailed specification of the data mining algorithm. Based on an extensive literature review of data mining the mining type (association rule mining) has been chosen and a supplementary literature on the specific type of algorithms was carried out to assist in the final identification. For the requirement analysis and the design of the system, the Unified Modelling Language (UML) methodology was employed. Finally the software platforms were chosen based on criteria related to implementation issues. These include Oracle 10g (version 10.2.0.3) for the database and Java 1.6.0 for the implementation. As a development environment Netbeans 5.5 was used.

Step 4: Data preparation and modelling

Data is the most important component in such systems and its efficient preparation and modelling would contribute towards the successful modelling of the problem. This step included all the initial data transformations and import to the selected software platforms. Once the data was cleaned and stored in the database the different datasets were integrated into one graph-theoretic data model. For the storage of the graph the Oracle Network data model was used as it provides a generic structure for the persistent storage of networks inside the database.

Step 5: System implementation

Based on the analysis carried out in previous step the database and the system were implemented. This involved the setting-up of all the components (software installation, configurations) and also the implementation of the algorithms.

Step 6: Analysis of the results

Finally a number of tests were carried out to assist in meeting the two last research questions. The tests were designed in such way to help in revealing the optimal configurations for acquiring the best possible results and also to perform some example case studies to demonstrate the use of the system. The trial tests have been performed on a small sample in order to assist in decisions regarding initial configurations. The case study involved two types of demonstrative test cases. The first group involved the locational influence tests while the second the classification tests. Both type of tests assist in the evaluation of the methodology against two requirements. Its ability to produce meaningful information about the way location influences the property price and its viability in a valuation process.

1.5 Main Contributions

The main contributions of this project are the following:

Use of knowledge discovery method to consider location into the valuation process

Knowledge discovery and in particular the task of mining association rules, has been traditionally applied to retail sector applications. In this work, a knowledge discovery method has been devised and applied to a different problem: that of property valuation.

A Residential Property Valuation Platform

Specification and prototype implementation of a residential property valuation platform using a knowledge discovery and data mining technology. This platform is designed to facilitate the location-aware methodology in an integrated manner rather than in an isolated function. The valuation engine is integrated within the database management system as opposed to a stand alone tool that requires additional data transformations.

Integration of multiple datasets

For the purposes of this research a number of the latest available datasets has been used. All the data had to be integrated into one meaningful data model. The task of integrating different datasets that have been produced for different purposes and from different sources is not trivial. It requires good knowledge of all their specifications in order to identify and handle errors, overlaps, etc. Difficulties associated with these datasets are reported in the data preparation section.

Real world application

One of the main contributions of this work is that the evaluation of the proposed method was not based on fictitious data, a common practice in algorithm development and research reported in literature. Instead, real-world data from three London Boroughs was used increasing in such way the degree of difficulty of designing and implementing the system but making also the results more meaningful. In literature, the need for real-world applications in the field of geographical knowledge discovery is stressed.

1.6 Structure of the Thesis

This section provides a guide through this document. Figure 1-1 provides the structure of the document in relation to the contents and outcomes in diagrammatic form. Brief summaries of each of the following chapters follow.

General Research Framework	Knowledge Discovery Methodology	Property Valuation Methods & Techniques	Graph-based modelling System Algorithms Requirements UML Diagrams	Software Platforms Datasets Data Preparation Study Area	Experiments Results	Discussion Future Work
Chapter 1	Chapter 2	Chapter 3	Chapter 4	Chapter 5	Chapter 6	Chapter 7
↓	↓	↓	↓	↓	↓	↓
Research Objective	Data Mining Technique	Background Knowledge	System Design	System Implementation	Case Study	Conclusions

Figure 1-1: Document Structure Overview

Chapter 2 presents a review of the areas knowledge discovery and geographical knowledge discovery and present the current state of the art in spatial data mining

technology. A particular emphasis is given in the methodologies adopted in this research.

In *Chapter 3*, some general concepts are presented followed by a review of the current valuation approaches and techniques. The location theory is then introduced coupled with a review of the way it affected the development of location-aware models. It concludes with a reference to up to date applications of GIS technology in the area of property valuation. This is further demonstrated by presenting some characteristic applications.

The knowledge-based system is presented in *Chapter 4* which consists of three parts. The first part deals with modelling concepts such as the graph model and also gives an overview of the general function of the data mining algorithm. The second part involves the design of the system. This is explored with the use of UML. Finally, the design of the database is described.

Chapter 5 focuses on the implementation aspects of the system. In the first section the employed software platforms are presented. The following sections deal with the datasets used in this research and the stages from data acquisition to the physical implementation of the database. In the final section of this chapter, the study area is introduced followed by a description of its characteristics.

The knowledge discovery process and how it is incorporated in the implemented system is presented in the first section of *Chapter 6*. The rest of the chapter focuses on the experiments carried out. The rationale for their design is presented followed by the description and analysis of the tests.

Chapter 7 concludes this thesis by summarising the work carried out. A discussion follows related to the initial research questions and how these were met through this research. Finally, the main research outcomes coupled with the opportunities for further research following the current work are discussed.

2 Knowledge Discovery in Geographic Information Science

This chapter provides a detailed literature review of the area of knowledge discovery in relation to the Geographic Information Science. It comprises of three main logical entities. The first, Section 2.1, presents the general concepts of the knowledge discovery methodology and provides a general overview of the data mining tasks and available techniques. It further emphasises the areas of association rule mining and associative classification as these techniques formed the basis of the developed spatial data mining algorithm. Section 2.2 discusses the application of knowledge discovery in the geographic domain. It presents the issues associated with spatial data and also reviews the existing ways to deal with these. Finally, Section 2.3 overviews the spatial data mining algorithms available in the literature for each of the main spatial data mining tasks. More emphasis is placed on spatial classification, spatial association mining and the hybrid method of spatial associative classification due to their relevance to this research. This chapter is concluded by a summary in Section 2.4

2.1 Knowledge Discovery Process

Developments in Information Technology (IT) have resulted in increasing volumes of data being collected and analysed by public or private organizations and researchers. The need for the development of new methods that could cope with the massive amount of the collected data became apparent and Knowledge Discovery in Databases

(KDD) that tend to be dynamic, incomplete, noisy, sparse and large became a very active research area (Matheus *et al.*, 1993; Fayyad *et al.*, 1996B). Over the last decade since KDD introduction, there has been a continuous growth in the field that resulted to the development of new techniques that are applicable to various research areas in industry and in academia. Major progress in fields such as biology and web/e-commerce, forced the knowledge discovery to go under an enormous transformation (Piatetsky-Shapiro, 2007).

In addition, recent advances in computer science highlighted the need to move from the confirmatory type of analysis to the knowledge discovery type. Confirmatory analysis requires a priori hypotheses that restrict the researcher and prevent the discovery of previously unknown information (Miller, 2004; Miller & Han, 2001).

In databases there is a lot more information in terms of hidden patterns, trends or relationships from what can be retrieved using traditional analysis and query methods. Where traditional analysis techniques fail to uncover hidden patterns from large and diverse datasets, knowledge discovery techniques succeed (Miller & Han., 2001).

In 1989, at the first KDD workshop (Piatetski-Shapiro, 1991) the term knowledge discovery in databases was introduced to describe the whole process of knowledge discovery. One of the most prevalent definitions of knowledge discovery is that proposed by Fayyad *et al.* (1996A).

They define knowledge discovery as “*the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data*” (p.6). According to that definition patterns must be valid (with some degree of certainty), novel to the system or to the user, offering tangible benefits to the user and finally understandable immediately or after post-processing.

Despite the fact that this definition was introduced over 10 years ago, it is not outdated. Its main points are still valid and capture the essence of knowledge discovery in large databases, although its interpretation has been broadened to cover recent challenges such as complex data types and it will continue to do so (Kriegel *et al.*, 2007). It is this need to cope with the recent advances that makes this field so highly active in terms of research even today. This is also underlined by the fact that back in 1996, there was only one relevant conference (KDD-96) and about 100

research papers while 10 years later there are more than 20 conferences dedicated to the area (Piatetsky-Shapiro, 2007).

In the literature, a number of methodologies have been proposed for knowledge discovery. Most of these approaches are mainly variations of the main scheme: data preparation, data mining and finally interpretation of the extracted knowledge. Since the methodology proposed by Fayyad and his colleagues is the framework used in this work it is presented in more detail.

Fayyad *et al.* (1996B) provide a broad description of the basic steps of the knowledge discovery process. As shown in Figure 2-1 the knowledge discovery process involves the following five general steps: Selection, Pre-processing, Transformation, Data mining and Interpretation / evaluation.

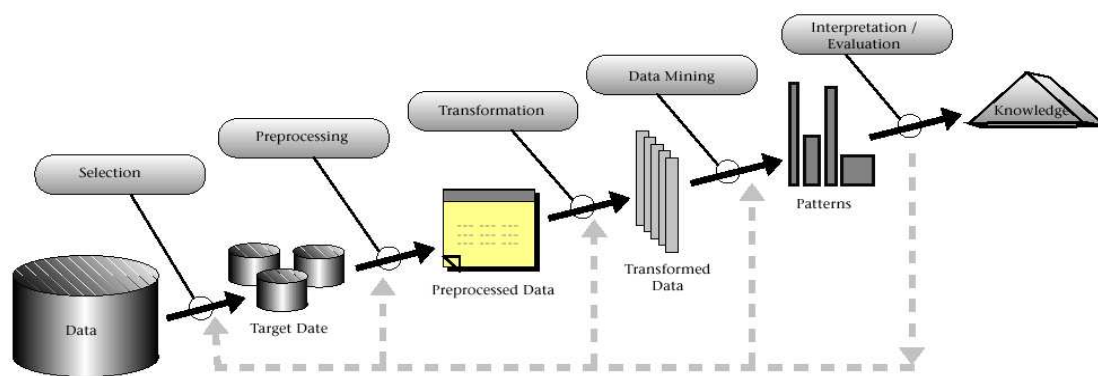


Figure 2-1: Knowledge Discovery Process

(Source: Fayyad *et al.*, 1996B)

A brief analysis of the knowledge discovery process follows (Fayyad *et al.*, 1996A; Miller, 2004; Miller & Han, 2001):

Selection of data involves the identification of the target dataset and further selection of the subsets or the variables, on which the discovery process will be focused. Reinartz (1999) and Barbara *et al.* (1997), both cited in Miller and Han (2001), offer automated techniques for data reduction or ‘focusing’.

Pre-processing includes, noise removal (e.g. incorrect data types, outliers), dealing with missing data fields and accounting for time-series information and known

changes. It can also include the enhancement of data by joining the selected datasets with external data.

The *transformation* step involves the reduction of data, its transformation and projection and finally its aggregation. This will assist in the best possible representation by using variables that capture the most variance. The use of dimensionality reduction techniques such as Principal Components Analysis or Factor Analysis, results in the further reduction of the number of the variables or even to the discovery of invariant representation of the data.

Data mining is the central component of the knowledge discovery process. This step can be further analysed to three more steps. The first involves the search and identification of the generic pattern type. The term pattern is used for an expression that describes a subset of data or a model applicable to the subset (Fayyad *et al.*, 1996B). Patterns are not random, casual or accidentally formed and are characterised by a high degree of repetition. Such patterns are: classes (data objects share similar characteristics), associations (data objects relate to or depend on each other), rules, clusters (data object groups), outliers (inconsistent or distinct data objects) or trends. The second step includes the identification of the specific data mining technique. Most of the algorithms are heuristics that employ intelligent search strategies using alternative approaches. There is a large variety of available techniques for each type of pattern. The final step is the application of the selected technique for pattern search.

Finally, the *interpretation / evaluation* step of the knowledge discovery process involves the interpretation of the discovered patterns usually through visualization and the consolidation of the discovered knowledge either by integrating it into a knowledge-base or by the generation of a report.

Knowledge discovery steps are not strictly sequential. Completed steps, for example, may be revisited after the incorporation of the previously extracted knowledge within the process resulting to an iterative process. It should also be noted the importance of the background knowledge of the application domain in the successful completion of each step. Background knowledge is crucial for directing the whole process (Fayyad *et al.*, 1996A; Fayyad *et al.*, 1996B, Miller, 2004; Miller & Han, 2001).

Task	Databases	Statistics	Artificial Intelligence	Visualization
Finding	Association Rules	Local pattern analysis and global inferential tests	Neural networks, decision trees	Exploratory visualization Visual data mining
Reporting	Rule lists	Significance and power	Likelihood estimation, information gain	A stimulus within the visual domain
Representing	Schema update, metadata	Fitted statistical models, local or global	Conceptual graphs, meta models	Shared between the scene and the observer
Validating	Weak significance testing	Significance tests	Learning followed by verification	Human subjects testing
Optimising	Reducing computational complexity	Data reduction and stratified sampling strategies	Stochastic search, gradient ascent methods	Hierarchical and adaptive methods, grand tours

Table 2-1: Academic communities and knowledge discovery

(Source: Yuan *et al.*, 2001)

One of the characteristics of knowledge discovery that adds to its complexity is that it is a multidisciplinary field. Different communities perceive it from their own particular perspective. Yuan *et al.* (2001) provides a non-exhaustive summary of the different perspectives and disciplines in the areas of data mining and knowledge discovery as shown in Table 2-1. The table shows where academic communities' perspectives and knowledge discovery tasks intersect.

The interest of different academic communities in knowledge discovery processes led to the development of a number of techniques that, depending on their origin, approach the problem from different perspectives. This variety of methods made knowledge discovery a very popular framework for problem solving. Another factor that contributed in the recent interest in knowledge discovery is the success stories in terms of applications that have been reported in the literature and in the press.

Such applications have been developed both for scientific and business purposes (Lavrac *et al.*, 2004; Washio, 2007). In science, one primary application domain that knowledge discovery has been successfully applied is astronomy (Fayyad, 1997). More recent developments led to the application of knowledge discovery

methodologies in areas such as biology (Page, 2003), genomics (Lee *et al.*, 2008) and web-mining (Kolari & Joshi, 2004). Other success stories include application domains in business such as customer marketing (e.g. target marketing, credit scoring), investment, fraud detection, manufacturing, telecommunications, data cleaning (Fayyad *et al.*, 1996B; Grossman *et al.*, 1999).

The application of such a methodology in real-world problem solving introduces issues that may not be apparent in experimental environment. Discovered patterns related to dynamic environments have potentially limited life-time. Although this is a valuable source to assist the rationale behind pattern changing it requires appropriate handling through the development of appropriate methods. Matheus *et al.* (1994) proposes as a solution the development of incremental methods for updating such patterns.

2.1.1 Data mining

The term data mining often has been used to describe the concept of pattern discovery in large datasets. This term appeared to be more popular within the fields of statistics, data analysis, Management Information Systems (MIS) and databases.

In the literature, the term ‘data-mining’ is interpreted in two different ways. The first is related to notions that can be found in traditional statistics. Data mining is often related to terms like ‘data grubbing’, ‘data fishing’ or ‘dredging’ (Lovell, 1983; Chatfield, 1995) or ‘data snooping’. This association is also responsible for the negativity associated with data mining in the past.

The second interpretation is linked to the introduction of the term knowledge discovery in databases (Piateski-Shapiro, 1991) to describe a whole process and places data mining as one of the components that comprise the knowledge discovery process. In such context, although data mining is a key component, it should not be used to describe the overall pattern discovery process (Fayyad *et al.*, 1996B). To further emphasize on the importance of the distinction between KDD and data mining they criticise the stand-alone use of data mining as a ‘dangerous activity’ that could lead to misleading or useless results. This also underlines the important role of the background knowledge based on which the whole process will be performed.

Fayyad *et al.* (1996B) define data mining as “*the application of specific algorithms for extracting patterns from data*” (p. 39). The data mining step involves either the fitting of models to data or the determination of patterns from data. In model fitting two are the approaches that can be adopted: statistical and logical. The statistical approach allows non-deterministic effects in the model where the logical is deterministic (Fayyad *et al.*, 1996B).

2.1.1.1 Data Mining vs. Statistics

It can be argued that data mining belongs to the area of statistics. Although a number of data mining techniques draw on techniques from statistics there are various facts that differentiate it from statistics. One of the major differences is that while statistics are used to validate the hypothesis, data mining ‘discovers’ patterns and hypothesis by data exploration (Chawla, 2000). Hence, data mining involves the automation of the generation of hypothesis process. The validation and verification of the generated hypothesis via statistical tools then may follow.

2.1.1.2 Relational Data Mining

Relational or Multi-Relational Data Mining is a branch of data mining that deals with knowledge discovery from tables stored in relational databases. Traditional data mining algorithms operate on a single table (attribute-value format) and therefore require a data preparation stage where data is transformed to a single table. This approach is also known as the propositional approach. On the contrary, relational algorithms overcome this ‘limitation’ and can operate directly on the original tables without the need for transformation. Most of the data mining tasks can be extended so that they can mine relational patterns. This approach is also known as first-order learning or relational learning (Dzeroski, 2003).

In practice, data mining algorithms designed to function on ‘single-table’ data, scale well (Knobbe *et al.*, 1999). Although scaling is important, this type of mining can be successfully applied only to simple problems. This is due to the description of the objects according to the attribute-value paradigm. Complex objects cannot be effectively described by a fixed set of attributes that only have a single (unstructured) value (Knobbe *et al.*, 1999).

A basis for the development of relational data mining approaches, that is commonly used, is the Inductive Logic Programming (ILP) paradigm (Dzeroski, 2003) which can be defined as the intersection of machine learning and logic programming (Muggleton & Raedt, 1994). ILP systems use a set of positive and negative examples in order to construct a theory. This is inherited from the field of inductive machine learning. From logic programming, ILP inherits its representational formalism, its semantically orientation and also extends logic programming by adopting induction rather than deduction as the basic mode of inference (Muggleton & Raedt, 1994).

2.1.2 Data Mining Tasks

The high-level aims of the data mining process are: prediction and description, although one cannot always distinguish between the two (Fayyat *et al.*, 1996B). In order for these aims to be achieved the use of one of the data mining tasks is required. The term task is used for the method that will help to achieve prediction or description. Due to its broad scope, data mining cannot be associated with only one task. Identification of the appropriate tasks for an application domain is of high importance since it will affect the quality of the results.

Various classification schemes regarding the data mining tasks can be found in the literature. These base the classification on the following criteria types:

- Type of database that is mined (relational databases, object-oriented databases, data warehouses, spatial databases, deductive databases)
- Type of knowledge to be mined (Predictive - Descriptive)
- Type of techniques to be utilized (generalization-based mining, pattern oriented mining, statistical mining)

For the description of the main data mining tasks Miller's classification is used. Miller and Han (2001) organise data mining tasks into the following five categories. The first is *Segmentation*. Segmentation involves the partitioning of the data into groups that share common characteristics. Segmentation can be further analysed into two sub-tasks: *Clustering and classification*. Those two sub-tasks are considered to be overlapping and therefore are grouped under the same term. Clustering describes data

through the examination of the relationships between data by grouping them into implicit classes. Classification on the other hand assigns data items into predefined classes.

The second is *Dependency analysis*. With dependency analysis one can define the value of a number of attributes based on the values of other attributes.

Deviation and outlier analysis is the third category and involves the identification of data that behaves differently from the standard, to determine further actions. Such outliers can be errors that need to be corrected or ignored, or can be unique cases that need further examination.

Trend detection deals with the fitting of lines and curves to the data in order to summarize them often over time.

Finally, *generalization* and *characterization* are compact descriptions of the database.

It should be noted that although these tasks can be applied separately they can also used in a combined way. This leads to hybrid solutions that can be used to tackle complex problems.

2.1.3 Data Mining Techniques

As knowledge discovery is a complicated and multidisciplinary process, there are various techniques one can use depending on which type of outcome is anticipated and the perspective the problem is approached. For example, to perform the segmentation task, statistics or Artificial Intelligence techniques can be used. Table 2-2 (Miller & Han, 2001) provides a list of relevant techniques for each of the data mining task categories discussed earlier. This table by no means is exhaustive but gives an indication of the possible alternatives.

As shown in Table 2-2 cluster analysis, Bayesian classification decisions or classification trees and artificial neural networks are all techniques for *clustering or classification*. Shekhar *et al.* (2002) also refers to the Logistic Regression Modelling technique for the classification problem.

Dependency analysis can be performed either by the use of graph theoretic models (Bayesian networks) or by mining association rules.

<i>Data mining task</i>	<i>Description</i>	<i>Techniques</i>
Segmentation	Clustering: Determining a finite set of implicit classes that describes the data. Classification: Mapping data items into pre-defined classes	- Cluster analysis - Bayesian classification - Decision or classification trees - Artificial neural networks
Dependency analysis	Finding rules to predict the value of some attribute based on the value of other attributes	- Bayesian networks - Association rules
Deviation and outlier analysis	Finding data items that exhibit unusual deviations from expectations	- Clustering and other data mining methods - Outlier detection
Trend detection	Lines and curves summarizing the database, often over time	- Regression - Sequential pattern extraction
Generalization and Characterization	Compact descriptions of the data	- Summary rules - Attribute-oriented induction

Table 2-2: Data mining tasks and techniques (After: Miller & Han, 2001)

For *Deviation and outlier analysis* clustering techniques can be used in the identification of data items that are inconsistent with the remaining set of data. Cluster analysis is not specifically designed for outlier detection and treats outliers as noise. However, when this is the case, the significance that they might have is ignored. Hence, the application of algorithms built for outlier detection purposes is considered to be a best practice since it might lead to unusual signals that reveal valuable information.

Trend detection usually involves the use of regression techniques both linear and logistic. In the case of time series data sequential pattern extraction techniques can be used.

Generalization and characterisation can be performed either by mining summary rules such as characterisation rules or by hierarchical aggregation of the data attributes by compressing data into generalised relations based on background knowledge (Attribute-oriented induction).

A more detailed overview of the association rule discovery and associative classification techniques is presented in the following section because of their relevance to this research.

2.1.3.1 Association Rules

Association rule mining was introduced in 1993 (Agrawal *et al.*) and immediately received a lot of attention from the academic community. Certain technical complexities in designing such algorithms, in association with the interesting outcomes that they could produce, contributed to their increasing popularity within the academic community. On the other hand, producing results that are easy to be interpreted at least at high levels made association rules appealing to practitioners. Although since their introduction they have been evolved to cope with the requirements imposed by new applications, association rule mining is still considered one of the most popular data mining approaches (Wu *et al.*, 2008).

Association rules are probabilistic statements denoting the co-occurrence of certain events within a database (Mannila & Smyth, 2001). They differentiate from other kinds of rules (e.g. classification rules) by aiming in discovering all the rules subject to given constraints, processing large training sets and finally allowing any item combination to appear to either in the antecedent or the consequent part of a rule (Webb, 2000). Mannila & Smyth (2001) added that the processing of very large datasets in an efficient way is the one of the main advantages of association rule mining.

The problem of association rule mining can be formally described as follows (Agrawal & Srikant, 1994): Let D be a set of transactions and $I = \{i_1, i_2, \dots, i_m\}$ a set of distinct items. Each transaction T is uniquely identified and is also a set of such items such that $T \subseteq I$. Let X, Y be subsets of set I and $X \cap Y = \emptyset$. A transaction T contains X , a set of items in I , if $X \subseteq T$. Using those definitions, an association rule can be defined as the implication of the form $X \Rightarrow Y(c\%, r\%)$ where $c\%, r\%$ are the confidence and the support of the rule respectively. Such a rule holds in the transaction set D with confidence c , if $c\%$ of transactions in D that contain X also contain Y . Support r of rule $X \Rightarrow Y$ means that $r\%$ of transactions in D contain $X \cup Y$.

In the case of the rule $X \Rightarrow Y$ the set X is called *antecedent* and the set Y *consequent*. Another property of association rules is the *itemset* which is a set of items and its size is based on the number of items that includes.

Confidence and *support* measures are user defined thresholds and are used as constraints in the rule generation process. Other constraints relate to the syntax of the generated rule (Agrawal & Srikant, 1994). These again are user-defined constraints that allow the generation of rules that comply with certain syntactic specifications. For example, the user may be interested in a specific subset of rules that have, as a consequent, only one item.

Association Rules are strongly related to the Market Basket type of analysis and they were also introduced in such context. Market Basket represents a bundle of goods or services consumer purchase, as a result of a decision making process based on the offered goods. Analysis of such data seeks to uncover existing inter-relationships and the outcome finds application on guiding the development of marketing strategies. Due to their volume and variety, basket type datasets were suitable for that type of mining and led to the unveiling of interesting associations that were not apparent by just presenting the data or by using simple queries.

One infamous example is the Diaper-Beer association pattern discovered by researchers working on behalf of a giant retail outlet (Chawla *et al.*, 2000). After analysing the sales products they discovered a significant association between the purchases of these products. This led to further investigations that resulted to the observation that these purchases were made by young fathers on Friday nights.

Process

Association rule mining is a two step process that involves initially the generation of all the sets of items (itemsets) that are above the support threshold (large itemsets) and then the generation of the rules that exceed the confidence threshold based on these large itemsets. The majority of the proposed algorithms follow the above schema but they differ on the way they ‘traverse’ the search space in order to generate the large itemsets (Frequent-Pattern Mining).

One of the most well known algorithms is the Apriori algorithm (Agrawal & Srikant, 1994). Most of the other algorithms adopt an Apriori-like level-wise approach in the generation of the frequent itemsets. Usually this method leads to the generation of a big number of frequent itemsets affecting the performance.

In 2004, a new algorithm (FP-Growth) was proposed by Han *et al.* (2000, 2004) which tackles this problem by employing a novel data structure (FP-tree) an FP-tree-based pattern-fragment growth mining and finally a partitioning-based, divide-and-conquer search technique. FP-Growth exhibits better performances comparing to Apriori especially when it deals with large datasets, long patterns and low support thresholds (Li *et al.*, 2001).

Performance

Apart from the conceptual design of the algorithm which affects the performance of an algorithm there are also implementation issues that can affect its performance too. The efficiency of the algorithm is also subject to different machine architectures and different compilers. Goethals (2003) compared his Apriori implementation with another commonly used implementation (Borgelt & Kruse, 2002) and found noticeable differences in performance.

Another factor that affects performance is data. Zheng *et al.* (2001) performed a comparative study by testing five well-known association rule mining algorithms. Amongst them were the Apriori and FP-Growth algorithms. The importance of their study derives from the fact that the study was based on real datasets instead of artificial ones. The outcome of their studies showed that performance improvements were strongly related to the artificial datasets and these improvements were not consistent when the algorithms applied to real world datasets. In the case of the latter, the choice of the algorithms mattered only at support levels that generate more than the necessary number of rules that can be used in practice.

Applications

The potential of association rules has been demonstrated through a number of example studies. Example applications include retail applications (e.g. targeting,

customer retention), medical applications (e.g. diagnosis enhancement), genomic applications and web applications (e.g. web-usage mining, web-retrieval).

2.1.3.2 Apriori

The Apriori algorithm belongs to the group of algorithms that employ a bottom-up Breadth-First-Search (BFS) search for the discovery of the frequent itemsets and is one of the most prevalent techniques. It was introduced in 1994 by Agrawal and Srikant in continuation to their previously proposed algorithm in 1993 by Agrawal *et al.* Although developed in the early 90's, the Apriori algorithm was identified as one of the 10 most important data mining algorithms by the IEEE International Conference on Data Mining (ICDM) in December 2006 (Wu et al., 2008).

One of the strengths of Apriori relates to its ability to generate the candidate itemsets in one pass without taking into consideration the transactions in the database. Given that any large itemset also produces large sub-items, the generation of k-itemsets can be based on the previously generated large (k-1)-itemsets by joining them and deleting those that contain any subset that is not large.

There are two main drawbacks of the candidate generate-and-test algorithm like Apriori (Liu *et al.*, 2004). The first relates to the generation of a large number of candidates that can be proven to be infrequent after scanning the database. The second is the need for multiple scans of the database which can reach the number of the maximal length of the itemset. Another difficulty is that Apriori does not take into consideration the underlying algebraic structure of the search space where such exists. As a result it does not benefit from partitioning the problem for parallel processing (Adamo, 2001).

The Apriori algorithm consists of two main functions: apriori-gen and subset. The apriori-gen function enables the generation of the candidate itemsets. The subset function determines the candidates that are contained in a given transaction.

As shown in Algorithm 2-1, on the first pass the algorithm counts the occurrences of the 1-itemsets in order to determine which of them are large (above the support threshold). Subsequently these 1-itemsets are used to generate the candidate itemsets

(Apriori-gen). Next, it performs a join to create the k-itemsets and then prunes those that are not included in the set of the large k-itemsets.

The data structure used for the storage of the candidate itemsets is that of a hash tree. Initially all the nodes are created as leaf nodes. Itemsets are stored in the leaves until the specified threshold (maximum number of stored itemsets in a leaf) is reached. In that case the leaf converts to an interior node. An interior node at depth d points to interior nodes at depth $d+1$. The root of the hash tree is set to depth 1. When a new itemset is added the tree is traversed starting from the root until it reaches a leaf. The direction that it takes when it reaches an interior node at depth d is based on a hash function applied to the d^{th} item of the itemset.

Apriori
Algorithm 2-1

```

D      Database
Lk    Set of large k-itemsets
Ck    Set of candidate k-itemsets

L1 = {large 1-itemsets};
for (k=2; Lk-1 ≠ ∅; k++) do begin
    insert into Ck
    select p.item1, p.item2, ..., p.itemk-1, q.itemk-1
    from Lk-1 p, Lk-1 q
    where p.item1 = q.item1, ..., p.itemk-2 = q.itemk-2, p.itemk-1 < q.itemk-1;
    forall itemsets c ∈ Ck do
        forall (k-1)-subsets s of c do
            if (s ∉ Lk-1) then
                delete c from Ck
        forall transactions t ∈ D do begin
            Ct = subset(Ck, t);
            forall candidates c ∈ Ct do
                c.count++;
        end
    Lk = {c ∈ Ck | c.count ≥ minsup}
end
Answer = ∪k Lk

```

(Source: Agrawal & Srikant, 1994)

Figure 2-2 shows the way the Apriori algorithm works when applied to the example database shown in the top left side of the figure. The initial scan of the database consists of two steps. The first step is a simple list generation of all the unique items (1-itemsets) in the database. At this initial stage all items considered as possible candidates. In the second step the occurrences of each 1-itemset are being counted and

based on the pre-defined threshold ($\text{minSup} = 2$) the algorithm selects the frequent 1-itemsets.

The second row in Figure 2-2 shows the second iteration. From the list with the 1-itemsets by performing a self-join the algorithm generates the list of 2-itemsets possible candidates. Based on this list, the algorithm scans the database to calculate the support for the candidate 2-itemsets and prunes the ones that are below the threshold.

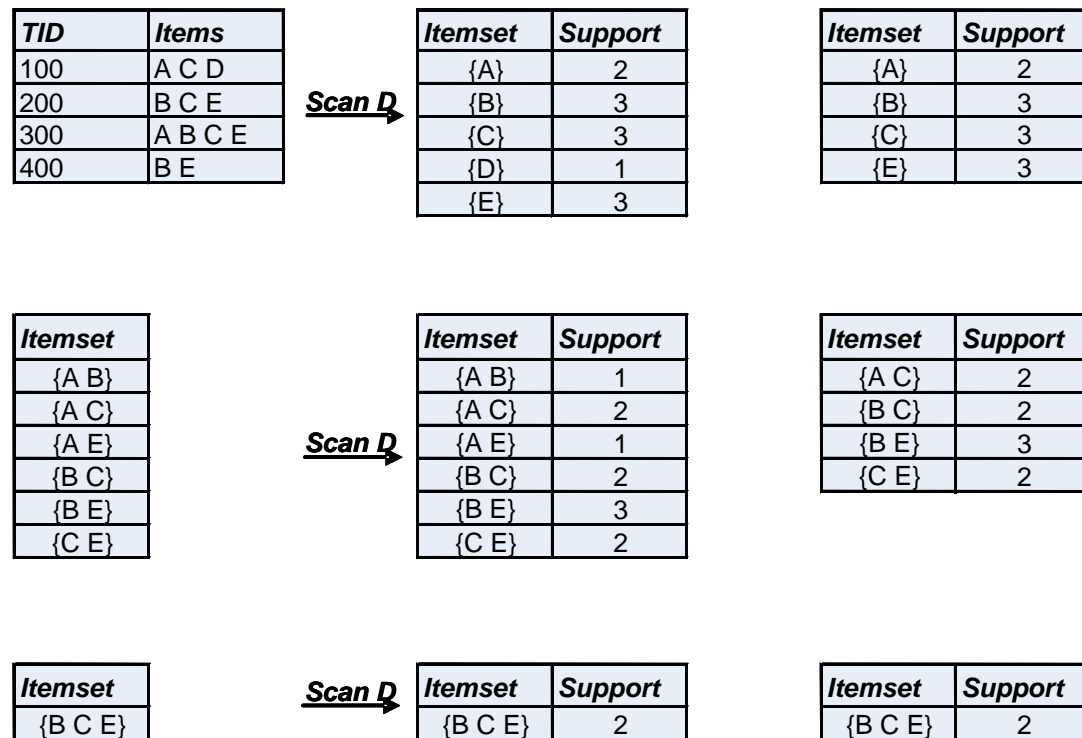


Figure 2-2: Example

In the third row the same procedure that results in the generation of 3-itemsets is shown. Based on line 6 of the Apriori algorithm (Algorithm 2-1) in order to create itemsets of k size a self join is performed with a join predicate that the first $k-2$ items of each $(k-1)$ -itemset are the same and the $k-1$ item of the first is smaller than the $k-1$ item of the second. In this case, to create 3-itemsets the join will be performed only to the 2-itemsets that their first item is the same. Therefore, only the {B,C} and {B,E} itemsets join to produce the 3-itemset {B,C,E} which is above the minimum support threshold. In this stage, the algorithm also terminates the process of discovering the large k -itemsets since there is no 4-itemset to be generated by the discovered large 3-itemsets.

2.1.3.3 Classifying Association Rules

Over the years, various association rule algorithms have been proposed that vary either in the way they perform the candidate itemset generation step or on the kind of the rules they are mining. Based on the latter, association rules can be classified based on the type of the extracted rules (Categorical-Quantitative-Sequential) or the level of the extracted rules (Generalised-Multi-level).

Categorical - Quantitative

Categorical association rules appear in the early studies where the datasets that were used for the testing of the algorithms comprised by categorical data (e.g. market basket data). In this case, both the antecedent and the consequent part consist of categorical attributes. Categorical attributes can quite easily be transformed into a set of pseudo-Boolean attributes and handled as such (Adamo, 2001).

Apart from categorical attributes, real-world datasets also include numerical attributes. This led to the introduction of the quantitative association rules (Srikant & Agrawal, 1996). Quantitative attributes can take a large number of values making their conversion to pseudo-Boolean attributes impossible. Therefore before the mining application, a process of *discretisation* must be applied. Although a number of discretisation techniques have been developed for classification, association rule mining presents certain difficulties that make the application of such methods unsuccessful (Adamo, 2001).

Srikant & Agrawal (1996) raise these issues and present the difficulties that mainly associate with the number of the intervals. Large number of classes may result in failing to reach the minimum support threshold. On the other hand, low number of classes cause loss of information that can also be translated to failure to reach the minimum confidence threshold.

Multi-Level

Revealing useful and meaningful associations cannot be always successfully restricted only to a certain concept level. In fact, the higher levels of hierarchy have more possibilities to generate strong associations since they are more general. The need for

rule mining based on some kind of taxonomy information led to the development of multi-level association rules.

Such an algorithm was first introduced by Han & Fu (1995) and involved the mining of association relationships at a level-by-level in a fixed hierarchy. This hierarchy was provided to generalise primitive level concepts to high level ones. Association rules at low level of hierarchy were examined only if their corresponding parents were above the support and confidence thresholds. Another characteristic of this algorithm was the use of non-uniform support and confidence thresholds throughout the hierarchy levels.

Generalised (Cross-Level)

In 1995, Srikand & Agrawal introduced the problem of mining generalised rules and proposed an algorithm and also an interestingness measure to tackle this. Generalised rule mining unlike multi-level rule mining is not confining the mining process in the same level of hierarchy. Instead, it allows the generation of rules between items that belong at any level in a taxonomy. Han & Fu (1999) also presented a modified version of their multi-level algorithm that also allowed the generation of ‘cross-level’ association rules.

Sequential

Sequential rule mining involves the discovery of frequent patterns in sequences of events where each event has an associated time of occurrence. Similar to the problem of mining association rules, the problem of mining sequential associations is to find the maximal sequences that are above the user-defined threshold where a sequence is an ordered list of itemsets (Agrawal & Srikant, 1995). Among the first algorithms for sequential association rule discovery are those proposed by Agrawal & Srikant (1995) and Mannila *et al.* (1997).

2.1.3.4 Measurement of interestingness

The need for the introduction of some kind of metrics that capture the significance of the generated association rules, derives from the fact that such a technique can

produce a lot of data itself, introducing a new knowledge management problem (Klemettinen *et al.*, 1994).

As mentioned in previous sections, one of the basic methods to decide on the importance of a discovered rule is the use of the support and confidence thresholds. Depending also on the efficient and careful preparation of the database, evaluation of association rules only on the basis of those two metrics can be misleading. Piatetsky & Shapiro (1991b) introduced the notion of ‘interestingness’ to deal with this issue.

Interestingness is defined in relation to the basic KDD definition (see Section 2.1) as a non-trivial process that aims to uncover valid, novel and useful patterns in large datasets. Based on this, interestingness is a metric used to differentiate between patterns that fulfil these requirements and those that do not.

Related studies can be classified into objective and subjective (Silberschatz & Tuzhilin, 1996). Objective ‘interestingness’ attempts to quantify the interest of a pattern based on its structure and the underlying data. Confidence and support belong to the group of objective measures. Other objective measures include: coverage, strength, statistical significance and simplicity (Liu *et al.*, 1996).

On the other hand, with subjective ‘interestingness’ the user plays an active role in the characterisation of a certain pattern as interesting or not. Among the proposed subjective measures are: object unexpectedness and actionability. The unexpectedness measure is an indication of how surprising a pattern is to the user. The actionability measure denotes the extend to which a pattern allows the user to directly act on it, turning it to a beneficial for the user result. Subjective measures rely on previously acquired knowledge of the domain and are application specific (Liu *et al.*, 1996).

2.1.3.5 Associative Classification

Rule-based classification models have been extensively applied in classification problems. Although traditional rule-based classification algorithms (e.g. C4.5 (Quinlan, 1993), FOIL (Quinlan & Cameron-Jones, 1993)) achieve high performance they lack in classification accuracy (Yin & Han, 2003). Associative classification is a technique that integrates two important data mining techniques, association rule

discovery and classification. This integration results in the production of more accurate classifiers (Liu *et al.*, 1998; Li *et al.*, 2001; Yin & Han, 2003).

The general steps of this approach are as follows (Liu *et al.*, 1998):

Discretisation: Classification datasets often contain continuous (numeric) data. Section 2.1.3.3 refers to the limitation of association rules to handle continuous (numeric) data. This step is dealing with this type of transformation. In the case of associative classification continuous attributes are being discretised based on the classification pre-determined target.

CAR Generation: Associative classification deals only with a specific subset of association rules, those in which the consequent part of the rule is restricted to the classification class attribute. This special case of association rules is called Class Association Rules (CARs).

Classifier: The final step of associative classification is the built of the classifier based on the extracted CARs. The selection is based on a number of CARs following a number of evaluation criteria.

Associative classification algorithms can be classified either depending on the way the CAR generation is performed or on the type of the rule evaluation measures used in the classification of the new (unseen) cases.

According to the first criterion, two groups can be distinguished. The first includes algorithms that generate rules in two stages following the general association rule paradigm. The second group consists of algorithms that are more close to the traditional rule-based classification algorithms where the rule generation is incorporated in the classifier determination.

Based on the way the classifier is determined, three main algorithm types can be identified (Coenen & Leng, 2004). Algorithms that base the classification of an unseen case on: Best Rule, Best k Rules and All Rules.

In the case of the 'Best Rule' selection, the rule that satisfies the unseen case is identified and used for the classification. This identification is based on some kind of sorting of the generated CARs. Four such sorting schemes can be identified in the

literature. The first, bases the ordering of the rules on confidence, support and the antecedent size, with the confidence to be considered the most important factor. The weighted relative accuracy is the second way and represents an interestingness measure. The third is the use of Laplace accuracy measures which is commonly used in the rule-based classification. Finally is the use of the χ^2 testing, a statistical technique used to test the independence of two variables.

The ‘Best k Rules’ method requires the identification of k rules that satisfy a new case. The selection of the rule is based on an averaging process on the selected rules. Similarly to the Best rule method, an initial sorting is required for the identification of the best k rules.

Finally, there is the ‘All Rules’ method. In this, the classification is based on the selection of all the rules that satisfy a given case followed by an evaluation that will result to the class identification. Hence, the classification of a new case is not based only on one rule. Instead, it is based on a set of highly correlated and high confidence rules (Li *et al.*, 2001). A common evaluation method in this type of algorithms is the weighted χ^2 testing.

In the majority of the associative classification algorithms, each rule in the classifier is associated with only one class label. The research concerning algorithms that perform multi-label classification is still limited and domain specific. Multi-label classification allows the human interaction for the refinement of the classes or the use of more information (e.g. dictionaries) (Thabtah *et al.*, 2005).

Thabtah *et al.* (2005) performed a comparative study about the predictive accuracy of 4 popular associative classification algorithms, CBA, CMAR, CPAR and MCAR. For the purposes of the study 12 benchmark problems have been used. The results indicated no prevalent algorithm in terms of predictive power. Table 2-3 shows the way the most popular associative classification algorithms operate.

Coenen & Leng (2004) tested different rule ordering and case satisfaction schemes against datasets from the UCI Machine Learning Repository (Asuncion & Newman, 2007). The variations tested included all the cases referred earlier in this section. The experiments showed that no method can be considered to be the best suited method for all the tested datasets. Nevertheless, the best rule coupled with the confidence-

support-size ordering scheme gave the best overall classification accuracy. In terms of performance there were little differences between the methods, with the weighted χ^2 testing to take slightly longer.

Algorithm	Rule Mining Method	Classification Method
CBA Liu <i>et al.</i> , 1998	Apriori – based	Best Rule / Sorting based on Confidence, Support, Size
CMAR Li <i>et al.</i> , 2001	FP-growth variant	All rules / Weighted χ^2 testing
CPAR Yin & Han, 2003	One step algorithm / FOIL	

Table 2-3: Associative Classification Algorithms

Advantages

Association-based Classification presents a number of advantages when compared with other traditional classification systems. Associative classification not only contributes toward a more accurate classifier but also successfully deals with existing issues in current classification rule mining systems. These are summarised as follows.

Understandable rules: One of the most important drawbacks of classification is that of ‘understandability’. In current systems, the formation of the classifier is based on a small set of rules. The generation of these rules is based on domain independent biases and heuristics that do not facilitate the generation of understandable or interesting rules to the user (Liu *et al.*, 1998).

Ability to deal with multiple attributes: Association rules can deal with multiple attributes. This helps to overcome one of the decision-tree induction limitations which examines only one attribute at a time (Li *et al.*, 2001) and hence contributes to classifiers with better predictive accuracy.

Performance: Associative classification systems do not require the loading of the whole database into the main memory. This is a requirement in the standard classification systems (Liu *et al.*, 1998) that makes them computationally expensive.

Accuracy: Extensive performance studies (Liu *et al.*, 1998; Li *et al.*, 2001) showed that classifiers build within such a framework have better accuracy when compared to those based on classic classification approaches e.g. C4.5 (Quinlan, 1993).

Issues

On the other hand, there are issues when the classification is based on association rules (Li *et al.*, 2001; Yin & Han, 2003). The first is scalability. Depending on the size of the database, a very large number of rules can be generated (Agrawal & Srikant, 1994). Therefore methods that facilitate the efficient storage and retrieving of these rules are necessary. The second relates to the identification of the appropriate rule that will lead to the effective classification. Approaches that use a confidence-based evaluation for the selection of the classifier rule, may lead to over fitting.

Another limitation derives from the fact that the classification achieved by this method is categorical therefore there is no information about any uncertainties in the classification (Ceci *et al.*, 2004)

2.1.3.6 Classification Based on Associations (CBA)

One of the first algorithms that dealt with the association rule mining and classification integration was proposed by Liu *et al.* in 1998. The CBA algorithm has two main components, one associated with the association rule mining task (rule generator) and the other associated with classification task (classifier builder).

The rule generator component (CBA-RG) follows the basic principles of the Apriori algorithm with some modifications to include the class attribute in the process of the rule generation. Therefore the problem is reduced to the discovery of the frequent 'ruleitems'. Ruleitems can be seen as a special case of the itemsets that apart from a set of items also include the class attribute (or class label). In fact, each ruleitem forms a rule where the antecedent part is the itemset while the consequent part is the class attribute.

Similarly to Apriori the characterisation of the ruleitems as frequent is based on the support constraint. Hence, a ruleitem that is frequent and its confidence is above the user defined threshold can be considered as possible CAR. In addition to these constraints Liu *et al.* (1998) considered an extra selection criterion that represents the accuracy of the rule. This was introduced to deal with the cases where the ruleitems have the same itemset but have different class attribute. According to this the rule

with the highest confidence is chosen to represent this ruleitem. In the cases of the same confidence the selection is random.

As is shown in Algorithm 2-2, CBA-RG function is analogous to the apriori-gen function. The difference is that CBA-RG calculates two support counts, one for the itemset and one for the ruleitem to be able to calculate the ruleitem confidence. In addition, the generated rules are subject to a pruning operation that can be optional.

The second and most important stage of the CBA algorithm involves the construction of the classifier based on the final set of CARs generated in the previous stage (CBA-RG). This is based on the covering method (Michalski, 1980, cited in Liu *et al.*, 1998) according to which for each class the best rule is being identified and the covered cases are being removed from the training set.

CBA-RG	Algorithm 2-2
D Database F_k Set of large k -ruleitems C_k Set of candidate k -ruleitems	
<pre> $F_1 = \{\text{large 1-ruleitems}\};$ $CAR_1 = \text{genRules}(F_1);$ $prCAR_1 = \text{pruneRules}(CAR_1);$ for ($k=2; F_{k-1} \neq \emptyset; k++$) do begin $C_k = \text{candidateGen}(F_{k-1});$ for each $d \in D$ do $C_d = \text{ruleSubset}(C_d, d)$ for each $c \in C_d$ do $c.\text{condsupCount}++;$ if $d.\text{class}=c.\text{class}$ then $c.\text{rulesupCount}++$ End End $F_k = \{ c \in C_k \mid c.\text{rulesupCount} \geq \text{minsup} \};$ $CAR_k = \text{genRules}(F_k);$ $prCAR_k = \text{pruneRules}(CAR_k);$ End $CARs = \bigcup_k CAR_k;$ $CARs = \bigcup_k prCAR_k;$ </pre>	

Source: Liu *et al.* (1998)

The CBA-CB works in a heuristic way and consists of three stages. Before the description of these stages, it is necessary to present the rule ordering criteria proposed by the authors. A rule r_i considers to have higher precedence compared to the rule r_j if

- i. The confidence of rule r_i greater than that of rule r_j
- ii. The confidence for both rules is the same but the support of r_i greater than that of rule r_j
- iii. Both the confidence and support are the same for the two rules but r_i was generated at an earlier stage than rule r_j

There are two main conditions such an algorithm should satisfy (Liu *et al.*, 1998). The first is to ensure that each training case $d \in D$ is covered by the rule with the highest precedence among the rules that cover this case. The second condition is that each rule in C should correctly identify at least one of the remaining training cases when selected.

The CBA-CB function builds the classifier in three stages. During stage 1 (Algorithm 2-3), for each training case d , the highest precedence rules that correctly (cRule) and wrongly (wRule) classify d are being identified. In the case that the rule that correctly classifies d , has higher precedence than the one which wrongly classifies d , the case d is covered by the cRule and the rule is marked that classifies a case correctly. In the case that the wRule is preceded the cRule, a collection A of the form $\langle dID, y, cRule, wRule \rangle$ is kept where dID is the unique id of the training case d , y the class label of d and $cRule, wRule$ the correct and wrong rules for the case d .

CBA-CB (Stage 1)

Algorithm 2-3

```

R = Set of Rules
D = Training data
C = Classifier
Temp = Temporary List
Q = Set of cRules that have higher priority than their corresponding
wRules
U = Set of all cRules
A = Collection of  $\langle d.id, d.class, cRule, yRule \rangle$ 

```

```

1  Q =  $\emptyset$ ; U =  $\emptyset$ ; A =  $\emptyset$ ;
2  for each case d  $\in$  D do
3      cRule = maxCoverRule(Cc,d);
4      wRule = maxCoverRule(Cw,d);
5      U = U  $\cup$  {cRule};
6      cRule.classCasesCovered[d.class]++;
7      if cRule > wRule then
8          Q = Q  $\cup$  {cRule};
9          mark cRule;
10     else A = A  $\cup$   $\langle d.id, d.class, cRule, wRule \rangle$ 
11 end

```

Source: Liu et al. (1998)

The second stage (Algorithm 2-4) of the CBA-CB function deals with the training cases that belong to the collection A, therefore the cases where the rule that covers them, could not have been decided during the first stage. For each training case d in A, if the wRule is marked, it means that this rule classifies correctly at least one case. In this case, this rule will cover the case d and the counters that hold the number of cases that are being covered by cRule or wRule will be updated. If the wRule is not marked, then the algorithm searches in set U to identify all the rules that wrongly classify the case d and have higher precedence than that of its cRule. All the returned rules can potentially replace the original cRule since they have higher precedence. This information is kept in the replace field of each rule in wSet and the counter in line 8 is being updated. The final set of the CARs that will be used for the classifier is the union of the sets U and wSet.

CBA- CB (Stage 2)*Algorithm 2-4*

```

R = Set of Rules
D = Training data
C = Classifier
Temp = Temporary List
Q = Set of cRules that have higher priority than their corresponding
wRules
U = Set of all cRules
A = Collection of <d.id, d.class, cRule, yRule>

```

```

1 for each entry <dID, d.class, cRule, wRule> ∈ A do
2   if wRule is marked then
3     cRule.classCasesCovered[d.class]--;
4     wRule.classCasesCovered[d.class]++;
5   else wSet = allCoverRules(U,dID.case,cRule);
6     for each rule w ∈ wSet do
7       w.replace = w.replace ∪ {<cRule,dID,d.class>};
8       w.classCasesCovered[d.class]++;
9     end
10    Q=Q ∪ wSet
11  end
12 end

```

Source: Liu et al. (1998)

The third stage (Algorithm 2-5) involves the final selection of the rules that will form the classifier. For each rule in Q that correctly identifies at least one training case the algorithm attempts to replace all the rules in r.replace by r because it precedes them. In the cases where the dID has been already be covered by a previous rule the replacement will not be performed. Counters in lines 8, 9 need to be updated accordingly. For each selected rule the ruleErrors (records the number of errors at

each stage) and classDistr (number of training cases in each class) needs updating. Also, a default class based on the majority class in the remaining training data is chosen and a default error is defined as the number of the remaining cases that will be wrongly classified by using the default class. The total number of errors is the sum of the rules and the default error.

The final part of stage three involves the selection of the rule that produces the lowest error and the pruning of all those that are after this rule. Finally a default class is associated to the selected rule and the classifier is returned without the total error and default-class.

CBA-CB (Stage 3)*Algorithm 2-5*

```

R = Set of Rules
D = Training data
C = Classifier
Temp = Temporary List
Q = Set of cRules that have higher priority than their corresponding wRules
U = Set of all cRules
A = Collection of <d.id, d.class, cRule, yRule>

1 classDistr = compClassDistr(D);
2 ruleErrors = 0;
3 Q = sort(Q);
4 for each rule r in Q in sequence do
5     if r.classCasesCovered[r.class] ≠ 0 then
6         for each entry <rul, dID, d.class> in r.replace do
7             if the dID case has been covered
by a previous r then
8                 r.classCasesCovered[d.class]--;
9                 else rul.classCasesCovered[d.class]--;
10                ruleErrors = ruleErrors + errorsOfRule(r);
11                classDistr = update(r, classDistr);
12                defaultClass = selectDefault(classDistr);
13                defaultErrors = defErr(defaultClass, classDistr);
14                totalErrors = ruleErrors + defaultErrors;
15                Insert <r, default-class, totalErrors> at end of C
16        end
17 end
18 Find the first rule p in C with the lowest totalErrors, and then discard
all the rules after p from C;
19 Add the default class associated to p to end of C;
20 Return C without total Errors and default-class;

```

Source: Liu et al. (1998)

2.2 Knowledge Discovery and Spatial Data

Pattern discovery and analysis has long been an important part of geography (Haggett *et al.*, 1977). Nowadays it assumes the use of computers however, the importance of ‘mining’ spatial data was recognised even before their invention and can be demonstrated in the following historical cases (Griffith, 1999; O’Sullivan & Unwin, 2002): John Snow’s work on the Asiatic cholera outbreak in London and the realisation of the water-borne nature of the disease (1855), the theory of Gondwanaland-all continents once formed a single landmass (1919) and finally the realisation that fluoride controls tooth-decay by observations in Colorado Springs (1909). In all the above cases the spatial data exploration resulted to hypotheses that later were scientifically confirmed.

A pattern can be defined as “*a geometrical expression of location theory*” (Rogers, 1969, cited in Haggett *et al.*, 1977). Following this, Haggett *et al.* (1977) define a pattern as “*a characteristic of spatial arrangement which describes the spacing of a set of objects with respect to one another*”. Given the recognised importance of pattern discovery a variety of related methods have been developed through the years. A number of them were developed even before the introduction of the concept of knowledge discovery in 1989. Most of these methods are used for point pattern discovery which mostly relates to the data mining task of clustering.

Another issue that makes the development of geographic knowledge discovery and geographic data mining techniques within Geographic Information Science necessary relates to the amount of geographic information available today. Technological advances in the areas of data capture and handling resulted to an explosion of digital geographic and geo-referenced data. Such amounts of data make their detailed examination expensive or even unrealistic hence commands the need for further investigation in the area of knowledge discovery (Lu *et al.*, 1993). Miller and Han (2001) emphasise the impact that these advances have on geographical research, and refer to it as the most dramatic shift in the information environment since the Age of Discovery in history. Data sources (e.g. remote sensing systems, GPS, sensors) and formats (e.g. imagery, video, sound) vary and the need for the development of tools that could handle such data is more apparent than ever (Ester *et al.*, 1999).

One of the first studies in the field of geographical knowledge discovery was conducted by Lu *et al.* (1993) where they proposed two algorithms for generalization. Since then a number of algorithms have been developed and are further discussed in the following sections.

2.2.1 GIS and Spatial Database Systems

Geographical Information Systems (GIS) are very powerful tools for capturing, storing, modelling, analysing, manipulating and visualising geographical data. In recent years, a number of developments made the use of GIS in handling geographical data widespread. As shown in Figure 2-3, the main elements of a GIS are: the Database element, Data processing element, Data storage and retrieval element, Data sharing element, Data presentation element, Spatial reasoning element and finally the spatio-temporal element (Worboys & Duckham, 2004). It is this wide functionality that made GIS technology applicable to a variety of sub-disciplines within geography from environmental to social geography.

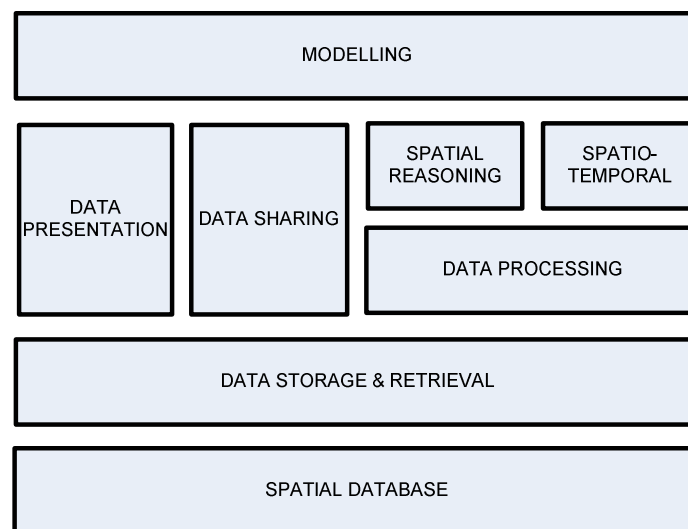


Figure 2-3: GIS elements

Given that spatial database systems provide the underlying database technology for a GIS, good design and understanding plays important role in the successful support of data mining functions.

Although several terms such as ‘pictorial’, ‘image’, ‘geometric’, ‘geographic’ have been used to describe a database system that handles geometric, geographic, or spatial

data, the term spatial database became the prevailing term to describe such databases. Such systems can be considered as databases dealing with objects in space rather than with images or pictures of space. Although database systems dealing with images may have analytical functionalities to extract objects from images and include some spatial database functions they are also capable to manage raster images as discrete entities (Guting, 1994). Gunther and Buchman (1990) and Frank (1991), (both cited in Guting, 1994), suggest two classes of systems, spatial database systems and image database systems.

Guting (1994) argues that there is no widely accepted definition of a spatial database system and defines spatial database systems as *“a database that offers spatial data types in its data model and query language, and supports spatial data types in its implementation, providing at least spatial indexing and spatial join methods”*(p. 357).

Guting further elucidate the above definition by deepening on the three main points of his definition:

i. By highlighting the fact that a spatial database system is based on a conventional database system implies that spatial information also relates to non-spatial data. Hence, there are systems capable of providing standard functions such as conventional data modelling and query support.

ii. Spatial data types, their relationships and operations are offered. Failing to capture fundamental abstractions of entities in space such as points, lines, polygons and, their relationships and operations such systems cannot deliver spatial modelling capabilities.

iii. Guting's definition considers spatial indexing mandatory. In addition, the ability to relate objects from different classes through some spatial relationship is considered important.

Spatial data mining algorithms use spatial computations such as spatial joins, nearest neighbour queries and others, therefore there is a need for efficient spatial access methods and data structures (Koperski *et al.*, 1996). Spatial access methods used to build indices on spatial data types (points, lines, polygons) are multi-dimensional trees

such as quad trees, k-d trees, R-trees, R*-trees (see Kuba, 2001). The later two methods received more attention in literature and have been widely implemented.

2.2.2 Special characteristics of Spatial Data

It is widely accepted that geographical data present a number of special features that differentiate them and also add complexity to their handling. Ignoring these unique characteristics of the data may result to erroneous and misleading results. Therefore effective modelling is required. Spatial dependency and spatial heterogeneity are two of the main features of geographical datasets (Miller & Han, 2001; Gahegan, 2001; Shekhar *et al.*, 2001; Chawla *et al.*, 2000; Openshaw, 1999).

Spatial dependency refers to the fact that spatial data tend to be highly self-correlated. It is common, for example, for people with similar socio-demographic profile to cluster together or economies within a region to be similar. This kind of dependency although less complicated exists in other domains as well, for example serial autocorrelation in time series data (Miller & Han, 2001).

Spatial dependency is so fundamental that led to the formulation of Tobler's first law of Geography (Tobler, 1970): "*Everything is related to everything else, but nearby things are more related than distant things.*"(p. 236). According to this law the values of the attributes of neighbour spatial objects have the tendency to affect each other. Furthermore Gould (1970, cited in Haggett *et al.*, 1977) refers to this lack of independence in spatial observation as the main cause for the substitution of spatial patterns.

Spatial heterogeneity can be a valuable source of information regarding phenomena under investigation. It refers to the variation in relationships over space hence to the non-stationarity of spatial data. Granger (1969, cited in Haggett *et al.*, 1977) defines stationarity as "*an assumption that the relationship between values of the processes is the same for every pair of points whose relative positions are the same*". Stationarity is completely unrealistic for spatial related variables (Granger, 1969, cited in Cliff and Ord, 1975) and also strongly related to the spatial dependence. Non-stationarity implies that data is spatially dependant (Haggett *et al.*, 1977).

Apart from these fundamental issues of geographical data, the literature references a number of other geographical data characteristics that command special handling that conventional data mining algorithms cannot offer.

Openshaw (1984) refers to an interpolation problem known as the *Modifiable Areal Unit Problem* (MAUP). MAUP is related to the fact that despite the growing data availability some data might be in spatial or temporal aggregated forms. Simply stated, the problem arises because different types and levels of aggregation can result in whole different representations of geographical phenomena (Cliff and Ord, 1975). MAUP causes problems particularly to cluster detection algorithms.

The different ways of spatial data representation add another constraint in the usage of traditional data mining algorithms. Spatial data types contain not only integers, dates and strings but also more complex data types such as *lines* and *polygons*. Coupled with the vector representation (points, lines, polygons) there is also *raster data* which is another common GIS data model.

Since a large number of datasets are collected from satellites, aerial photographs Digital Elevation Model (DEM) etc., the volume of data that is stored as qualitative and categorical raster data is increasing. Additional problems are caused by the enrichment of geographical databases with ill-structured data such as imagery and geo-referenced multimedia (Miller & Han, 2001).

Therefore extracting meaningful and useful information from spatial data is not as easy as from traditional numeric and categorical data. Although one popular way to overcome this difficulty is to convert spatial components to non-spatial via feature selection there is the alternative of finding new models and new patterns more suitable for spatial data and their unique properties (Shekhar *et al.*, 2001).

Spatial relationships also introduce another level of complexity. Basic spatial relationships between objects such as topological (adjacent, inside, disjoint), directional (above, below, north of) and metric (distance) add more to the complexity of mining spatial data (Shekhar *et al.*, 2001; Miller & Han, 2001). Spatio-temporal relationships also present complexities compared to other databases. Data in non-geographic databases can be represented in an information space meaningfully as points. Geographic data represented as points could affect the data mining process

adversely due to measurement artefacts. Further problems relate to the complexities that arise in object transformations over time and in geographical relationships such as distance, direction and connectivity (Miller & Han., 2001).

In relation to the spatial relationship complexity, Miller and Han (2001) introduces another issue, that of *geographic measurement frameworks*. Although the most common measurement framework is topology and geometry consistent to Euclidean space there are geographical phenomena that display properties that behave according to other topologies and geometries. Such cases include travel-time relationships in an urban area, or disease patterns in space and time. Therefore searching for patterns and trends sometimes benefits from the projection of data into a different information space. An example of the later can be derived from transportation systems where pattern extraction benefits from the projection of the data to an information space whose spatial dimensions are non-metric. Gahegan (2000, cited in Miller & Han, 2001) argues that the useful information implicit in the geographic measurement framework is ignored by a number of induction and machine learning tools. Hence, there is a need to incorporate scalable versions of the available analytical cartographic techniques for estimating appropriate distance measures and projecting geographic information into geographic knowledge discovery (Miller & Han, 2001).

Finally, geographical datasets can comprise a *large volume* of data. This is more evident nowadays due to the trend to record almost every transaction across one or more database management systems depending on the application. These comprise of data sources that can provide valuable data for analysis. However, the volume of this data expands exponentially (Koperski *et al.*, 1998A). To extract valuable information and perform intelligent analyses from datasets that are large and sometimes found across distributed data sources, new type of algorithms and techniques are required which capable of handling such volumes of data efficiently.

Conventional data mining algorithms often make assumptions that do not comply with the special features of spatial data. One of the fundamental assumptions, especially in statistical based data mining methods, is that of independence. Statistical theory usually demands independent observations (Haggett *et al.*, 1977) while in the geographical domain that is not the case. This assumption contradicts spatially

dependency and ignoring that might result to inaccurate or inconsistent hypotheses or models (Shekhar *et al.*, 2001; Chawla *et al.*, 2000).

Regression modelling is an example of such a technique. It is commonly used in econometrics for prediction and complementary modelling is required to handle the spatial autocorrelation when applied to spatial data since the basic assumption in this method is that error vector ε is independent.

Dealing with the problem of spatial dependency involves two main directions (Haggett *et al.*, 1977). The first is to modify existing models so that they can allow the existence of autocorrelation in the data process which is extremely lengthy. The second is to transform the input data in such way that the correlation is removed. After this correction, conventional models could be employed.

Spatial stationarity is difficult to accomplish. Because of this global parameters fail to provide a description of the geographic phenomenon at any specific location. One way to account for spatial non-stationarity is by using spatial differencing techniques (Cliff and Ord, 1975).

2.2.3 Modelling Spatial Dependencies

In the literature, various ways to handle the spatial dimension of the examined data in a data mining model have been proposed. Based on the origin of a given technique, two main approaches can be identified. The first is the statistical approach and the second is based on the materialisation of spatial relationships.

Statistical Approach

Following the statistical approach to model the spatial dependency the most common practice is the modelling with relation to spatial autocorrelation. This is accomplished by initially quantifying the spatial autocorrelation of the spatial variables and then incorporating the outcome to modified traditional statistical models (e.g. Regression models).

A number of techniques that attempt to quantify spatial autocorrelation exist. Two of the most commonly encountered in spatial statistics literature are Moran's I and Geary's c measures (Chawla, *et al.*, 2000). Moran's I is considered as one of the older

measures to test for autocorrelation and is expressed in the form of Equation 2-1 where N is the total number of areas, w_{ij} the spatial weights and finally x_i , x_j and \bar{x} are the attribute values for areas i and j and mean attribute value respectively (Wang, 2006).

$$I = \frac{N \sum_i \sum_j w_{ij} (x_i - \bar{x})(x_j - \bar{x})}{(\sum_i \sum_j w_{ij}) \sum_i (x_i - \bar{x})^2} \quad \text{Equation 2-1}$$

Moran's I ranges from -1 to 1, denoting negative and positive autocorrelation respectively. Values that approach to 0 indicate the absence of spatial autocorrelation.

An alternative to Moran's I statistic is Geary's c that instead of basing the calculation on the deviation from the mean it uses the deviations of each observation with one another and is expressed as follows (Wang, 2006).

$$C = \frac{(N-1) \sum_i \sum_j w_{ij} (x_i - x_j)^2}{2(\sum_i \sum_j w_{ij}) \sum_i (x_i - \bar{x})^2} \quad \text{Equation 2-2}$$

Geary's c ranges from 0, for strong positive autocorrelation, to a value that approximates 2 for strong negative autocorrelation. In the absence of spatial autocorrelation the expected value is 1.

Both Moran's I and Geary's c have been implemented in several GIS or GIS-related packages such as ArcGIS and Geoda.

Spatial Relationships Approach

Methodologies of this type do not benefit from the use of statistical techniques to model the spatial dependencies. Instead the modelling is based on the direct modelling of basic spatial relations (Ester *et al.*, 1999). Spatial relations can be grouped in three broad categories (Egenhofer & Franzosa, 1991): topological, metric and relations based on the partial or total order of spatial objects.

A *topological relation* between spatial objects is a relation that holds irrespectively of transformations of the reference points such as translation, rotation or scale (Egenhofer, 1991). Among the several approaches to model these topological

relationships within a GIS context the most prevalent is considered the 9-intersection model proposed by Egenhofer & Franzosa (1991).

Their modelling was based on point-set topology and the notions of interior and boundary. Within this framework a topological relationship between two pointsets A and B can be derived from a set of intersections between their boundaries, interiors and complements. Therefore for the two pointsets A and B , based upon the comparison of A 's *interior* (A°), *boundary* (∂A) and *complement* (A^{-}), with B 's *interior* (B°), *boundary* (∂B) and *complement* (B^{-}) the 9-intersection (I_n) takes the form of the Equation 2-3.

$$I_n(A, B) = \begin{pmatrix} \partial A \cap \partial B & \partial A \cap B^\circ & \partial A \cap B^{-1} \\ A^\circ \cap \partial B & A^\circ \cap B^\circ & A^\circ \cap B^{-1} \\ A^{-1} \cap \partial B & A^{-1} \cap B^\circ & A^{-1} \cap B^{-1} \end{pmatrix} \quad \text{Equation 2-3}$$

The 9-intersection model realises 8 types of topological relations between pointsets that belong to 2-dimensional space. These relations are: A disjoint B , A meets B , A equals B , A inside B , A covered by B , A contains B , A covers B and A overlaps B (Egenhofer, 1991).

This model is an extension of their previously proposed 4-intersection model to include the relationships of the objects with respect to the embedded space. This addition allows the detection of objects that are or not completely included by other objects (Egenhofer, 1991).

Distance and direction belong to *metric spatial relations*. The main difference from the topological relations is that both direction and distance remain invariant under a smaller group of transformations. Due to this differentiation Frank (1996) characterises the topological relations as first level qualification. That is, topological relations provide a first level classification of spatial relationships which can be further explained by metric relations.

One characteristic of metric relationships is that their definition can be model specific. The distance metric, for example, can be defined in a number of different ways. The most simple is the Euclidean distance, or when referred to curved spaces Geodesic distance which is the shortest distance between two points. Depending on the nature

of the model, other types of distances might be more suitable like Spherical Manhattan Distance, Lexicographic Distance or even use distances that belong to other spaces like travel time distance which belong to the quasimetric space (Worboys and Duckham, 2004).

In Worboys and Duckham (2004) the following formal definition of a metric space and therefore a definition of a metric distance is provided. Let S a point set and $s, t, u \in S$. S defines a metric space when there is a distance function d that takes ordered pairs (s,t) and results a distance that is subject to three conditions.

- i. $d(s,t) > 0$ when s and t are distinct points and $d(s,t) = 0$ when s and t are identical.
- ii. $d(s,t) = d(t,s)$
- iii. $d(s,t) + d(t,u) \geq d(s,u)$ (Triangle inequality)

Thus, when referring to a distance that belongs to a metric space then those three conditions must be fulfilled. Other distances may comply with some of the conditions but not all of them. In the case of *travel time* distance conditions 1, 3 are satisfied but condition 2 may not be.

The determination of a metric relation in the case of extended objects is not straightforward. The definitions that apply in the case of point to point relation need to be extended to express metric relations between line and point, line and line and line and area (Frank, 1996). This is more apparent in the case of direction. Since we are not referring to one point but to a set of points some kind of reference must be introduced in the source object based on which, the directions to the reference objects will be determined.

There are several proposed definitions for the direction and distance relations. For illustration purposes here the definitions provided by Ester *et al.* (1999) are presented. The choice was based on the fact that along with the topological relations these relations have been used in spatial data mining algorithms that are reviewed in the following sections. Using a generic representation of spatial objects (set of points) to cover all types of spatial objects (points, polygons etc.) the above relationships can be formally defined as follows (Ester *et al.*, 1999).

Let a set of points $Points = \{p_1, p_2, \dots, p_d\}$ that belong to a d -dimensional Euclidean vector space. Spatial objects can be represented by $O \in 2^{Points}$. In the case of a 2-

dimensional space and a given point $p = (p_x, p_y)$, p_x and p_y stand for the x,y co-ordinates respectively. Let $\Delta x(O) := \max\{|o_x - p_x| \mid o, p \in O\}$ be the x-extension of object O and $\Delta y(O) := \max\{|o_y - p_y| \mid o, p \in O\}$ the y-extension of object O .

For a distance function dist between two spatial objects O_1, O_2 the distance relation $A \text{ distance}_{\sigma c} B$ holds iff $\text{dist}(O_1, O_2) \sigma c$ where σ one of the arithmetic predicates $=, <, >$ and c a real number.

The directional relationship between the source object O_1 and a reference or destination object O_2 is defined based on a representative point $\text{rep}(O_1)$ of the source object O_1 . This is the centre of a virtual coordinate system which will be used for the determination of the direction. For example, in Figure 2-4 the directional relation $B \text{ northeast } A$ holds iff $\forall b \in B: b_x \geq \text{rep}(A)_x \wedge b_y \geq \text{rep}(A)_y$

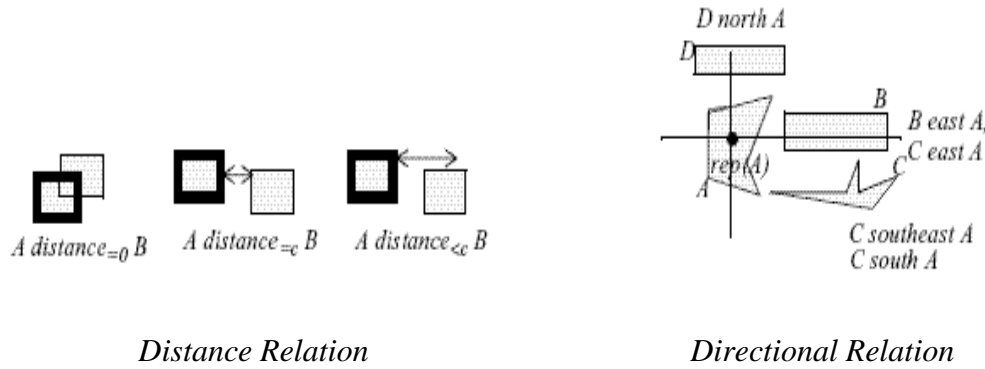


Figure 2-4: Metric Spatial Relationships

(Source: Ester et al., 1999)

Since algorithms of this type rely on the materialisation of the topological and metric relationships, a distinction can be made based on the way this materialisation is performed. Two main directions can be distinguished regarding the incorporation of the spatial relationships. In the first direction, this is included in the pre-processing step allowing in that way the application of traditional data mining algorithms for the pattern discovery. The second approach incorporates this as part of the algorithm. The calculation is on-the-fly and usually is performed in two steps. The first step involves the determination of spatial relationships at a coarse level. The second step refines the results of the initial step by identifying spatial relationships at a finer level.

2.3 Spatial Data Mining

As presented in the previous sections, spatial data is associated with unique characteristics that justify its differentiation from non-spatial data in relation to data mining. These differences relate both to the different needs and anticipated outcomes when investigating spatial phenomena and also to the need for modification or redesigning of data mining algorithms to deal with these special characteristics.

This section focuses on spatial data mining and presents an overview of spatial data mining techniques. A further emphasis is placed on techniques that belong to spatial data mining tasks relevant to this research that is Spatial Classification and Spatial Dependency Analysis.

2.3.1 Tasks

In this section, the data mining tasks are revisited through a spatial data perspective. Miller and Han (2001) provides the following description of the spatial data mining tasks:

Spatial segmentation includes the tasks of *spatial clustering* and *spatial classification*. *Spatial clustering* involves the grouping of spatial objects into classes or clusters where objects of the same group have similar characteristics. Such grouping can be based on any combination of spatial or aspatial attributes of objects or on the proximity of objects in space or time or both.

Spatial clustering attracted a lot of research interest. Research in computer science led to the development of a number of scalable algorithms along with methods for finding proximity relationships between clusters and spatial features. On the other hand, spatial analytical approach is focused on finding theoretical conditions for appropriate clustering in space and time (Miller & Han, 2001). Han *et al.* (2001) classify clustering algorithms that perform reasonably well on large geographical databases into four general categories: partitioning, hierarchical, density-based and grid-based (see Section 2.3.2.1). These algorithms mainly work with numerical attributes.

Spatial classification involves the organization of spatial objects into pre-defined classes based on a relevant set of attributes and attribute values (Miller & Han, 2001; Ester *et al.*, 1997).

Spatial dependency analysis includes the discovery of rules that predict the value of some attributes, which is based on the value of other attributes. One or more of such attributes is spatial.

Spatial outlier analysis involves the identification of outliers and the analysis of their properties. Shekhar *et al.* (2003) define the spatial outlier as “*a spatially referenced object whose non-spatial attribute values are significantly different from those of other spatially referenced objects in its spatial neighbourhood*” (p. 140).

Spatial trend detection identifies change patterns in relation to the neighbourhood of some spatial objects.

Finally, *geographic characterisation* and *generalization* involves the compact description of a selected subset of the database (Ester *et al.*, 2001). It is a very important task in data mining since geographic phenomena often present complex hierarchical dependencies (Miller & Han, 2001).

2.3.2 Techniques

There is a variety of techniques that can be used in order to perform the spatial data mining tasks outlined above. An overview of such techniques according to their applicability to those tasks follows. Spatial classification and spatial dependency techniques are reviewed in depth since they form the basis for the spatial data mining algorithm used in the developed system.

2.3.2.1 Spatial Clustering Techniques

Spatial cluster analysis is one of the most commonly used techniques in analyzing spatial data. Due to its popularity and wide applicability in geographical research this area has been widely researched. This popularity is also underlined by the development of a large number of algorithms. Clustering algorithms appear irrespectively of the data mining technology. These methods usually involve statistical

approximation or heuristics, due to the requirement to deal with large volumes of high dimensional data. This type of analysis involves the definition and assignment of a set of classes based on the data's relative proximity within the information space (Miller & Han, 2001). According to Han *et al.* (2001) clustering algorithms can be classified in the following general groups: partitioning methods, hierarchical methods, density-based and finally Grid-based methods, which are described below.

Iterative Relocation
Algorithm 2-6

Input:	The number of cluster k , and a database containing n objects
Output:	A set of k clusters which minimizes a criterion function E
Method:	<ol style="list-style-type: none"> 1. arbitrarily choose k centres/distributions as the initial solution 2. repeat 3. (re)compute membership of the objects according to present solution 4. update some/all cluster centres/distributions according to new memberships of the objects 5. until no change to E;

Source: Han et al. (2001)

Partitioning algorithms had been very popular even before the appearance of data mining. A partitioning algorithm for a given set D of n objects in a d -dimensional space and an input parameter k , organises the objects into k clusters such that the total deviation of each object from its cluster centre or from a cluster distribution is minimised. The deviation point is usually called *similarity function* (Han *et al.*, 2001A). Representative algorithms of this type are the *k-means*, the *Expectation Maximization (EM)* and the *k-medoid* algorithms. Although the three algorithmic methods differ in the way they represent clusters, they share a similar general approach for their computations. All three methods adopt an iterative relocation technique (Algorithm 2-6) to find a local optimal k centre. However, they differ in the way they perform steps 3 and 4 and in the criterion function. One weakness of partitioning methods is their requirement to specify the parameter k , and that they fail to find arbitrarily-shaped clusters (Han *et al.* 2001).

Hierarchical methods are based on the decomposition of a given dataset by structuring a dendrogram that can be formed either using the bottom-up or top-down approach (Han *et al.*, 2001A). The bottom-up, or agglomerative approach initially creates

separate groups for each object. It then merges them according to some measures (e.g. distance of the two centres of two groups). On the other hand, a top-down or divisive approach initially considers every object in the same group and then splits them, based on some measures, until either each object is in one cluster or a termination condition is satisfied. Depending on which approach an algorithm follows they are classified into two categories: Agglomerative and Divisive respectively. Examples of agglomerative algorithms include CURE (Guha *et al.*, 1998), CHAMELEON (Karypis *et al.*, 1999) and BIRCH (Zhang *et al.*, 1996). One of the primitive divisive algorithms is DIANA proposed by Kaufman and Rousseeuw (1990).

Unlike most of the partitioning methods that are based on the distance between two objects, density-based algorithms are based on the notion of density. Such algorithms consider clusters as dense regions of objects, which are separated by regions of low density. Advantages of this kind of algorithms include their ability to filter out noise and discover arbitrary shape clusters (Han *et al.*, 2001A). DBSCAN (Ester *et al.*, 1996), OPTICS (Ankerst *et al.*, 1999), DENCLUE (Hinneburg and Keim, 1998) are some of the density-based algorithms.

Grid-based algorithms differ in the way they handle data. Unlike algorithms that are index-based, these algorithms adopt a grid-based clustering approach using grid data structures. This results to an increase in efficiency especially in the case of high dimensional data, since processing time depends only on the number of cells in each dimension in the quantized space. On the other hand, although summarizing information increases efficiency it increasingly loses effectiveness as the number of dimensions increases. Typical examples of this category are the STING (Wang *et al.*, 1997), the WaveCluster (Sheikholeslami *et al.*, 1998) and the CLIQUE (Agrawal *et al.*, 1998) algorithms.

Despite the fact that spatial clustering is an extremely popular technique, cluster reasoning attracted less attraction. Explaining the ‘why’ behind a cluster formulation is extremely important and usually remains unanswered. In an attempt to enhance spatial clustering algorithms Knorr and Ng (1996) propose an algorithm that seeks to explain the reasoning behind the formulation of a spatial cluster. They based their method in the calculation of the aggregate proximity relationships between the input clusters and related features. In addition, based on the extracted relationship they

identify possible existing ‘commonalities’ among the various clusters. By the term commonalities they referred to the discovery of frequent specific feature-cluster relationships in the dataset.

2.3.2.2 Spatial Classification

Spatial classification in its simplest form is to find the function: $f : D \rightarrow L$ where D is the n -dimensional space of attribute data and L represents the set of labels (Shekhar & Chawla, 2003). Classic data mining techniques for classification such as regression or Bayesian classifiers fail to cope with data that are not independently generated. This failure affects the overall classification accuracy (Shekhar *et al.*, 2002).

Hence, for spatial data that are characterised by dependency, classification techniques that model spatial dependency are needed. Shekhar *et al.* (2002) refers to a number of studies that applied classic classification techniques to spatially dependant data with not satisfactory results, fact that highlights the need for spatial handling when dealing with spatial data.

The logistic Spatial Autoregression (SAR) Model and Markov Random Field-Based Bayesian Classifiers are two approaches that incorporate spatial dependence into classification models (Shekhar *et al.*, 2002). Additionally Ester *et al.* (1997) propose a classification rule algorithm for spatial data mining. Koperski *et al.* (1998B) also proposed another classification rule algorithm. An overview of the three approaches along with overviews of the available algorithms follows.

Logistic Spatial Autoregression Model (SAR)

The Logistic Spatial Autoregression Model (SAR) is an extension of the classic regression model that incorporates spatial dependence. Based on a SAR model the class label of a location is depended both on the class label of the neighborhood and on the feature values (Shekhar *et al.*, 2002).

In this case, the equation incorporates the spatial dependencies of the error term or the dependent variable. In order for this direct modeling of the spatial dependency to

be achieved, the regression equation can be modified as follows (Shekhar *et al.*, 2002):

$$y = \rho W_y + X\beta + \varepsilon \quad \text{Equation 2-4}$$

where W is the neighborhood relationship contiguity matrix and ρ stands as an indicator of the strength of spatial dependency between the elements of the dependent variable.

Markov Random Field-Based Bayesian Classifiers

This classification model, bases its estimation on Markov Random Fields (MRF) and Bayes' rules. Li (1995, cited in Shekhar *et al.*, 2002) defines an MRF as a set of random variables, the interdependency relationship of which is represented by an undirected graph.

Bayesian classifiers calculate the probability of the class label for a given dataset by using Bayes' rule (Shekhar *et al.*, 2002):

$$\Pr(c_i|X) = \frac{\Pr(X|c_i) \Pr(c_i)}{\Pr(X)} \quad \text{Equation 2-5}$$

where c_i stands for the class labels for given data X .

Classification Rules

Classification rules are sets of rules within a database that work as a classifier. There are several techniques to derive such classification rules such as entropy-based, statistical or artificial neural networks (Miller & Han, 2001). Ester *et al.* (1997) and Koperski *et al.* (1998B) provide methodologies for mining classification rules out of spatial data. An overview of their algorithms follows.

Classification Algorithm by Ester et al. (1997)

This algorithm presented by Ester *et al.* (1997) is based on the ID3 algorithm, an inductive learning algorithm introduced by Quinlan (cited in Ester *et al.*, 1997), and discovers classification rules in order to determine the class of an object based on the values of its attributes. ID3 algorithm was designed for relational databases, taking into consideration only the attributes of the object to be classified. It adopts the

strategy ‘divide and conquer’ and bases the selection of attributes upon the information entropy (Quinlan, 1993, cited in Li *et al.*, 2000).

This extension for spatial databases considers also the attributes of the neighbouring objects introducing the concept of neighbourhood graphs to explicitly represent those implicit neighbourhood relations relative to the classification task.

Ester *et al.* (1997), define a generalized attribute for a neighbourhood path $p=[o_1, \dots, o_k]$ as a tuple (attribute-name, index). In their definition, index refers to a valid position of an object o_{index} within the neighbourhood path that has the particular attribute. Attribute-name is the name of that attribute.

Furthermore, this classification algorithm allows the input of a parameter max-length and a predicate. Max-length limits the length of neighbourhood path since the influence of neighbourhood objects and their attributes decrease with the increase of distance. The input of the predicate focuses the search of classification rules on the objects fulfilling this predicate.

Classification Algorithm by Koperski et al. (1998B)

Koperski *et al.* (1998B) propose a methodology that enables the classification of spatial objects based on aggregated values of non-spatial attributes for neighbouring regions and spatial relations between objects that are represented as spatial predicates. The objective of this methodology is to mine rules that group together objects that share the same class label.

The classification is based on four different types of data: Non-spatial attributes of the data objects, spatially related attributes with non-spatial values, spatial predicates and spatial functions. Along with the set of objects and other spatial objects with non-spatial attributes, two other inputs are also required: (a) geo-mining queries that specify: the objects that will be used in the classification, the predictive attributes, predicates and functions and finally the attribute, predicate and function used as a class label; and (b) a set of concept hierarchies.

Deriving spatial predicates and functions from data can be quite time consuming. To overcome that, a two-step approach is adopted. Initially, rough computations have

been performed and then refined calculations involving machine-learning methods applied only on the promising patterns. After the completion of this step, predicates, functions and attributes are defined for each data object of the sample.

OID	high_profit	Predicates
1	Y	sum_population(x, MEDIUM), avg_income(x, SMALL), close_to(x, park), close_to(x, water)
2	Y	sum_population(x, LARGE), avg_income(x, MEDIUM), close_to(x, park), close_to(x, water)
3	N	sum_population(x, MEDIUM), avg_income(x, LARGE), close_to(x, park), close_to(x, water)
4	N	sum_population(x, SMALL), avg_income(x, MEDIUM), close_to(x, park), close_to(x, water)
5	N	sum_population(x, LARGE), avg_income(x, LARGE), close_to(x, park), close_to(x, water)

Figure 2-5: Example of generalised predicates

(Source: Koperski et al., 1998B)

The next step involves the search for the optimum buffer size to calculate the aggregation values for all the relevant attributes. After the completion of this step the building of a set of predicates that describe all the objects based on relevant predicates, functions and attributes follows. When every object is described, the generalisation of these sets of predicates, based on a hierarchical concept, is performed. Finally, the binary decision tree is created using an ID3 algorithm (Quinlan, 1986, cited in Ester *et al.*, 1997). Figure 2-6 presents a sample output decision tree based on data shown in Figure 2-5.

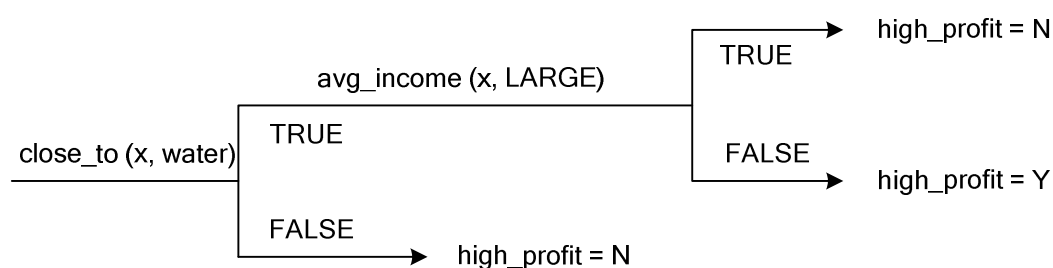


Figure 2-6: Example decision tree

(Source: Koperski et al., 1998B)

2.3.2.3 Spatial Dependency

The most prevalent technique to perform the spatial dependency task is through the discovery of spatial associations. In the literature, two main directions in the

association rule mining algorithms can be identified: Co-location mining and Reference feature based mining. An overview of these two types of algorithms follows.

Co-location Algorithms

Co-location mining involves the identification of spatial events that frequently co-occur within the geographic space. Figure 2-7 is an example dataset where the problem of co-location mining can be applied. There are two subsets of events that frequently occur together within their neighbourhood that can be identified in this sample and are indicated by the blue and red ellipses.

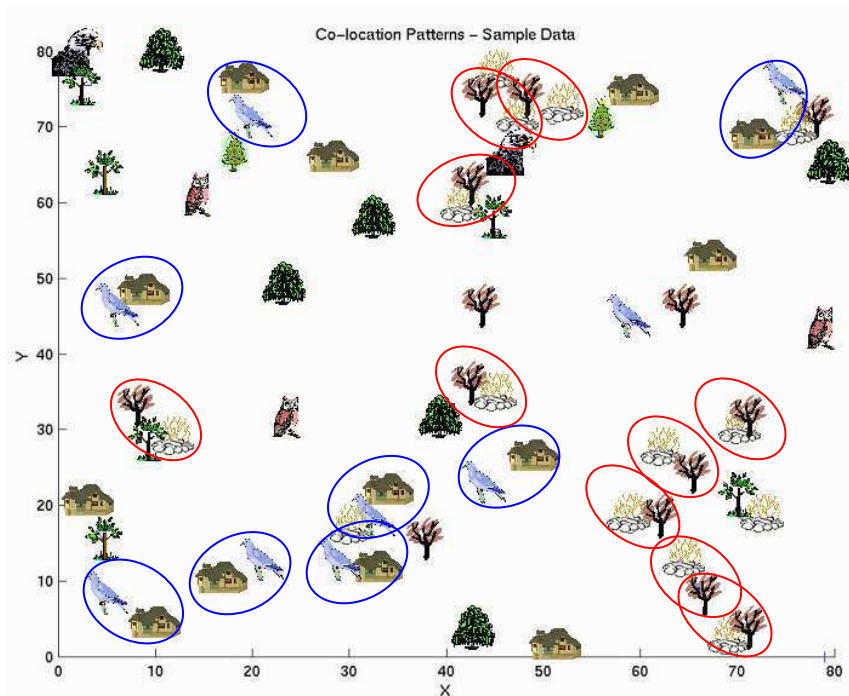


Figure 2-7 : Example of co-location patterns

(After: Shekhar & Chawla, 2003)

The problem of mining spatial co-location rules can be formally described as follows (Yoo & Shekhar, 2004): Let $E = \{e_1, \dots, e_k\}$, be a set of Boolean spatial events, $S = \{i_1, \dots, i_n\}$ a set of their instances and R a neighbour relation over S . A co-location C is defined as the subset $C \subseteq E$ whose instances $I \subseteq S$ form a clique based on a relation R . Let C_1, C_2 be subsets of set C and $C_1 \cap C_2 = \emptyset$. Using these definitions a co-location rule is an expression of the form $C_1 \rightarrow C_2(p, cp)$ where p and cp represent

the prevalence measure and the conditional probability respectively. The conditional probability $Pr(C_1/C_2)$, is the probability of the instance C_2 to be found in an instance of C_1 . As a prevalence measure of the co-location the participation index $Pi(C)$ is used and is defined as the $\min_{e_i \in C} \{Pr(C, e_i)\}$ where is the participation ratio for event type e_i in a co-location C and $Pr(C, e_i)$ is the fraction of instances of e_i which participate in any instance of co-location C .

As described in Section 2.1.3.1, association rule mining is applied to data organised in transactions. In the case of co-location mining transactions are not clearly defined, therefore a way of partitioning the spatial database must be found and applied. Geographical space is continuous and unless imposed by the problem, there are no reference points to relate with and form the transactions. As a result, research on co-location mining has been focused on the issue of partitioning the spatial database without compromising the accuracy of the outcome.

Unless it is embedded in the problem itself, partitioning continuous space into transactions is challenging. The partitioning of the continuous space must be performed in such way that the splitting of co-location patterns across different transactions minimizes the risk of splitting patterns across transactions (Yoo & Shekhar, 2004).

Two main issues arise: find an efficient way to partition the space and also keep track of the relationships that are across partitions. Several partitioning methods can be adopted and used in the case of the neighbourhood transactions. Such methods include the use of grids, maximal cliques, max-clique agglomerative clustering, min cut partitioning etc. (Yoo & Shekhar, 2004).

Morimoto (2001) proposed a methodology for co-location pattern discovery or in his term, discovery of frequent neighbouring class sets. His approach makes use of a 'nearest' grouping function based on Euclidean distance for the space partitioning and the identification of the neighbourhoods by constructing a Voronoi diagram. The problem with this partitioning method is that it may omit relationships that exist across the partitions.

Yoo & Shekhar (2004) propose a partial join approach that reduces the number of joins in order to identify those instances where their relationship is cut apart across transactions. For the partitioning of the space and the materialisation of the neighbourhoods the clique partition method proposed by the authors is used.

In 2006, Yoo & Shekhar proposed another algorithm for co-location pattern mining that does not require a join. In this work, they propose a star neighbourhood partition model for the materialisation of the neighbourhoods. Their co-location mining algorithm operates in three stages. In the first stage the input spatial data form a set of disjoint neighbourhood graphs. In the second phase the star instances (candidate co-location instances) are gathered and coarsely filtered based on their prevalence values. Finally, the third phase filters the co-location instances from their star instances and finds prevalent co-locations and co-location rules. The second and third phases are repeated for each increment of the co-location pattern size.

Reference feature based

Unlike co-location algorithms, reference feature based algorithms follow the general structure of association rule mining and base the rule extraction on transactions created in relation to reference points. Algorithms of this type have been developed by Malerba and Lisi (2001), Koperski and Han (1995) and Savinov (2003). An overview of these algorithms follows.

SPADA (SPatial PAttern Discovery Algorithm)

Introduced by Malerba and Lisi (2001), SPADA algorithm is based on inductive learning programming (ILP) that enables the extraction of multi-level association rules. Multi-level association rules allow the spatial objects to be at different granularity levels.

Koperski and Han (1995)

This algorithm introduced by Koperski and Han (1995) is designed to uncover strong spatial association rules in geographical information databases. To do so first they introduce the concepts of spatial association rules, support and confidence. They

define a spatial association rule as:

$$P_1 \wedge \dots \wedge P_m \rightarrow Q_1 \wedge \dots \wedge Q_n \text{ (c\%)}$$

where at least one of the above predicates is spatial and a c% of the objects that satisfy the antecedent of the rule will also satisfy the consequent of the rule.

The algorithm consists of five steps and requires as input a database (spatial, relational, concept hierarchies), a query and finally two thresholds (minimum support, minimum confidence). These five steps can be summarised as:

Step 1: Collection of all task relevant objects into one database (spatial query)

Step 2: Extraction of neighbourhood objects and store of predicates that describe the spatial relationship into relational database.

Step 3: Computation of the support for the predicates (from step 2) and filtering of those that are below the thresholds.

Step 4: Refined computations on the predicates obtained in step 3

Step 5: Generation of the association rules at multiple concept levels.

According to Koperski and Han (1995), among the strongest advantages of their approach is the fact that the mining process is directed by the user and its efficiency. Malerba *et al.* (2002) on the other hand, they argue that this approach suffers from limitations related to the method's single-table assumption.

This algorithm has been implemented in the Geo-associator module of GeoMiner spatial data mining system (see Section 2.3.3). An extension of this algorithm to deal with uncertainties that may exist in spatial data can be found in Clementini *et al.* (2000). Ester *et al.* (1997) also use this methodology to apply their neighbourhood graphs for the discovery of association rules.

Optimist (Savinov, 2003)

The Optimist algorithm has been introduced by Savinov (2003) and has been implemented into SPIN! spatial data mining system (see Section 2.3.3) as one of its

components. The input of data is facilitated by a SPIN! query component and rules are stored in the rule base component.

Savinov argues that the strength of this algorithm lies on the extraction of highly expressive multiple-valued rules by only one pass over the data and also on its efficiency.

2.3.2.4 Spatial Outliers

Shekhar *et al.* (2003) broadly classify the outlier detection methods into two categories: One-dimensional (linear) and Multi-dimensional. Figure 2-8 demonstrates the proposed classification.

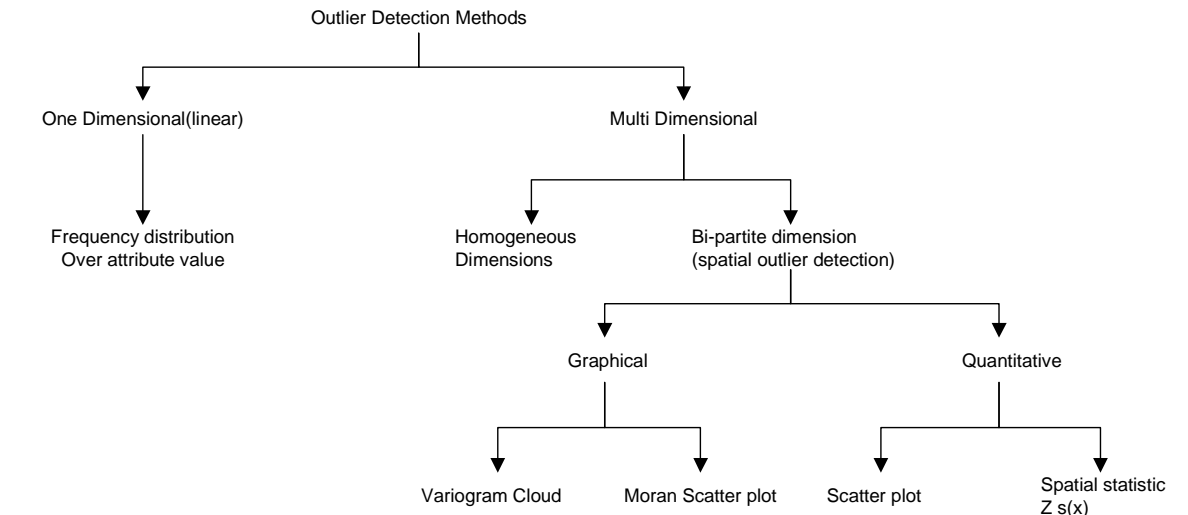


Figure 2-8: Classification of Outlier Detection Methods

(Source: Shekhar *et al.*, 2003)

One-dimensional methods include distribution-based algorithms also known as discordancy tests. Multi-dimensional methods are further classified as Homogeneous and Bi-partite dimension (Shekhar *et al.*, 2003). In addition to the above classification, Ng (2001) classifies outlier detection techniques into four categories: noise-based, distribution-based, depth-based and distance-based.

The above classifications are overlapping. What follows is a combined review of the work of both Shekhar *et al.* (2003) and Ng (2001).

Noise-based Algorithms

This category includes algorithms that, although not exclusively designed for outlier detection, are nevertheless capable to designate certain data objects as noise and therefore as outliers. Hence, detection of outliers is not the main task of these methods but a by-product.

Algorithms of this type can be drawn from several technique groups, such as from robust estimation and data clustering methods. For example, in case of partitioning clustering algorithms (k-means, k-medoids etc.), an object is characterised as an outlier if its removal results in more tight clusters.

Identification of outliers through their characterisation as noise underestimates the importance of outliers. Techniques that are designed to carry out a specific task, base the definition of an outlier on their task, resulting to outliers that may not be suitable for the application.

One-Dimensional

Unlike noise-based algorithms, distribution-based algorithms are specially designed for outlier detection, basing outlier definition on the value distribution within a dataset. Considering that values of a dataset follow a standard distribution, outliers can easily be defined and can also have an indication of their strength.

For a normal distribution value v is labelled as an outlier when it is out of the range $\mu \pm 3\sigma$ where μ and σ are known mean and variance respectively. Distribution-based outlier definitions are also known in statistics under the term discordancy tests.

Advantages of distribution-based algorithms are that they are straightforward and their ability to give indication on the strength of the identified outliers. The main disadvantage of this method is related to the assumption that the distribution within a given dataset is known. In cases when a distribution is not known, the application of distribution-fitting usually contributes in overcoming the problem. Although distribution-fitting might solve problems of unknown distributions, it has the following shortcomings. The first relates to the fact that not all distributions fit on standard distributions. The second is that it is time-consuming in terms of conduct.

Another disadvantage of distribution-based approaches relates to their inability in general terms to handle high dimensional data such as geographic data.

Multi-Dimensional

This group of algorithms can be further classified into Homogeneous Dimensions and Bi-partite Dimension methods. Homogeneous dimensions methods include the depth-based and distance-based algorithms. Depth-based approaches are based on the notion of data depth. Data is organised in k-dimensional space and is being represented as points within that space. A depth is applied to its data object based on the location of the data. In such approaches, data objects with small depth are more likely to be characterised as outliers.

Due to the fact that outlier definition is based on data depth, these approaches can be further classified depending on the definition of data depth they use. Examples of such definitions include convex hull peeling depth (Preparata and Shamos, 1988) and Tukey depth (Tukey, 1975, 1977) both cited in Ng (2001). Tukey depth is more sophisticated and popular than peeling depth.

Distance-based algorithms are based on distance and distribution. Knorr *et al.* (2000, cited in Ng, 2001) characterise an O as a DB(p, D) outlier if at least p per cent of the other objects are of distance $\geq D$ from O. Distance-based algorithms are more efficient than other algorithms even in the case of large dimensional data and are computationally more tractable than depth-based outliers. Furthermore more distance-based algorithms can be conducted in such way that ensures that the dataset is read no more than three times. That is very important in case of large datasets. Another advantage of these algorithms is that apart from the identification of outliers they also provide an explanation of why a specific outlier is exceptional, contributing to the validation of the resulted outliers and to the understanding of the data. A main disadvantage of distance-based algorithms is their strong dependency on the existence of an appropriate distance function e.g. weighted Euclidean distance.

Similarly, Bi-partite dimension methods are of two kinds: Graphical and Quantitative. The Graphical techniques are based on the data visualisation for the identification of the outliers. Among them are the variogram-cloud and Moran scatterplot. Quantitative

techniques provide a precise test to distinguish the spatial outliers. Quantitative techniques include scatterplots and spatial statistic $Z_{s(x)}$ (Z-score).

2.3.2.5 Spatial Trend Detection

Trend detection usually involves the use of regression techniques or the use of sequential pattern extraction techniques. Regression techniques include linear and logistic regression analysis and are usually combined with filtering techniques such as stepwise regression. One can note that often data violate the strict regression assumptions. When that is the case, violations are less crucial if estimated parameters are used to predict, rather to explain phenomena. (Miller & Han, 2001). Pattern extraction methods are used to explore time series data. Hence, they search for temporal correlations or pre-defined patterns in a single temporal data series (Miller & Han, 2001).

An example of a spatial trend detection algorithm is that proposed by Ester *et al.* (1997). This algorithm is designed to discover patterns of change of some non-spatial attributes in the neighbourhood of some database object starting with an object o . For a specified attribute both the local changes moving to the neighbours and the distance to these neighbours are calculated. The trend for object o is identified by applying a linear regression to the pairs of changing values-distance. For correlation coefficients larger than a specified threshold, the trend of object o is the slope of the resulting linear function. For smaller coefficients, no trend is applied on o .

2.3.2.6 Geographic Characterisation and Generalisation

For generalization and characterisation, techniques such as summary rules and attribute-oriented induction are used. Summary rules are relatively small sets of logical statements that condense information contained in the database (Miller & Han, 2001). One type of such a rule is the characterization rule. Klosgen and Zytkow (1996) cited in Miller & Han (2001) define characterisation rule as: an assertion that data items belonging to a specified concept have stated properties, where *concept* is some state or idea generalised from particular instances.

The characterisation algorithm proposed by Ester *et al.* (2001) is designed to mine characterisation rules from spatial data based on the concepts of neighbourhood graphs and neighbourhood paths. Initially the algorithm selects a small set of target objects based on a criterion condition and then expands this selection around the target points by selecting also the regions for which the distribution of values differs from the distribution in the whole database. After the final selection the generation of the characterisation rule that describes the target data objects follows. The generated characterisation rule is expressed as:

$$\text{Target} \Rightarrow p_1(n_1, \text{freq-fac}_1) \wedge \dots \wedge p_k(n_k, \text{freq-fac}_k)$$

Where for all target objects extended by n_i neighbours, property p_i is freq-fac_i times more or less frequent than in the database.

Attribute-oriented induction is another powerful technique that can be applied to the Geographic characterisation and generalisation tasks of spatial data mining. It involves the hierarchical aggregation of data attributes by compressing data into generalized relations (Miller & Han, 2001). Data aggregation is based on background knowledge. Background knowledge is represented in the form of a concept hierarchy (Miller & Han, 2001), which can be derived, either from experts or from data cardinality analysis (Han and Fu, 1996).

An attribute-oriented induction has been implemented in DBMiner data mining system on top of which GeoMiner (see Section 2.3.3) is implemented. Lu *et al.* (1993) provide such technique extended for generalisation-based knowledge discovery in spatial databases. For the construction of the background knowledge two types of concept hierarchies are created: thematic and spatial. Based on those hierarchies, an induction is performed by summarising the relationships between a spatial and aspatial attributes at a high concept level.

2.3.2.7 Spatial Associative Classification

Following the example of developing hybrid data mining techniques there is also some research on the integration of spatial association rule mining and classification towards the development of spatial associative classification algorithms.

Representative algorithms of this type are the SPARC algorithm (Han *et al.*, 2001B) and the algorithm proposed by Ceci *et al.* (2004).

SPARC algorithm operates in two steps. Spatial relationships are pre-calculated and stored in an information-associated spatial join index structure that is used as an input for the algorithm. The first step involves the discovery of the classification rules while the second step results the classification model based on the extracted rules in the first step. The classification rule mining is based on an Apriori-like algorithm while the construction of the classifier is based on a classification rule sorting scheme that removes rules that have lower precedence and do not cover any additional cases.

Unlike SPARC, the algorithm proposed by Ceci *et al.* (2004) is not based on Apriori for rule discovery. Instead, this spatial associative classification algorithm is based on the previously developed SPADA algorithm (Lisi & Malerba, 2004) and a multi-relational naïve Bayesian classifier.

2.3.3 Existing Systems

The need for the development of new tools that will efficiently handle spatial data have been underlined, in this chapter. Although a number of data mining tools exist (Intelligent Data Miner, MineSet etc) they mainly focus on the prediction and modelling of customer-buying behaviour through analysing commercial datasets. Therefore their usefulness will be limited, if applicable, in a GIS context where pattern recognition is more often the case (Openshaw, 1999).

Apart from the fact that most of the developed tools specialise in prediction and not in pattern discovery, there is an additional reason why there is a need for new methods and tools that take into consideration the data origins. That is the nature of the geographical data itself. Ignoring the special characteristics of geographic data is risky and might produce misleading results.

To date two spatial data mining prototypes have been developed: GeoMiner and SPIN! These packages are examples of a holistic approach in the design of such systems. It should be noted that limited type of spatial data mining functionality (mainly spatial statistics procedures) is readily available and is also incorporated into

mainstream GIS packages. An example of such functionality is the Spatial Analyst module of ArcGIS software.

GeoMiner is a knowledge discovery system prototype that was designed and implemented at the Database Systems Research Laboratory, in Simon Fraser University in 1997. It is an extension of their relational data mining system (DBMiner). It is developed on the top of DBMiner and its general architecture consists of:

- i. Graphical User Interface for interactive mining
- ii. Seven discovery modules: Geo-characteriser, Geo-Associator, Geo-comparator, Geo-classifier, Geo-cluster analyser, Geo-Predictor and Geo-pattern analyser.
- iii. Spatial Database server which includes MapInfo Professional 4.1
- iv. Data cube mining engine
- v. Data and knowledge-base

Algorithms implemented in this prototype such as the Koperski and Han (1995) algorithm for association rule mining in the geo-Associator module, have been reviewed in previous sections.

The SPIN! Project (2001-2004) was funded by the European Commission and involved the development of a web-based spatial data mining system by integrating GIS and data mining functionality. The general architecture of SPIN! is a n-tier Client/Server-architecture based on Enterprise Java Beans that includes client, application server, database server (s) and optionally compute servers. SPIN! can be considered as an integrator of already existed technologies. These include: Descartes (visualisation based data mining system), Lava/Magma (GIS), GAM (Geographical Analysis Machine- Exploratory Spatial Analysis Tools) and Geoprocessor and Kepler.

2.4 Summary

In this chapter the general concepts of knowledge discovery and especially their application to the geographic domain have been discussed and identified as an appropriate research area. Knowledge discovery in conventional databases is a well

documented and recognised area. Although knowledge discovery in geographical databases is a relatively new area, it has enjoyed attention from the academic community. Today, the academic community has identified Geographical Knowledge Discovery as a new but important, exciting, and dynamic field that is quickly becoming a useful tool in geosciences (Miller, 2004; Gahegan, 2001; Ester *et al.*, 1997; Fayyad *et al.*, 1996B; Koperski *et al.*, 1998A; Ester *et al.*, 1999).

Many researchers argue that this is an emerging research domain which can potentially lead to compelling results. In the literature, several areas that need elaboration have been identified (National Research Council, 2003; Turner, 2002; Miller & Han, 2001; Battenfield *et al.*, 2001; Ester *et al.*, 2001; Openshaw *et al.*, 1999; Koperski *et al.*, 1998A). Among them is mainly the handling of the special characteristics of spatial data and the successful integration of knowledge discovery and Geographic Information Science stand out as very promising areas (Koperski *et al.*, 1998A; Yuan *et al.*, 2001; Miller, 2004).

A typical knowledge discovery process involves five basic steps: (1) Selection, (2) Pre-processing, (3) Transformation, (4) Data mining and (5) Interpretation / evaluation. Data mining is a key component in this process and involves the application of algorithms that enable pattern discovery. Common data mining tasks include: Segmentation, Dependency analysis, Deviation and Outlier analysis and Generalisation and characterisation. As knowledge discovery is a complicated and multidisciplinary process, there are various techniques one can use, depending on which outcome is anticipated and the perspective the problem is approached.

The hybrid technique of associative classification has been selected as being most appropriate in the context of this research. It combines two data mining techniques, association rule discovery and classification. As discussed earlier, recent research has shown that such rule-based techniques present a number of advantages, such as improved accuracy and understandability, when compared to other classic classification approaches.

Given that these techniques have been developed for non-spatial data their applicability to spatial data has been investigated. In the case of geographical knowledge discovery additional considerations were discussed related to the special

features of spatial data. Spatial dependency and spatial heterogeneity are two of the main features of geographical datasets that mainly affect statistical based techniques. Other complexities include the MAUP, spatial data representation issues, spatial relationships and the geographic measurement frameworks.

Conventional data mining algorithms often make assumptions that do not comply with the special features of spatial data. Ignoring these unique characteristics of the data may result to erroneous and misleading results. Therefore effective modelling is required.

A number of spatial data mining techniques that apply to the various tasks of data mining have been identified and reviewed. Two main directions in the handling of the spatial dimension, within the spatial data mining models, have been identified. The first is the statistical approach that deals with the modelling of spatial autocorrelation. The second approach involves the direct modelling of spatial relationships such as topological and distance relationships. The latter was selected as being the most appropriate approach to model the spatial dependencies within the chosen data mining technique.

3 **Property Valuation**

In the previous chapter, the crucial role of background knowledge of the application area as guidance to the whole process of knowledge discovery was underlined. In this chapter, a detailed account of the property valuation area is given. In the first part some general property related concepts are presented. This is followed by a reference to the main factors that affect the property prices. After introducing these general aspects, in the following section the common property valuation methods and techniques are summarised. The following section presents the theoretical background that formed the basis for the incorporation of the spatial element into the valuation modelling. In Section 3.6, the most prevalent techniques are being revisited by examining the location element and how it is being handled in the literature. The final section examines the way GIS technology can contribute in the modelling of location and concludes with the presentation of five examples of its application in property valuation research.

3.1 Property Market

The property market consists of a number of different submarkets in which different operations are taking place. Submarkets can be defined as sub-areas within the broader market area that stand out in some important way (Thrall, 2002). A general rule is that members of each submarket should share similar characteristics and therefore have a large degree of similarity. Accordingly, members of different submarkets should differ at least on the segmentation criteria. Following this, the physical definition of a submarket is accomplished in such way that minimises the variation amongst the members.

Examples of submarkets include submarkets based on the type of landuse such as residential and industrial or submarkets based on geographical location. It is apparent that the smaller the sub-market is, the greater the similarity - although extremely specialised criteria may lead to an inadequately small number of members that cannot support valid analyses. Therefore identification of submarkets is heavily dependant on the type and scale of analysis.

Property markets are characterised by imperfection. Basic conditions required by economists to be satisfied so that a market to be characterised as perfect include: willingness to buy and sell (many buyers and sellers), perfect and complete information availability; and homogeneous product (Evans, 2004). In that aspect, property market differs from many other markets (e.g. a stock exchange market) in a number of ways that stem from the unique features of property.

Property has unique characteristics that reflect on the complexities associated with property-related decisions. These characteristics have been depicted by a number of researchers as follows (Stapleton, 1989; Anselin, 1998; Meen, 2001):

Heterogeneity: A property can be of several types (residential, commercial) and also have a number of operations associated with it (e.g. personal accommodation or letting).

Locational fixity: Properties are immobile and permanently fixed to a location. This introduces extra considerations in the form of external factors that need to be taken into account to adjust the price difference of otherwise identical properties.

Durability: Property stock is characterised by longevity. Its life-time cycle is far longer than other commodities.

Supply: Supply is associated with high cost, while the response rates to sudden changes in demand are quite low.

3.2 Property Value

In economic theory there are three criteria that a product or a factor of production must satisfy in order to have a cash value (Turner, 1977): it must have utility, it must

be capable of ownership and finally it must be limited in supply. It is apparent that a 'property' fully complies with all these requirements.

Two main directions can be distinguished in economic theory, related to the value formulation: the classical and the neo-classical. Classical economists such as Smith and Ricardo considered value in relation to labour. Smith (1776) refers to that aspect by stating that "*labour is the real measure of exchangeable value of all commodities*". Ricardo (1817), in accordance to the classical school, in his theory of explanation of the rent formulation, states that the exchange value of all commodities is not regulated by the less quantity of labour suffice for their production due to highly favourable circumstances but by the greater quantity of labour required for their production in less favourable circumstances. Another supporter of this theory that relates the value of a commodity to the labour was Karl Marx (Marx, 1867).

On the other hand, neoclassical economists reject this relation and argue that the value of a commodity is a measure of its desirability. Hence, the value is directly depended on market forces and the supply and demand mechanisms.

To make the abstract term 'value' more tangible, in the case of property value, several definitions have been proposed. It has to be noted that due to the high dimensionality of the property there is no one general definition that applies to every case. In their majority, definitions are case specific and in relation to the general functions of a property e.g. personal accommodation. Based on that, two main types of property values can be distinguished: market and rental value.

A number of definitions regarding the 'market value' of a property exist. Lawrance *et al.* (1971) define market value or price value of a particular interest in landed property as the amount of money a willing and able purchaser is giving to obtain that interest at a particular time. Moreover, Pagourtzi *et al.* (2003) define market value in relation to the assumptions that have been made in estimating the exchange price of a property if it were to be sold in the open market. Such assumptions include the nature of the legal interest, the physical condition of the property and also all the potential purchasers in the market.

What is apparent in both definitions is that value is subjective. In the second definition for example, the estimation is based on assumptions regarding a number of

factors. That alone introduces subjectivity in the valuation and strongly relates the accuracy of the valuation to the accuracy of these assumptions. Given that valid valuations assume accurate estimations of a market value, a commonly accepted definition of market value is necessary in order to ensure consistency.

The International Valuation Standards Committee (IVSC) set a conceptual framework and introduce the following definition of market price. *“Market value is the estimated amount for which a property (or Asset for more general cases) should exchange on the date of valuation between a willing buyer and a willing seller in an arm’s length transaction after proper marketing wherein the parties had each acted knowledgeably, prudently and without compulsion” (RICS, 2003).* To clarify the above definition a description of the conceptual framework of each element of the above definition is presented in Table 3-1.

On the other hand, rental value relates to another property characteristic that not only can be used for personal accommodation but also it can be a source of income. Following this, the rental value can be defined in relation to the expected income a property will produce.

Property value estimates are required for various purposes. Among the occasions that might require the estimation of the market value are (Mackmin, 1994): mortgage purposes, auction sales, compulsory purchases and tax purposes. Although market value based valuations are common practice among a number of jurisdictions, there are others that they adopt a rental value approach. An example of such an approach is the UK’s property local taxation system for businesses and non-domestic properties based on rating. Rating bases the assessment on an estimated rental value (Dale & McLaughlin, 1988).

Finally, apart from market and rental value there is another estimate that is commonly used within the context of valuation, that of the asking price. Mackmin (1994) defines asking price as the price suggested by a seller guided by an agent in order to stimulate the market and provide the basis for negotiations. Therefore asking price is often higher in comparison to property’s market value because it is subject to a seller’s personal interests. In cases that the valuation has been made by a qualified valuer, asking price and market value are equal (Mackmin, 1994). Cheshire &

Sheppard (1995) refer to the asking price as a good approximation to the market price in a stable market.

Term	Description
<i>'The estimated amount'</i>	Is the amount payable for the property expressed in the local currency. The Market Value, according to the standard, is the most probable price that can be obtained in the market at the date of valuation. This price is the best price that the seller can reasonable obtain and the most advantageous price the buyer can reasonable obtain. Inflated or deflated estimations due to special arrangements or circumstances are excluded from this estimate.
<i>'a property should exchange'</i>	Emphasises the fact that this is an estimate and not a predetermined or actual sale price.
<i>'on the date of valuation'</i>	Market Value is time-specific and subject to any market changes. Therefore this estimate reflects the market state at a given time. Consequently this definition assumes that there is simultaneous exchange and completion of the transaction.
<i>'between a willing buyer'</i>	Refers to a buyer who although motivated is not determined to buy at any price but to buy in accordance to the real state of the market and not higher than it requires.
<i>'a willing seller'</i>	Accordingly a willing seller is one who is motivated but not forced to sell at any price and neither prepared to wait for a price that is not reasonable in the current market. The willing seller agrees to sell a property at the best obtainable price after appropriate marketing.
<i>'in an arm's-length transaction '</i>	Refers to transactions that the involves parties have no particular or special relationship that may result to a price that doesn't reflect the market.
<i>'after proper marketing'</i>	Refers to the pre-valuation period where the property would be presented to the market in such way that can obtain the best price. The duration of this exposure may vary according to the market conditions but must be enough so that can attract sufficient number of potential buyers.
<i>'wherein the parties had each acted knowledgeably, prudently'</i>	Assumes that both parties act for self-interest and are informed about the characteristics and potential uses of the property and the state of the valuation on the date of the valuation. Prudence relates to the state of the market at the time of the valuation. A prudent seller or buyer acts based on the best market information available at the time.
<i>'and without compulsion'</i>	Refers to the fact that both parties are motivated and willing to go ahead with the transaction but not forced.

Table 3-1: RICS Valuation Framework (After RICS)

3.2.1 Property Value Determinants

The market value of a property reflects a range of physical, locational and neighbourhood factors (Longley *et al.*, 1996). These factors that affect the value of a property can be classified into two categories: external and internal (Goodall, 1977,

cited in RICS 1999; Dale & McLaughlin, 1989). Extending this classification, internal factors include physical attributes and legal factors that relate to the property while external factors include location, economic and socio-economic factors (Dale & McLaughlin, 1989; Wyatt & Ralphs, 2003). One of the first references to this distinction can be found in Marshall (1890) when he refers to the site and situation value of an industry and how industries' advantaged situations (e.g. proximity to road network, proximity to labour market) increase the value in the case of similar sites. Similarly, Wilkinson (1973) speaks about dwelling-specific and location-specific factors. Also Lawrance *et al.* (1971) refer to two main factors that affect the value of the residential properties: accommodation and situation. They further analyse the impact of location to the value of the property by classifying location factors into concrete and uncertain. Concrete factors can be considered time to travel to work and proximity to amenities, while an example of uncertain factor is fashion. A more detailed account of both the internal and external value determinants depicted in Table 3-2 follows.

Internal Factors	
Physical Factors	Legal Factors
Structure Condition Design / Character Facilities Topology / Geology	Leasehold Freehold
External Factors	
Location	Economy
Topology / Geology Proximity to Transportation / Amenities / Public Services / Non-residential Landuses General Infrastructure Environment (e.g. Pollution/Noise levels) Socio-Economic Profile (e.g. Crime levels, Deprivation)	Local-Central Government policies General Economy State

Table 3-2: Value Determinants

Physical attributes include information about the site and the building, in other words are attributes that describe the specific land parcel. Site characteristics such as shape, size along with site qualities like the presence of green space are important to be taken into account. Building related attributes that affect value involve design, accommodation, construction and current condition (Mackmin, 1994). Follain and

Jimenez (1985) classify physical characteristics into two categories: living space attributes and structural quality. Living space attributes include attributes such as size and number of rooms while the structural quality attributes refer mostly to quality measures.

The second set of internal factors relate to the legal status of the property. The type of legal title of a property is considered to be one of the strong determinants of property value, hence legal considerations are of first priority (Mackmin, 1994). Freehold and Leasehold are the two principal interests on land or buildings in England and Wales although there is also the option of a life interest in a particular property or parcel of land.

Freehold is the largest legal estate in land one can hold. The main characteristic of this title is perpetuity. Owners of such title have the right to occupy and use the land, transfer the title in whole or partially and finally to create interests such as periodic tenancies, leaseholds and interests. Leasehold, on the other hand is for a definite term of years, subject to the payment of an annual rent and to the covenants contained in the lease (Lawrence *et al.*, 1971). Traditionally, a lease holds for 99 or 999 years. In such cases, key role in the formation of the value play also the terms and conditions in the lease (Mackmin, 1994).

External factors are equally, if not more, important factors comparing to internal factors in the sense that they can affect the value of a property in numerous ways but because their influence is externally driven, there are no ways to avoid or modify them. In the case of economic factors, one of the most important issues involves the state of property market at the time of the transaction. Property market can be easily affected by actions of both the National and Local Government. Common governmental acts that have impact on the value usually relate to planning (zoning), public goods, environmental regulations and taxation policies (Thrall, 2002).

Another example of external influence is socio-economic factors that relate to neighbourhood quality aspects in terms of ethnicity, crime levels, culture etc. Property values in areas that are targeted as high-profile areas are higher when comparing to properties with similar physical and legal characteristics located in low-profile areas. The most commonly used source for the socio-economic

characteristics of an area is the Census. It offers measurements and descriptive characteristics of the population organised at several geographical levels. Variables that are considered important in an urban environment are population counts, race and income (Thrall, 2002). Although information about income is important it is not available in the UK Census. Alternate sources include lifestyle and geo-demographic datasets, such as Experian's Mosaic, that offer area profiling based on data mainly sourced from Census 2001 integrated with other customer related databases.

The impact of external factors such as proximity to non-residential landuses, in the case of residential property, can be grouped under the term externality. Externalities can have positive or negative effect on the property. Another characteristic is that the impact of externalities can vary in intensity and type as one moves to different submarkets (Thrall, 2002).

For example, the impact of a public open space in a heavily dense urban area is not the same as that of a similar park in a suburban area. In the first case, the park can be considered as a positive externality that will add value. In the second case, the park most probably will not contribute positively and depending on the residents it may even be perceived as a negative externality (noise).

It is apparent that location is a very important factor that affects the value of a property. As one of the main focuses of this research project is the modelling of location by using knowledge discovery, location issues are being further discussed in a separate section.

In general, internal factors are more easily captured and hence incorporated into the valuation models. This is not the case for the external factors which are not always tangible. Such an example is fashion. Fashion is an external factor that is quite difficult either to predict or describe in such a model. Another factor can be the personal or sentimental interest of a buyer on a specific property that may result to value estimates that are not realistic reflections of the market.

3.3 Property Valuation

The estimation of the market value of a property is called property valuation. Property valuation for a specific purpose is a non-trivial process since it involves the consideration of a variety of underlying factors of the market and the way these affect the value of the property at a given time. As discussed in Section 3.2.1, such factors may include governmental policies, geographical factors or even factors such as fashion or season.

Additional considerations stem from the fact that property valuation is case specific. Its correct exercise requires the a priori knowledge of the purpose that commences the valuation and also the type of the property (e.g. residential or commercial). Examples of different types of valuations include valuations for purchase and sale, transfer, tax assessment, expropriation, inheritance or estate settlement, investment and financing, insurance and property development (Pagourtzi *et al.*, 2003; Mackmin, 1994).

Residential Property Valuation

Residential appraisal involves the process of value estimation that is exercised on properties that are suitable for residence. According to Jenkins *et al.* (1998), residential valuation has received less attention compared to commercial appraisal from the scientific community. That mainly happened because researchers realised that the development of better and scientifically proven valuation methods that tackle the higher complexity in commercial valuation would add significant value to practitioners. This led to the concentration of the scientific community to addressing commercial markets. This complexity and the added value, primarily regarding the vendor of the property, in commercial valuations is also reflected in the higher fees charged by the commercial valuers compared to those charged in relation to residential valuations (Jenkins *et al.*, 1998).

However, there are complexities involved when dealing with residential appraisals as well. Apart from characteristics that are common amongst all type of properties (see Section 3.1) there are additional characteristics that originate from the nature of the housing. Housing is a commodity but unlike other commodities is a complex combination of provisions (Orford, 1999). It is a necessity and cannot be substituted.

Apart from a shelter provided by the dwelling itself, housing meets essential requirements for living (Jenkins *et al.*, 1998). Knox (1995, cited in Orford 1999) states that housing is the major determinant for protection, security, autonomy, comfort, well-being and status while the ownership of housing permits access to resources such as educational, medical, financial and leisure facilities. In addition, “.....it has various forms of value to the user and above all it is the point from which the user relates to every other aspect of the urban scene” (Harvey, 1972 (p16), cited in Orford, 1999).

3.3.1 Issues in Property Valuation

Property valuation as a non-trivial process not only involves the consideration of a variety of factors of the market but it is also performed by a variety of actors. What follows is an overview of issues related to data and to valuers that dictate the need for the development of new approaches that will benefit from new technologies that can potentially lead to more consistent valuations.

Data related issues

Information is considered to be ‘the hub of the wheel’ driving the property market. It is considered to be the fourth resource, along with land, labour and capital (McCluskey *et al.*, 1997). Despite the important role of information in the property industry there were several issues regarding the data especially those relating to information about the sales.

In the recent past, a major issue that had a direct impact on the quality of a valuation was data availability. Information related to sales or to the transaction itself was not usually publicly available. Institutions holding information on transactions such as the Land Registry and the Inland Revenue in England and Wales did not provide such information readily and when they did there was an associated cost with it.

Therefore, valuers usually relied upon their own source of information, which are not always reliable enough (Almond *et al.*, 1997). That also led to a secrecy practice from the data holders since access to the right information is what make the difference in property markets. Stapleton (1989) speaks for a ‘gaining advantage’ policy due to the imperfect nature of the property market.

Today, information is becoming more readily available. Various data related to property analysis can be accessed on the Internet. This has led the major vendors of property data to follow this trend and hence property information such as transactions is now available. The use of internet as a data distributor had also an impact on the pricing of these products which is declining. This was achieved by reducing the distribution costs, having access to larger markets and also competition (Thrall, 2001). Despite this breakthrough, legislation such as the Data Protection Act still limits the access to certain types of data. Such example is the access to the structural information of the property.

Another issue relates to the lack of standardised data that lead to limitations, uncertainties and errors (McCluskey, 1997). Example of such limitations is the Address-Point product of Ordnance Survey that is used for the geo-referencing of the properties. This dataset is based on the Postal Addressing File (PAF) of Royal Mail meaning that properties without a postal delivery point will not have spatial references even though they have rateable values (Vickers, 2003). This is common in cases of functions that spread across multiple buildings (e.g. Universities).

Limited access and lack of standardisation are not the only sources of inconsistencies. As discussed in Section 3.2.1, there are various special attributes that are linked to the transaction and have an impact on value. Their incorporation in the valuation involves a number of considerations such as their identification and the decision whether they should be included in an appraisal. This introduces another problem, which is the lack of a consensus among practitioners on which variables affect value (Almond *et al.*, 1997).

Apart from data availability / accessibility / quality issues, there were also other issues that related to the management of information. Ineffective gathering, storage and access of information affected the use of available data in the appraisal process. Today technological advances led the way to a number of changes into the property analysis and also to the information distribution. Information Technology is used to make the most of the available information. Nevertheless that was not always the case. Almond *et al.* (1997) have undertaken a survey involving major leading institutions which shows that there are indeed issues relating to IT. In this survey, 53% of those institutes responded to the questionnaire and their response led to the

following observations:

- absence of comparables databases or universal comparable databases
- limited use of IT for inspection and appraisal
- limited use of field computers
- limited use of statistical analysis
- no use of advanced techniques such as Artificial Intelligence is made

Valuer related issues

Data issues have a direct impact on the quality of the valuations by affecting the procedure followed by the valuers. In the case of applying the comparison method (see Section 3.4), the most experienced valuers in the absence of comparable data, draw on comparables from memory. However, if the selection process is based on memory it might affect the quality of the valuation in terms of completeness and accuracy. It is essential for valuers to possess the ability to perform skilled analysis and interpretation of the results. A GIS approach in valuation (see Section 3.7) would increase the efficiency by assisting the valuer enhance his or her skills (RICS, 1998).

Valuers could also be subject to biased valuations. Wolverton and Diaz (1996) and Gronow *et al.* (1996) both cited in Almond *et al.* (1997), based on their research in US and UK respectively, argue that revealing to the valuers the tentative sale price, agreed between the buyer and the seller, introduces bias into the selection of comparables and also to the resulting valuation. Furthermore, pressure from clients can also affect the valuers' estimation. Although researchers argue that is not the case with residential appraisers, practitioners in the UK suggest that client pressure exists (Almond *et al.*, 1997). Anon (1996, cited in Almond *et al.*, 1997) supports this view, by arguing that there is a common practice in the UK for valuers to estimate value based on the tentative price where the resulting value is within 10-15% of the tentative sale price.

The need to deal with these issues, the variety of the property stock and also the need to provide a commonly accepted framework for property appraisal led to the development of several methods and techniques that can be applied to the problem of property valuation. The following section provides an overview of the most prevalent methods and techniques.

3.4 Property Valuation Methods and Techniques

The most commonly applied methods to valuation can be broadly classified into two categories: traditional and advanced methods. Traditional methods include all the standard five main methods of valuation while advanced methods include techniques that mainly benefit from computational developments.

The five main standard recognised valuation methods are (Lawrance *et al.*, 1971; Scarret, 1991; Millington, 2001): Comparative Method (Comparison), Contractor's Method (Cost Method), Residual Method (Development Method), Profits Method (Accounts Method), Investment Method (Capitalization/Income Method). Advanced methods include techniques such as Hedonic Price Modelling, Artificial Neural Networks (ANN), Case-based Reasoning and Spatial analysis methods.

An overview of the most prevalent methods and techniques in respect to their applicability to property valuation follows.

3.4.1 Traditional valuation methods

Comparative method

The Comparative method, also known as Direct Capital Comparison (DCC) method, is the underpinning technique for all the other valuation approaches. It is the most widely used in practice, the most reliable (Turner, 1977) and it is used for sale, purchase and rental property valuations. It is mainly used for the appraisal of residential properties.

The market price in this case is based on recent transactions of comparable properties that are used as value indicators. The selling price of its comparable property must be then adjusted to account for the differences between the property under consideration and the comparables. Adjustments are based on differences in the properties' physical characteristics, neighbourhood profile, transaction date etc. The valued price is derived from the adjusted prices. Valid valuations using the comparative method depend heavily on the availability of correct, up to date and complete transactional data (Millington, 2001).

Contractor's method (or Cost Method)

This method is usually applied when there is no evidence of a similar property to that under consideration, 'changing hands' in the open market. That is, the case of very specialised properties where evidence of comparable sales is hard to find. Therefore, this method can be applied for valuations involving properties strongly linked to the business that is carried out in the property. Examples of such properties include hospitals, schools, specialized factories etc.

However, the valuer must consider the contribution of the building in the whole business when determines the market value by reference to its replacement costs. For example, machineries of a factory might contribute more to the value of a business than the building itself (Pagourtzi *et al.*, 2003). The main assumption of this approach is that the value of the property equals the cost of the premises.

This valuation approach is also common in non-investment markets. That is in countries where property investment is not a common practice and owner-occupation is the dominant property utilisation. However, when the occupational market is dominant by renting companies and there is a scarcity in the market, the price will be determined by the supply and demand characteristics of the market instead (Pagourtzi *et al.*, 2003). In the UK this method is used in rating and occasionally in cases of compulsory purchases (Turner, 1977).

Residual Method (or Development Method)

This method is used in cases where there is a potential of higher income from property improvement, alteration or redevelopment. Given that this method considers development costs it can also be used to assess the price for plots or sites that can be developed. Because the value is related to the level of profitability of the development or the improvement of the land or the property (Scarrett, 1991), this method is also suitable for the valuation of residential properties purchased for renovation purposes.

With the residual method the valuer determines the Present (capital) Value of the estimated future income deducting all the required costs to transform the property in a particular form that will command the estimated price. Also in most cases there is

an allowance for risks related to the speculative nature of the estimated increased income. Costs that should be deducted from the estimated value can include: demolition costs, development costs, professional fees and finally developer's profits. To apply this method effectively one should base all calculations upon the maximum utilisation of the property. The residual method heavily depends on the valuer's judgement when they consider the factors that affect the value of the property (Lawrance, 1971; Scarrett, 1991; Pagourtzi *et al.*, 2003).

Profits Method (or Accounts Method)

This profit-based method is used to assess the value of a property when it is heavily linked to the business that is carried out in that property. That is, the capital value is estimated in relation to the volume of the trade or business carried out in that property (Scarrett, 1991). The profits method, also called the accounts method, is not very direct and usually is applied to the valuation of special types of property. Examples of such properties include hotels, public houses, cinemas and theatres.

It is a two-step process (Lawrance, 1971). First it involves the estimation of the gross earnings or receipts and then it deducts all related costs. Such costs can include working expenses, interest upon the capital and an amount for remuneration to the tenant for tenant's risks and enterprise. The balance is the expected rent to be paid. The estimated value of the rent is then capitalised at a Year's Purchase (YP). This value is based on sale analysis of other similar properties.

Investment Method (or Capitalization/Income Method)

This method is applied in cases where the property under valuation is considered as an investment. The objective of this method is the estimation of the capital value of a future income discounted at an appropriate rate of interest. Therefore, this method is based on the knowledge or the ability to estimate, first the income the property will produce and then the interest that will be discounted. The easiest case of this method is when a comparable model is applied resulting to a direct capital value estimation (Lawrance, 1971; Pagourtzi *et al.*, 2003).

Investment method is commonly used for commercial property valuations although it is also applicable to residential valuations purchased for an investment. According to Turner (1977) this method is a five-step process:

1. Find evidence of recent transactions of similar interests in land.
2. Analyse the evidence to find the appropriate rate of interest.
3. Convert the interest rate to a figure of years' purchase (YP).
4. Determine the net income derived from the interest in question.
5. Multiply the net income by the figure of YP to arrive at the capital value.

An alternative approach to the traditional investment valuation method is using discounted cash flow approaches (DCF) instead of the YP (Isaak and Steley, 2000).

The successful application of direct capital comparison depends on the degree of heterogeneity in the market. To cope with markets that are characterised by a high degree of heterogeneity a comparison is made between the returns attained by rental or the outright sale of the property. Distinction between rental and yield reflects the interaction between the occupational and investment sub-markets (Lawrance, 1971; Pagourtzi *et al.*, 2003).

3.4.2 Advanced techniques

Several techniques have been applied to the problem of property valuation. One can broadly distinguish between the statistical approach (regression) and those that belong to heuristics and Artificial Intelligence (e.g. CBR, ANN). Both regression based techniques and ANN can be considered as data mining techniques (see Section 2.1.3).

Figure 3-1 gives a high-level description of the three most prevalent approaches to property valuation. These modelling techniques have been widely adopted by the property valuation community in the appraisal of the residential property. As the comparative methodology has been described in a previous section what follows is a discussion about the remaining most prevalent practices and the way they have been employed in the residential property modelling.

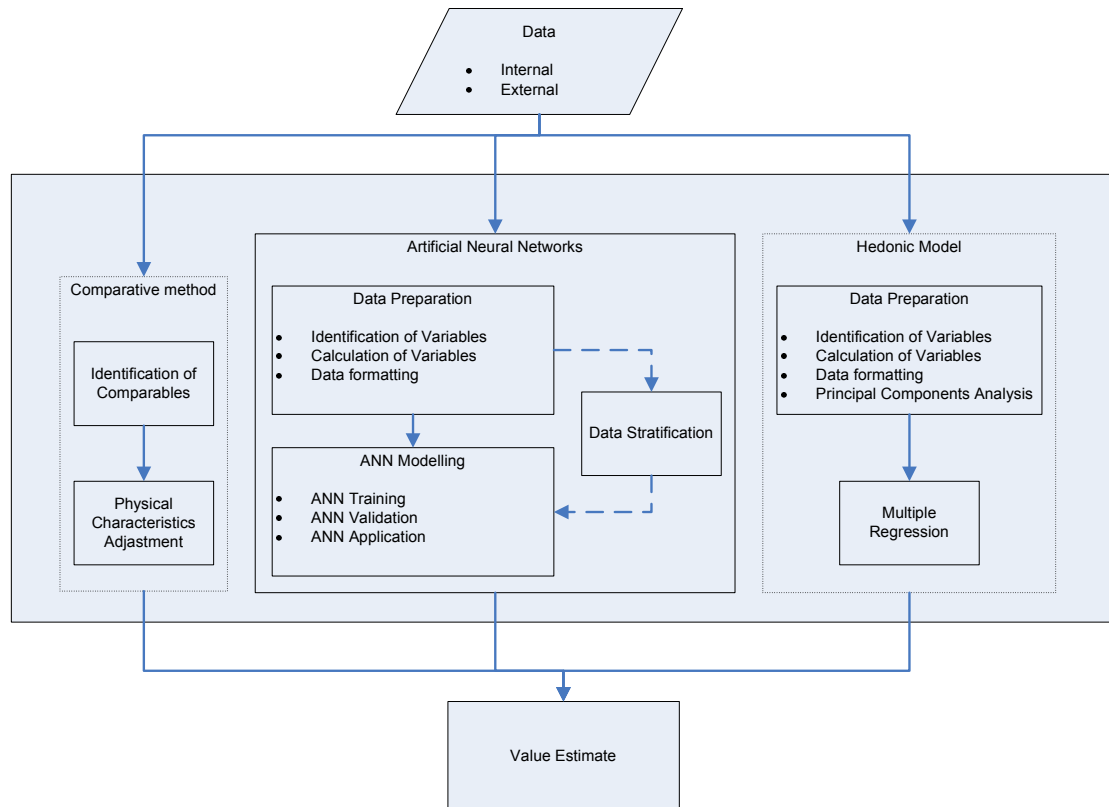


Figure 3-1: Property valuation approaches

Artificial Neural Networks and Property Valuation (ANN)

Artificial Neural Networks (ANN), first introduced in late 1980's, belong to the wider group of Artificial intelligence and were designed to simulate the operation of the human brain. ANNs were named after the network of nerve cells in the brain and are often referred in literature as neurocomputers or connectionist networks.

The basic elements of ANNs are called neurons or nodes. The connections, so called 'intelligence of the network', between the nodes are determined by the application of weights. Each neuron typically sums the weighted signals of each connection (or synapse). The result is the new signal of this node which is transmitted to the node or nodes this node is connected to. The combined result of these elementary (or unit) operations, leads to advanced functions such as: learning, induction and pattern recognition.

ANN classifications can be made based on the type of the network architecture employed or on the nature of the learning process. Although, the architectural structure of the network is strongly related to the learning algorithm employed,

hence, the learning process. According to the former classification an ANN can be: feedforward (Single or Multi-layer) and feedback or recurrent networks. According to the type of learning can be either a supervised or an unsupervised ANN (Haykin, 1998).

Feedforward networks, in their simplest form, consist of at least two layers, the input layer and the output layer. In the case of the multilayer feedforward networks between the input and the output layer one or more hidden layers intervene. On the other hand recurrent networks may consist of only one layer, the input layer. The main difference between these two types of architecture is that while the first type is strictly feedforward the second type allows the existence of feedback loops (Haykin, 1998).

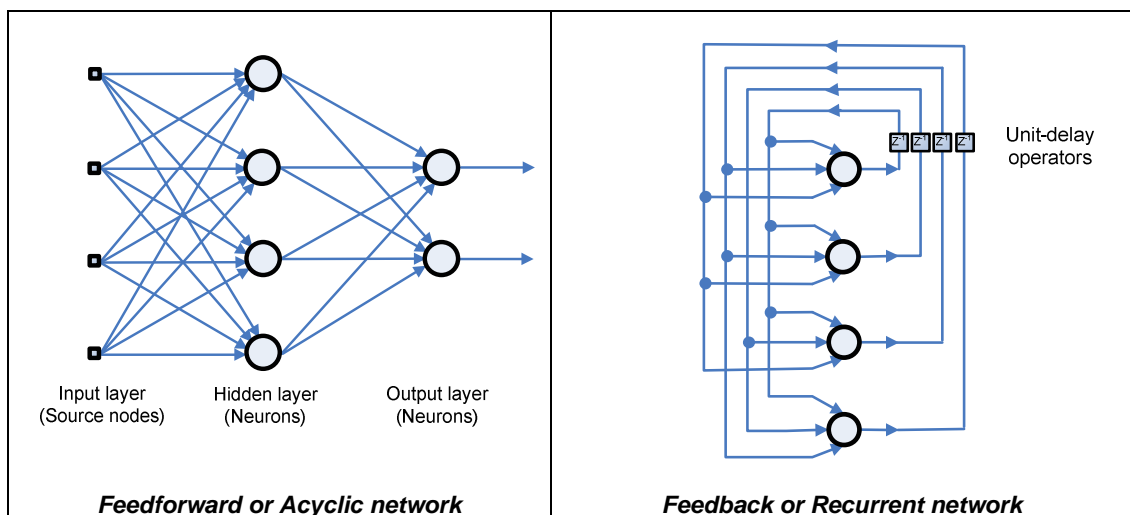


Table 3-3: Examples of feedforward and recurrent networks

(After: Haykin, 1998)

Examples of ANN methods include the multilayer-perceptron network (MLP), also called the backpropagation algorithm and the self-organising map (SOM). The backpropagation algorithm, introduced by Rumelhart *et al.* (1986) is considered the most popular method for the training of multilayer perceptions (Haykin, 1998). Its architecture is a feed forward network and it is based on the principle of supervised learning process. On the other hand, SOM introduced by Kohonen (1982), is the mapping result from high dimensional data space onto a one or two-dimensional lattice structure. Its architecture is a competitive network and it is based on the principle of unsupervised learning. Although Kohonen's model was not the first self-

organised model proposed, it was this model that enjoyed the most attention in the literature and has been widely applied becoming a benchmark in its field.

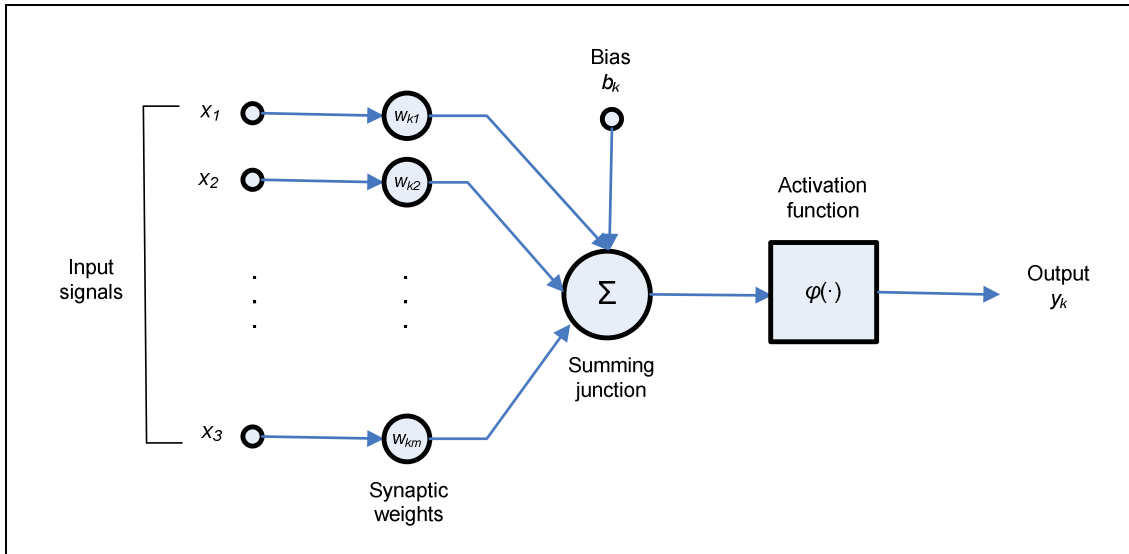


Table 3-4: Example model of a neuron

(After: Haykin, 1998)

The most typical weighted summation function used in a feedforward/feedback propagation neural network model can be mathematically expressed as (Haykin, 1998):

$$u_k = \sum_{j=1}^m w_{kj} x_j$$

and

$$y_k = \varphi(u_k + b_k)$$

where x_j are the input values, w_{ij} are the synaptic weights of the neuron assigned to the input values, u_k is the linear combiner output due to the input values, b_k is the bias and $\varphi(\cdot)$ is the activation function.

The activation function links the transformation values to the output variable values or y_k . There are several forms such an activation function can take. Examples of such functions include (Haykin, 1998): linear functions, linear threshold functions, step linear functions, piece-wise linear functions and sigmoid functions or Gaussian functions.

As discussed previously, roughly the main components of an artificial neural network model are: the input data layer, the hidden layer(s) and the output layer. As shown in Figure 3-1, ANN modelling process can be broken down into three steps: training, validation and application. In the case of property valuation an initial training data set is required in order for the model to give the estimation of the prices of new properties from the same market. More specifically, the input data may include property attributes such as number of bathrooms, parcel size and age of house. In the case of location aware modelling it can also incorporate locational variables (e.g. distance). All these variables have been identified, calculated and put into the appropriate format in the data preparation step (see Figure 3-1). The output layer includes the property prices. Finally, the hidden layer includes two processes that link the values from the input data to the output measures. These processes involve the application of weights through weight functions and the activation functions.

The main drawback of ANNs is that they are characterised by lack of transparency. This is also known as the *black box* problem and has been stressed in the literature (McCluskey & Anand, 1999). Jenkins *et al.* (1998) also comment on that, and present this lack of transparency as a challenge that cannot be avoided. According to Gopal *et al.* (2001) there are three main ways to provide valuable insights into the way ANN behaves: Visualisation, Rule Extraction and Statistical Methods. They also suggest a set of visualisation tools implemented in MATLAB for the interpretation of the ARTMAP Neural Network dynamics and statistics. Other issues include model scalability (McCluskey & Anand, 1999) and subjectivity (weights, variable selection, scaling) (Almont *et al.*, 1997; Carlson, 2002; E., McCluskey & Anand, 1999).

Although such alternative approaches present drawbacks mainly related to the black box problem, they also offer solutions to problems that traditional approaches fail to cope with. Kauko (2003) highlights that the advantages of the alternative modelling approaches against the traditional are in the way they deal with the notions of multiple equilibrium, fuzziness, non-linearity and residual price effects.

Hedonic price modelling

Hedonic price modelling is an econometric technique that is used to analyse complex commodities whose individual attributes do not have observable market prices. It has its theoretic basis on the hedonic hypothesis where the value of goods is attributed in relation to their ‘utility-bearing’ attributes or characteristics (Rosen, 1974). Accordingly, although individual attributes do not reflect an observable market price the sum of their values is equivalent to the market price of the commodity (Orford, 1999). In this context, a property price can be considered as the sum of the price of each attribute that comprise the property. Attributes of that type include property structure, environmental quality, accessibility/proximity; neighbourhood amenities etc. (see Section 3.6).

In the hedonic model, the value of the property along with the factors that is pre-assumed that affect the property value are the dependent and independent variables in a regression-based equation and in each simplest form is presented in Equation 3-1 (Meen, 2001).

$$PH_{it} = \sum_{j=1}^k \beta_j X_{ijt} + v_{it} \quad \text{Equation 3-1}$$

where : i = property

t = time

PH_{it} = property price at given time t

X_{ijt} = vector of characteristics

B_j = implicit prices of the k characteristics

V_{it} = error term

A large number of studies on pricing involve hedonic price modelling for isolating the various value determinants. According to Kauko (2003) there are two main reasons for the use of hedonic modelling for such type of analysis. The first is the theoretical foundation on microeconomic theories that involves a mathematical rigour resulting to more ‘scientific’ analysis. It is a generalisation of the locational theories to include more exploratory variables (Meen, 2001). The second reason is that their foundations are quite straightforward to the end users.

Although these modelling techniques are preferable when a relationship between a price and various characteristics is desired they present certain limitations. Such limitations involve the consideration of aspects like outliers, non-linearity, spatial dependence, discontinuity and fuzziness into the appraisal (Kauko, 2003). Further limitations derive from their close link to theories and the requirement for a priori assumptions. Also there is loss of theoretical elegance and explicit predictions (Meen, 2001).

As discussed, both ANN and regression based hedonic models present certain properties that justify their application in property valuation. Therefore, regardless of certain weaknesses, both techniques have been extensively used in property valuation analyses. This popularity triggered a number of comparison studies. Most comparative studies evaluate the MRA which is the traditional approach to hedonic modelling and ANN on a technical basis such as performance and accuracy. On that basis, results acquired from MRA proved more consistent compared to those from ANN. ANNs were proved sensitive to model changes and also dependant on the software used (Worzala, 1995).

3.5 Location theory in property valuation

So far, the most prominent methods and techniques in the area of property valuation have been examined. No reference to the variables and to the reasoning process behind their identification has been made so far. Variables are included in the model in order to represent in the best possible way the value determinants (see Section 3.2.1). Hence, better representation of these factors results in more accurate and realistic models.

The selection process involves a good understanding of the value determinants. In the case of internal factors this is relatively straight foreword. An example could be the incorporation of the structure of the property in the model. This could be easily attributed through the number of rooms, information which can be easily gained and assessed. Variable selection becomes more complicated in the case of external factors. Location is one of the most important representative of external influence yet it is quite abstract fact that hinders the variable identification process.

Before going into a detailed account of the ways that location has been materialised and incorporated into valuation models it is necessary to examine the theories that offered the foundations for this. These theories belong to the broad area of Human Geography and have been seeking to explain the economic value of location (location theory). These theories developed mainly by economists, have their roots in the main classical and neo-classical economic theories.

One of the first economic theorists that laid the ground for the study and analysis of land values was Ricardo. Ricardo developed a theory that relates the formulation of the land value to the relative productivity of the site (Thrall, 2002). According to his theory, the most productive land has the highest value. The land value of a less productive land would be equal to that of the most productive land less the amount of investment required for that land to reach the higher productivity level.

Before proceeding to the presentation of the main location theories, it is necessary to refer to a number of assumptions that rule these models. Since these theories seek to explain the spatial arrangement of the different landuses in relation to the variable distance other parameters should be kept as constants. Although 'we cannot stop the world' (Lloyd & Dicken, 1972), by using a number of pre-defined simplifying assumptions this can be achieved.

These assumptions are referring to two main types: land surface characteristics and population characteristics. The basic assumption is the reference to an isotropic plain populated by economic men that act in a perfectly competitive market. A summary of these assumptions follows (Lloyd & Dicken, 1972):

Characteristics of the plain:

All activities take place on a completely flat and featureless plain, where all the physical characteristics are invariant with respect to direction (*Isotropic Plain*). In addition, all points are equally easily accessible by the free movement at any direction which is enabled by the absence of barriers.

The second assumption about the plain regards the transport conditions. There is a single uniform transport system enabling equal transport throughout the plain. The

transportation cost is dependant on the distance hence, proportional to it and given the isotropic plain, invariant to direction.

Finally, the quality of the land is evenly distributed and costs associated with raw materials are the same.

Characteristics of the population:

The population is evenly distributed across the plain. This results to identical density at every point on the plain. In addition, there are no variations within the population. All members are identical in terms of income, demands and tastes.

The last assumption defines the behaviour of the basic actors. According to it Producers and Consumers act in a perfectly rational way according to their knowledge hence they have the ability to behave in an ‘optimum’ fashion (*Homo Economicus*).

One of the most well known and influential landuse models is the von Thunen model developed in 1826, designed to analyse agricultural location patterns. It seeks to explain the variations of the farm product prices and also how these variations affect the agricultural land use. It assumes that a farmer will produce the commodity that will result to the maximisation of their net returns (Johnston *et al.*, 2000).

The model compares the relationships between the production cost, the market price and the distance from the market centre and is expressed in the form of Equation 3-2 (Dunn, 1954, cited in Johnston *et al.*, 2000):

$$L = E(p-a) - Efk \quad \text{Equation 3-2}$$

where: L = Land Rent

E = Yield per unit of land.

p = market price per unit of yield.

a = Average production costs per unit of yield.

k = Distance from market (in kilometres or miles).

f = Freight rate per unit of yield and unit of distance.

Figure 3-2 shows a graphical representation of this model (location-rent curves) and the arrangement of three example landuses in the form of *concentric rings* or *zones* (Lloyd & Dicken, 1972) according to the model.

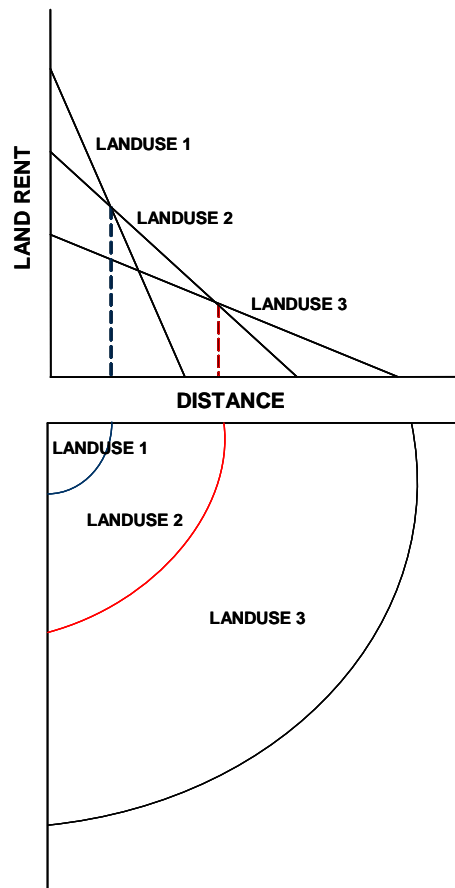


Figure 3-2: von Thunen model

There are certain limitations associated with the von Thunen model that relate mainly to the level of simplification in the underlying assumptions that result from the structure of society in the pre – industrialisation period when it was developed.

Alonso builds up on the von Thunen model resulting into a bid-rent theory and moves from the agricultural land use patterns to patterns of land use within the intra-urban environment such as residential and urban firms. The similarities between the urban landuse theory and agricultural location theory have been also identified by others (Isard, 1956, cited in Lloyd & Dicken, 1972). According to the model, residential land users bid for utility maximisation under certain constraints such as budget and profit for firms.

For the case of an individual with an income y , the basic model can be expressed in

the following equation (Alonso, 1965):

$$y = p_z z + P(t)q + k(t) \quad \text{Equation 3-3}$$

where: y = income

p_z = price of the composite good.

Z = quantity of the composite good;

$P(t)$ = price of land at distance t from the centre of the city;

q = quantity of land;

$k(t)$ = commuting costs to distance t ;

t = distance from the centre of the city.

Figure 3-3 presents the bid-rent curves and also the spatial patterns of the landuses within an intra-urban environment.

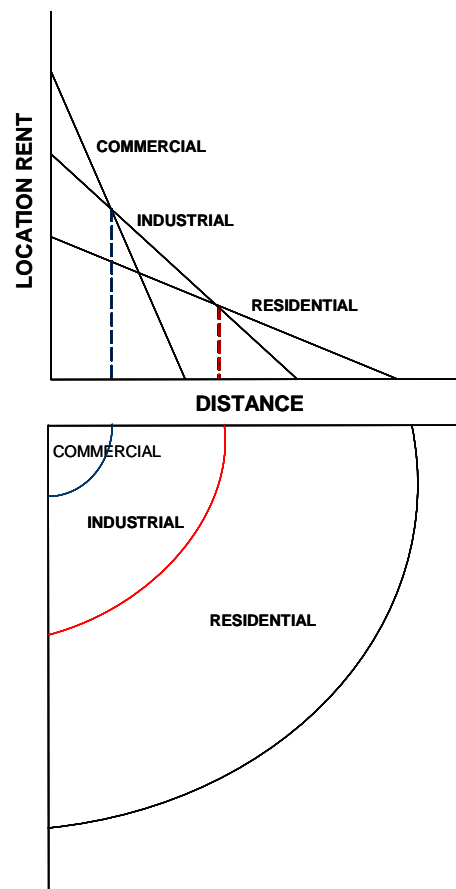


Figure 3-3: Alonso's Model

Key elements to this model are accessibility and its relation to transport costs (Johnston *et al.*, 2000). As shown in Figure 3-3 landuses are formed in a similar way to the Von Thunen model. The way that landuses form concentric circles around the city centre is a result based on the following relationships (Alonso, 1960, cited in Lloyd & Dicken, 1972):

- Land values are determined by the landuses after the users competitively bid with each other.
- Landuse distribution is based on the land values and the ability to pay which corresponds to the level of location rent of the particular product at a particular place.
- Central locations are captured by the steeper curves by depicting those landuses that benefit more by locating close to the centre.

Similar to the von Thunen model, Alonso's model is based on certain assumptions related to the land characteristics. More specifically, referring to the characteristics of the plain Alonso states: "*...it does not have such features as hills, low land, beautiful views, social cachet or pleasant breezes*" (Alonso, 1964, p. 17).

Another assumption relates to the centre of the plain. It is assumed that proximity to the centre of the plain is desirable from all the three examined land uses, agricultural, urban firm and residential, and therefore they all bid for a location close to centre.

Similarly to the concept of the different crops in the agricultural landuse model, empirical studies have been conducted that focused on the study of intra-residential structures. Examples include Wilson (2000): Burgess's ring structure (1925), Hoyt's sectoral structure (1939) and Harris and Ullman's multi-nuclei structure (1945).

Most of the attached criticism to these models relates to the simplifying assumptions by making its relevance to the real-world distant. Some of the key points are (RICS, 1999; Meen, 2001):

In the real world there are no areas that comply with the specifications of the isotropic plain. The existence of various geographical features result to a unique morphology that by no means can be considered consisted across the plain. This also

violates the assumption of the free and equally easy access – travelling to any location by the absence of physical obstacles. The assumption that these model make regarding the operation of a free market and perfectly informed market players, ‘homo economicus’ paradigm, is criticised as unrealistic. Hence, there is a distortion of the perfect market assumption. The final point related to the model assumptions that raises criticism is that of the presence of one central point that coincides with the location of all the employment. This mono-centric paradigm is not representative of today’s modern cities where usually more than one centres exist.

Another source of criticism relates to the fact that in these models there is no account of *neighbourhood quality*. Restriction only to distance relationships overlooks other important factors such as social and physical factors of the neighbourhood in the form of positive or negative externalities. Meen (2001) provides a reference to a number of studies that highlight the importance of neighbourhood quality that brings it second in importance after size.

Furthermore, changes in income, distribution and spatial pattern of the demand will cause a change in urban land values and the pattern of uses. Also, changes in transportation costs will have greater effect on the uses that heavily depend on transportation. These models also do not account for spatial interdependencies and ignore spillover effects. Finally it is difficult to test these theories due to the durability of property.

3.6 Location in Property Valuation Research - Techniques

The theories reviewed in the previous section offered the basis for the incorporation of locational variables in the valuation models. This section provides an overview of the way each modelling technique is handling location and relevant research in that field is presented.

Hedonic Price Modelling and Location

Early hedonic modelling research studies mainly involved only the incorporation of structural information about the property, implying continuity of their effect over space. Assuming this, leads to the overlooking of the fact that demand for specific

property characteristics such as garage space may also vary spatially (Thériault, 2003). This was identified and stressed by a number of researchers (Li & Brown, 1980; Cheshire & Sheppard, 1995; Orford, 1999) initialising a new phase in the hedonic modelling literature. Anselin (1998) also notes the absence of an explicit 'spatial' treatment of the property market in the empirical research and underlines the need to deal with issues like spatial autocorrelation. Cheshire & Sheppard's (1995) results also showed the importance of location.

This realisation resulted in the incorporation of several spatial variables in the model formulation. This is an area where the technological advances played a major role and contributed a lot in the improvement of the quality of the locational variables. In the early regression based empirical studies, accessibility was mainly included as a relation of the distance to the city centre. The measurements were rarely made at individual property level and in most of the cases were involving distances for a group of properties based on similar characteristics. Such characteristics included, similar distance from the Central Business District (CBD) (e.g. Evans, 1973), the use the administrative boundaries such as boroughs (Wabe, 1971), or neighbourhood (Wilkinson, 1973) who also used individual property level. An observation based on these empirical studies is that the dominant category of variables is related to the structure of the properties followed by the neighbourhood quality in the form of socioeconomic variables. Coarseness characterises the locational variables and also the sample in the studies is kept at a low number.

Through the years, advances in computer science led to the development of new technologies such as GIS, that enabled the move to more detailed and large scale studies. Recent studies are not restricted only to simple distance measures but also incorporate various other variables such as visibility measures and drive times. In general, the variables that are used in the majority of the location-aware hedonic models involve accessibility measures and neighbourhood quality measures (Follain & Jimenez, 1985).

The definition of location-aware hedonic models varies upon the different submarkets. Since these models require an a priori determination of the independent factors, the variables that will be incorporated are problem specific and unique. This

applies more to location – related variables although the will to pay for specific structural features can also vary geographically.

Accessibility measures directly relate to the location theories (see Section 3.5) where distance from the centre of the market is the dominant variable. Hence most of the hedonic modelling research has been based on the monocentric models such as Alonso's. In these models accessibility is addressed in relation to distance from the CBD. This monocentric model has been criticised as unrealistic and non-applicable in the case of modern cities although there are studies that suggest that the performance of monocentric models can perform well when location-specific variables are appropriately captured (Cheshire & Sheppard, 1995).

Recent research addressed the issue of multi-centric or polycentric models. Accessibility to the CBD cannot be considered as the only measure for valuing access to employment and consumption since households also value access to various other locations. For example access to Higher Educational Institutes or parks is also considered important. Dubin & Sung (1987) emphasised on the effect the existence of sub-centres have. Based on the findings of their study they concluded that the effect that employment and amenity centres, whether CDB or suburban, have on property prices is similarly limited and restricted within a 1 to 1.5 mile radius.

Accessibility measures include calculations of travel time, walking distance or straight line distance from the property to various locations depending on the model. Table 3-5 shows representative hedonic studies that included locational variables.

Neighbourhood quality is quite abstract and there is not a definite way of defining it and hence to justify the modelling by the use of the appropriate variables. However, there is a tendency since the early studies to materialise the neighbourhood quality through the use of three types of variables (Dubin & Sung, 1990): socio-economic status, quality of services and racial composition.

<i>Research</i>	<i>Case Study</i>	<i>Level</i>	<i>Modelling Technique</i>	<i>Internal Factors</i>	<i>External Factors</i>	<i>GIS</i>
Li & Brown (1980)	Area: Boston Suburban areas Sample size: 781	Property	Hedonic Model	Structural	Accessibility: Distance to: CBD, Distinct Landuses Neighbourhood Quality: Socioeconomic variables Environmental Quality: Measurements of: Views, Noise levels	
Dubin & Sung (1990)	Area: Baltimore Sample size: 486 Period: 1978	Property	J-tests	Structural	Accessibility: Distance to: CBD Neighbourhood Quality: Measures of Services: School quality, Police Protection (Crime levels)	
Cheshire & Sheppard (1995)	Area: Reading Sample size: 490 Period: 1984	Property	Hedonic Model	Structural	Accessibility: Distance to: Bus Network, Road Network(Classified) Neighbourhood Quality Socioeconomic factors, Services (Schools) Environmental quality: Local Topography, Landuses 1Km	
Lake <i>et al.</i> (1998)	Area: Glasgow Sample size: 4000 Period: 1986	Property	Hedonic Model	Structural	Accessibility measures: Distance to: CBD, Distinct Landuses Neighbourhood quality: Socioeconomic factors Environmental quality: Measures of: Noise level, Visual Impact	√
Orford (1999)	Area: Cardiff Sample size: 1500	Multi-level	Hedonic Model	Structural	Accessibility measures: Distance to: CBD, Distinct Landuses Neighbourhood quality:	√
Lake <i>et al.</i> (2000)	Area: Glasgow Sample size: 3456 Period: 1986	Property	Hedonic Model	Structural	Accessibility measures: Distance to: CBD, Distinct Landuses Neighbourhood quality: Socioeconomic factors Environmental quality: Measures of: Visual Impact of various landuses	√
Din <i>et al.</i> (2001)	Area: Geneva Sample size: 285 Period: 1978-1992	Property	Hedonic Model & ANN	Structural	Accessibility: Distance to: Individual landuses Neighbourhood quality: Socioeconomic variables Environmental quality: Measurements of: Views, Quietness	√
Thériault <i>et al.</i> (2003)	Area: Quebec Sample size: 4040 Period: 1990-1991	Property	Hedonic Model	Structural	Accessibility measures :Distance to: CBD, Services (Regional-Community levels) Neighbourhood quality: Socioeconomic Factors	√

Table 3-5: Location aware Hedonic Studies

Variables related to the *socio-economic status* are indicators of the socio-economic urban status. This for a neighbourhood can be defined in terms of unemployment rates or other characteristics of the households. Such indicators include the average level of qualifications and average income.

Provision of *high quality services* within the neighbourhood area gives a competitive advantage to a neighbourhood in relation to other neighbourhoods that are characterised by lower standards. Measures that are commonly used to capture the quality of the services relate to the local school quality and crime rates that reflect the level of police protection. School performance is usually measured by the proportion of the students that obtain more than five GCSEs, where crime rates are measured by the percentage of notifiable offences per resident population (Meen, 2001). Another type of measures relates to those used to describe the quality of the existent housing stock.

The final group of measurements to model neighbourhood quality relate to the *racial composition*. Incorporation of descriptive measurements about population in models is a common practice. Racial composition of the neighbourhood can be used as a surrogate measure for preferences since similar ethnic groups tend to cluster. The main source for these variables is the Census.

Although *environmental quality* can be considered as an aspect of the general quality of the neighbourhood, here is handled as a separate one. Amongst the most popular measures for environmental quality are pollution and noise levels measures. These measures can be incorporated either as direct measures based on measures held by environmental agencies or as proxies (e.g. proximity to highways).

Another group of environmental quality indicators include measures of the visual impact that the presence of distinct landscape features has on the property value (e.g. view to a beach or a river). Example studies include the impact of beach view (Pompe & Rinehart, 1995), ocean, mountain and lake view (Benson *et al.* 1998) and the impact of river view (McLeod, 1984).

Capturing in the form of variables and incorporating the whole location effect on the price is not trivial since parameters may vary upon location and market segments

(Adair *et al.*, 1996). That requires the modeller to deal with issues such as multicollinearity, autocorrelation and heteroscedasticity (Thériault *et al.*, 2003; Anselin, 1988; Orford, 1999) associated with the spatial nature of the property market.

A number of ways have been proposed to handle spatial dependency (Thériault *et al.*, 2003): considering wider range of spatial attributes, especially those related to environment, using information on the socio-economic status, improving measurements of interactions by adjusting trend surfaces over principal components and developing flexible ways to measure spatial dependency.

The latter leads to another main research direction that deals with issues that result from the special characteristics of the spatial data (see Section 2.2.2). It includes the study and development of a number of techniques that deal with spatial effects such as spatial dependence and spatial heterogeneity. As mentioned, it is impossible to identify and incorporate in the model all the spatial variables that may have an impact on the price of a property. This results in the spatial autocorrelation of the residuals due to the omitted variables.

Examples of research that deals with this include spatial linear models proposed by Anselin (1988), Pace *et al.* (2000), Dubin (1998) and Basu & Thibodeau (1998). Models of this type can be further classified into lattice and geostatistical models (Militino *et al.*, 2004). Models of the first type handle the spatial dependence in the form of a weights matrix, by modelling the spatial process as an autoregressive model in order to get the estimation of the covariance matrix of the error terms. Models of the second type are not so common in the property literature and they involve the direct estimation of the covariance matrix that represents the dependency between the errors.

One way to overcome the heterogeneity in a regression model is to explicitly model the 'parameter drift'. Models that account for spatial heterogeneity are: the spatial expansion model (Casetti, 1972) and the geographically weighted regression (GWR) (Fotheringham *et al.*, 1998). In Casetti's spatial expansion method the parameters in the regression model are expressed as explicit functions of locations. In such way, parameters for the variables (e.g. structural variables) are allowed to vary

spatially. GWR bases the parameter estimation on a location-based weighting function. An alternative to those methods is the use of a multi-level approach to model the effects at appropriate levels.

Artificial Intelligence Approaches and Location

Similarly to the regression-based applications, the first studies that facilitated ANN for property valuation were mainly based on structural data. Increasingly some implicit locational information was added in the models (McCluskey & Anand, 1999). Jenkins *et al.* (1998) provide a research example of application of ANNs to model residential appraisal. Their approach initially involved the uncovering of the property sub markets, followed by the modeling of each one of them separately. For the completion of the first phase they used a ‘Self Organising Map’ approach while for the second phase they used Multilevel Perceptron (MLP) Networks. In order to further refine their model they also used Census data.

In the case of the Self-Organising Maps (Cohonen Maps) two main trends can be identified. The first is their use in conjunction with another technique (e.g. ANN) for unveiling the spatial segmentation of the housing market. The second is their use as stand-alone tools either for the identification of the comparables based on a number of exploratory variables or to measure the impact of certain externalities used as exploratory variables. Table 3-6 shows a number of relevant studies

Research	Case Study	Level	Modelling Technique	Internal Factors	External Factors	GIS
McCluskey & Anand (1999)	Area: N. Ireland Sample size: 412 Period: 1995-1997	Property	ANN (hybrid)	Structural(8)	Ward	
Carlson E (2002)	Area: Helsinki Sample size: 4750 Period: 1985-1998	Property	SOM	Structural (6)	Environmental factors (Distance from roads, railways, HVT lines)	√
Kauko (2002)	Area: Jyväskylä Järvenpää Period: 1993-1997	Property	SOM	Structural()	Distance from power lines	
Kauko (2002)	Area: Helsinki Period: 1993-1994	Property	SOM	Structural ()	Neighbourhood Quality (Socio-Economic Factors, Amenities, Public Services etc.)	

Table 3-6: Location aware ANN studies

Spatial Statistics

An alternative approach that makes use of spatial statistics is the employment of surface modeling as an analysis tool. For the generation of the surfaces, spatial interpolation is employed. This procedure involves the estimation of the values at unknown locations within an area covered by existing discrete observations. There are a number of techniques that perform interpolation. One can broadly classify them to those that base the estimation on sample control points (local interpolation) and to those that use the whole population of control points (global interpolation) (Wang, 2006). Amongst such techniques are: the trend surface analysis, geographically weighted regression (GWR), kriging and inverse distance weighting (IDW).

An example application of such an approach is given by McCluskey *et al.* (2000). They based their analysis on the building of three types of models. The first included a MRA model only on regressors referring to the physical characteristics of the property. The second employed an interpolation technique on sample points (includes location). The third was a hybrid approach where for the surface building the MRA residuals expressed in percentage term were used. All models performed relatively well but the best performing was the hybrid model. Other studies of that type are shown in Table 3-7.

<i>Research</i>	<i>Case Study</i>	<i>Level</i>	<i>Modelling Technique</i>	<i>Internal Factors</i>	<i>External Factors</i>	<i>GIS</i>
Gallimore <i>et al.</i> (1996)	Stafford (218) (1992-1993)	Property	MRA (Value response Surface)	Structural	-	√
Deddis <i>et al.</i> (2002)	Londonderry (650) (1998-1999)	Property	MRA	Structural	Submarkets (Value Response Surface)	√

Table 3-7: Spatial Statistics Studies

3.7 GIS in Property Valuation

In order to be consistent, valuation techniques have to rely on the analysis of diverse data. Data may also vary in terms of format, type and volume. Adoption of a GIS-based approach in property valuation presents a number of advantages. Among the strongest are its analytical capabilities, the visualisation and finally the ability to

integrate data from a wide range of sources. To emphasise on the importance of GIS technology within the valuation process, in terms of locational analysis, the Appraisal Institute (1992, cited in Wyatt, 1995) compares the importance and projected benefits of the application of GIS to valuation with those of the computerised discounted cash flow modelling to financial analysis. Table 3-8 gives a summary of the GIS use in the property-related area in relation to the main elements of a typical GIS (see Section 2.2.1).

<i>GIS Element</i>	<i>Functions</i>	<i>Example Applications</i>
<i>Database</i>	<i>Data Modelling</i>	<i>Property Database (Geo-reference)</i>
<i>Data Processing</i>	<i>Geometric Algorithms</i>	<i>Measure impact of location based on distances from significant features (Distance)</i>
	<i>Topological Algorithms</i>	<i>Identify comparables (Point-in-Polygon)</i>
	<i>Data Conversion Algorithms</i>	<i>Data integration</i>
	<i>Network-based Algorithms</i>	<i>Measure impact of location based on road-network distances from significant features (Network Analysis)</i>
	<i>Statistical Algorithms</i>	<i>Data exploration Pattern analysis Autocorrelation measures</i>
<i>Data Sharing</i>	<i>Interoperability</i>	<i>Web-based Applications</i>
<i>Data Presentation</i>	<i>Base Maps</i>	<i>Presentation-Justification</i>
	<i>Visualisation</i>	<i>Identify Comparables (Visual Analysis)</i>
<i>Spatiotemporal element</i>	<i>Temporal Information Systems</i>	<i>Temporal Property Transactions</i>

Table 3-8: GIS use in Property Related Research

GIS technology also enabled the moving from small scale studies to large scale studies by facilitating the easy generation of variables which was the main difficulty in the past - generate variables for a large number of properties (see Section 3.6). Apart from the self-evident use of a GIS for the calculation of locational variables

there is a number of ways where a GIS, given an appropriate dataset, can also contribute in the calculation of structural information (e.g. floor space). This ability, although not as precise as the information coming from on-site inspection, can prove important in the absence of other data. Examples include the use of OS Land-Line dataset to calculate floor space and determine the type of property (Lake *et al.*, 1998).

Considering the advantages of GIS technologies, a number of applications relevant to property valuation have been developed. A large application area where GIS can significantly contribute and has been widely recognised is that of Mass Appraisal Valuations. Additionally, a number of researchers have applied GIS to property valuation using various aspects of it.

A review of research in the application of GIS in property valuation follows. This is an indication of the alternative ways GIS technology has been applied in property valuation research.

Predictive role of GIS (Longley *et al.*, 1993, 1994)

Longley *et al.* (1993) developed a methodology for GIS-based predictions of the capital value of property for the Inner Area of Cardiff. This involved the development of a street-based GIS, which was employed in the modelling of capital values. Additionally, they appraised their model by carrying out a comparison between the predicted values and those of the official valuations (Longley *et al.*, 1994).

Their methodology initially involved the conduct of a survey of asking prices of property and then the capital value prediction. The survey involved the asking price of all the properties that were on sale during December 1991. This resulted in 796 properties, which represented the 2.1% of the properties within the Inner Area. The sampled properties they selected were deemed evenly distributed across the Inner Area.

The other source of data used for the prediction of the capital values was the actual valuation data (April 1991). Those were the rateable property values. The house

prices in the 8-month period between the conduct of the valuation dataset was characterised as stable with a slight fall at the worst case.

The output of this study was a geographical model that enabled the assignment of the capital value of the properties that were not on sale based on the assumption that the asking price indicates the capital value. Finally, the assignment of the capital value was based on rateable values, dwelling type, the House Condition Survey area and aggregate regression relationships between asking prices and capital values. Figure 3-4 illustrates the assignment process for the modelled capital values.

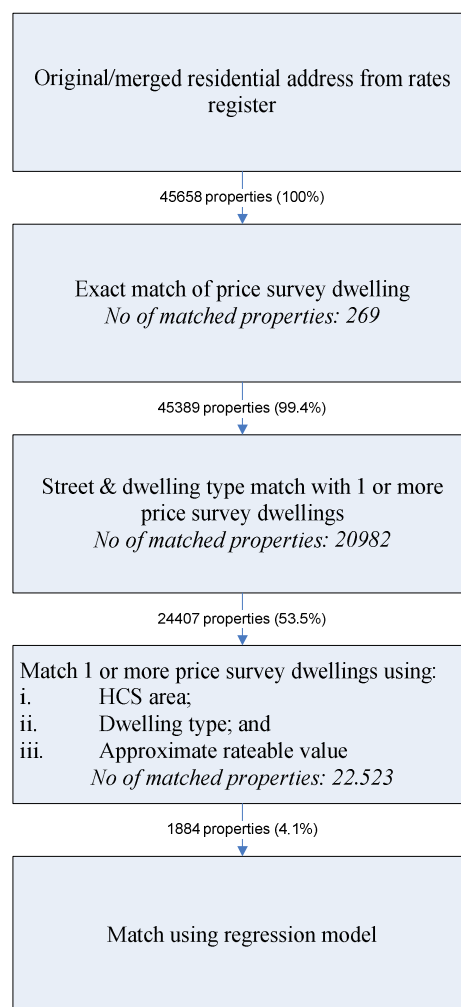


Figure 3-4: Assignment process for modelled capital values

(After: Longley *et al.*, 1994)

Longley *et al.* (1994) proceeded to the evaluation of their model and they highlighted the ways in which a GIS approach can contribute in monitoring the different local taxation regimes. According to their evaluation, in nine of the twelve communities

the percentages of the cases that presented zero mismatches between the observed and the predicted values exceed the others. The cases that presented mismatches were mainly in communities with large degree of heterogeneity in dwelling types and house prices. The GIS software package used in this study was ESRI's ArcInfo.

3D Value Surfaces for Location Modelling (Gallimore et al., 1996)

Gallimore *et al.* (1996) investigated the use of combined multiple regression analysis (MRA) and 3D response surface modelling in residential property valuation. The study area was the town of Stafford USA and the data used in the implementation of their methodology was extracted from mortgage valuation reports. The final dataset consisted of 218 properties and physical and structural information has been recorded. There was also information about the actual selling price where it was available.

Their methodology involved two stages. The first was the development of an MRA model, which enabled the prediction of the residential property value if that was going to be on sale. This model was not location aware therefore the second stage of their study involved the generation of a 3D location value-response surface. The main assumption in this approach was that the variance between the actual selling price and the predicted from the MRA model price reflects the influence of location. The value-response surface was generated from an interpolation grid, which was modelled to reflect the influence of location on each property. Finally, the input from the surface was then used for the MRA model refinement.

For statistical analysis the SPSS-PC software was used while for the value-response surface generation the GIS software IDRISI, developed by Clark Labs, was employed.

Accessibility Index (Wyatt, 1996)

The aims of this research that has been carried out by the University of Brighton were twofold. The first was the development of a spatial property information system that would demonstrate the potential benefits the development of a National Land Information System could offer to property valuation. The second was to develop a

more explicit approach to modelling the spatial influences on commercial property value.

In this study, the physical analysis of the property took place through a user interface. The user inputs the general structural profile of the property based on which, a set of comparables are being selected by the system. This selection is refined using another set of more detailed information entered by the user and then the system adjusts the value factors between the property in question and the comparables. Then, the final set of comparables is selected. By the completion of this stage the physical differences of the comparables are adjusted and therefore any variations in value can be explained by differences in location (i.e. locational values).

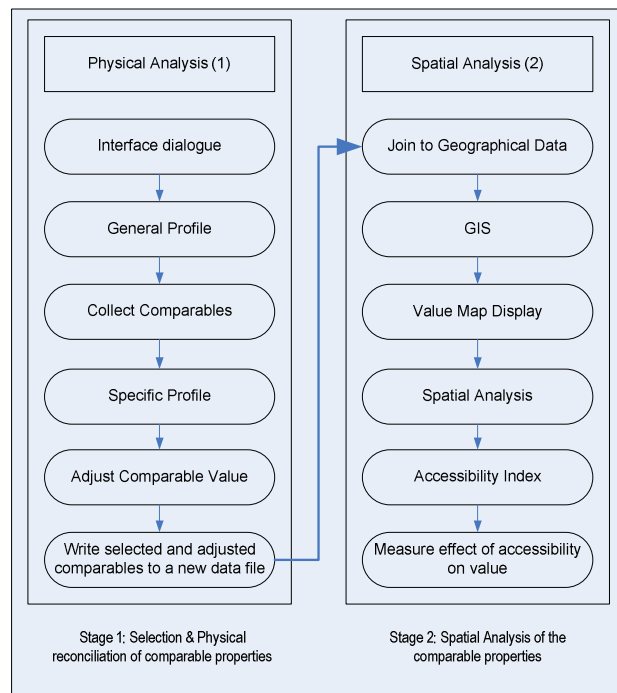


Figure 3-5: Methodology Overview

(Source: Wyatt, 1995)

The second part of the work involved the spatial analysis of the selected comparable values. That resulted to an accessibility index at an intra-urban level using a gravity modelling technique on a GIS platform. The basic assumption here is that the point of maximum accessibility is not the centre of the urban area as traditional urban theories do. The use of an accessibility index to measure the effects of location on value is also supported by the Appraisal Institute (1992, cited in Wyatt, 1995) when they argued that the quality of a property's location could be quantified by

calculating the time-distance relationships or linkages between the property and all the possible destinations. This analysis resulted to index values that were highly correlated with the locational values identified in stage one.

The GIS software platform that was used for the implementation of this study was ESRI's ArcInfo. Figure 6-3 is an illustration of the proposed methodology.

Structural Variables

As mentioned in previous section (see Section 3.6), when variable calculation is required, the role of the GIS is usually restricted in the calculation of locational measures such as distance measures. Lake *et al.* (2000) present an example of how GIS can be used in the materialisation of structural property data. Their approach involved the extraction of structural measures such as ground floor and plot areas and property type using the OS Land-Line.Plus dataset.

Value maps

Value maps is another application area that can benefit significantly from GIS. Value maps show the geographical variations of property values or land values and their applicability varies from planning and development to taxation and valuation (Wyatt & Ralphs, 2003; Vickers & Thurstain-Goodwin, 2002).

When first appeared they were paper maps but with the advent of GIS nowadays the digital format is most common. GIS provide all the necessary technologies for the creation of such maps, from input data and calculations such as spatial interpolation to visualisation and further analysis. Vickers & Thurstain-Goodwin (2002) present a number of potentials from the use of land value maps in assisting monitoring, planning and assessment.

3.8 Location aware property valuation in a knowledge discovery setting

Previous sections discussed a number of complexities in connection to residential property valuation modelling. These mainly arise from the complex structure of housing that is composed not only of structural characteristics such as number of

rooms but also of locational. Valuation involves the translation of these housing attributes into monetary value. That assumes knowledge of the impact these attributes have on price.

In the case of the structural characteristics this is relatively straightforward and this is reflected in a number of techniques that use these. On the other hand, identification and quantification of the spatial influences on property price is not trivial and involves geographical analysis to discover complex spatial configurations. As a result, most of the valuation models account for the structural characteristics and handle location indirectly. Considering that location is a primary influence on property value there is a need for adopting spatial methodologies that tackle this.

As discussed earlier in this chapter, two are the main types of valuation model that location is taken explicitly into account. The first is the hedonic regression model and the second is the Artificial Neural Network model. Both rely on a deductive approach which uses a set of assumptions on which predictions are based. These assumptions relate to the way location affects property prices and have their roots in early locational theories or have been proved empirically. These assumptions are captured by the variables in these models that eventually predict property values. The predictions are tested by comparing with real known transactions.

Initial studies have explored how location affects the value of properties in terms of accessibility. Based on these studies, theories that try to explain these effects have been developed. Although focus has been placed on the impact upon prices caused by accessibility to the city centre there are other types of relationships that can have positive or negative impact on property prices. Therefore these models are considered dated and fail to explain a price pattern within modern cities where the city centre does not necessarily coincide with the only centre of employment and consumption.

Proximity to non-residential landuses such as parks and schools is thought to increase the price of a property. On the other hand, proximity to industrial landuses is associated with low prices. It is apparent that locational factors, impact upon property prices not in an isolated manner but in a combined way and also that vary

upon case by case. Hence, there is a need for a model that accounts for these effects to determine a value to a property based on its location in an inductive manner.

Such approaches are usually rule based spatial classification algorithms such as algorithms that use decision trees. A general format involves first the identification of the spatial relationships and then the application of a rule induction algorithm. Decision trees are easy to explain. However the fact that decision-tree classifiers examine one variable at a time limits their ability to support satisfactorily a purely location oriented valuation approach. On the other hand, associative classification algorithms form classifiers that are explainable. These are based on classification rules that explore highly confident associations among multiple variables at a time.

Spatial configurations should be modelled in such way that facilitate the inductive approach and support the associative classification algorithm. An efficient way to do so is by using a graph theoretic approach to model the way locational externalities affect the property price. Graphs are used to analyze relations between units and have been traditionally used in geography to represent flows between locations. Nodes represent locations and edges represent flows between points, such as roads. In a property valuation case nodes can represent residential and non residential locations while edges can represent information about spatial relationships.

3.9 Summary

In this chapter the general concepts of property valuation and in particular of location –aware property valuation were presented. Property valuation is an estimate on the value of a property. It is a complex process, based on a variety of factors that affect in various ways the determination of a property's price. These factors can be broadly classified in internal and external. The most prominent representative of the external factors is location. Property valuation is also subject to the type of the property and the purpose of the valuation.

To tackle these issues and perform consistent valuations a number of methodologies and techniques have been proposed. These can be classified to traditional and advanced methods. Traditional methods include the Comparative Method, the

Contractor's Method, the Residual Method, the Profits Method and the Investment Method. Advanced techniques include the hedonic price modelling and the ANN.

Both hedonic modelling and ANN have been widely applied to property valuation research. In early studies, the focus was mainly on the structural data ignoring the effect of location on the price formulation. With the realisation of the importance of location in the price formulation, research work based on locational theories started to emerge. This trend was strengthened by the technological advances and the development of appropriate tools.

The incorporation of locational variables in the modelling is less straightforward when compared to the internal factors. A number of studies that dealt with this have been identified and the locational measures used, have been reported. Amongst the most prevalent variables are these that attempt to capture accessibility, neighbourhood quality and environmental quality.

As the models extended to include more complicated variables, the need for new technologies to assist in their materialisation became more apparent. Such technology is GIS which can assist in a number of ways in the property-related research. To demonstrate this, a number of representative studies have been reviewed. Finally, a link between location aware modelling in property valuation and knowledge discovery has been established.

4 **Design of a Property Valuation System**

So far, the existing data mining and spatial data mining techniques along with the importance of the adaptation of knowledge discovery techniques in geographical problem solving has been reviewed. In addition, an introduction to property valuation and detailed review of the way the locational influence is accounted in the valuation models have been provided. In this chapter, issues related to the design of the developed system that facilitates the proposed methodology are presented.

The present chapter is divided into three main parts. The first part gives an overview of the adopted methodology to meet the objectives of this research and the rationale behind its formulation. The second part presents the modelling and the design of the knowledge discovery algorithm. The third and final part covers all the aspects of the design phase of the whole system.

4.1 Research Opportunities

Although knowledge discovery in conventional databases is a well documented and recognised area, its application in geographical databases is a relatively new area and it has enjoyed a lot of attention from the academic community. Over the past decade, the academic community has identified Geographic Knowledge Discovery as an important, attractive, emerging and dynamic field that can prove to become a useful tool in the field of Geographic Information Science (Miller, 2004; Gahegan, 2001; Ester *et al.*, 1997; Fayyad *et al.*, 1996B; Koperski *et al.*, 1998A; Ester *et al.*, 1999).

Many researchers argue that it is an emerging area which can potentially lead to compelling results. Several areas that need elaboration have been identified (National Research Council, 2003; Turner, 2002; Miller & Han, 2001; Buttenfield *et al.*, 2001; Ester *et al.*, 2001; Openshaw *et al.*, 1999; Koperski *et al.*, 1998A). Among them, the successful integration of knowledge discovery and Geographic Information Science and the representation of the background or the extracted knowledge stand out as very promising areas (Koperski *et al.*, 1998A; Yuan *et al.*, 2001; Miller, 2004)

In this study, a knowledge discovery approach has been adopted to model the contribution of location to the value of a property. There is a common adage within the property industry that says there are three crucial factors that determine the success of a property: 'location, location and location' (Britton *et al.*, 1989, cited in Wyatt, 1995). It is true that in the literature location is considered as a major value determinant. Fraser (1993) argues that location is "*of dominant importance in understanding the demand for any urban property is its location, both in a regional and a local sense*". Gelfand *et al.* (2004) also talks about the "*axiomatic importance of location on selling price*" (p. 150).

Despite the wide recognition of the importance of location, it is considered to be the most neglected factor in valuation models (Kauko, 2003) and its incorporation is rather implicit and mainly based on a valuer's local knowledge (Wyatt & Ralphs, 2003; Wyatt, 1995). Studies are concentrated on identifying and quantifying the effects of physical attributes of a property on value rather than those of its location. (Wyatt, 1997). Therefore the transition from spaceless valuation models (Dubin *et al.*, 1999) to models that attempt to measure the impact of location on the value is necessary. Wyatt (1996) highlights this need by arguing that the development of a methodology that attempts to measure the impact of location on value is an important addition to valuation theory.

There have been some attempts to measure the impact of location on value. However, complexities related to modelling of location resulted that in most computer-based valuation models, the incorporation of the location becomes insignificant. These models are either structured in such a way that only homogeneous areas are considered or use oversimplified heuristic rules such as the distance from a city centre, which again diminishes the role of location.

Furthermore, computer-based valuation models (e.g. AVM) in their majority are statistical models. They base their estimation on various techniques such as indexation or comparables identification but mainly on regression. Their accuracy relies on the volume of data but although there are a variety of developed data models, valuation is based mainly on property characteristics. On the other hand, valuers factor in location intuitively based on their background knowledge of the market and intuition.

As shown, both the areas of geographical knowledge discovery and location-aware property valuation although highly active, there are still issues that remain to be investigated. In the next section, an overview of the proposed methodology is provided.

4.2 A new approach

In the proposed methodology, knowledge of the impact that existing locational features may or may not have on the property price is considered absent. It is a data-driven approach that is not tied to theories that attempt to explain the role of location in the property value therefore it does not require a priori assumptions about the variables. Since no a priori selection of the variables is required, the whole process relies on pattern discovery based on the topology of the area. The whole model is entirely location-oriented and aims to investigate the validity of such an approach.

The use of a knowledge discovery methodology will help in the extraction of this missing information and its incorporation in the valuation model by using an appropriate representation. The selection of this approach was based on two facts. The first is related to the successful application of knowledge discovery techniques to complex problem solving to reveal previously unknown information. Considering the complexities associated with location the application seems ideal. The second reason relates to the research challenges that pose the adoption of such an approach not in conventional databases but in spatial databases. The complexity of geographical phenomena (Gahegan, 2001) along with the large size of spatial datasets not only justifies the application of knowledge discovery to spatial datasets but they also make it highly attractive.

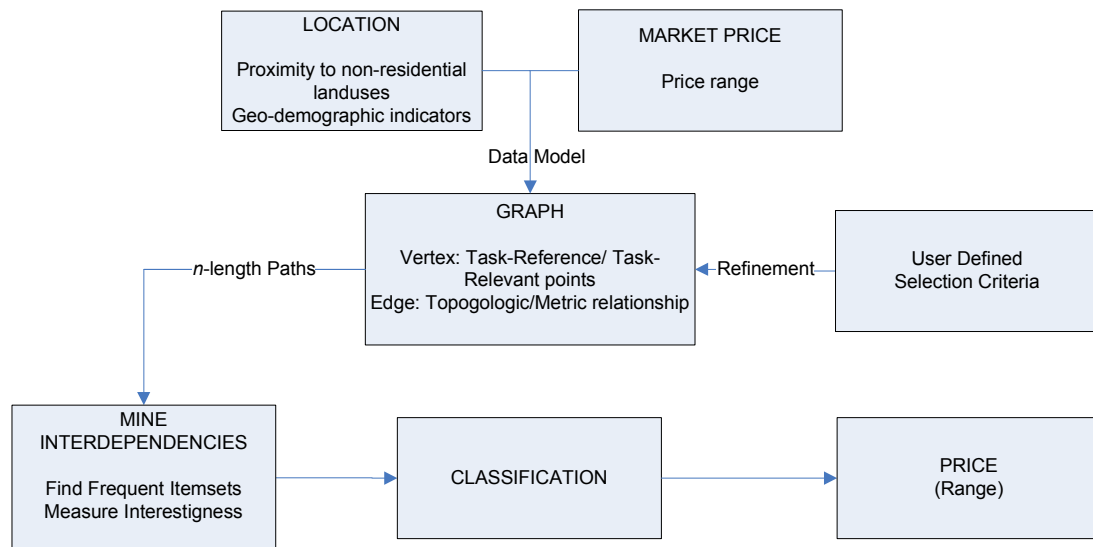


Figure 4-1: Proposed Methodology

Figure 4-1 shows a diagram of the proposed methodology. As shown, the process has been designed so that location can be taken into account implicitly during the property valuation process. Initially, data related to this study is structured as a multi-layered graph.

Since the classification is based entirely on the arrangement of the spatial objects, the accurate modeling of their inter-relationships is of utmost importance. As discussed in Chapter 3, the value of a property is closely related to its surroundings. Empirical research has demonstrated that the presence of certain spatial objects such as schools or parks can have an effect on the price. Although this effect is documented in the literature there are certain issues.

The first is that the type and magnitude of the effect that these spatial entities have on the property is context dependant. For example proximity to a park in a heavily urban environment is highly priced. On the other hand, the same park in a suburban environment may not have such positive impact on the price since the morphology of the areas is different.

The second issue is that the different spatial entities do not affect the price in an isolated way. Especially within urban environments where multiple landuses are in a close proximity to a property, prices cannot be determined based on isolated spatial entities alone. In that case, it is more valid to consider their combined effect that is the result of their existing inter-relationships. A graph-based modelling approach

tackles these issues as it effectively models all the relationships and provides information about their connectivity at multiple levels.

Property valuation is performed by accessing this graph and applying data mining techniques. A graph traversal algorithm is applied to calculate the paths that form the input to the data mining algorithm. For the description of the dataset and the extraction of the knowledge in terms of patterns that will lead to the modelling of location, the dependency analysis task has been chosen and in particular the mining of association rules. Since the purpose of this method is to perform valuation based on limited, if any, information about the kind of locational influence on the price, it is necessary to primarily describe all the possible dependencies between the price and the several locational features. Dependency analysis in the form of multi-level association rule mining enables the discovery of higher-order interactions between locational features. It also takes into account any existing interdependencies resulting in more accurate modelling of location's contribution to property prices since spatial factors influence value not in an isolated way.

The second task that belongs to the predictive data mining is the classification task. Based on the output of the dependency analysis, classification will divide the data into classes and hence perform the valuation.

4.3 The modelling and Knowledge Discovery Algorithm

Before going into the detailed design issues of the knowledge-based system that accommodates the proposed methodology, some aspects of the general design are discussed. These include the data model, the graph traversal algorithm and the general function of the data mining algorithm, which is the central component to the system.

4.3.1 Graph-theoretic approach for modelling location

As stated in the previous section the modelling of the spatial interrelations was based on a graph-theoretic approach. Graphs are very important modelling tools and have been successfully applied to various problems in several domains. Graphs, unlike

trees, base their architecture on the problem they model aiming at the best possible representation. Problems may belong to the physical world or can be abstract. Example applications include from optimisation problems in engineering and modelling of social networks in social sciences, to the modelling of the physical world (e.g. road network).

Since the introduction of graphs as formal models to represent networks, they have been applied in various geographical problems. Although graphs are considered as highly abstracted models of spatial relationships that represent only connectivity, they can be proved useful modeling tools when applied to specific problems (Worboys & Duckham, 2004). One apparent application within the geographic domain is the use of directed graphs to model networks such as roads, rivers and so forth (see Cliff *et al.*, 1979).

Network connectivity is not always enough to capture the spatial arrangements. Other relationships such as topological, metric and directional may be of interest and hence should be represented within the data model. On the other hand, the use of a graph theoretic model presents a number of advantages. Graphs are extensively used as modeling tools resulting to the development of various operations and efficient algorithms that one can base their analysis on. Furthermore, graphs enable the investigation of relationships that extend further to those that exist within the immediate neighborhood. The latter characteristic is what makes graphs extremely relevant to this research.

In this research no previous knowledge of the area is assumed. Therefore it is necessary to model the data in such a way that all the inter-relationships are represented. A graph-theoretic approach fulfils such a requirement and also allows the representation of higher order relationships. These inter-relationships can be expressed in the form of topological, metric or directional relationships. For this purpose, the graph should be extended to include such relationships in order to facilitate their connectivity analysis. In such context, paths represent a sequence of spatial relationships. This extension can be achieved by implicitly model these spatial relationships within the graph (Ester *et al.*, 1997).

Before presenting the conceived graph, it is necessary to introduce the following definitions about graphs and their properties (West, 2000; Lafore, 2003; Evans & Minieka, 1992).

A simple *graph* can be formally described as a set $G = (V, E)$, where V is a set of vertices (or nodes) $V = \{v_1, v_2, \dots, v_n\}$ and E is a set of edges (or arcs) $E = \{e_1, e_2, \dots, e_n\}$ where $V \neq \emptyset$ and $E \neq \emptyset$ and $E = \{(v_i, v_j) \mid v_i, v_j \in V\}$. For an edge $e_i = (v_i, v_j)$ vertices v_i, v_j and edge e_i are incident to one another. The vertices v_i, v_j of the edge $e_i = (v_i, v_j)$ are adjacent. The degree of a vertex v_i denoted as $deg(v_i)$ stands for the number of edges incident with it.

A *directed graph* (or digraph) is a graph $G = (V, E)$ in which the set V is a set of ordered vertices hence each edge has a direction assigned to it. In the case of directed $(v_i, v_j) \neq (v_j, v_i)$.

A *weighted graph* is a graph $G = (V, E, W)$ in which each edge has a weight assigned to it. Weights can represent either the physical distance or costs associated with the edges.

A *path* $v_1, v_2, \dots, v_n \in V$ within a graph $G = (V, E)$ between $v_1, v_n \in V$ is defined as the connected sequence $(v_1, v_2), (v_2, v_3), \dots, (v_{n-1}, v_n)$. A path is called *simple* when none of its nodes is incident with more than two of each edges hence the path does not contain circles. Graphs in which for every set of vertices there is a path that connects them are called *connected graph*. Paths have specific *length* that denotes the number of edges traversed.

Figure 4-2, shows the structure of the graph G that models the spatial relationships in the current research. It is a directed weighted graph that has two levels of hierarchy. The first level models the spatial relationships at property level. First level nodes represent two types of spatial entities: the *Reference spatial entities* and the *Task-relevant entities* that are entities relevant to the association. More specifically, the reference spatial entities represent the properties for which the sale price is known (P_1, P_2). The task-relevant entities represent all the non-residential landuses and other spatial objects relevant to the study such as bus stops (L_1, L_2).

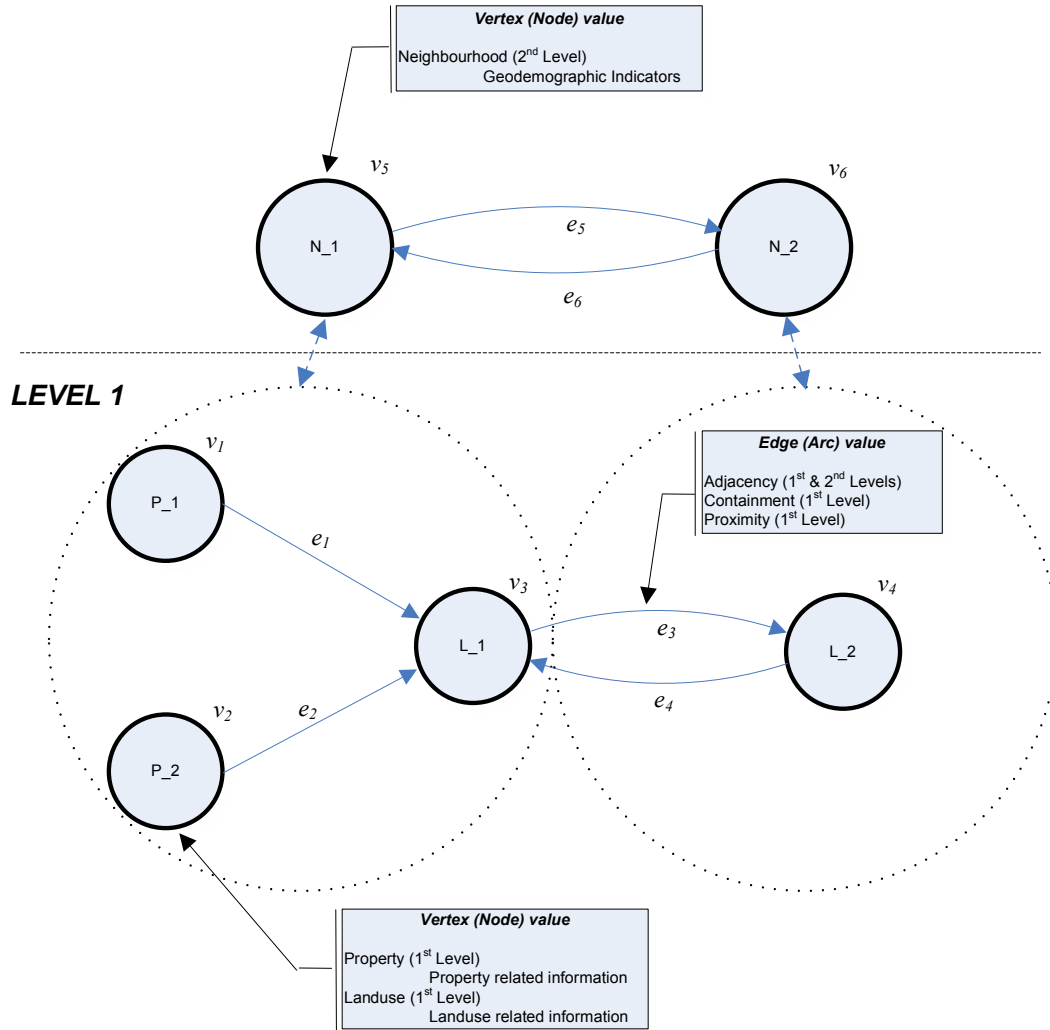
LEVEL 2

Figure 4-2: Structure of graph that models spatial relationships

The second level of the graph, models the spatial relationships at neighborhood level. In this case, nodes represent individual neighborhoods (N_1, N_2). This distinction was dictated by the need for the system to also facilitate the investigation of the way neighborhood quality affects the price of the property.

As discussed in Chapter 3 (see Section 3.6), the neighborhood quality has an effect on the property price. Neighborhood quality is commonly expressed in terms of socio-economic status, racial composition and quality of services. Most of this information can be accessed through the Census where it exists at certain levels of aggregation such as wards, output areas etc. Since we are not only interested in the neighborhood quality effect of the neighborhood that the property belongs to but also

in the sequential effect caused by its adjacency to other neighborhoods this adjacency should be modeled.

The edges between two first level nodes denote that a certain type of spatial relationship holds between them. The modeled spatial relationships include *Adjacency*, *Containment* and *Proximity*. An adjacency relationship holds when two spatial objects of polygon type share a common boundary. The containment relationship is used to model two types of spatial arrangements. The first is to model activities that share the same polygonal reference. The second spatial arrangement reflects the cases where activities exist within broader landuses. Such examples are the sport facilities that are part of big parks. Finally, proximity was introduced to model all the relationships that are within close range to the property but are not directly connected to it. As mentioned, the second level nodes represent the neighborhood. The adjacency relationship was used to model the spatial relationships at that level.

The direction of the edges was determined based on the type and level of the node. Edges that are incident to reference nodes (e_1 , e_2) have only one direction that leads away from them. This is to ensure that the reference points are not included within the extracted paths of other reference points. On the other hand, the edges that are incident only to non-residential landuses or belong to the second level have both directions (e_3 , e_4).

The weights were applied in the form of costs according to the type of spatial relationship the edges model. Containment, adjacency and proximity relationships were assigned with costs 1, 2 and 3 respectively. The use of the weights in conjunction with the use of a shortest path type algorithm described in the following section, ensures the correct order of the nodes within a path. That is, from a starting node v_i and for a given length n ensures that the nodes are visited by accessing first the links that denote the closest relationships.

4.3.2 Graph traversal algorithm

The relationship of a given reference object with respect to the task relevant points is expressed in the form of paths that have the object as a starting point. It is apparent that we are interested in simple paths. That is, we are interested in paths that consist of a sequence of links that do not visit the same vertex twice. This constraint is to ensure that no cycles are included in the calculated paths in the cases that both directions exist. In that way, the calculated paths will only lead away from the starting object and there will be no redundant paths returning to the start. In the case of the graph G illustrated in Figure 4-2, the location of vertex v_2 is described by the paths:

$$Path_1: v_2 (e_2) v_3$$

$$Path_2: v_2 (e_2) v_3 (e_3) v_4$$

$Path_1$ and $Path_2$ are paths with length 1 and 2 respectively. In the parenthesis, the edge that was used to access each node is shown.

Paths are extracted from a graph by the use of fundamental operations that access the graph from a given starting point. This search is usually performed according to two common approaches, the *depth-first-search* (DFS) and the *breadth-first-search* (BFS). Another very useful traverse method of graphs is the *Shortest Path* algorithm. It was proposed in 1959 by Edsger Dijkstra and it is one of the most common operations applied to weighted graphs. For a weighted graph $G = (N, E)$ the Dijkstra algorithm performs as follows (Worboys & Duckham, 2004):

Distances are represented as a weighting function w ($w: E \rightarrow \mathbb{R}^+$) in addition to a target weighting function t ($t: N \rightarrow \mathbb{R}^+$) that is used to store the minimum distances between the starting point to each node. Dijkstra's algorithm starts by initialling the target weights to infinity (a very high number) except for the starting node and the nodes adjacent to that. Then the algorithm traverses the entire graph from the starting node. At each step it sorts any unvisited nodes in ascending order of their target weights. It then recalculates the minimum target weights t .

This algorithm can calculate the best (in terms of cost) path between two nodes. It can also calculate all the shortest paths that start from a given point, termed a single-

source shortest path algorithm (Worboys & Duckham, 2004). This ability makes its application meaningful in this research.

In this research, the search algorithm facilitates the retrieval of the necessary spatial information that is then used as an input in the association-based analysis. In particular, the interest is in representing the spatial arrangement of different landuses or activities in relation to known reference points in the form of paths.

Since the graph is not based only on one spatial relation type, it is essential to ensure that the higher order nodes are accessed through links that denote closer relationships. For example, in the case where two nodes are connected via an ‘adjacency’ link, given the way the graph was realised, there is also a ‘near’ link that connects them. Simple graph traversal will result in two paths, one that accesses the next level node through the adjacency link and another that accesses the same next level node through the ‘near’ link. Both paths are valid since they represent true relationships but they also result to the creation of redundant information. This can be easily avoided by adopting a ‘shortest path’ logic in the design of the search algorithm based on the weighting of the different spatial relationships.

Although the shortest path algorithm effectively deals with the above issue, also it presents an additional consideration based on the way shortest path algorithms traverse graphs. Such algorithms, search for a single path out of all possible paths in a given graph which has the least length, based on a cost function, connecting the given starting node with sequentially adjacent nodes. In case of more than one paths with equal costs, these algorithms select one arbitrarily. For example when traversing the sample graph shown in Figure 4-3 from node v_1 , the node v_4 will be accessed either from node v_2 or from node v_3 since both paths ($v_1-v_2-v_4$, $v_1-v_3-v_4$) are equal in terms of cost.

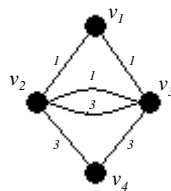


Figure 4-3: Example

For spatial applications and in particular in this property valuation application, key spatial relationships can be lost as a result of this. Hence, the application of a shortest path algorithm alone for the calculation of paths connecting more than one adjacent nodes is not appropriate. For this reason, a search algorithm that finds all least cost paths of any length, that is connecting any given number of adjacent nodes in a single path, has been developed to include all important spatial relations. The result is a set of all such paths for each transaction, or reference object in the dataset. These are the paths that are then used as inputs in the data mining algorithm for determining associations.

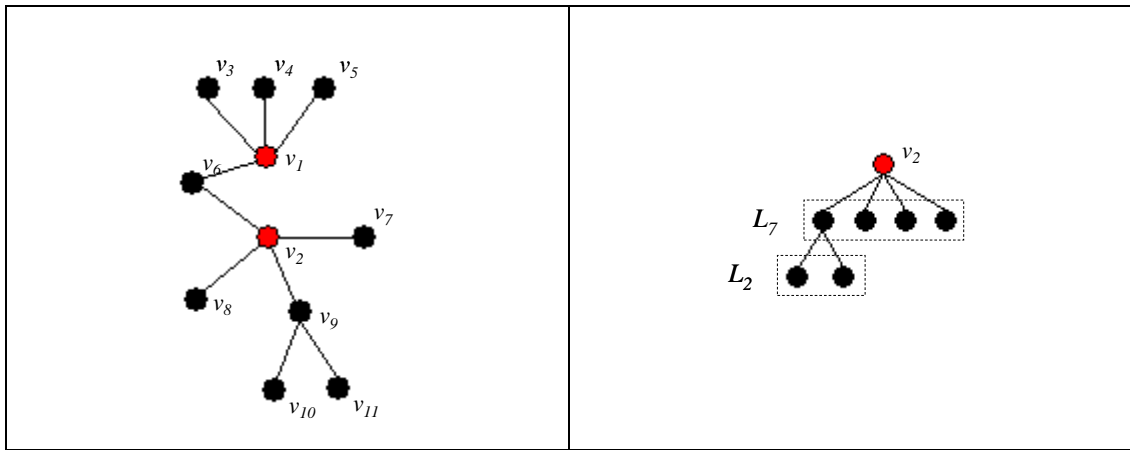


Figure 4-4: Data Structure

This search algorithm starts the graph traversal from each given reference object (transaction). The algorithm progresses by finding the least cost path from that object to all immediately adjacent objects in the dataset. These can be landuses. These links are stored in an array with the transaction as the start node and these objects as end nodes. The algorithm progresses by using as a starting point the newly discovered spatial objects, for example these landuses, to discover their nearest neighbouring objects and determine least cost links between them repeating the process until the specified path length has been considered. This extension follows certain constraints that ensure the leading-away direction of the path and also the exclusion of cycles. In Figure 4-4, the data structure is shown while the whole procedure is illustrated in Algorithm 4-1 below.

Search Algorithm*Algorithm 4-1*

Finds all paths from a transaction to all connected nodes at path

findShortestPaths (criteria, pathLength)

Step 1 - Get all transactions t from spatial database based on spatial or user defined criteria and set as starting points and store in array T .

Step 2 - Generate all paths of length = pathLength

```

for each  $t \in T$ 
    set next  $n_i = \text{next } t$ 

procedure openNearNeighbours (node  $n$ )
    repeat
        create  $l_y$  and add to  $L$ 
        for each  $n_i$ 
            find all  $k_j$  for  $n_i$  excluding parent node, self, peer node
            add  $n_i$  to  $l_y$ 
            for each  $k_j$  in  $n_i$ 
                set next  $n_i = \text{next } k_j$ 
                openNearNeighbours (next  $n$ )
             $y=y+1$ 
    until level  $y = m$ 

```

Step 3 - Extract paths

Starting from final nodes at $l_m \in L$ to $l_i \in L$ extract each link (k_j, n_i) and add to path P until t is reached

Step 4 - Save in database

Save paths into *ghu_paths*, *ghu_plinks* tables

end for

Where:

- Set of levels $L = \{l_1, \dots, l_y, \dots, l_m\}$ and m is the given path length
- $l_y \in L$ is array of nodes containing links between a head node and end nodes as shown in Figure 4-4. Thus $l_y = \{n_1, n_2, \dots, n_i, \dots\}$ and $i \in S$, where S is the set of spatial objects
- each $n_i = \{k_1, k_2, \dots, k_j, \dots\}$ where $k_j \in S$, k_j represents end nodes which are closest neighbours with 1st degree relationship with a head node n_i and $k_j \neq i$
- a head node n_i is connected with its closest neighbours k_j via a path $P_{ik_j} = \min\{d, a, c\}$, where d is the 1st degree link representing proximity between i and k_j , a is the 1st degree link representing adjacency between i and k_j and c is

the 1st degree link representing containment between i and k_j . The values of these links are $\{d, a, c\} = \{3, 2, 1\}$.

4.3.3 Data mining algorithm

Figure 4-5, presents a high-level overview of the data mining process that consists of two tasks, the mining of interdependencies and the classification. This is implemented in the data mining component of the system.

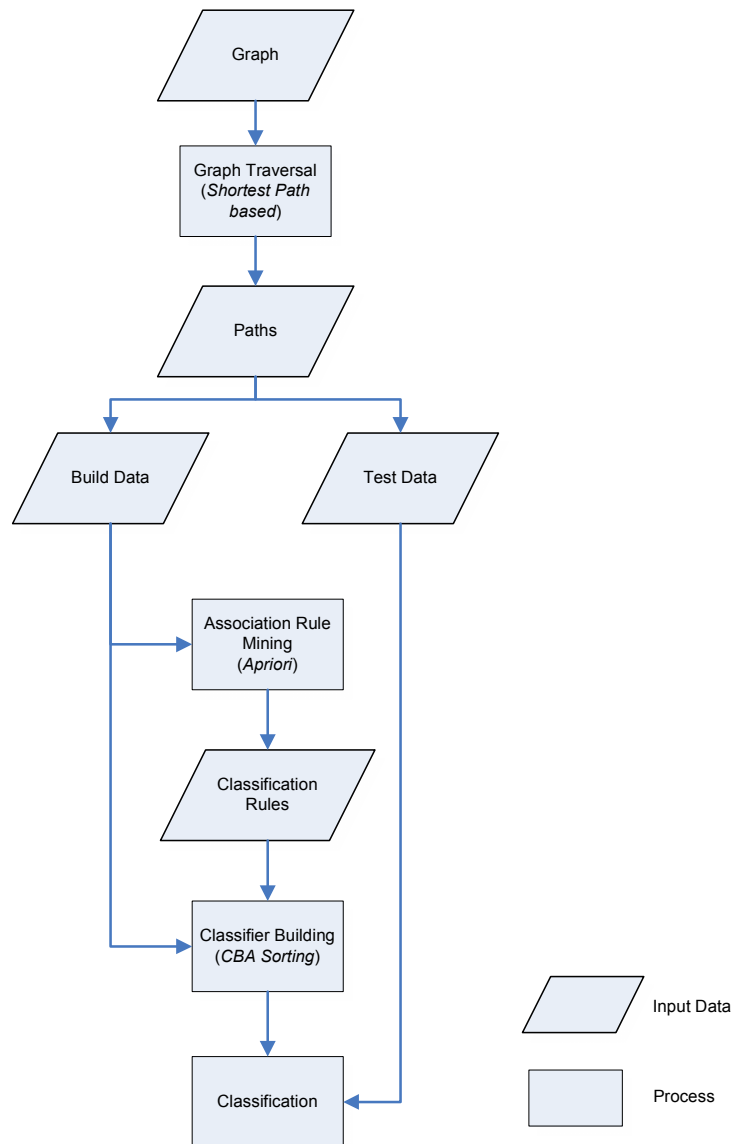


Figure 4-5: Data mining process overview

As discussed in the previous sections, the data mining process is based on a weighted and directed graph. The input of the data mining algorithm is a set of shortest paths that describe each of the reference nodes, extracted with the application of the

shortest path based algorithm presented in the previous section. The length of the extracted paths varies from 1 to n , with paths with length 1 representing the immediate relationships of the reference node to its surrounding task relevant nodes.

The computed paths are then reformatted appropriately for the association rule mining to be applied. In classification problems, data has to split in order to form the build and test data. Build data is used for the training of the model that result to the building of the classifier. The testing of the classifier is based on the test data. Similarly in this case, data split into build data and test data. The build data is used for the association rule mining while test data is used for the evaluation of the classifier. The separation criterion used, is based on the random sampling (Build Data – Test Data) which is a common practice in classification problems (Pyle, 1999). For the discovery of the association rules the Apriori algorithm (see Section 2.1.3.2) is used. Since there is not enough evidence in the literature in favour of a certain association rule mining algorithm (see Section 2.1.3.1), the Apriori algorithm was chosen as it is considered one of the most important representatives of its type.

Given the limitation of association rules in dealing with numerical data (see Section 2.1.3.1) the classification is not directly based on the individual prices. Instead, a discretisation operation is applied that results in a number of classes. The range within these classes and the number of the different classes depends on various factors such as the geographical size of the examined area. These considerations are further discussed in the case study (see Chapter 6).

In the case of association based classification, the classifier comprises of a set of classification rules (see Section 2.1.3.5). In this context, since we are interested in classifying the test cases based on their spatial association relationships, the classification rules take the following generic form:

Spatial relationship \rightarrow Non-Spatial Classifier

where the spatial relationship is in the form of an association relationship while the non-spatial classifier is a price range. To ensure this, further to the support and confidence thresholds, two additional constraints guide the rule generation. The first is that the descendant part of the rule must contain *only* the classifier. The second is that the antecedent must also contain a non-spatial description of the reference node

such as property type. This can be relaxed by restricting the mining process on homogeneous property types. An example classification rule is the following:

Commercial Services= NEAR \rightarrow PRICE_RANGE= [1610000-2800000]
(support=100, confidence=100)

This rule implies that properties that are located near commercial services belong at the price range of 1610000 – 2800000. It is quite general and gives the estimation in relation to a single-member antecedent. This is also a strong rule since both the support and the confidence are 100% which means that this rule satisfies all the training cases.

The construction of the classifier is based on the three-step procedure of CBA (see Section 2.1.3.6) in conjunction with the build data. The selection of this method was based on the fact that the best-rule sorting consistently performs well irrespective of the type of the dataset is applied to (see Section 2.1.3.5). The evaluation of the classifier in terms of accuracy is performed using the test data by calculating the percentage of the cases that are correctly classified from the classifier. Finally, all the results such as classification rules, classifier, test classification and accuracy are reported through the system's logger.

4.4 A Procedure for the design and implementation of the system

4.4.1 Analysis and Design Methodology

For the analysis and the description of the detailed design of the system the Unified Modelling Language (UML) concepts and diagrams were used.

UML is an object-oriented methodology. As discussed by its creators (Rumbaugh *et al.*, 2005), UML is a general-purpose visual modelling language that includes semantic concepts, notation and general guidelines that enable and guide the specification, visualisation, construction and documentation of the components of a software system. It consists of static, dynamic, environmental, and organisational parts that intend to capture the information about the static and dynamic behaviour of

a system. For organisation purposes, all these parts are divided into views that are expressed in the form of different diagrams. Views can be further organised into four major areas: structural, dynamic, physical and model management (Rumbaugh *et al.*, 2005).

Structural classification captures the structure and organisation of operations of the data quantized in classes. Key elements are the classifiers (actor, class etc.) and their relationships (association, dependency etc.). Classifiers can either represent objects (e.g. class) or represent behavioural concepts (e.g. actors). Views that belong to this area are depicted with diagrams such as class diagram, internal structure, collaboration diagram, component diagram and use case diagram. On the other hand, dynamic behaviour describes a series of changes regarding the components of the system described with the structural views over time. Views of this type are illustrated through diagrams such as state machine diagrams, activity diagrams, sequence diagrams and communication diagrams. Finally, the physical layout (deployment diagram) and the model management (package diagram) contain views that describe the computational resources and the organisation of the models in the form of hierarchical units respectively.

Diagrams that were used to fully describe the proposed system are: the class diagram, the component diagram, the use case diagram and the sequence diagram. These are further analysed in the following sections.

4.4.2 Conceptual Architecture and Non-functional Requirements of the System

As already stated, one of the primary objectives of this research is to design and implement an integrated system that incorporates knowledge discovery functions. Figure 4-6 illustrates the conceptual architecture of such a system.

This architecture is based on three main components. The first is the *data loader* that is responsible for getting the data from the various sources. The second is the *valuation engine* which is an implementation of the data mining algorithm and the classification algorithm. The main functionality of the data mining algorithm is to extract the knowledge from the data and place it back in the classification engine that performs the valuation. The third and last component is the *visualisation component*.

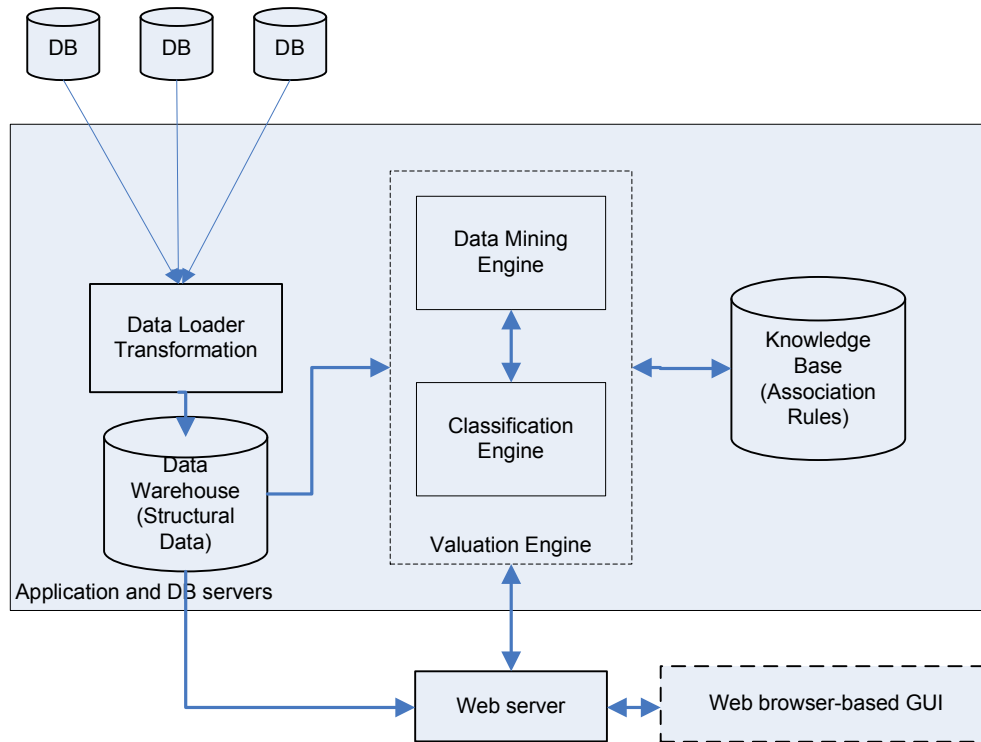


Figure 4-6: Conceptual System design

The focus of this research was on the implementation of the two first components: Data Loader and the Valuation Engine. A number of requirements for the proposed system have been set. Such a system:

Should support spatial data

The system must be built in conjunction with a Database Management System that supports spatial data. Although spatial relationships are modelled as graphs, all the pre-calculations are based on spatial indices. Hence the Data Management System must support such operations.

Should support graphs and graph operations

The system must support graph-based operations or provide the required tools for the development of these operations. Additionally, it must facilitate the efficient storage of the graph. It should be noted that there are two directions one can follow. One is the on-the-fly calculation of the graph. This stores the graph in the memory for the whole duration of the operations. The other approach is the persistent storage of the graph in the database. The second approach increases the efficiency of the data

mining algorithm since it does not involve the spatial calculations step. This approach is preferable in this case since the graph is consisted mostly from the task-relevant objects that can be considered relatively stable over a period of time. Therefore the whole graph can be considered relatively stable with the set of reference points being the only possible source of frequent updates. In the case of new reference points, the update of the graph is quite straight forward. It first involves the calculation of the spatial relationships of the new objects in relation to the task-relevant objects. Then these objects should be appended in the node and link tables.

Should support the integration of the data-mining algorithm within the DBMS

In the literature, a number of advantages are presented that support the integration of the data mining algorithms with the database management system, as oppose to a stand alone tool. Specifically, Netz *et al.* (2000) comments on the need to integrate data mining with database systems in order to make this analytical technique stronger, by backing it up with technologies such as data warehouses in order to deal with issues such as data integrity and effective data management. Furthermore, Ester *et al.* (1999) further emphasise on the advantages of such integration related to better storage management, avoidance of inconsistencies and finally, use of already existing functions (e.g. indexing) without the need for further implementations. Finally, integration with the database is considered a more unified approach compared to designs where the data mining component exists separately and the input is a flat file.

4.4.3 System Design

What follows is the description of the system in terms of its structure and dynamic behaviour.

4.4.3.1 Use Case Diagram and Functional Requirements

Use case diagrams are used to describe the external behaviour of the system as this can be seen from outside users. Although it does not provide a structural guide for the implementation of the system, it provides a logical description of the

functionality required. Hence, it can be used as a first presentation of the usage requirements of the system.

Figure 4-7 presents a high-level overview of the usage requirements of such a system. This use case diagram illustrates the system's functionality in relation to idealised users. This functionality is provided by the classifiers and expressed in terms of their interactions. Classifiers shown in this diagram are actors and use cases. Relationship types such as generalization, usage and association, model the interactions between the classifiers. It should be noted that these actors denote roles that do not necessarily coincide with real persons but they can represent processes or other systems.

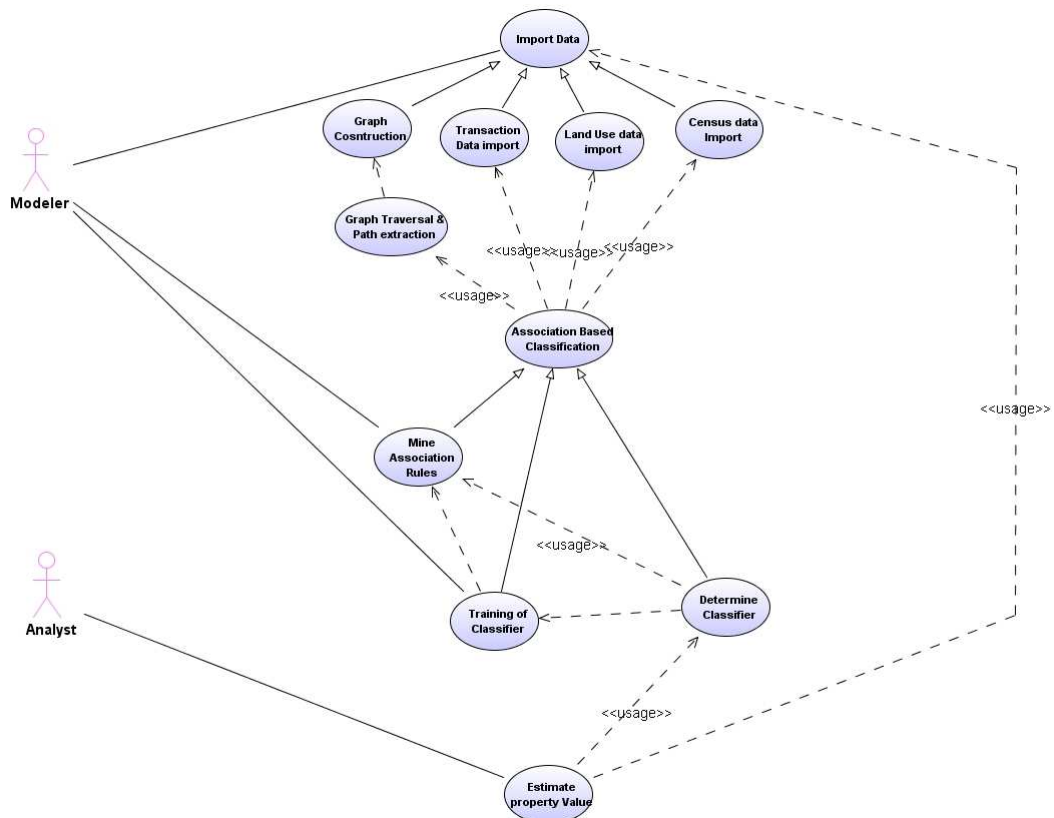


Figure 4-7: Use Cases Diagram

The functionality provided by each classifier should be in accordance with the main objective of the system, which is to support property value estimations by utilising spatial data mining algorithms. Hence, each of the use cases represents a piece of functionality that can either be autonomous or can be mixed with that of other use cases.

In the Figure 4-7 use cases are drawn as ellipses while different types of links denote the type of relationships between actors and use cases. Two main actors can be distinguished, modeller and analyst. Modeller associates with three use cases: Import data, Mine Association Rules and Classifier Training. Analyst associates with the Estimate Property Value.

A generalisation relationship holds between the import data use case and the Graph Construction, Transaction Import, Landuse Import and Census Import use cases denoting a parent-child relationship. Association Based Classification is a generalisation of the Mine Association rules, Training of Classifier and Determine Classifier denoting again a parent-child relationship. Use cases that are linked with dashed arrows denote a usage dependency.

Use case *Import data* is a key use case. Its purpose is to deal with the construction of the input dataset. This case is invoked when the actor *Modeller* initialises the application. The Modeller defines a number of parameters that relate to the size and dimensions of the input dataset. On completion, the input data is prepared to be further processed.

Use case *Mine association rules* is also a key use case. Its purpose is to mine the interdependencies in the input data, extract association rules and store them in the database. The *Modeller* invokes this use case when a valid input dataset based on user specified criteria has been generated. Association rules are determined, stored in the database and displayed on the screen.

Use case *Training of Classifier* is responsible for the determination of the classifier. It is invoked by the *Modeller*, after the completion of the *Mine association rules* use case. It uses a dataset that has been generated and the association rules to determine the appropriate classifier. On completion it displays the classifier.

The final use case is the *Estimate property value* use case. It is invoked by the *Analyst* and is responsible for the property valuation function. It uses the classifier determined, in conjunction with a test dataset or a test case. It calculates the accuracy of the classification and finally it displays the classification of the test dataset or the test case.

Packages of the CAPV system

Figure 4-8 shows the six main packages of the system in the form of a package diagram. This was used to model the organisation of the whole system (model). A brief description of them follows.

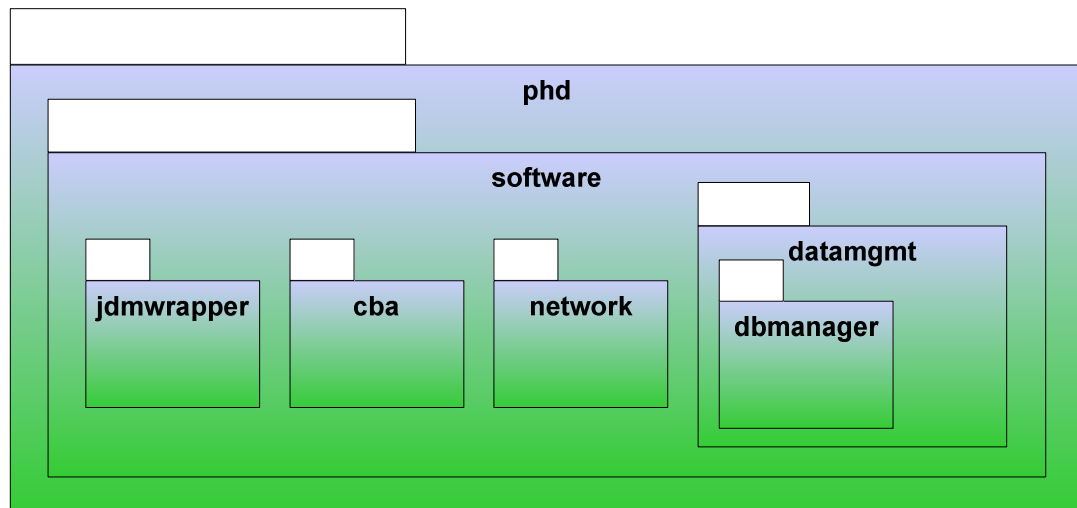


Figure 4-8: Package Diagram

phd.software

This is a container package (component), within which all components of the system are located. It also contains the main start-up class and the main server class which is a container of the systems services.

software.datamgmt

Contains classes used for accessing the spatial database (Data Warehouse containing the structural data of the systems) and also the knowledge database. Classes that are located in this package (component), use the standard JDBC API to connect and access the Oracle database. It is also a wrapper class to Oracle Spatial's API that manages the graph and maps relational data into objects by creating network objects that are instances of classes located in the network package.

datamgmt.dbmanager

The dbmanager package (component) is located within the datamgmt package and contains classes for managing connections to the database using a connection pool.

The reason for this is the optimal management of the life cycle of connections and increased performance of the database and access to the data.

software.jdmwrapper

This package (component) contains a wrapper class to Oracle Data Mining (ODM) Java API libraries. This wrapper provides methods to the system at the appropriate level of abstraction related to the Apriori algorithm.

software.cba

This package (component) contains classes that implement the CBA algorithm. In particular, the construction –training of the classifier and the classification stages of the algorithm.

software.network

This package (component) contains classes that map relational data into objects. These are created and supplied to the system through the classes contained in the *datamgmt* package.

Class Diagrams

In the class diagrams shown below, a graphical representation of the model's static elements (classes, relationships) of three of the main packages is provided. In those packages, the algorithms described are implemented in the classes discussed below. Classes are drawn as rectangles while their inter-relationships as arcs. For the presentation of the class diagrams, the packages shown in the package diagram were used as criterion for the creation of individual diagrams.

to the spatial database. The CAPVSPatialRepository class contains methods that handle the network stored in the database. It includes methods of three main types: methods that manage the graph and retrieve components of the graph (e.g. **getNodes()**), methods that traverse the graph and calculate the paths (e.g. **findShortestPaths()**) and finally methods that update the database.

The CAPVJDMRepository class mainly manages the knowledge base. It contains methods that handle the extracted association rules (e.g. **getRules()**) and also methods that set up the training and test datasets. Similar to CAPVSPatialRepository, this class maps data mining information that is in relational data structures into object oriented data structures for the classification task.

The last class within the phd.software.datamgmt package is the CAPVRepositoryBase. This is a superclass that contains common methods for the accessing and managing of the database (e.g. **closeResultSet()**) used in both CAPVSPatialRepository and CAPVJDMRepository classes. This relationship between classes CAPVSPatialRepository and CAPVJDMRepository and CAPVRepositoryBase is reflected in the generalisation relationship shown, which denotes a parent-child relationship or a superclass – class relationship.

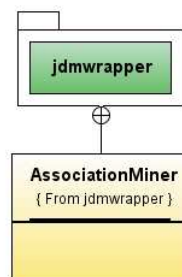


Figure 4-10: JDM Wrapper class diagram

Class AssociationMiner is the only class contained in the jdmwrapper package (Figure 4-10). This is a class that “wraps” the data mining functionality available in Oracle to interface it with the CAPV system. The methods in this class are used to initialise, configure and execute the Apriori algorithm. In particular there are three types of methods. The first provides the model settings, that control the model contents by filtering criteria such as min/max support/confidence thresholds, rule length and rule syntax. The second deals with data preparation tasks such as

discretisation and finally the third type relates to the model application and the presentation of the results.

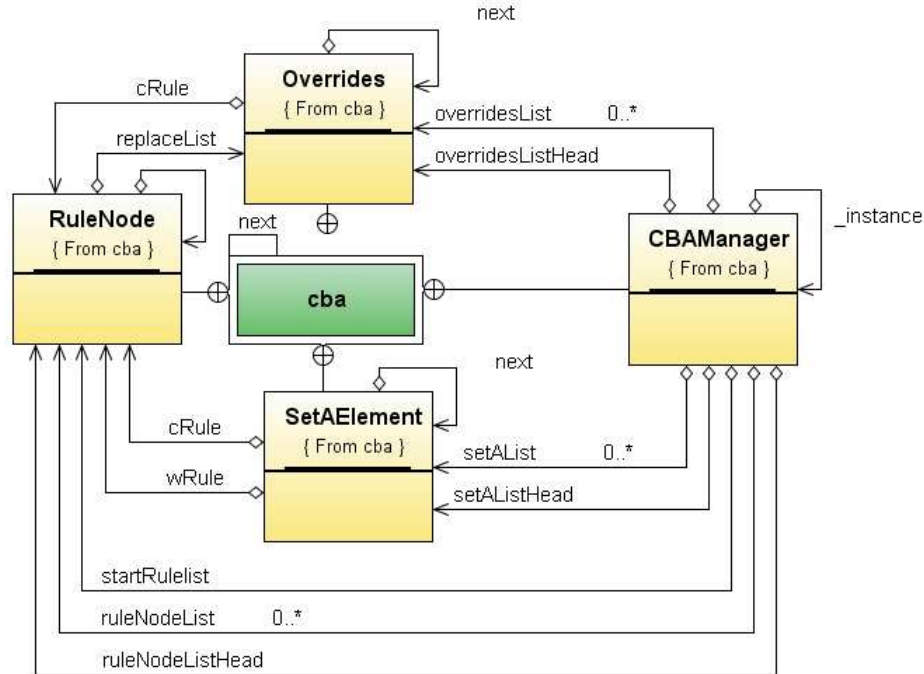


Figure 4-11: CBA Manager class diagram

Cba package (Figure 4-11) contains the class **CBAManager** that implements the CBA algorithm. This algorithm uses three linked lists (see Section 2.1.3.6). These linked lists are of type **Overrides**, **SetAElement** and **RuleNode** and are created and handled in **CBAManager**. Identifier **_instance** is a singleton of **CBAManager**.

Sequence Diagram

Figure 4-12 shows the sequence diagram that illustrates the behaviour sequence of the system within a certain timeframe. It is organised as a two-dimensional chart. The horizontal dimension represents the individual objects and the vertical dimension represents the time axis. The vertical dashed lines represent the lifeline of one object while the double filled line represents that an execution specification of a procedure on that object is active. Solid arrows denote calls while dashed arrows the returns. In the above diagram it is assumed that all the switches are on.

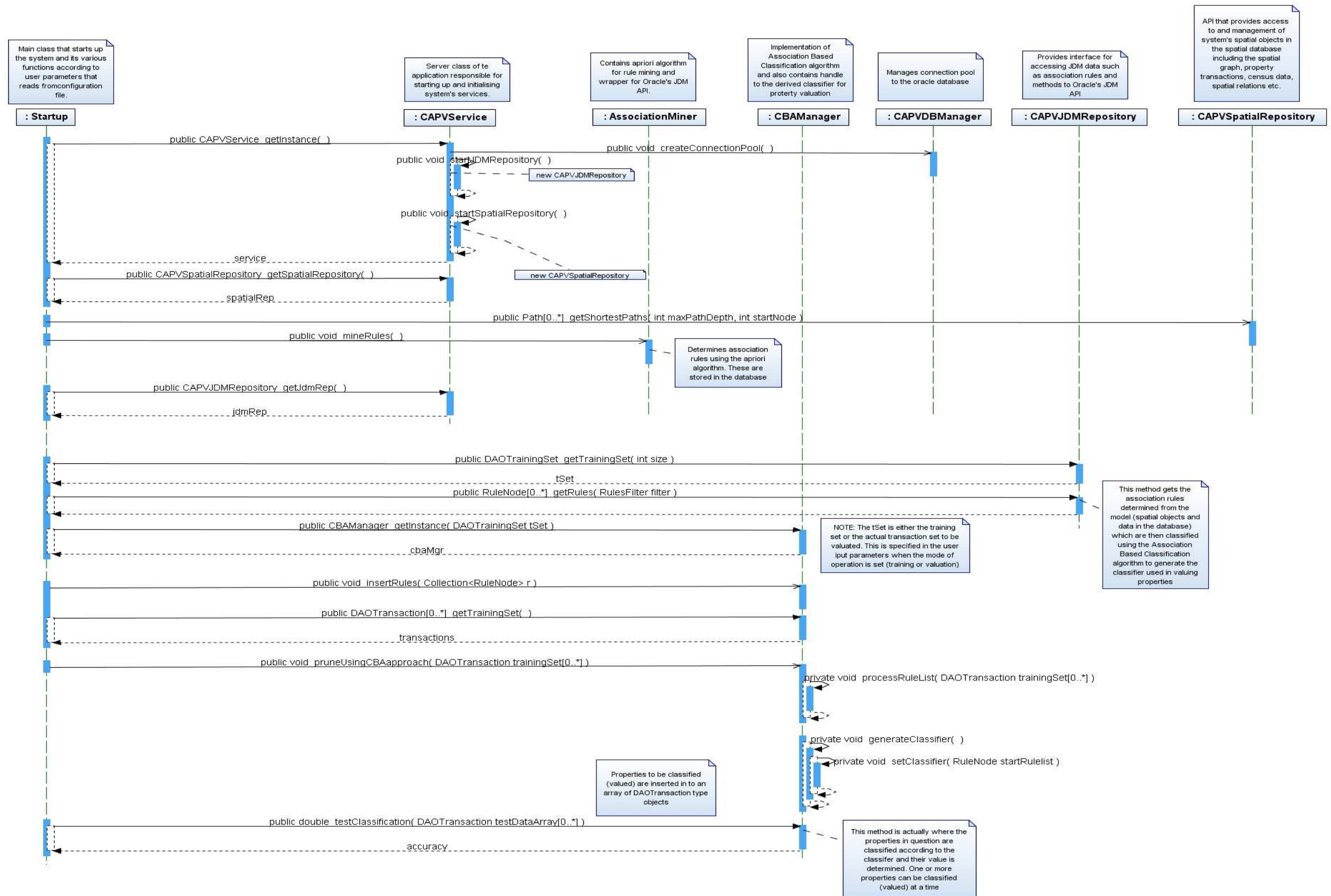


Figure 4-12: Sequence diagram

4.4.4 Data Requirements

Knowledge discovery is a data-driven approach. Therefore it is heavily dependent on good quality data that describe in the best possible manner the real-world problems. One of the primary objectives of this research is to develop a valuation model that takes location explicitly into account. Therefore, one essential requirement is that the data models the spatial arrangements in the best possible way.

In Section 3.2.1, there is a reference to a number of factors that could have an impact on the price of a property, classified as internal and external. Consistent with this classification, selected data must contain direct or indirect information that relates to the surroundings of the property in terms of its geographical location and also its structural characteristics.

Further to this, the volume of the data on which the analysis will be based must be high. For the results to be significant, the data employed in the analysis should be considered as a good representative sample of the whole population. Since the method employed is based on frequent pattern mining, it is apparent that the higher the volume of the data the better the quality of the results in terms of validity.

Another requirement relates to the format and standardisation of the data. As collection of the data was also included as task of this project, it is apparent that acquiring data in digital format would extremely reduce the time spend in data capture. Although such datasets may be found in several governmental organizations or companies, their completeness and specifications vary upon them. Additionally, access to them is not always possible. This relates to the two other objectives of the project. The first is to use datasets that are publicly available. The second is the use of datasets that can be considered as standards within the area, investigating in such way how much information can be extracted from them.

4.4.5 Database Design

One of the main requirements of the system is that it should offer data mining functionality in an integrated manner in respect to the DBMS. Hence, the efficient design of the database is of great importance. In the following sections aspects of the database design phase are presented.

4.4.5.1 Requirements Analysis

In this stage of the design, information related to the part of the system that will be supported by the database system was analysed. Information involved database related requirements of the main system and also data requirements.

System imposed requirements relate to the efficient storage and management of the graph – based model. As stated in Section 4.4.2, to increase the efficiency of the system the persistent storage of the graph is preferred over the on-the-fly calculation. Hence, the database should provide the adequate structure for the modelling and storage of graph models.

The database also should comply with the main data requirement which is the appropriate representation of location (see Section 4.4.4). Therefore, the database should be structured in such way that holds all the appropriate locational information. These requirements are reflected in the structure of the database schema that is presented in the following sections.

4.4.5.2 Conceptual Database Design

Figure 4-13 shows the conceptual schema of the persistent database. The proposed model includes seven main components that consist of two types: components that relate to the graph structure (graph components / entities) and components that refer to additional information about the nodes (descriptive components / entities). The structure of the components related to the graph is based on the Oracle Network data model.

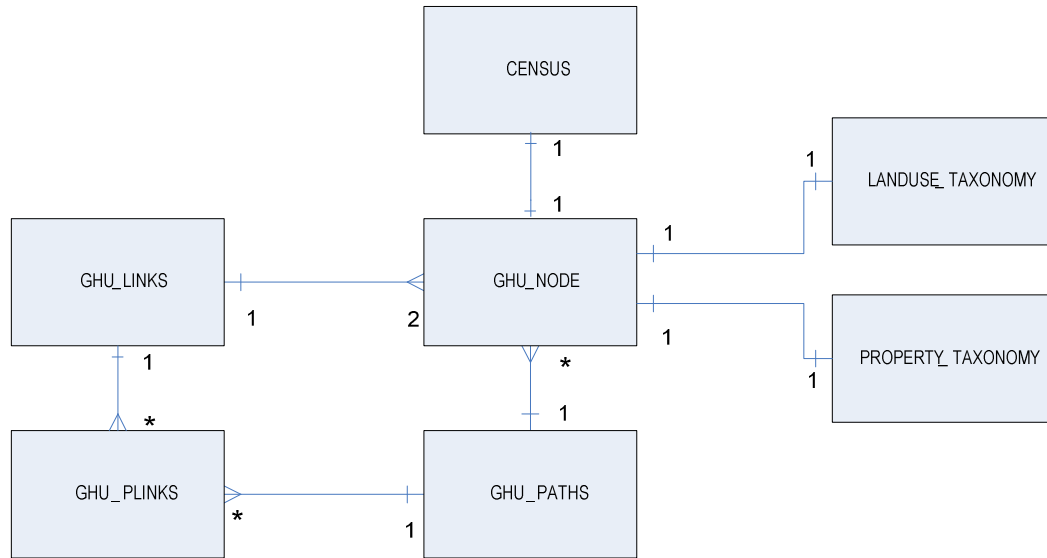


Figure 4-13: Conceptual Database Schema

Oracle Network data model provides a generic structure for the persistent storage of the network inside the database. Given its generic nature, there is a great degree of flexibility in the designing of the tables. The only requirements relate to the obligatory existence of certain columns within the tables that form the network. This structure does not necessarily require the creation of new tables. Views over existing tables can be used instead. This is extremely useful since the original schema remains intact. In this research, since the graph works as an integrated structure for the different datasets, new tables have been created. A more detailed analysis of the components follows.

Graph components

The four components that relate to the storage of the graph are the GHU_NODE (NODE_ID), GHU_LINKS (LINK_ID, START_NODE_ID, END_NODE_ID), GHU_PATHS (PATH_ID, START_NODE_ID, END_NODE_ID, COST, SIMPLE) and GHU_PLINKS (PATH_ID, LINK_ID, SEQ_NO) tables. GHU_NODE and GHU_LINK are used for the storage of the nodes and edges of the graph (see Section 4.3). Each of the network tables must include certain columns that required for the correct operation of the network. They can also include further descriptive information about the type of data they include. The remaining two entities, GHU_PATHS and GHU_PLINKS are not used to store information about the structure of the graph. They are used to store information about the computed paths. In the parenthesis, next to the component's name, the required columns are shown.

Descriptive components

The components that are used to hold the descriptive information about both the reference and the task-relevant points are: the CENSUS, LANDUSE_TAXONOMY and PROPERTY_TAXONOMY.

The Census component holds information that is used to create a general geo-demographic profile at neighbourhood level. Landuse_Taxonomy and Property_Taxonomy provide a detailed description at multiple level of abstraction associated with the two types of the nodes. Figure 4-13 also shows the cardinality between the entities.

4.4.5.3 Logical Database Design

At this stage the conceptual model described in the previous section was translated into the logical data model by deriving the relational schema from it. Figure 4-14 shows the conceptual data model showing all the attributes.

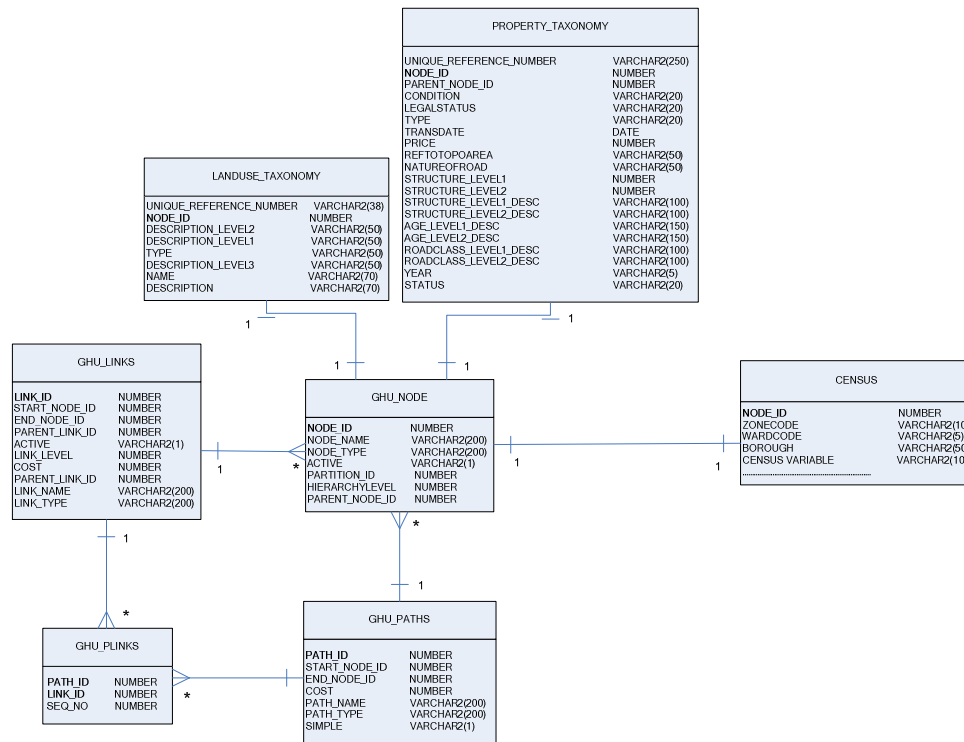


Figure 4-14: Logical Schema

4.5 Summary

This chapter covered the aspects of modelling and design of the system that accommodates the knowledge discovery based methodology to account for location into property valuation. This approach does not require the development of a fixed, monolithic and hard coded mathematical model where variables are fixed. Hence, prior knowledge of the geographical area and its characteristics is not needed.

To meet this, an appropriate data model that will be the basis for the data mining algorithm is proposed. It is a graph-theoretic data structure that captures the location of the known points in relation to its surroundings. This relationship is expressed in terms of topological and metric associations. This data structure enables the investigation of higher order relationships without restricting the analysis at the immediate neighbours.

For the graph traversal, a shortest-path based algorithm was developed that calculates all the paths of any length. The computed paths are then stored in the database and are used as an input for the data mining algorithm. Associative classification was chosen to be implemented in the data mining component of the whole system. The selection was based on the high accuracies reported in the literature achieved by this method when compared to other traditional classification approaches. Additionally, this approach is transparent in the way classification is performed fact that enables the better evaluation of the method.

All these are components of an integrated system. The proposed system was designed in such way that accommodates the investigation required. The design was based on UML and is presented in the form of various UML diagrams. The main requirements in the identification of the datasets were: adequate representation through the appropriate datasets of the factors that affect the price; quantity of the data and this is imposed from the technique; finally the datasets should be publicly available and considered as standards. Finally, this chapter concludes with a presentation of the database design. The design of the database was based on requirements imposed by the system and the data.

5 **Implementation**

In the previous chapter an integrated spatial data mining system has been presented. As it was designed to facilitate a complete data-driven approach to property valuation, data is of outmost importance. In this chapter, implementation issues related to the system and the database used in the evaluation of the method are presented.

Initially, the software platforms employed in the implementation phase are discussed. The next sections are focused on the identification of the datasets based on the requirements set in the previous chapter, accompanied with a description of the acquisition phase.

The core section of the chapter covers the aspects related to the database implementation. This is broken down into four sub-sections that coincide with the main steps of data preparation in terms of decisions and transformations. In the final section, the study area is presented and a brief profile of it is given.

5.1 Software Platforms

The implementation of the system involved the selection of a Database Management System and also the programming language to meet the requirements set in the previous section. For the database Oracle 10g (version 10.2.0.3) has been selected. Java 1.6.0 was selected for the implementation. Netbeans 5.5 was used for the development, as it is an easy to use Integrated Development Environment (IDE).

Requirements related to the datasets and also to the system underpinned the selection of Oracle. As the datasets are quite large, a robust and efficient data base management system should be used. Further to this, Oracle Spatial is used as it provides a robust data structure for modeling and managing spatial datasets and generating and storing graph structures. The development of open standards for spatial data (OpenGIS Simple Feature Specification & SQL/MM Part 3) enabled the efficient manipulation of spatial data in a unified manner through their implementation in Oracle Spatial. Spatial data types (point, line, polygon) are modeled in the SDO_GEOMETRY data type which is internally represented as an Oracle object data type. The population of this data type can be accomplished by the use of the corresponding objects constructor like any other object type.

Other advantages include the use of a standard language (SQL) eliminating the need for software specific language. Finally, all the advantages that associate with Oracle such as scalability, integrity, security, recovery and advance user management features that are not necessarily provided in other spatial management tools. Oracle Spatial is available through the standard installation of Oracle Database Server.

Oracle Spatial has a two-tier architecture (Database Server – Application server). The analytical functionality is available in either a Java API or a PL/SQL API provided in the form of PL/SQL functions. Both options have to offer advantages and disadvantages. Due to the fact that Oracle Data Mining (ODM) is a database technology, PL/SQL is considered the main access API hence the available functionality is more complete when compared to that of the ODM Java API.

On the other hand, ODM Java API is compliant to the Java Data Mining standard API for data mining developed through the Java Community Process. JDM provides interfaces that support data mining functions such as classification, regression, clustering and association. The ODM Java API, replaces the previously developed Java API for data mining available in Oracle 10.1 and implements Oracle-specific extensions to that standard. The ODM Java API enables the development of data mining tools in development environments such as Netbeans through a set of defined classes. In addition, use of the ODM Java API also achieves results in better performance since it is executed as a client/server model. Since performance is

important when dealing with voluminous datasets and the omitted functions were not relevant to the project, the Java APIs have been used.

In Oracle Spatial, data visualization is supported through a java server-side component included with the Oracle Application Server, Mapviewer. One limitation of this application is the lack of support for non-spatial graphs. Due to this limitation for the visualization of the graph the UCINET (Borgatti *et al.*, 1999) software was used.

5.2 Data Sources

An initial survey that involved the identification of possible sources of relevant datasets has been carried out. Table 5-1 presents a list of the identified potential data providers together with the type of information they provide. In this, both commercial and public vendors are listed.

Based on the data requirements presented in the previous chapter, this list has been filtered down to the final datasets. This was also affected by additional constraints such as budget restrictions and time limitations.

For the modeling of the spatial relationships the Ordnance Survey MASTERMAP, POINTX Points of Interest, Cities Revealed and Census 2001 datasets are used. MASTERMAP dataset fulfils the requirements for completeness and standardisation since it is the most commonly used dataset in spatial applications in the UK where parcel based modeling is required. Additionally, in the MASTERMAP data set a unique identifier is used and that makes the MASTERMAP data set particularly useful for property applications (Wyatt & Ralphs, 2003).

More specifically, the MASTERMAP layers relevant to this work are the Address layer, Integrated Transport Network layer (ITN) and the Topography layer. Although there is a large amount of information which can be sourced from these layers, there are certain limitations that dictated the use of additional datasets. One of these limitations is based on the fact that information is presented in a very detailed form. Another relates with the absence of landuse information.

Provider	Product
Ordnance Survey	MASTERMAP <ul style="list-style-type: none"> • Address Layer • Topography Layer • ITN Layer POINT-X
The GeoInformation Group	Cities Revealed <ul style="list-style-type: none"> • I2I • Imagery Data
Valuation Office Agency	Council Tax Valuations Rating Lists Dwelling Details Database
Land Registry	Transactional Price Dataset Average Prices
Web-based Services	Transactional Price Dataset Asking Prices
City Council	Planning information and housing surveys, School Catchments Areas
Environmental Agency	Complementary environmental information about air quality, flood risk etc.
Office for National Statistics (ONS)	UK 2001 Census Survey of English Housing English House Condition Survey Crime Statistics

Table 5-1: Potential Data Providers

Although it is possible to acquire some sort of such information, for example by using the Cartographic Text layer, this by no means can be considered complete. Therefore, to complete the picture in terms of non-residential land-uses supplementary information is needed.

For this purpose, two other datasets are used. The first contains detailed information about non-residential landuses in the form of points of interest (Point of Interest, POI). Supplementary information for these landuses came from the Cities Revealed dataset. This was necessary due to limitations associated with the point representation of the POI dataset. Both datasets have a direct link to the Mastermap

dataset. The 2001 Census dataset is used for the construction of the geo-demographic profile at Output Area (OA) level.

Property price information and additional information (new/old, type, tenure, date of transaction) of the properties are based on Land Registry transactional data information. However, this information was sourced through a third party web-based service provider and not directly from Land Registry. This was imposed by certain budget limitations. These third party services are entitled to manage property prices information that is extracted or derived from information produced by the Land Registry.

Finally, for information about structural characteristics such as the type of the property, the Cities Revealed dataset is used. Although other sources that contain detailed structural information have been approached, due to legal considerations, access to such data was restricted. A detailed account of these datasets is provided in a later section (see Section 5.4.1).

5.3 Data Acquisition

Following the identification of the final datasets, the datasets have been acquired. Since a number of datasets used came from various providers, there was a need to be imported into an intermediate schema. These manipulations were different for each of the datasets and are briefly summarised as follows.

Land Registry Data

As mentioned, the provider for the price information was a web-based service. PROVISER holds and manages transactional information. The property prices information on PROVISER is extracted or derived from information produced by Land Registry. The data is displayed on their web-site in a tabular format shown in Figure 5-1 and hold information about the address and basic characteristics of each property. Queries based on selection criteria such as year of purchase and postcode, are also available and were used to collect all the appropriate information.

PROVISER

PROVISER > Individual Prices

Tell a Friend!

CHANNELS

[UK Individual House Prices](#)
[UK Average House Prices](#)
[Mortgage Centre](#)
[Street Maps](#)

SEARCH

Individual Property Prices

Enter a Town/Postcode:

w1c

Refine Search

Road:

New/Old:

Type:

Time:

Until: /

Find Information

BROWSE BY

[National Region](#)
[Local Authority](#)
[Town](#)
[County](#)
[Postcode](#)

MORTGAGE CENTRE

[First Time Buyer](#)
[Home Mover](#)
[Remortgage](#)

[First BTL Property](#)
[Building Your Portfolio](#)
[Remortgage](#)

[Credit Problems](#)
[Self Employed](#)
[High Income Multiple](#)
[Right to Buy](#)

[Calculators: Max Borrowing \(Res\)](#)
[Max Borrowing \(BTL\)](#)
[Monthly Cost](#)
[Affordability](#)

[UK Property](#)
[London Rents](#)
[UK House Price](#)
[Estate Agents](#)

2007 House Price News

Today's latest news & info on the housing market at Guardian Property

Land Registry Documents

Title deeds, ownership, plans All supplied next working day

Ads by Google

GET PROVISER TO HELP ARRANGE YOUR MORTGAGE!

[First Time Buyer](#)
[Home Mover](#)
[Remortgage](#)

[First BTL Property](#)
[Building Your Portfolio](#)
[Remortgage](#)

[Credit Problems](#)
[Self Employed](#)
[High Income Multiple](#)

[High Loan to Value](#)
[Self Certify](#)
[Right to Buy](#)

[Calculators: Max Borrowing \(Res\)](#)
[Max Borrowing \(BTL\)](#)
[Monthly Cost](#)
[Affordability](#)

The table below shows the properties that have sold in W1C. You need to be [LOGGED IN](#) as a [REGISTERED MEMBER](#) to be able to see the actual sale price figures for properties sold less than 2 years ago.

Use the search box below to refine your search. Try entering a road name in the 'Street' box below to filter your results. Also, if your account has been Upgraded you can sort the results by the 'Date Sold' or 'Price' by clicking on the column titles below

Individual Property Prices For Houses and Flats Sold in W1C

Address	Map	New Build	Type	Tenure	Date Sold	Price
12 STRATFORD PLACE CITY OF WESTMINSTER LONDON W1C 1BB	Map		Flat-Mais	Free	15/04/2002	£700,000
12 STRATFORD PLACE CITY OF WESTMINSTER LONDON W1C 1BB	Map	Mr. Captor evaluation copy	Flat-Mais	Free	27/06/2002	£1,025,000
295 OXFORD STREET CITY OF WESTMINSTER LONDON W1C 2DY	Map		Flat-Mais	Lease	23/06/2004	£900,000

Report Error In Price Info

Source acknowledgement: The property prices information on PROVISER is extracted from or derived from information produced by Land Registry

© Crown copyright material is reproduced with the permission of Land Registry. This material was last updated on 31 July 2007. It covers the period 1 April 2000 to 30 June 2007.

a) **Permitted Use.** Viewers of this Information are granted permission to access this Crown copyright material and to download the Crown copyright material onto electronic, magnetic, optical or similar storage media provided that such activities are for private research, study or in-house use only. Any other use of the material requires the formal written permission of Land Registry which can be requested from Land Registry at Lincoln's Inn Fields, London WC2A 3PH.

b) **Restricted Use.** Viewers must not copy, distribute, sell or publish any of the Crown copyright material

By using this site you are confirming that you agree to abide by our [Terms of Use](#)

MEMBER LOGIN

User:

Password:

LOGIN

REGISTER

FORGOT PASSWORD

Ads by Google

[House Price Crash](#)
Tracking the state of the housing market the UK, USA and Australia.
[www.housepricecrash.co](#)

[2007 House Price News](#)
Today's latest news & info on the housing market at Guardian Property
[money.guardian.co.uk/p](#)

[Land Registry Search](#)
Access Land Registry deeds, plans covenants, boundary & owner details
[www.MonkeyMove.com](#)

[Map Postcode Boundaries](#)
Map postcode boundaries as sectors, districts or areas in any GIS
[www.beacon-dedworth.u](#)

TOP - TERMS OF USE & DISCLAIMER - PRIVACY STATEMENT - CONTACT US

© PROVISER 1997-2006 PROVISER is a registered trademark of TSI Consulting Ltd; Reg. in England: 3242182; Reg. Office: 30-38 Hammersmith Broadway, London, W6 7AB

[UK-Car-Discount Ltd](#)
We are a discount new car retailer supplying a wide range of...

[SolveMyDebt](#)
SolveMyDebt can arrange immediate Debt Solutions from Debt M...

[The Insurance Surg](#)
Specialise in providing life and travel insurance to clients...

Figure 5-1: PROVISER Website

The resulted dataset comprises of 50,000 transactions within the study area (see Section 5.5) over a period of 6 years (2000-2006). This represents 60 percent of the registered transactions for this period. The raw data was transformed using a Perl script into a suitable format and stored into comma separated ASCII files to facilitate the geo-referencing process (see Section 5.4). The resulted ASCII files were imported into the intermediate Oracle database schema via Oracle's SQL *Loader.

SQL *Loader runs from DOS prompt and enables the load of large data into an Oracle database.

Ordnance Survey

The Mastermap dataset was provided by the Ordnance Survey in GML format and included the Topographic, Address and ITN layers. Figure 5-2 shows an extract of this dataset. The thematic map was created based on the Description Group and Description Term fields. Data was imported into the Oracle Spatial database schema using the GO Loader software. GO Loader is a software solution that enables the loading of such data into an Oracle Spatial by translating the GML format into that of an Oracle database. The POINT-X data was provided in a point text file format.

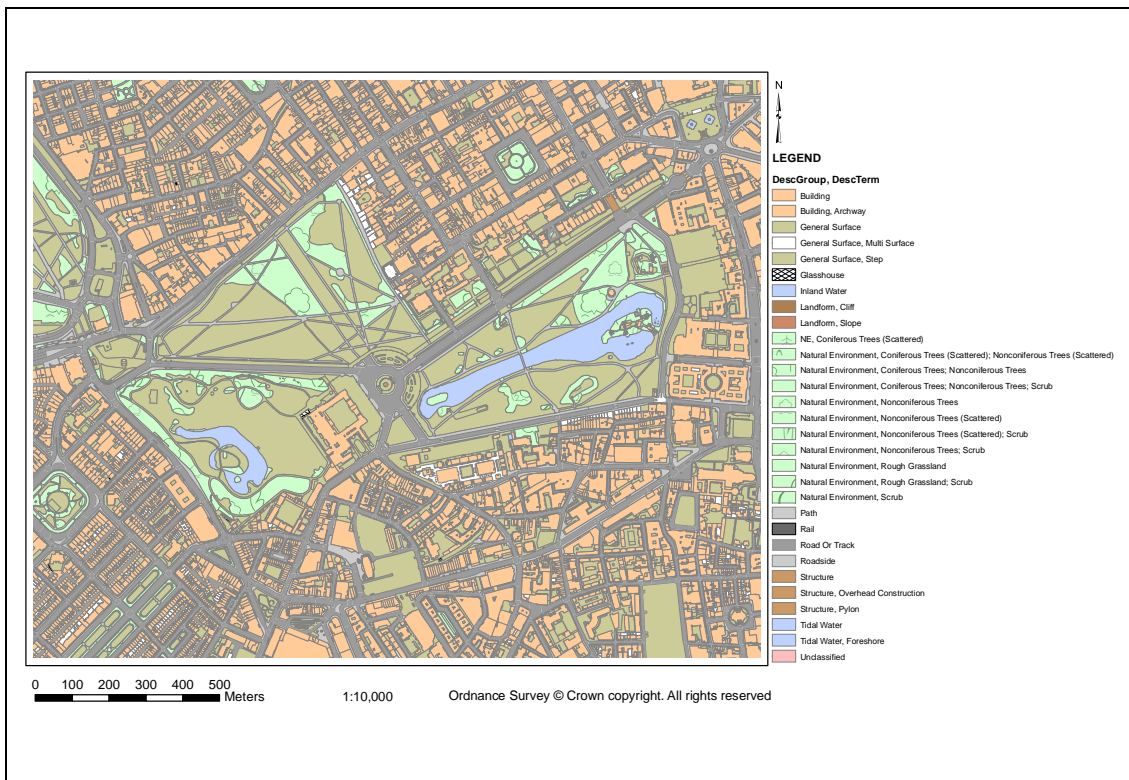


Figure 5-2: OS Mastermap dataset

Cities Revealed

The Cities Revealed dataset (Figure 5-3) was provided in a Mapinfo coverage format and was imported into the intermediate Oracle Spatial database schema through the Easy Loader Mapinfo option.

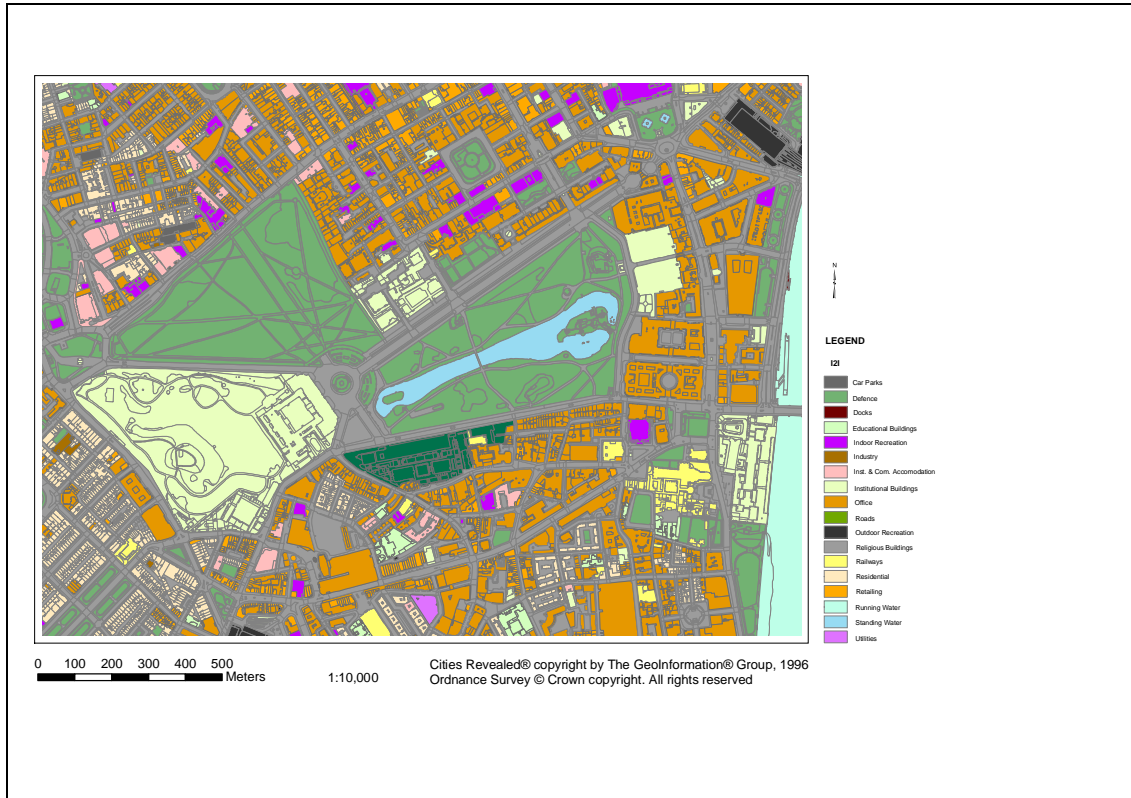


Figure 5-3: I2I Dataset (Landuse Classification)

Census

Census dataset was available in Excel format. The required variables involving the Output Areas within the study area were gathered and then percentages were calculated for each one of them. All the produced variables were imported into the temporal Oracle database schema and stored into a table via Oracle's SQL *Loader.

5.4 Database Implementation

The database implementation involved the preparation of the data and the population of the designed database (see Section 4.4.5). The methodology followed involved six main steps.

1. Examination of the imported datasets and identification of all the existing relationships (Section 5.4.1)
2. Preparation of the datasets

- Setting the level of representation in relation to nodes and edges (Section 5.4.2)
 - Data preparation (Section 5.4.3)
3. Creation of the import tables to populate the final database (Section 5.4.4)

Those steps are further analysed in the following sections.

5.4.1 Description of the Initial Datasets

The OS MasterMap *Topography Layer* is organised in such way so that each feature is represented as a point, a line or a polygon. Point features are stored into the topographic point, cartographic symbol and cartographic text tables. Apart from the first one which refers to topographic details and spot heights, these provide graphic information (e.g. text placement information). Topographic information such as topographic area boundaries and administrative boundaries is represented as lines in the Topographic line table. The topographic area table stores topographic information that is represented as a polygon.

The OS MasterMap *Address Layer* is the replacement of ADDRESSPOINT product and provides a georeference for the GB delivery points in Royal Mail's postcode address file. Address Layer differentiates in a number of aspects. One of the main differences is that in Address Layer the link to the OS MasterMap Topography Layer is explicitly defined through the Topographic Identifier (TOID) of the building the address relates to. Other information in the AddressPoint feature is organised in attributes related to a unique identifier, a postal address, positioning information and quality information for the coordinate.

The *ITN Layer* provides digital information about the road structure coupled with routing information. All public roads and most of the private roads are included in the database. Each road segment is individually represented by road link features that represent the general alignment of the road. Attribution attached to road links includes the road type and nature classifications.

PointX is a national Points of Interest database that provides positioning and descriptive information about features such as shops, schools etc. Each feature is uniquely identified and classified following a three level classification scheme.

Finally, the *I2I* databases from Cities Revealed provide information for the major urban areas across the UK and are organised in 5 layers: Historical aerial photography, Modern high resolution aerial photography, Land Use mapping, Building Class. Relevant to this work are the LandUse and BuildingClass Datasets and both correspond to the OS Mastermap polygons. The LandUse information is classified according to the National Land Use Database (NLUD) v3.3. This version is a hybrid classification that uses both landuse and landcover classes. Building information related to the age and structure is available for the residential buildings.

Figure 5-4 gives a high-level representation of the cardinality of relationships between the main objects in the intermediate schema.

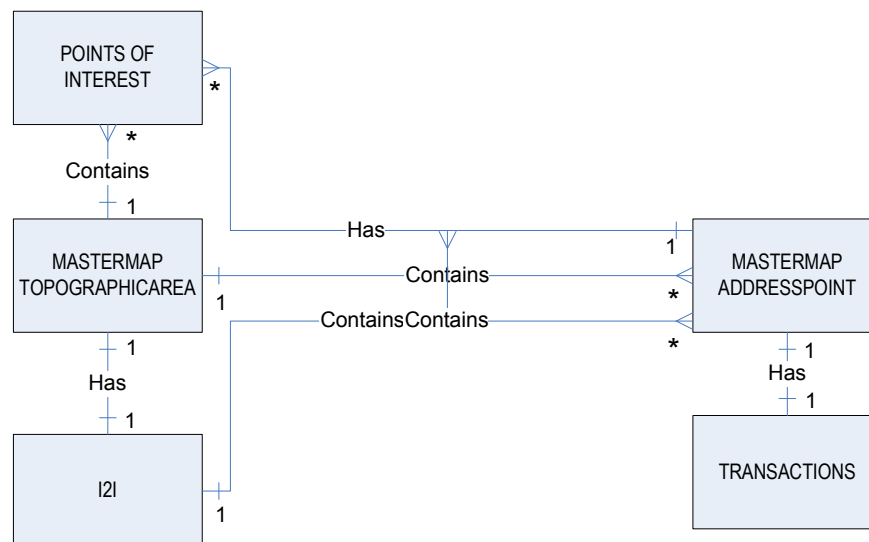


Figure 5-4: Initial Datasets Relationships

Polygons that form the Topographic area feature may *contain* one or more address delivery points. It may also *contain* one or more points of interest. Each polygon *has* a landuse classification. Each Point of Interest may *have* an address delivery point and *belong* to a topographic area polygon. Each address point *belongs* to a Topographic area polygon and to a landuse polygon. Each point may *have* a non-residential activity (POI) and may *have* transaction information. This is further

5.4.2 Level of Representation

One of the first considerations in the implementation of the database was to decide on the appropriate spatial level of representation which would form the basis for all the spatial relationship calculations for the first level spatial objects. The first approach would have been to base everything on a 'point' type representation given that half of the input databases original geometry is that of point geometry. Although this approach would have been less time consuming it was not considered appropriate for the needs of this project since the interest is not only focused on the impact caused from distance type relationships but also on the impact caused from topological relationships.

The reduction of the locational information to x, y coordinates results in the loss of important information such as information about the shape and size. Even in the case of distance calculations this approach is not adequate. Since in this study distance is used to represent the proximity of the known transaction to the different landuses, this proximity would have been wrongly represented (if represented at all). In the case of landuses that have large coverage, for example, in the case of a park such as Hyde Park, the measured distance would have been between the transaction based on the x,y coordinates from AddressPoint and the centroid of the park polygon. This distance is not a realistic representation since properties benefit from the proximity to the boundaries of such an amenity. Therefore, polygon geometry was chosen to model the spatial entities apart from those where size was not an issue (e.g. Bus Stops).

As explained in Section 4.3.1, the graph also consists of a second level of spatial objects that relate to the modelling of the neighbourhood. Similarly, the level of spatial representation of the neighbourhood had to be decided. Defining neighbourhood in terms of its physical boundaries is not straightforward since it can be perceived in different ways. For example, an entirely spatial view of a neighbourhood leads to its definition based on natural boundaries.

Since neighbourhood is considered a key unit in the analysis of small area phenomena, several implementations have been used based on the needs. These include the realisation using the postal and census geography or the physical boundaries or using a combined approach. In this study, since the neighbourhood

variables were sourced from the 2001 Census, the materialisation of the neighbourhood boundaries had to be based on census geography.

Output Areas or in the case of previous Censuses Enumeration Districts, have been commonly used in property analysis studies to denote the neighbourhood. Due to their size, Output Areas resemble more to the common conception of the immediate neighbourhood compared to the wards that cover a larger geographic area. Hence, the 2001 Census Output Area boundaries have been selected to model neighbourhoods and all the calculated variables refer to that level of aggregation.

5.4.3 Data Preparation

Data preparation involved data manipulations at both the spatial and the attribution levels of the modelled objects. Spatial alterations that led to the extraction and synthesis of the required pieces of information from each dataset were necessary. These were dictated by differences in formats and spatial references due to the fact that these datasets were designed to fulfil the needs of different users. Manipulations at the attribution level mainly included reclassification of the attributes into classes that were more meaningful to this project and also, the creation of taxonomies where applicable.

An overview of the integration approach is shown in Figure 5-7. Spatial data manipulations can be grouped into 4 broad types involving tasks such as Generalisation, Data Cleaning, Geo-Referencing, Graph Construction and finally attribute-based manipulations such as creation of taxonomies and other variable calculations. Each of these subtasks is discussed in the following sections.

Generalisation

Mastermap polygons formed the basis for the polygon geometry acquisition. Mastermap layers have a hierarchical structure. Each layer consists of a number of themes that include a number of features. Features correspond to the geographic

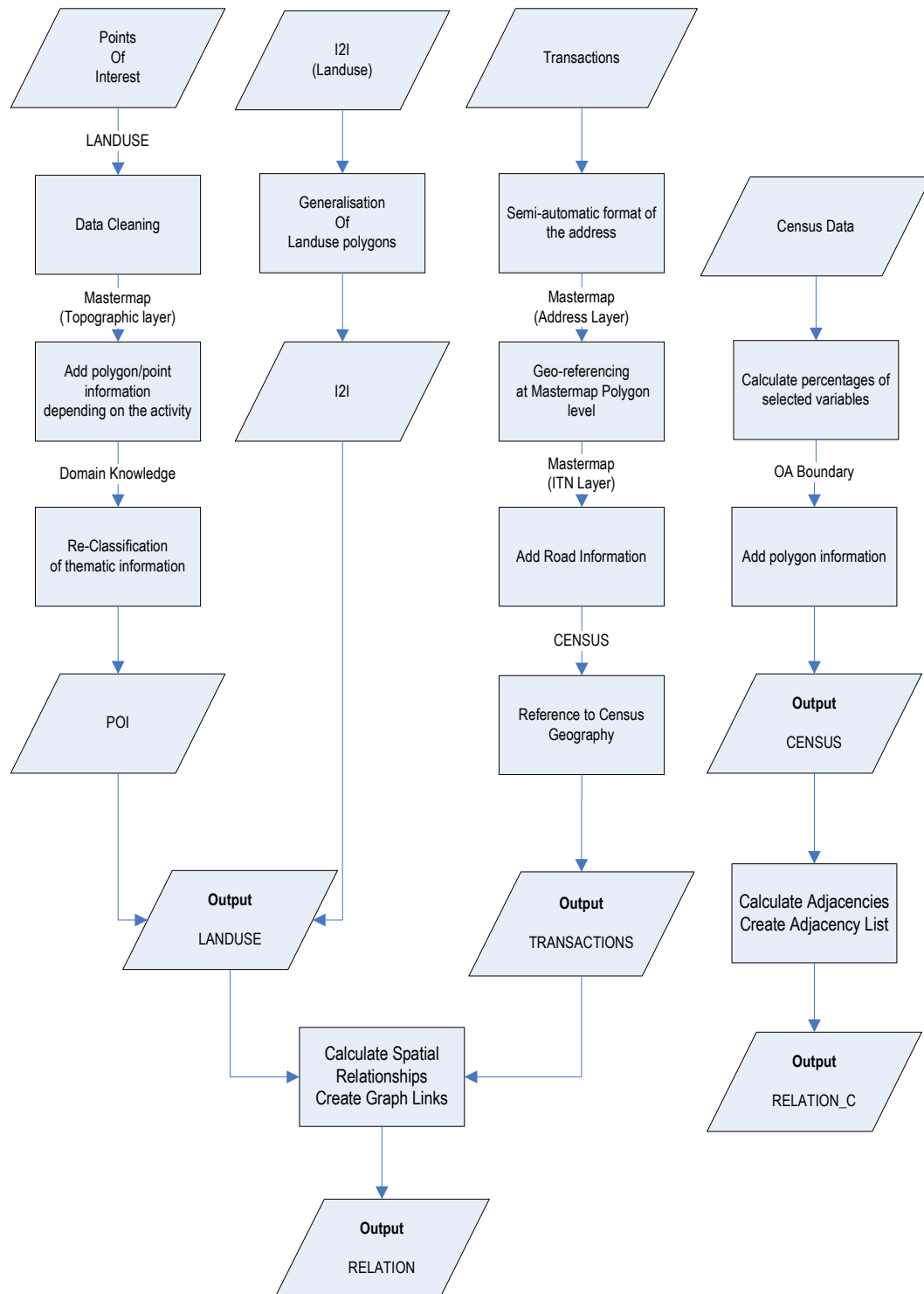


Figure 5-7: Data Integration methodology

entities that can be captured and represented in the data. As a general rule, polygonal features are adjacent in a way that one completes each other as oppose to one being on top of the other. This characteristic introduces problems when the identification of geographic entities that are formed from more than one polygon is required.

To further illustrate this difficulty, the example of park polygon identification is used. As it can be seen in Figure 5-8 (Left image), where two major parks are illustrated, more than one polygons is used to define them. Since there is no explicit landuse information in this product, queries such as ‘Select all parks within study area’ cannot be implemented. One solution could be to base the polygon selection on the attribution associated with each feature. Such attributes include: featureCode, descriptiveGroup, descriptiveTerm, make and theme.

Although this might be effective in some cases, in the majority of the cases and in particular in the case of big parks such as Hyde Park fails. This is mainly due to the fact that attributes contain landcover type of information. In the example of Hyde Park, approximately 1130 polygonal features are used to define it. These represent spatial entities that belong to the building, land, roads tracks and paths, structures and water themes. It is apparent that the identification of a generic type of query that applies in every case is not possible to be achieved.

Therefore, due to the heavily detailed nature of the data a generalization of the datasets was necessary. This was achieved in two steps. The first involved the use of the landuse information provided in the I2I dataset to acquire a more appropriate polygon shape for the landuses. For this, polygons based on a particular landuse type were merged to form one spatial object per landuse type (Figure 5-8). For the merging, Mapinfo Professional (version 7.8) was used. Information was available only at a NLUD level (see Appendix A) hence information for the further classification of the objects was not available.

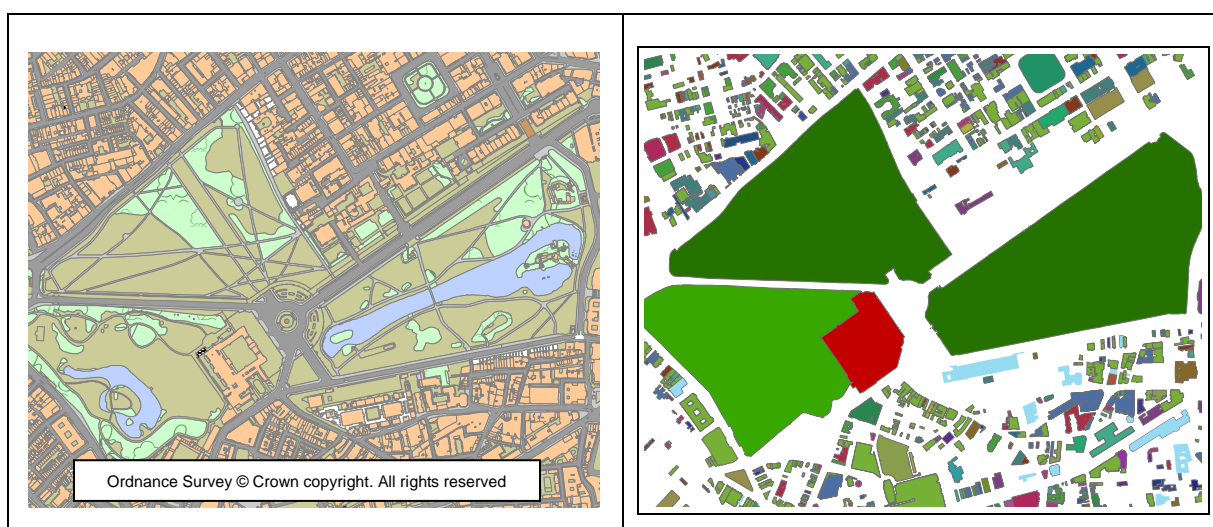


Figure 5-8: Generalisation

Data Cleaning

Data cleaning was necessary in order to avoid duplicates that would result in multiple representations of the same spatial entity. There were two sources of this. The first was the existence of the same activity in more than one Point of Interest categories. For example, London Eye exists in both the Tourists Attractions and the Viewpoints categories. The second was related to the multi-functional places. For example in the case of shopping centres, a point exists under the Shopping Centres And Retail Parks class but also the individual shops within the shopping centres exist as separate points.

The identification of these duplicates was based on an SQL query based on the combined criterion TOID and Name. The action taken, varied upon the case. After surveying the data the possible categories that may be affected have been identified and the appropriate correction applied to Shopping Centres, Parks, Cemeteries, Hospitals, Palaces and Universities.

Geo-referencing

Geo-referencing was a two step process. Firstly, the address was formatted in such a way so that the best possible join result would be achieved. Geo-referencing is performed by calling the stored procedure geoReference (Figure 5-9).

<i>Procedure</i>	<i>Description</i>
geoReference	Associates transactions with TopographicArea and adds polygon geometry (Appendix C)
getRelations	Uses the Oracle spatial operators to create and stores the graph links in the form: from_id, to_id, 'Relation Type' (Appendix C)

Figure 5-9: Implemented PL/SQL procedures

This whole procedure had to be repeated a number of times since differences in the address format between the Land Registry data and the Addresspoint resulted in a quite large number of records without polygon references after the initial join.

Another problem that encountered was the missing polygon reference in the Addresspoint layer.

Spatial Relationships retrieval

Spatial relationships retrieval between the spatial objects was based on a set of spatial operators that Oracle Spatial supports by utilising an R-tree index. An R-tree index is created by the execution of a simple SQL statement on the geometry column of the table. Prior to this, spatial metadata information for the spatial layer must be inserted in the USER_SDO_GEOM_METADATA. Figure 5-10 provides both the spatial indexing and the metadata updating SQL statements for the table landuse.

```
INSERT INTO user_sdo_geom_metadata
(table_name, column_name, diminfo, srid)
VALUES
('LANDUSE', 'GEOMETRY',
MDSYS.SDO_DIM_ARRAY(MDSYS.SDO_DIM_ELEMENT('X', -10000000, 10000000, .001),
MDSYS.SDO_DIM_ELEMENT('Y', -10000000, 10000000, .001)), 81989 *)
```

```
CREATE INDEX lu_sp_idx ON landuse(geometry) INDEXTYPE IS MDSYS.SPATIAL_INDEX;
```

** British National Grid*

Figure 5-10: Spatial Metadata & Indexing SQL Statements

The valuation of the spatial operators is a two stage process which is not open to the user (Kothuri *et al.*, 2004). The first involves the evaluation of the operator by the use of the spatial index (primary filter). Based on the approximations in the index, a potential set of rows that satisfy the conditions of the spatial operator is identified. The identification of the final and correct rows is based on the Geometry Engine (secondary filter).

Table 5-2 shows the two types of spatial operations used in the calculation of the spatial relationships between the spatial objects. These were stored into a table (relation) by calling the stored procedure getRelations (Figure 5-9).

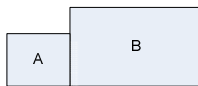

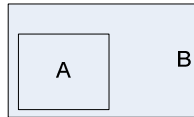
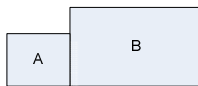

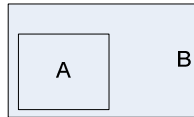
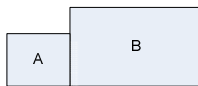

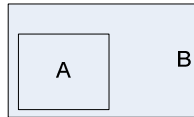
<i>Operator</i>	<i>Description</i>	<i>Parameters</i>						
SDO_RELATE	This operator returns all the objects that satisfy a required relationship/interaction	<div>MASK<table><tr><td>TOUCH</td><td></td></tr><tr><td>EQUAL</td><td></td></tr><tr><td>CONTAINS</td><td></td></tr></table></div>	TOUCH		EQUAL		CONTAINS	
TOUCH								
EQUAL								
CONTAINS								
SDO_WITHIN_DISTANCE	This operator is used for proximity analysis and returns all the objects within a user-specified distance.	<div>DISTANCE UNIT (Optional)</div>						

Table 5-2: Oracle Spatial Operators Used

Taxonomies

As discussed in the previous sections, the level of abstraction of the parameters involved in the association rule mining is very important. Strong rules may not be evident at lower levels but may exist at higher levels or even at cross-levels. For the purposes of this study, two types of taxonomies were developed regarding the attributes of the two types of data involved (reference & task-relevant). Taxonomies were developed partly on the existent hierarchy imposed in the data and partly based on the knowledge domain.

Figure 5-11 shows the three level taxonomy related to the non-residential landuses (task-relevant). For the syntax of the two first levels the Point of Interest classification was followed. In cases considered of important significance landuses were further classified forming a third level of hierarchy. The different pattern indicates the classes that were created either by merging or splitting the original (POI) classes.

Property taxonomy (see Figure 5-12) was based on the information available in the I2I dataset. It also includes taxonomy of the roads since proximity to road is not explicitly modelled in the landuse dataset. In this case the first level classification was based on the natureofroad field of the table Roadlink (ITN layer). The second level is after the descriptiveterm field of the same table.

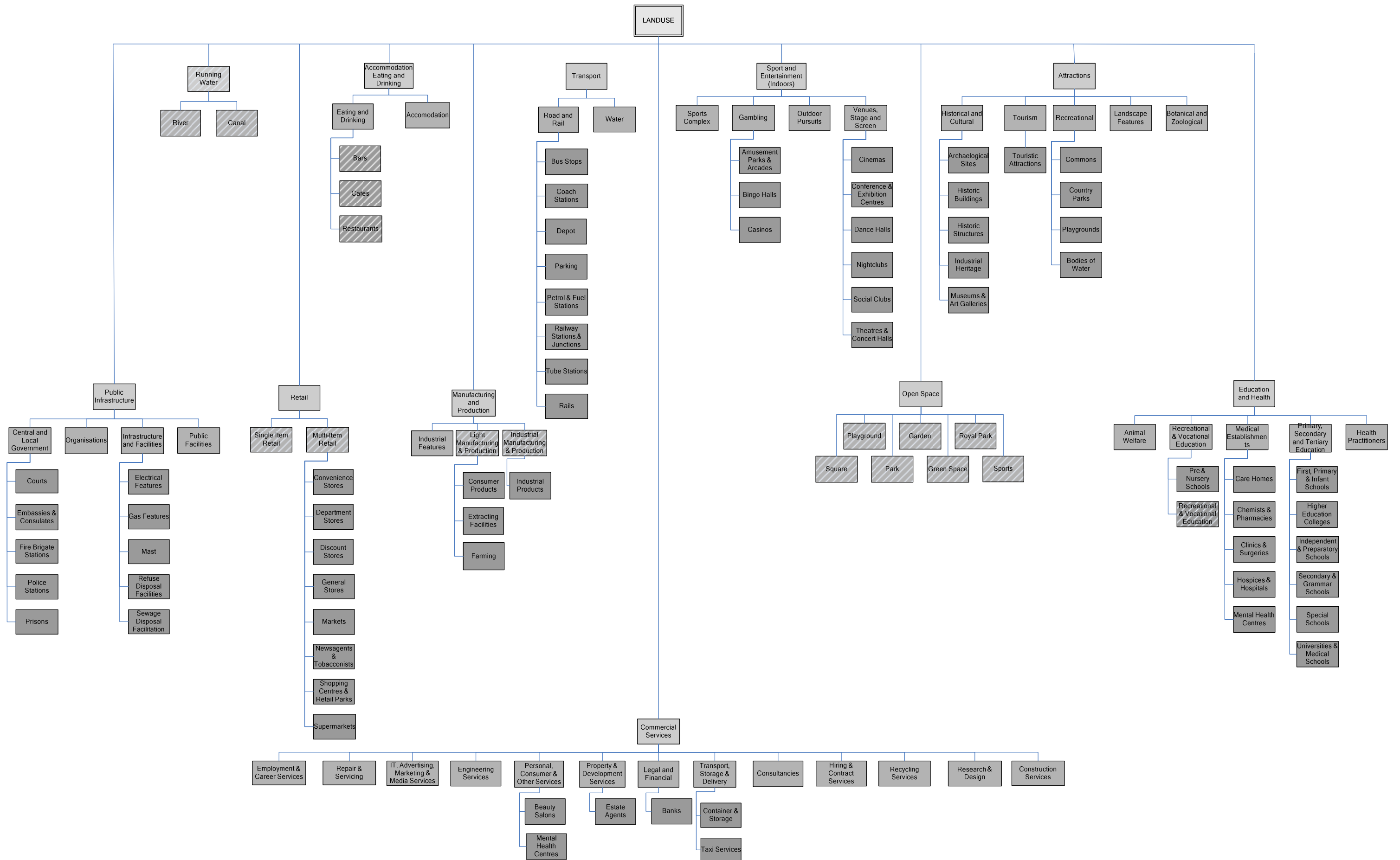


Figure 5-11: Landuse Taxonomy

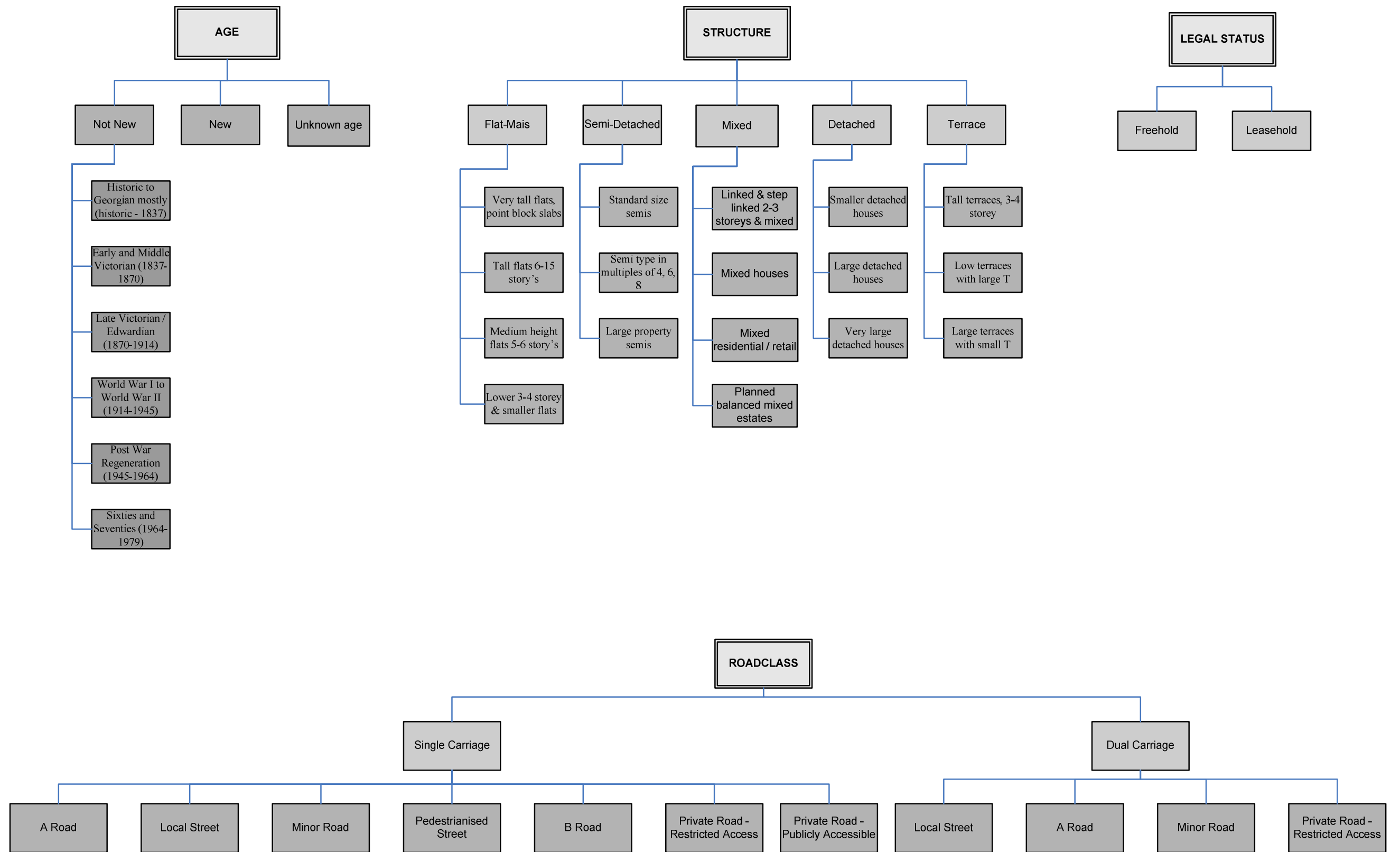


Figure 5-12: Property & Road Taxonomy

2001 Census variables

The identification of the Census variables used in this research was mainly based on two criteria. The first was their ability to give a general description of the area both in terms of location and socio-economic status. The second criterion was based on knowledge already extracted in previous property valuation related studies.

<i>Census Table</i>	<i>Variables</i>
<u>KS002</u> : Contains information about age structure of the OA.	Census variables have been aggregated into variables representing three age groups (Under16, 16-74, Over 74)
<u>KS006</u> : Contains information about the number of people in ethnic groups	Census variables have been aggregated into variables representing five ethnic groups (White, Mixed, Asian/Asian British, Black/Black British, Chinese/Other)
<u>KS013</u> : Contains information about the qualifications	The Census variables that represent the 4 levels of qualifications and the no qualification counts (No Qualifications, Level1, Level2, Level3, Level4, Level4/5)
<u>KS14A</u> : Contains the Census socio-economic classification	The Census variables represent the type of employment (Large employers and higher managerial occupations, Higher professional occupations, Lower managerial and professional occupations, Intermediate occupations, Small employers and own account workers, Lower supervisory and technical occupations, Semi-routine occupations, Routine occupations, Never worked, Long-term unemployed)
<u>KS015</u> : Contains information about the way people travel to work	Census variables have been aggregated into three types (Public Transport, Private, Other)
<u>KS017</u> : Contains information about the households with / without cars	The Census variables have been aggregated into two variables that represent the possession or not of a car (No Cars, With Cars)
<u>KS018</u> : Contains information about the households and the type of tenure	The Census variables have been aggregated to form three variables (Owner Occupied, Rented Local Authority, Rented)
<u>KS019</u> : Contains information about the amenities available in households	Census variables have been aggregated to two variables that represent the existence or not of central heating (With Central Heating / Without Central Heating)
<u>UV008</u> : Contains information about resident's country of birth	The Census variables have been aggregated into six new variables (Europe, Africa, Asia, N. America, S. America, Other)

Table 5-3: 2001 Census Variables

Table 5-3 shows the Census tables and the final variables chosen to reflect the neighborhood quality followed by a short description.

5.4.4 Population of the database

In Section 4.4.5.2, the two main components of the data model, graph and descriptive, were presented. Graph components relate to the graph structure while the descriptive components are used to keep additional descriptive information.

Graph-related tables can be created either manually or automatically by calling the stored procedure `CREATE_<TYPE>_NETWORK`. This procedure creates all the basic structure for the required type of network. It also updates the network metadata (`USER_SDO_NETWORK_METADATA`) based on the parameters entered when the procedure is invoked. It is this that defines the network in terms of type and structure. Due to this, multiple networks can be defined based on the same node and link tables.

The Census, Landuse_Taxonomy and Property_Taxonomy tables have been created according to the logical design of the datatabase. The Census table holds processed information derived from the 2001 Census. The other two tables hold additional information about the nodes (taxonomy). Figure 5-13 shows how the resulted tables produced in the integration procedure, were used in the population of the database tables.

The node table (`GHU_NODE`) holds information about the two level nodes. The first level nodes are of two types: nodes that represent polygons where the transactional information is available (reference spatial entities) and nodes that represent the aggregated landuse polygons (task-relevant entities). The second level nodes represent the 2001 Census Output Area polygons.

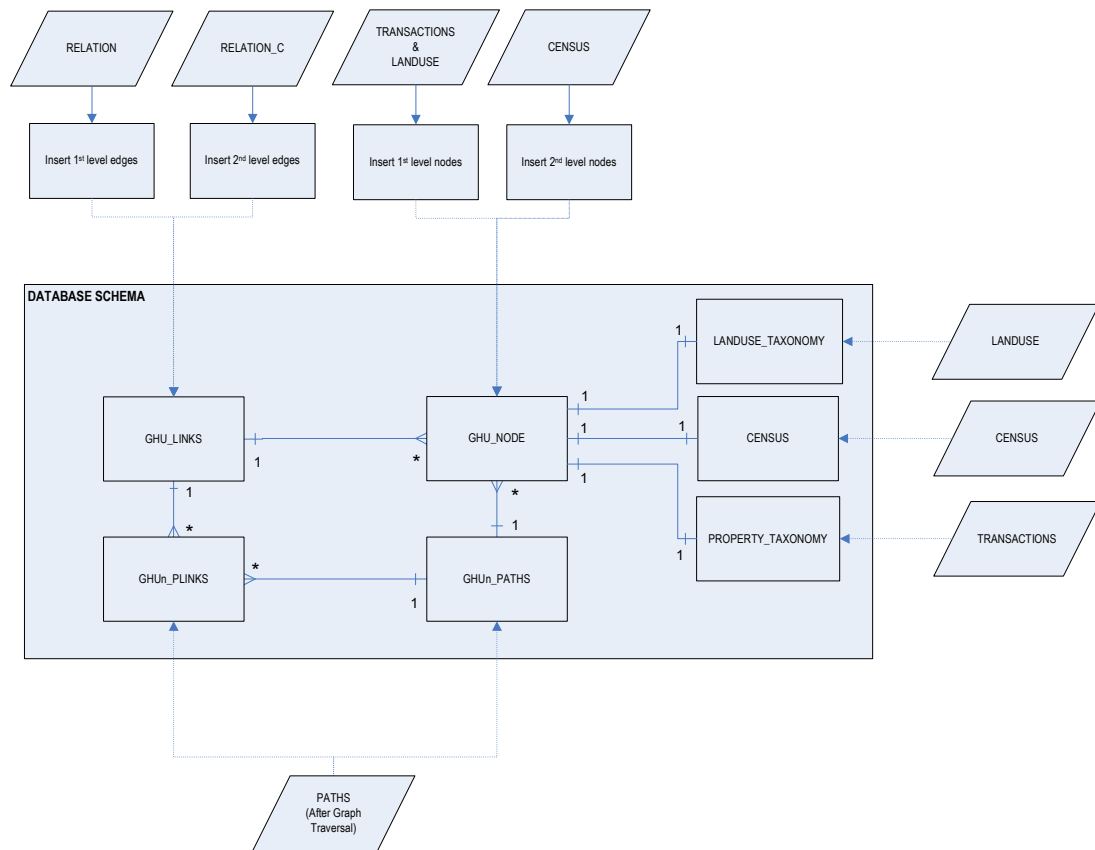


Figure 5-13: Database schema population

The link table (GHU_LINKS) is populated where one of the relevant spatial relationships holds between the nodes. Again, these are of two types: relationships between the first level nodes and relationships between the second level nodes. As described in Section 4.3.1 the examined relationships are: adjacency, containment and proximity. For the first level nodes the adjacency, containment and proximity has been calculated to denote the spatial arrangement (see Section 4.3.1). It should be noted that the containment relationship in most of the cases means that the property is located above a certain activity. Because of the way the Mastermap building representation was captured, polygon boundaries are not necessarily coincide with distinct buildings. Hence, in certain cases, it can also represent adjacency.

For the second level nodes the adjacency relationship has been calculated. In this case, adjacency is enough since the way Output Areas have been designed is to ensure continuity.

In Figure 5-14 an extract of the realised graph for the case study area is illustrated. At the left the 1st order relationships of one reference point with the immediate

connections to the non-residential landuses is shown. On the right all the nodes that have first & second order relationship with the specific reference point are presented. For the visualisation of the graph the Ucinet software was used.

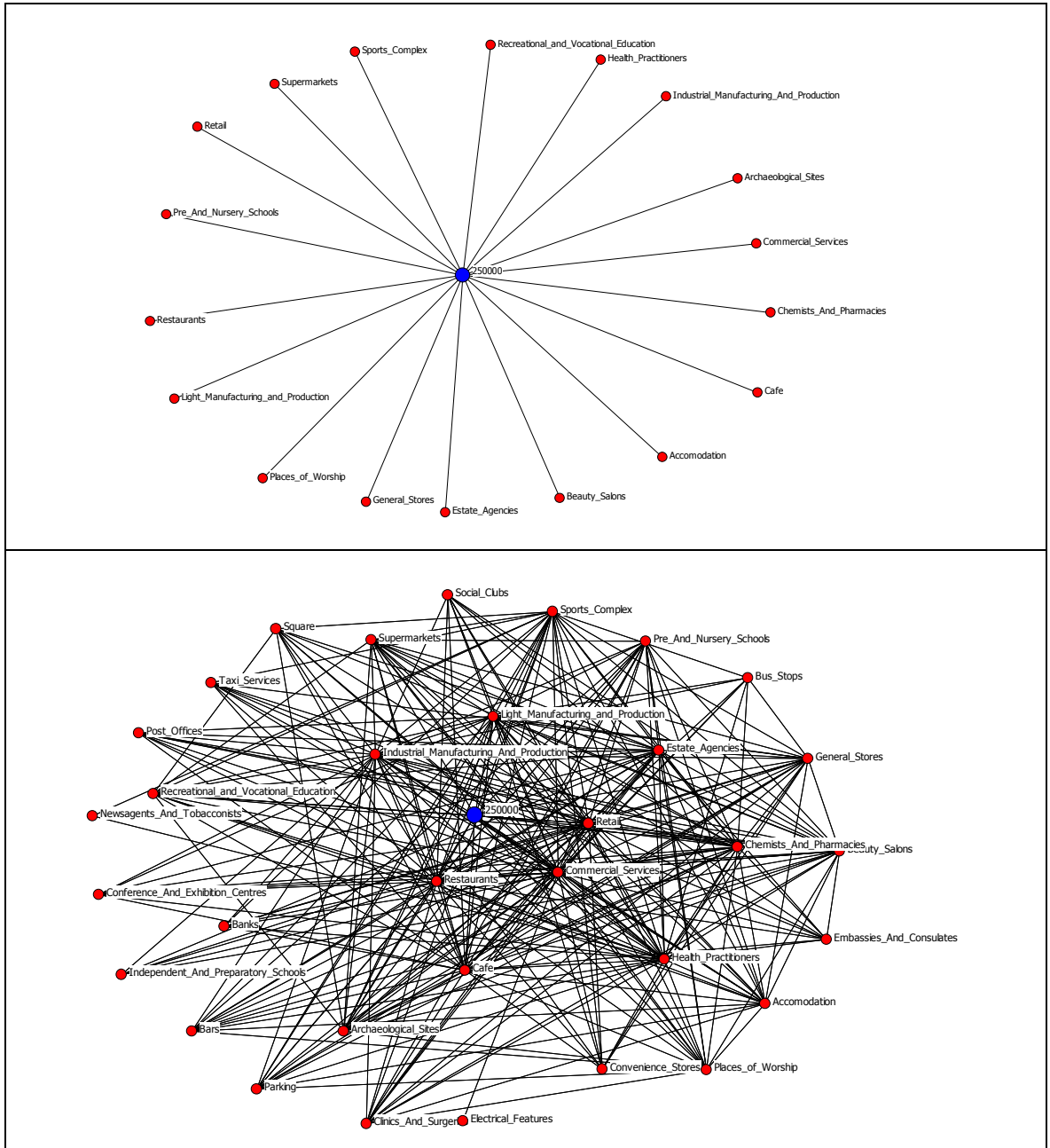


Figure 5-14: 1st & 2nd Order Spatial Relationship Graph

5.5 Study Area

The study area selected for the purposes of the implementation and testing of the proposed system comprises of three London boroughs: Hammersmith & Fulham, Kensington & Chelsea and the City of Westminster (Figure 5-15). These boroughs are centrally located within the wider London area and comprise the core of the London city centre.

This choice was based on a number of criteria that mainly related either to the requirements imposed by the proposed methodology or data availability constraints. Data mining algorithms are data hungry and for their successful application a reasonable database size is required. Additionally, as association rule mining is a pattern recognition technique it had to be ensured that the selected area had increased chances of including repetitive and diverse associations. Hence the area of London was chosen since it complies with the two main requirements of the method: volume and diversity.

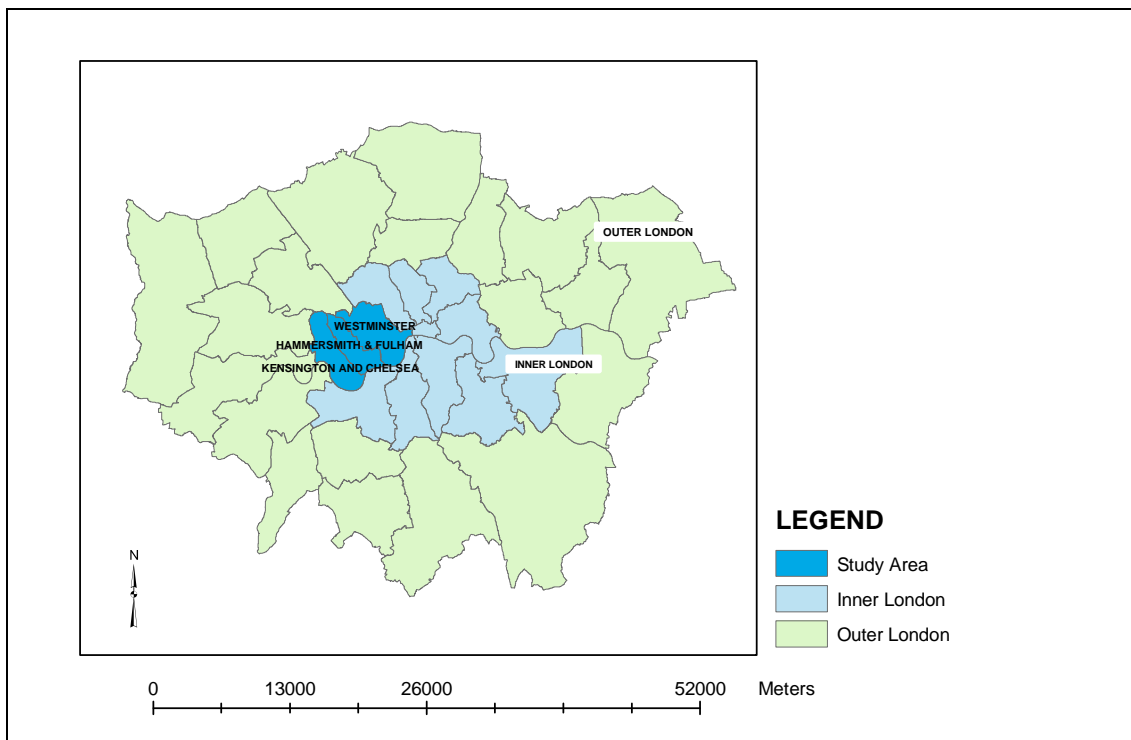


Figure 5-15: Study Area

As this research progressed the initial study area had to be restricted to three boroughs. Data availability and also time constraints were the main reasons for this reduction. Although the main topographical background of the whole area enclosed

by the M25 was available, the detailed information for the various landuses inferred from the POINT-X dataset was not available. Due to its commercial value and use, the POINT-X dataset license was not available for this project for the whole London area. Furthermore the collection and positioning of the transactional data for such a wide area would have been hugely time consuming without necessarily adding any value to the research.

For all the above reasons the final study area was defined as the area enclosed by the Administrative Boundaries of the boroughs of Hammersmith & Fulham, Kensington & Chelsea and the City of Westminster. A brief description of these areas in respect with their physical characteristics and also their housing stock follows.

Hammersmith & Fulham

Hammersmith & Fulham is located on the west side of the Inner Area of London and its size approximates 17.2 km². Its population, according to the most recent census (Census 2001), is 165,242 and a majority of this belongs to the age group of 30-44 years old. Figure 5-16, shows the ethnic composition of borough based on the ethnic group table from 2001 Census.

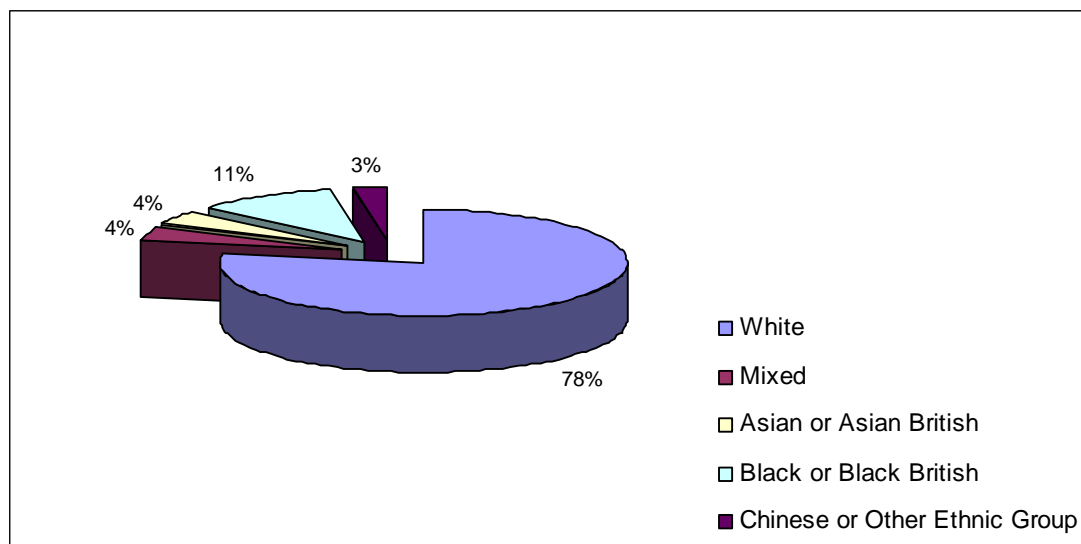


Figure 5-16: Ethnic Composition (Hammersmith & Fulham)

With its industrial past declining the Hammersmith and Fulham borough became a major sub-regional office location (Local Economy, 2006). That was helped by the fact that it is traversed by the A4 – a major artery of the Central London.

Similarly to the other two boroughs of the study, it is dominated by flat type accommodation (Figure 5-17, based on the Household Spaces and Accommodation Type from 2001 Census: Key Statistics) but the percentage is low compared to Westminster and Kensington & Chelsea.

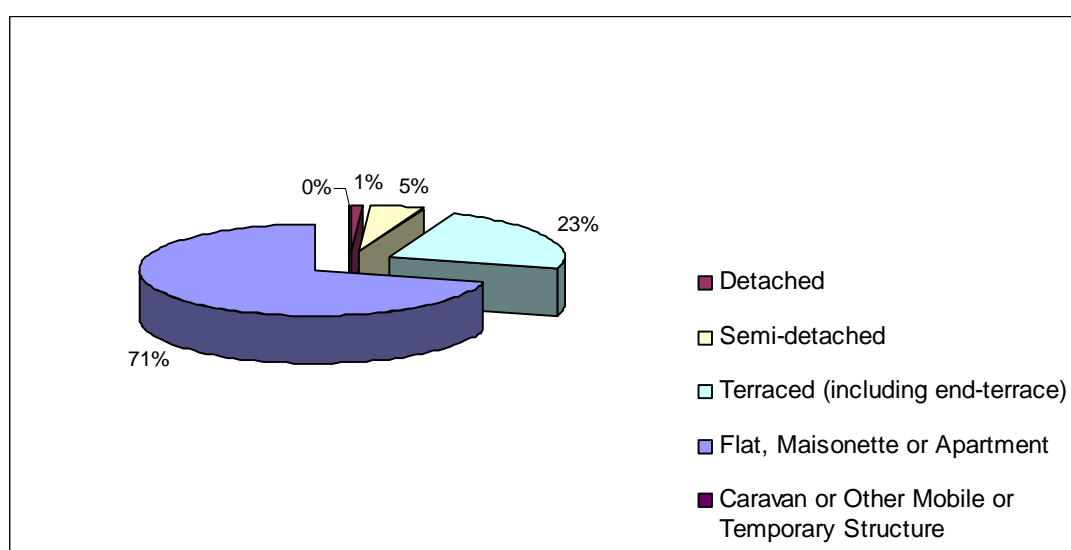


Figure 5-17: Accommodation Type (Hammersmith & Fulham)

Registered transactions for the five year period between January 2000 and December 2006 by Land registry were 22,860 transactions.

Kensington & Chelsea

Kensington and Chelsea is located within the Inner London area and shares a common border with Westminster on its west side. The total population according to the 2001 Census is 158,919. Similarly to Westminster a majority of this population belongs to the age group of 30-44. The ethnic composition of the borough, based on the ethnic group table from 2001 Census, is shown in Figure 5-18.

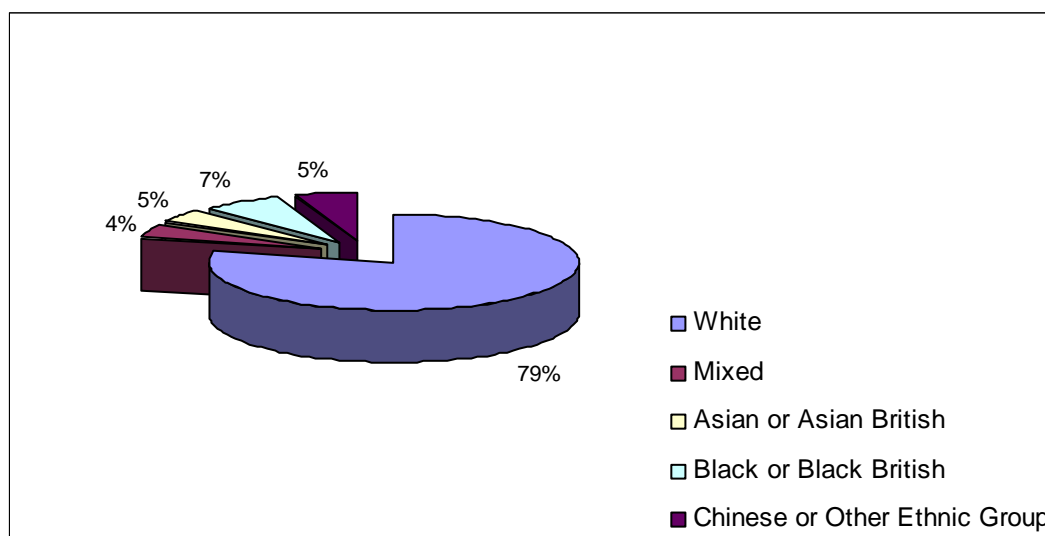


Figure 5-18: Ethnic Composition (Kensington and Chelsea)

Kensington and Chelsea is one of the wealthiest boroughs, a prime residential area with a prestigious past. It benefits from its proximity to the centre, yet it is quite secluded. It contains big parks such as Holland Park and other open space areas, museums, universities and exclusive retail market.

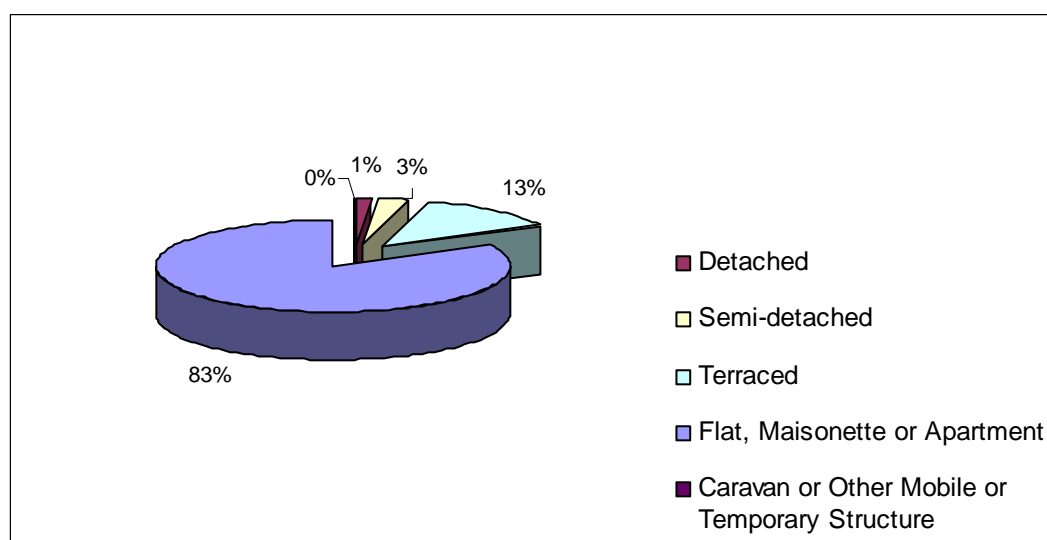


Figure 5-19: Accommodation Type (Kensington & Chelsea)

Figure 5-19 (based on the Household Spaces and Accommodation Type from 2001 Census: Key Statistics), shows the prevalence of the flats as an accommodation type with second the Semi-detached houses. In the period between January 2000 and December 2006 the Land Registry registered 24,655 transactions in this borough.

City of Westminster

The City of Westminster borough covers approximately an area of 22 km² and is located within the Inner area of London. It is considered one of the most densely populated boroughs in the UK. Its population according to the 2001 Census is 181,286 and a majority of this belongs to the 30-44 age group.

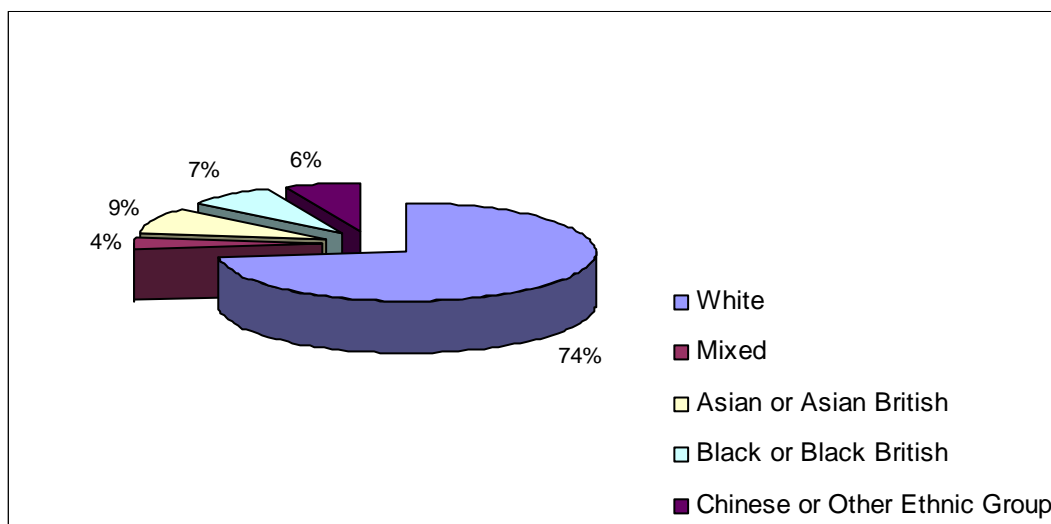


Figure 5-20: Ethnic Composition (Westminster)

Figure 5-20 illustrates the ethnic composition of the borough based on the ethnic group table from 2001 Census: Key Statistics. Diversity characterises this borough at both ethnic and cultural level. This is reflected in the fact that Westminster although overall is quite prestigious, it includes some of the most deprived areas in the UK.

Westminster is unique in the sense that includes a great number of famous landmarks including Buckingham Palace, Westminster Abbey, Big Ben and major parks such as Hyde Park and Green Park. It also has some of London's main gateways such as Paddington and Charring Cross. It is also the base of a number of Universities and Colleges.

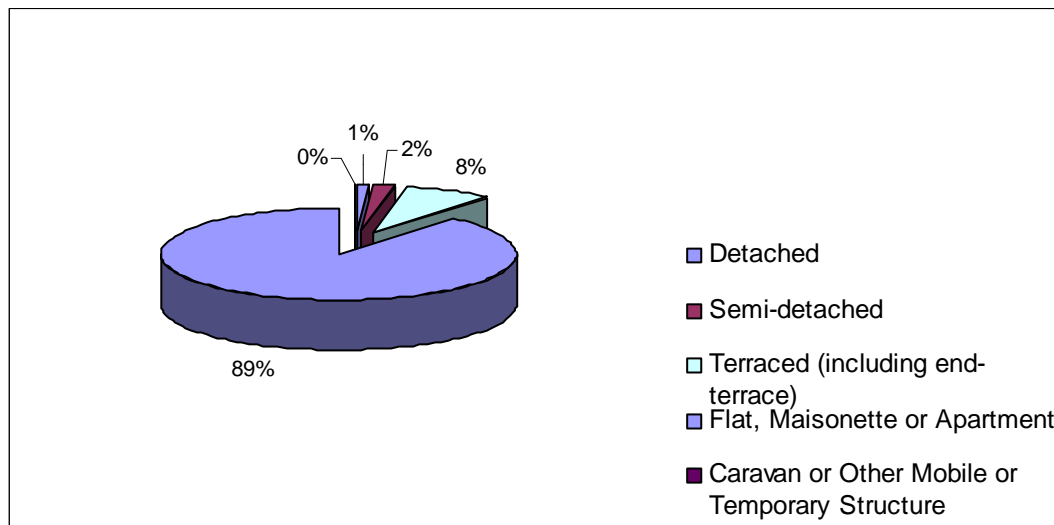


Figure 5-21: Accommodation Type (Westminster)

Flats consist the most prevalent type of accommodation within the borough (Figure 5-21 based on the Household Spaces and Accommodation Type from 2001 Census: Key Statistics). According to the land Registry records there were 34,380 transactions registered for the period between January 2000 and December 2006.

5.6 Summary

For the implementation of the system Java 1.6.0 was selected and for the database management Oracle 10g. For the purposes of development and test of the spatial data mining methodology followed in this project, a database was created that formed the test environment.

The data requirements set in the previous chapter guided the process towards the dataset identification. For the modeling of the spatial relationships the OS MASTERMAP, POINTX Points of Interest, Cities Revealed and Census 2001 datasets were used. To complete the picture in terms of non-residential land-uses, supplementary information from two other datasets were used. These were the Point of Interest and Cities Revealed datasets. Property price information and additional information was based on Land Registry transactional data.

After the data was sourced and imported in a temporary schema, a data preparation methodology has been followed. For the modelling of the spatial relationships a

polygon geometry was chosen and the neighbourhoods were represented by the Census OAs. The database was realised for three London boroughs: Westminster, Kensington and Chelsea and Hammersmith and Fulham.

6 Case Study

In the two previous chapters, a detailed account of the implemented system and the data used for this research has been given. Here, the focus is on the rationale for the case study and also on the presentation of the results. In the first part, an account on the whole knowledge discovery process is provided as it is used in this project and acts as a bridge to the previous chapters. The introductory section is followed by the description of the design of the experiments as well as some initial observations regarding the optimal parameter configuration of the data mining algorithm. In the concluding section, the case study results are presented and analysed.

6.1 KD Process

As already discussed in Section 2.1, dealing with data mining techniques outside the whole knowledge discovery framework may lead to undesired and erroneous results. Although data mining is associated with automation, since it involves the analysis of vast amounts of data, the role of the analyst is of equal importance. Analysts have a leading role in the whole process by making decisions regarding issues relating to data preparation, interpretation and presentation. Hence, their role is to coordinate the whole procedure by utilising the appropriate tools for the completion of each of the knowledge discovery steps.

Especially in the case of spatial data, where diversity is one of the main characteristics, its representation in a fully automated knowledge discovery system is difficult to accomplish. Data preparation of spatial data is case specific and involves

the execution of a number of tasks beyond the common data preparation tasks such as discretisation.

Before proceeding to the testing of the methodology, an overview of how the knowledge discovery process has been implemented is presented. To assist this, Figure 6-1 illustrates the knowledge discovery process that has been carried out throughout this research. It also shows how the implemented platform fits within this process (highlighted process steps).

Application Domain

This phase (Step 1 in KDD process) involves the understanding of the application domain and the identification and setting of the procedure. This step was crucial and resulted in stating the problem and in the identification of the goals. Since the broad application area (Land Management) was known beforehand, to meet the requirements of this step a literature review has been carried out (see Chapter 3). This leads to the determination of the aim of the data mining process and involves the modelling and incorporation of location into the valuation model.

Data Selection

Initially, a number of potential data sources were identified and based on availability and other limitations the final selection of the initial datasets was made (see Section 5.3). Once the data was sourced, the second phase of this step involved the auditing of the datasets. This required the examination and exploration of the acquired data in terms of fields, format types and existing relationships (see Section 5.4.1). Furthermore this step dealt with data representation issues such as level of representation (see Section 5.4.2)

Cleaning / Pre-processing

As illustrated, initial datasets from the previous step were anticipated to be large, heterogeneous and incomplete. This step was quite time consuming, fact which is quite common in this type of projects. Pyle (1999) states that it can require up to 60% of the effort in the whole knowledge discovery process and that, was also confirmed in this project.

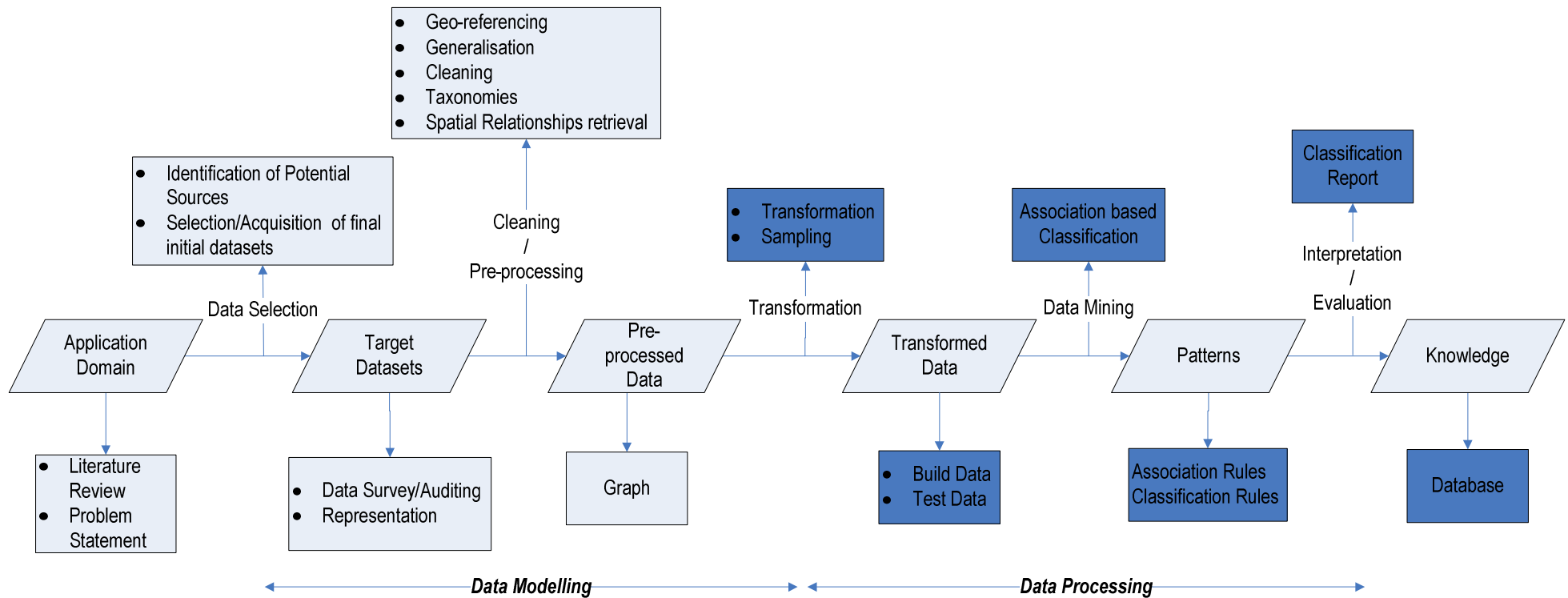


Figure 6-1: Knowledge Discovery process

This step, apart from tasks such as removal of noise, handling of outliers in the data and dealing with the incomplete data, that are common to data mining processes in general, it also involved a number of other tasks. These were related to the nature of the specific data involved in the study and included: Geo-referencing, Generalisation, Cleaning, Taxonomies, Spatial Relationships retrieval (see Section 5.4.3). This phase was finalised with the correction of mistakes and the data coding where necessary. It is apparent that all these procedures were closely related to the chosen data mining algorithm, that is the association rule mining procedure described in Section 2.1.3.1.

Data transformation

This is the second phase of data preparation and involves the reduction of the dimensionality of the datasets. Here, the final number of variables that will be used to represent the data will be defined in order to avoid invariant representation of the data and also data split to form the build and test data. This step is included in the implemented system where the user is responsible for the inclusion (or not) of a number of available parameters by specifying the ones of interest in the system's configuration file.

Data mining

The data mining engine is a core component of the implemented system (see Chapter 4). The user has access to the tuning and parameter configuration of the algorithm through a configuration file. This step is further explored in the next section in relation to the experiments.

Interpretation and Evaluation

The results (Association Rules or Classification results) are reported in the form of a text file (see Appendix B). Each file is named according to the area which refers to, followed by a code that corresponds to the experiment guide (Figure 6-3). It consists of seven parts: data transformation, build model, display classification rules, sort classification rules, print classifier, classify and display accuracy. The two first parts display information related to the data transformations and the set-up of the association rule model. The following two parts display the mined classification rules without and with the CBA sorting scheme. The classifier is displayed and consists of

the classification rules selected from CBA and also the default class based on the majority class of the unclassified training data. Finally, details about the classification including the accuracy conclude the output file.

6.2 Design of Experiments

The experiments have been designed in such way that put into test the two main steps of the data mining component: the association (-classification) rules builder and the classifier in relation to their application to the property valuation area. Additionally, some initial tests based on a sample have been carried out to assist in initial decisions.

The first step in the designing process involved the identification of all the parameters that have an active role in the process. In the identified parameters two main categories can be distinguished. The first includes parameters that relate to the input and mainly relate to the identification of the appropriate level of detail needed in the analysis. The second set of parameters is the tuning parameters. The tuning parameters are algorithm specific and relate to the tuning of Apriori algorithm that affects the accuracy of the algorithm.

Figure 6-2 shows the key parameters in a diagrammatic form. The aspects that have a direct affect on and hence will form the evaluation are also shown. The input related parameters are of two types. There are parameters that affect the size of the sample and parameters that affect the dimensionality. Examples include the geographical level of the analysis and the path length respectively. The algorithm specific factors are also of two types, those that are explicitly related to the tuning of the algorithm and those that relate indirectly and involve the data transformation.

Before carrying on with the main experiments, to investigate the optimum parameters in order to have an accurate classifier based on valid classification rules, a number of initial tests had to be performed. These were necessary in order to assist with decisions that had to be made about the parameters that control the input property-related dataset.

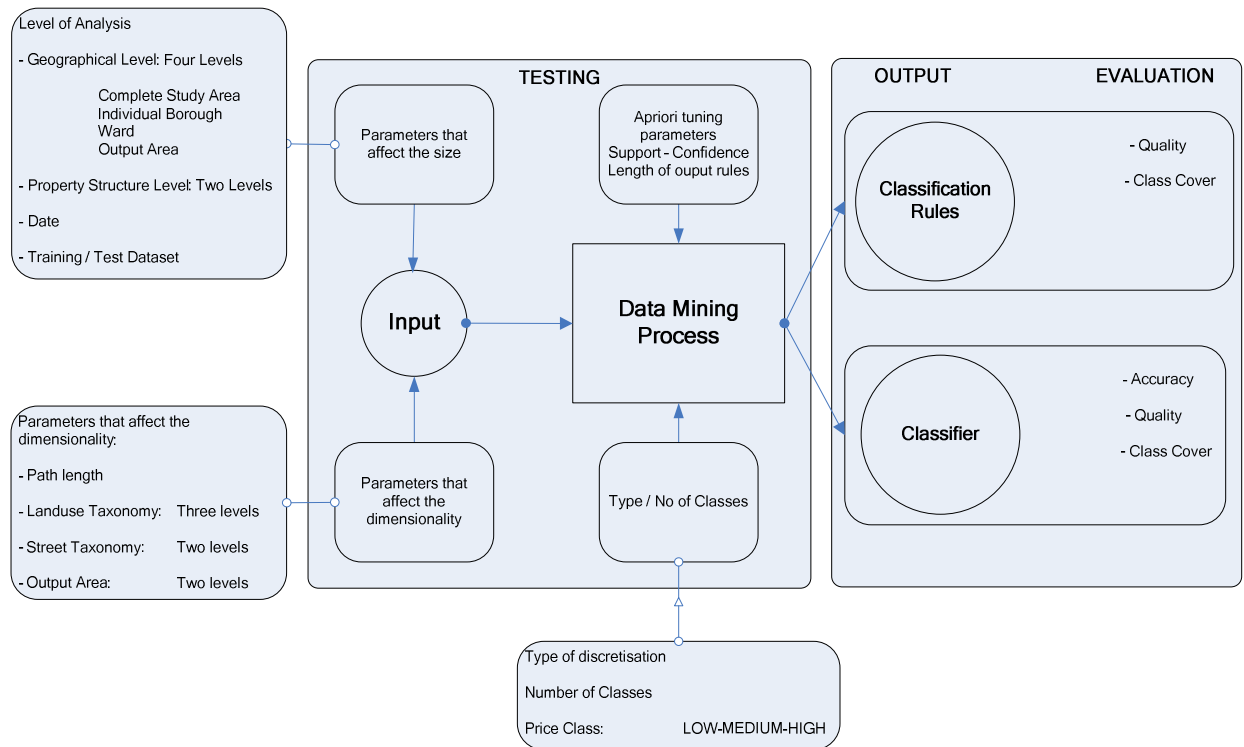


Figure 6-2: Study Parameters

The first was whether the study should be performed on the complete sample irrespective of the transaction date. The study involves transactions that spread over a period of 6 years. The property market in the area of central London is highly active hence assuming stability over such a long period is not suitable. Nevertheless, including the entire available sample for a given area may still produce some meaningful results. Therefore, although in general both the tests and the case study prices were treated per year some tests were performed based on the whole available sample.

The second, related to the nature of the building stock in central London that resulted in Mastermap polygons (which were the spatial reference for the properties) with more than one different transaction per year. Keeping these transactions in the study could introduce bias towards these polygons given the nature of the method. On the other hand, by making the assumption that the dataset comprises from all the registered transactions at a certain period, keeping them is justified by the fact that they reflect a favourite trend regarding a specific polygon (building). This could be

the result of the existence of unique structural or locational characteristics or it may just reflect availability at a given period.

In order to tackle the second issue, it was decided to perform key tests using both cases. The first case was to use a dataset where polygons were unique and in the cases of more than one transactions per year, their average price was used. The second case was to take everything into consideration and see which of the two approaches gives better logical results.

The final consideration involved the discretisation method and the number of classes the data would be classified into. Both decisions have a direct impact on the success of the output and the methodology.

Figure 6-3 gives a diagram that illustrates all the possible combinations that can be used to test the method. It is apparent that there is an extremely large number of experiments reflecting different parameter combinations that can be performed but not all of them significantly contribute to the analysis or significantly differ from each other. Therefore, in order to decide which combinations would be used in the demonstrative case study analysis to maximise the possibilities of meaningful results and minimising the number of redundant experiments and also to tackle the above issues it was decided to perform the majority of all the possible combinations for a dataset subset based on its outcome the pruning of the tests would be performed. The tests were named after this guide and denote the parameters included in the model.

For the creation of the test dataset, the year constraint have been used which resulted in a dataset of all the available transactions that relate to year 2004. As mentioned above, to investigate the impact of the existence of multiple transactions in the same polygon (Block of Flats) a second test dataset has been created in which transactions have been aggregated at Mastermap polygon level by using the average price. The first dataset comprised 9104 records while the second 5930 records.

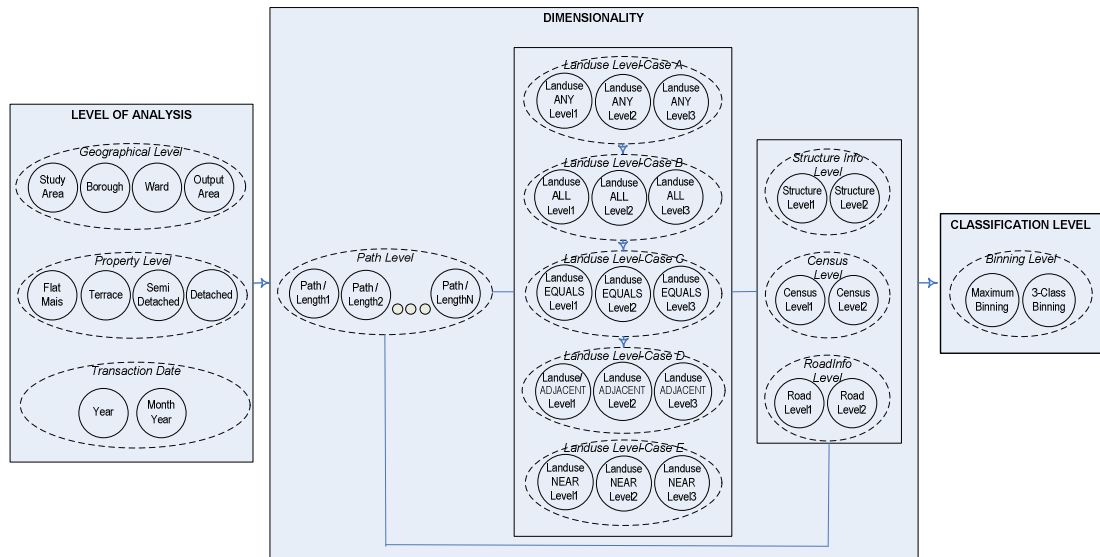


Figure 6-3: Experiment Guide

The parameters that control the Apriori algorithm, such as the confidence and support thresholds, have been initially set in such way as to enable the capture of the best possible result, irrespective of computational cost. Based on the results the user can further select the optimum settings for each level of analysis. As a starting point the threshold 10% for support and confidence have been used. Based on this, a rule of length n is valid when at least 10% of the records in the dataset contain both its antecedent and consequent parts and at least 10% of the records in the database when contain its antecedent also contain its consequent. The rule generation has been additionally constrained by setting a maximum length of 20 items per rule.

Unique polygons vs. All transactions

A number of experiments, at different geographical levels, have been performed to examine the impact of the multiple occupancy in the outcome. In Table 6-1 the performed experiments are presented. The “u” in the models name denotes that it was based on the unique polygons sample.

The outcome of this investigation varied upon the geographical levels of the analysis. In the case of the Borough level, the results achieved by using the unique polygon and the multiple polygon approach presented no particular differences. Ward level tests demonstrated better results when the aggregated version was used. In particular, it resulted in better class distribution coverage compared to the multiple polygon

approach. When limiting the testing into the OA level the multiple polygon approach performs better and this can be attributed in the radical reduction of the sample size.

Input Dataset Description		Classification Rules			Comments
Model_Name	No of Records	No of Rules	Rule Type	Class Cover	
SA_Flat_P11_Lall1_3Class_2004	9104	229	Generic	3 / 3	-
SAu_Flat_P11_Lall1_3Class_2004	5930	180	Generic	3 / 3	Not important changes compared to SA_Type_P11_Lall1_3Class
Westminster_Flat_P11_Lall1_3Class_2004	4069	227	Generic	3 / 3	Interesting results
Westminsteru_Flat_P11_Lall1_3Class_2004	2192	198	Generic	3 / 3	Not important changes compared to Borough_Type_P11_Lall1_3Class
BKGG_Flat_P11_Lall1_3Class_2004	474	198	Mixed	3 / 3	-
BKGG_Flat_P11_Lall2_3Class_2004	474	1000	Not Generic	2 / 3	Most of the coverage is for Class 1
BKGG_Flat_P11_Lall3_3Class_2004	474	1000	Not Generic	3 / 3	Most of the coverage is for Class 1
BKGGu_Flat_P11_Lall1_3Class_2004	167	146	Mixed	3 / 3	Better Class Distribution Coverage (Compared to Ward_Type_P11_Lall1_3Class)
BKGGu_Flat_P11_Lall2_3Class_2004	167	1011	Not Generic	3 / 3	Better Class Distribution Coverage (Compared to Ward_Type_P12_Lall1_3Class)
BKGGu_Flat_P11_Lall3_3Class_2004	167	700	Not Generic	3 / 3	Better Class Distribution Coverage (Compared to Ward_Type_P12_Lall1_3Class)
00BKGG0007_Flat_P11_3Class_2004	4	5628	Mixed	3 / 3	Better distribution compared to the 'u' version
00BKGG0007u_Flat_P11_3Class_2004	2	8446	Mixed	2 / 3	-

Table 6-1: Unique vs. All transactions experiments

Number of Classes

As mentioned in the previous chapter, one of the limitations that association rule mining present is the poor handling of numeric data. To overcome this limitation a discretisation process must be applied prior to the mining procedure. This introduces two main considerations. The first is the discretisation method applied and the second regards the number of classes. Both directly affect the quality of the outcome in terms of accuracy.

The methods tested were the *equal-interval (equiwidth)* binning and the quantile binning. In the first method, which is based on the bin size, the division of the whole data range is performed in a user specified number of ranges of equal size. In the second approach, each of the ranges are not equal and they are defined in such way that include equal number of data values.

For example, when the equal-interval binning is applied to a sample of the property dataset where the prices range from 60,000 to 5,350,000 and 3 bins are required, the resulted ranges are: [60,000 – 1,823,334), [1,823,334 – 3,586,667) and [3,586,667 –

5,350,000]. The percentage of records that belong in each of these classes is 95.06%, 3.85% and 1.09% respectively. In the case of the quantile binning, for the same sample and number of bins, the classes are: [60,000 – 270,000), [270,500 – 490,000] and (490,000 – 5,350,000]. The ranges in this case are not of equal width and are adjusted in such way that ensures similar number of cases within each one of them. The percentages of the cases within each class are 33.4%, 33.89% and 32.71%.

Association rules are sensitive to the number of cases belonging to each range given the way the support and confidence metrics are being calculated. Hence, it is essential for the accuracy of the results that each range has equal chances in participating in rules. In that way, participation denotes a trend and not only membership because of being in the range where the majority of the cases belong. As it was expected at the borough level the equal-interval discretisation performed poorly, resulting to a limited representation of the ranges into the rules with the predominance of the ranges that included the most data cases. Increasing the number of bins educes that problem but still results in a limited representation of the ranges in the rules. This is due to the fact that the confidence and support thresholds are not reached by every category.

In the case of quantile binning, by keeping the number of bins low, a meaningful result can be still achieved from, while there is full coverage of the ranges in the resulted rules. As expected, in the higher geographic levels, less bins result in wide ranges. As the study area is confined into smaller geographical groups, where price variations are not so dramatic, the variation can be captured by this limited number of bins.

After a number of tests, it was decided that the maximum number of bins that gives valid rules and maintains full coverage of the classes is in the majority of the cases 10 bins for the Borough level. This can be achieved without sacrificing the level of accuracy by lowering the thresholds too much. By exceeding this and keeping an extremely low threshold the full coverage can still be achieved but the resulted rules are useless. An example of a 15 bin classification for the area of Hammersmith & Fulham is shown in Figure 6-4.

```

Rule 1557: Commercial Services= NEAR Retail= NEAR ==> PRICE_RANGE= (173000-190000]
                                                (support=5.5564, confidence=8.6587)
Rule 1554: Commercial Services= NEAR Education and Health= NEAR ==> PRICE_RANGE= (173000-190000]
                                                (support=5.3058, confidence=8.6372)
Rule 1197: Retail= NEAR ==> PRICE_RANGE= (173000-190000] (support=5.6006, confidence=8.5779)
Rule 1150: Public Infrastructure= NEAR ==> PRICE_RANGE= [59950-150000] (support=5.1584, confidence=8.5158)
Rule 1195: Education and Health= NEAR ==> PRICE_RANGE= (173000-190000] (support=5.3943, confidence=8.441)
Rule 1551: Commercial Services= NEAR Retail= NEAR ==> PRICE_RANGE= [59950-150000]
                                                (support=5.1584, confidence=8.0386)
Rule 1191: Retail= NEAR ==> PRICE_RANGE= [59950-150000] (support=5.2027, confidence=7.9684)
Rule 1193: Commercial Services= NEAR ==> PRICE_RANGE= (173000-190000] (support=7.2366, confidence=7.8284)
Rule 1123: Commercial Services= NEAR ==> PRICE_RANGE= (218000-230000] (support=6.6912, confidence=7.2385)
Rule 1189: Commercial Services= NEAR ==> PRICE_RANGE= [59950-150000] (support=6.4996, confidence=7.0312)
Rule 1224: Commercial Services= NEAR ==> PRICE_RANGE= (202500-218000] (support=6.4407, confidence=6.9675)
Rule 1125: Commercial Services= NEAR ==> PRICE_RANGE= (267000-285000] (support=6.4259, confidence=6.9515)
Rule 1226: Commercial Services= NEAR ==> PRICE_RANGE= (332000-365000] (support=6.3228, confidence=6.8399)
Rule 1171: Commercial Services= NEAR ==> PRICE_RANGE= (150000-173000] (support=6.308, confidence=6.824)
Rule 1201: Commercial Services= NEAR ==> PRICE_RANGE= (249999-267000] (support=6.2049, confidence=6.7124)
Rule 1129: Commercial Services= NEAR ==> PRICE_RANGE= (308000-332000] (support=6.0722, confidence=6.5689)
Rule 1175: Commercial Services= NEAR ==> PRICE_RANGE= (241500-249999] (support=6.0575, confidence=6.5529)
Rule 1228: Commercial Services= NEAR ==> PRICE_RANGE= (365000-445000] (support=6.0133, confidence=6.5051)
Rule 1230: Commercial Services= NEAR ==> PRICE_RANGE= (445000-2500000] (support=5.8954, confidence=6.3776)
Rule 1173: Commercial Services= NEAR ==> PRICE_RANGE= (230000-241500] (support=5.5122, confidence=5.963)
Rule 1199: Commercial Services= NEAR ==> PRICE_RANGE= (190000-202500] (support=5.4237, confidence=5.8673)
Rule 1127: Commercial Services= NEAR ==> PRICE_RANGE= (285000-308000] (support=5.3353, confidence=5.7717)

```

Figure 6-4: Hammersmith&Fulham_Flat_P11_DescL1_All_15Class

Each rule consists of two parts. In the antecedent part, the conditions of the rule are presented. In the first rule, the conditions include proximity to commercial services and retail facilities. In the consequent part of the rule, the resulting class based on these conditions is displayed. Each rule is also accompanied by the rule id (unique identifier in the database) and the support and confidence metrics which are percentages. As we observe, after the 7th rule where different conditions result in different classes, the same condition (proximity to commercial services) results in all the 15 classes. Although a full coverage is achieved, these rules fail to capture the variations in the conditions that result in different classes. Hence, their contribution in the quality of the classifier is limited.

Figure 6-5 shows the distribution of the sample used in the main case study. Areas that appear not covered by the sample are the main open space areas (e.g. Royal parks) and other spacious non residential landuses. In an ideal case study the whole population would be necessary. A number of factors make that impossible in most of the cases.

In particular in this case, in order for the whole population to be part of the analysis a transaction for each and every property should have been known over the examined period. Another source of information loss comes from the properties that although

were sold over the specific period, were not in the land registry records at the time of the data collection.

As mentioned in Section 5.3 the participated sample consists of the 60 percent of the registered transactions. This 40% loss occurred partly due to the way the data was collected and partly during the georeferencing phase where either the address description did not match or no polygonal reference in the AddressPoint data was found. Additionally, entries that referred to transactions of multiple-floor flats and garages have been excluded.

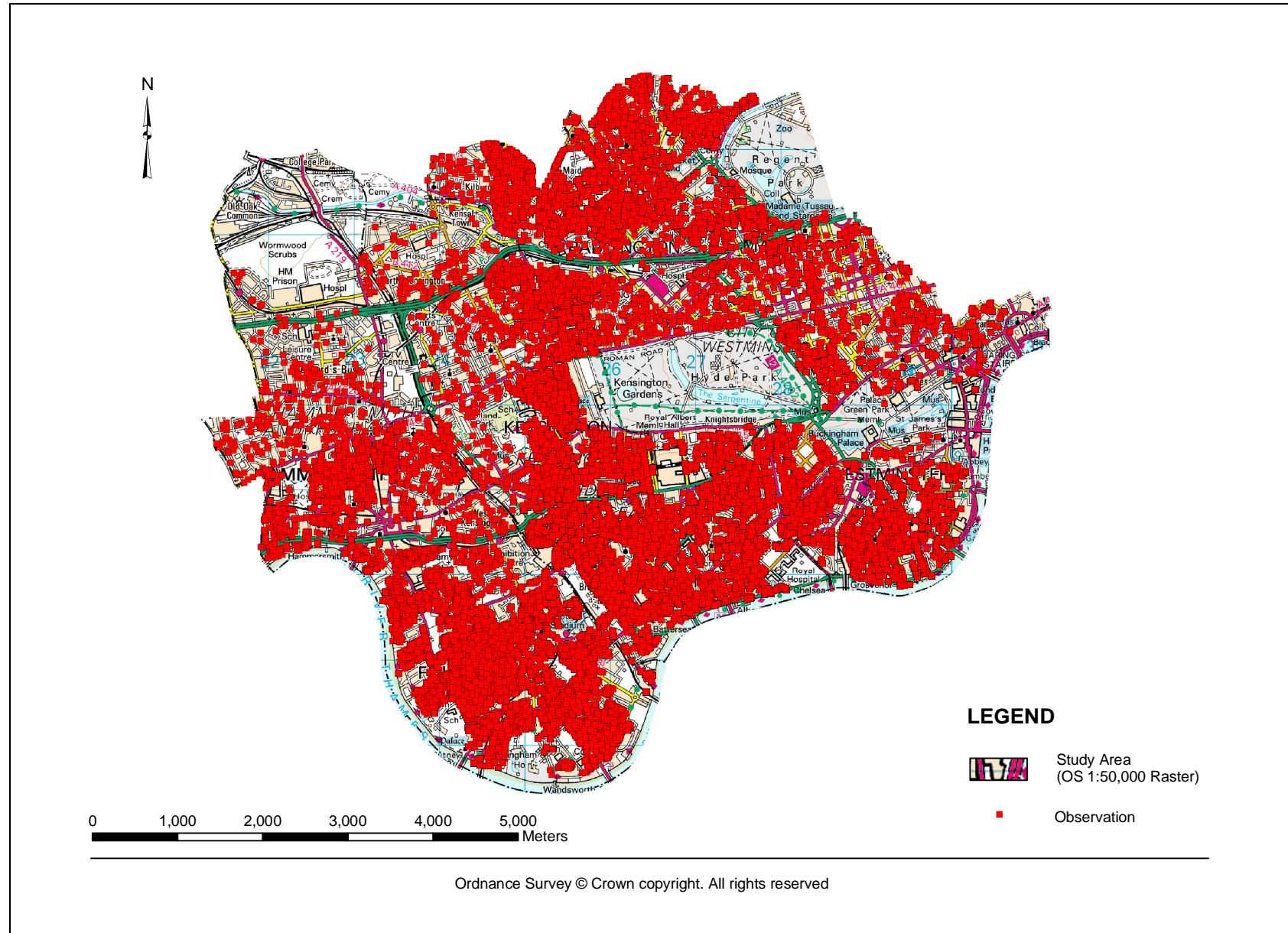


Figure 6-5: Sample Geographic Distribution

6.3 Test cases

As mentioned in the previous section, the methodology is tested against two requirements. The first is its ability to produce meaningful information about the way location affects the price. The second is to test the viability of such an approach in the valuation process. To evaluate the results of the first, a number of tests have been performed and the extracted association rules have been analysed. The evaluation of the second was based on the acquired classification accuracy. The presented tests are named after the experiment guide (Figure 6-3). The output files of these experiments can be found on the attached CD.

Locational influence tests

The *Borough_PropertyType_DescLevel_Pl_ClassNo* was performed for each of the four general property types. The path length was set to 1 and individual prices were classified into three price ranges. The number of the bins (3) was kept small to accommodate the clarity and demonstration of the results. This test was performed for all the three levels of the landuse description (Figure 5-11). The year constraint has not been used in this case in order to make use of the highest available number of known cases. The association rules generation was based on support and confidence thresholds of 10%. The maximum number of rules was set to 20, due to the high density of the input dataset that results in a large average number of associations per transactions. This configuration was chosen to capture the most information possible at the least time by reducing the number of experiments.

The extracted rules were further processed and the percentages of the examined landuses within each of the three classes (low - medium - high) were calculated. The results are presented in the form of bar charts, where for each landuse the proportion of its association with each of the classes is shown with different colour. It has to be noted that in the following charts the presentation of the locational features is given in an isolated way while their effect is a combined result depicted in the resulted rules. This approach was chosen due to the volume of the extracted rules that would make their direct presentation in the document impossible.

As mentioned in the previous section, the property data acquired cover 60% of the actual number of transactions for the 2000 - 2006 period. Although this is a satisfactory proportion of the whole data population any potential bias that may arise as a result of the 40% not considered here should also be examined. An unbiased data set will ensure the results of the case study are meaningful. To eliminate the possibility of such a bias and justify the sample employed two potential problem areas needed further investigation. The first area to examine was whether the sample had a bias towards certain geographic areas. Exclusion of areas from the sample would lead to results that are not representative of the whole area under investigation.

Although Figure 6-5 shows an aggregate distribution of the data across all geographical areas considered, since the analysis is performed for each of the four property types it is necessary to ensure that even geographical distribution of the sample for each property type exists. In the following figures the geographical distribution of the data per property type in relation to the property stock according to Census 2001 is shown. The maps produced, use the Household spaces and accommodation type table (KS016) from the Census 2001 key statistics dataset. For the classification the Natural Break (Jenks) scheme was used. The structure of the housing stock recorded in Census 2001 for the three case study Boroughs mapped by Census Output Area is presented in the following figures.

Figure 6-6 shows the distribution of the detached houses for which transactional information was available and therefore taken into account in the analysis. The stock of detached houses in the study area is quite low reflecting the general situation in London where other types of housing e.g. flats are more common. As expected, the number of known transactions is relatively small but it is distributed in accordance to the detached houses stock. An example of this is the Abbey Road ward which is at the north end of the Westminster Borough. Abbey Road ward has the higher percentage of detached houses compared to all the other wards within the study area fact that is reflected in the higher density of the sample there.

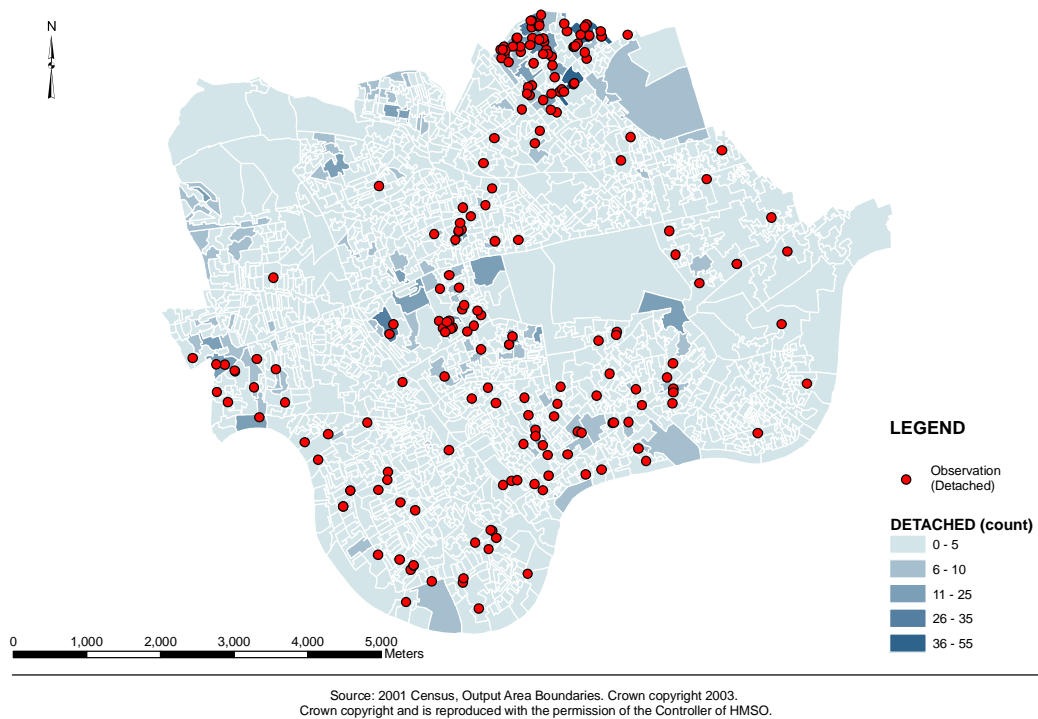


Figure 6-6: Geographic distribution for the detached properties sample

A similar situation is presented in Figure 6-7 where the distribution of the semi-detached transactions in relation to the existing semi-detached stock is shown. As the figure suggests the data sample satisfactorily covers the area under investigation for this property type. It should be noted that large output areas that appear with relatively high numbers of housing stock are mostly covered by parks or other spacious landuses. As a result, it may seem that large geographic areas are not covered by the sample resulting to missing patterns. In reality, the housing stock in these cases exists at small areas usually at the borders of these output areas. Such an example is the area covered by the output areas that are located at the top left part of the Hammersmith and Fulham Borough. These output areas belong to the College Park and Old Oak ward. An approximate 90% percentage of this ward's total area is covered by several landuses with large land requirements such as the Old Oak Common depot, Wormwood Scrubs park, St Mary's cemetery and the Wormwood prison. In this specific case the residential sections are limited to a number of few streets and are located at the upper and lower parts of the ward.

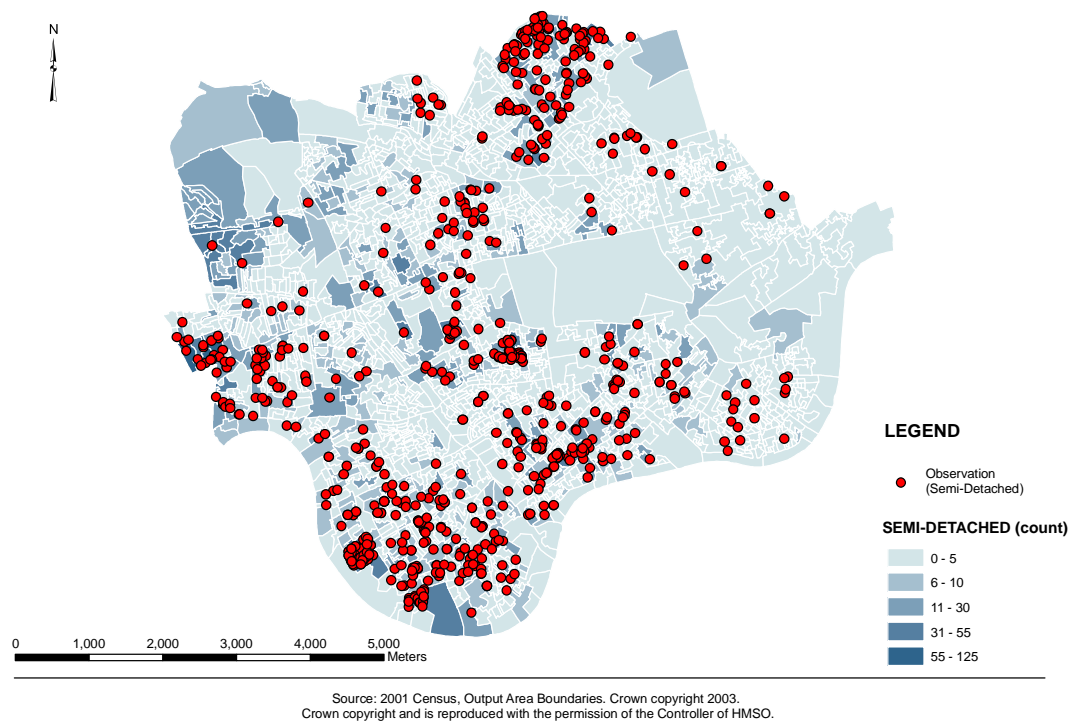


Figure 6-7: Geographic distribution for the semi-detached properties sample

After flats, terraced houses are considered the most common accommodation type within these three Boroughs. Consequently, the sample (Figure 6-8) is more densely

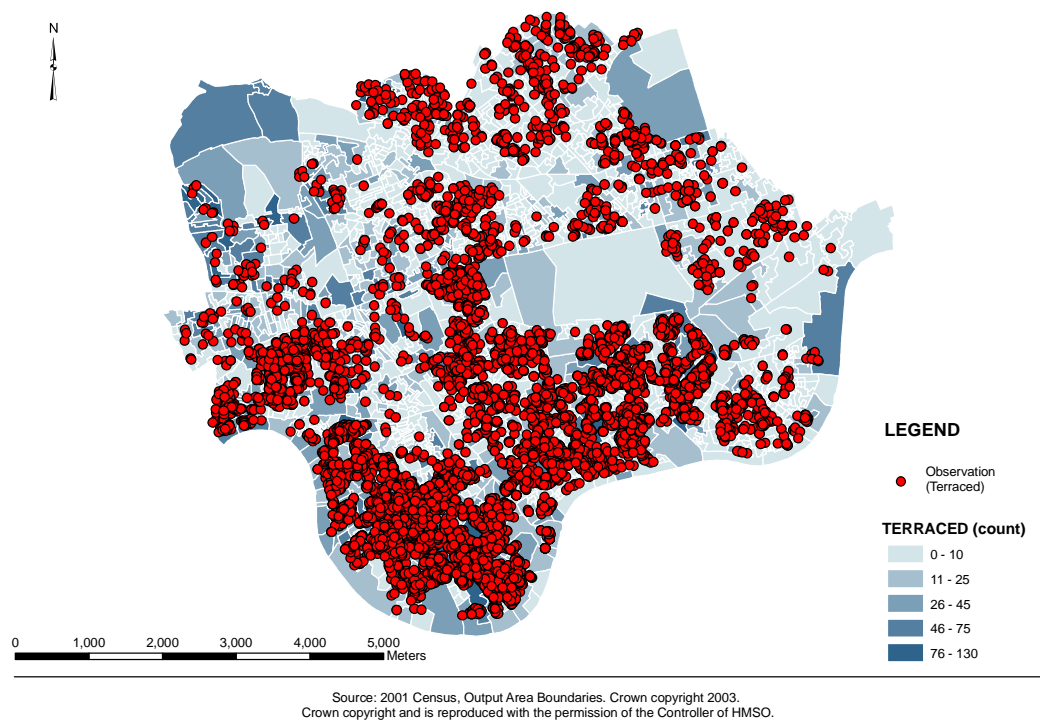


Figure 6-8: Geographic distribution for the terraced properties sample

populated compared to that of the detached and the semi-detached properties and covers all the homogeneous, with respect to landuse types, output areas. Again, the density of the sample varies in proportion to this type of accommodation within each output area.

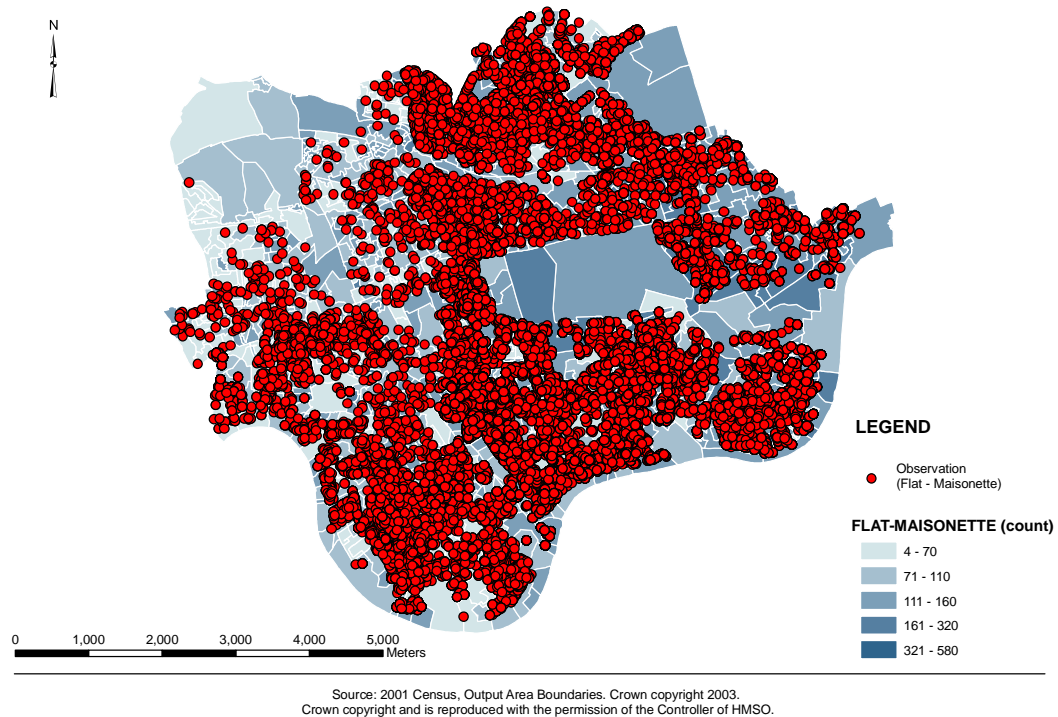


Figure 6-9: Geographic distribution for the flats and maisonettes sample

Depending on the Borough, flats account for around 80% of the housing stock of the examined area. Based on this, the majority of the housing transaction records regard flat sales (Figure 6-9).

So far the geographical distribution of the sample and also its analogy to the existing housing stock has been presented. This alone is not enough to validate it since there is no indication of the percentage of the actual transactions per property type that is represented within the sample. This is the second area that had to be investigated further. Bar charts that compare the number of transactions per property type with that of the sample for the three Boroughs have been produced and are presented here. The information about the volume of sales for the examined period is based on the Land Registry quarterly reports of residential property price (Land Registry, 2007).

In Figure 6-10 the comparative charts for each of the three Boroughs are presented. For the Kensington and Chelsea Borough almost the 70% of the recorded transactions have been used in the analysis.

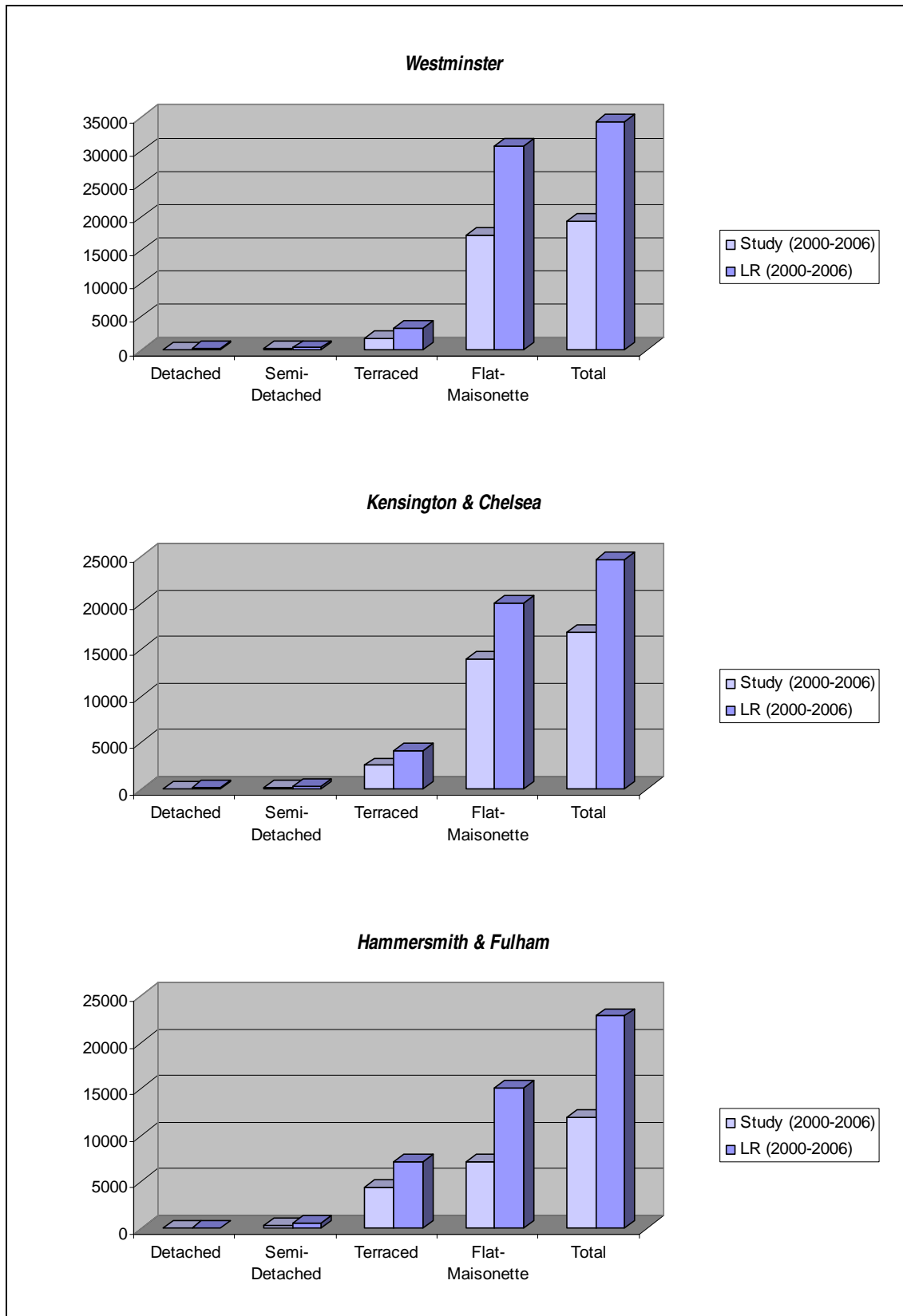


Figure 6-10: Actual and study volumes of sales comparison per property type

The Westminster Borough has been represented by a 57% of the actual transactions while the least percentage of matched records relates to the Hammermith and Fulham Borough (52%). In all the three cases the percentage of the missing transactions is consistent among the different property types within the same Borough. Therefore there are no cases where the loss is associated with a certain property type fact that would affect the quality of the results for that specific property type.

Although in the case study time periods have not been explicitly taken into account, Figure 6-11 shows a comparison of the number of the study transactions to that of the registered transactions for each year of the examined period (2000-2006).

It can be observed that the percentage of the matched records within each year per Borough follows that of each Borough for the whole period. An exception is year 2006 where in all three cases the percentage is noticeably lower than the average. This can be explained by the fact that the collection process has been completed within that year since to validate the algorithm and the approach a dataset of 50,000 transactions was deemed to be adequately.

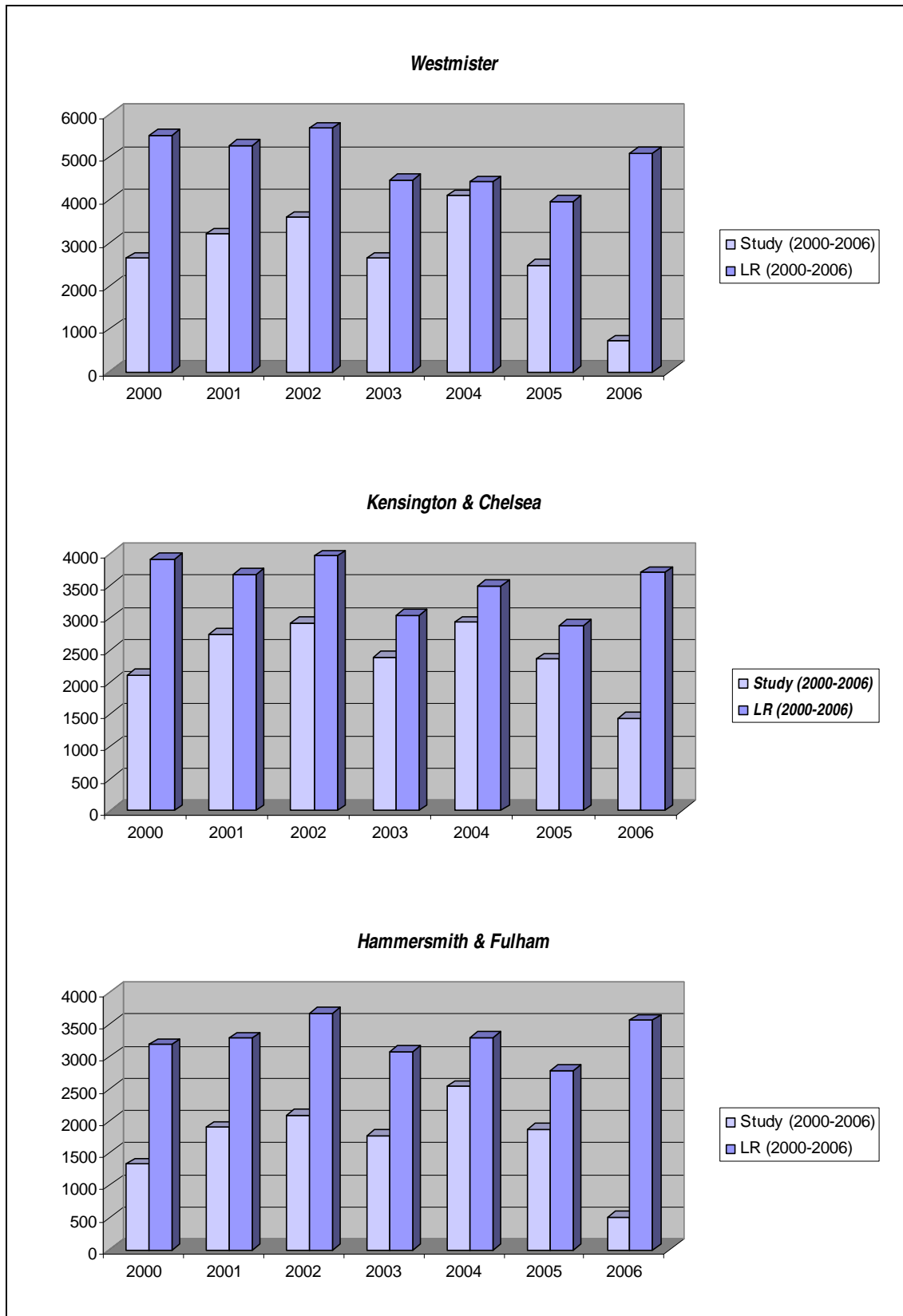


Figure 6-11: Actual and study volumes of sales comparison per year

Westminster_PropertyLevel_PL1_DescL1_3Class

Figure 6-12 shows the proportion of the contribution of each individual landuse that the extracted rules are consisted of and classify properties either at the lower, medium or the higher price band for the case of the Westminster flats. As it is shown, although there are landuses that associate with both the lower and higher classes there are landuses such as Open Space where its membership to the high and medium price rules is predominant. It is also shown that proximity to Sport & Entertainment facilities contributes positively resulting in medium and high prices. Proximity to Transportation, on the other hand, equally relates to the three price bands. An interesting result is that of the proximity to Attractions (e.g. historic places, tourist attractions) where is associated only to the low prices range.

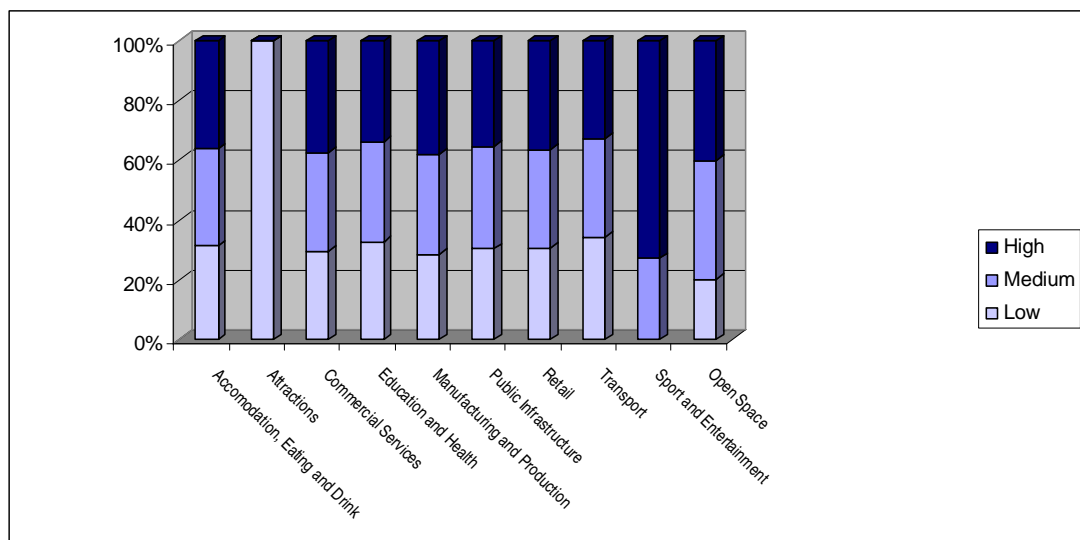


Figure 6-12: *Westminster_Flat_PL1_DescL1_3Class*

At the same level of detail, Figure 6-13 refers to the market of terraced houses. As shown, the results present great similarities to those of the previous test. The only noticeable difference relates to the proximity to the transportation system where here in the majority of the rules contributes to lower price.

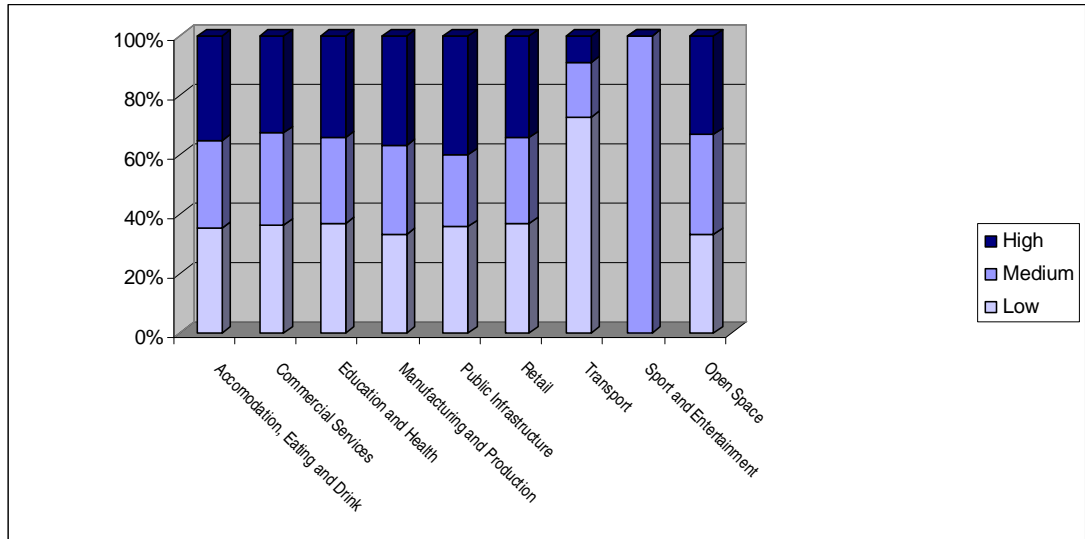


Figure 6-13: Westminster_Terrace_PI1_DescL1_3Class

In the case of Semi Detached properties (Figure 6-14), the results present differences when compared to that of Flat and Terrace markets. An observation one can make is that proximity to non-residential landuses in the majority of the cases exists in rules that result to low price ranges. Additionally, proximity to Manufacturing and Production activities is always associated with low prices.

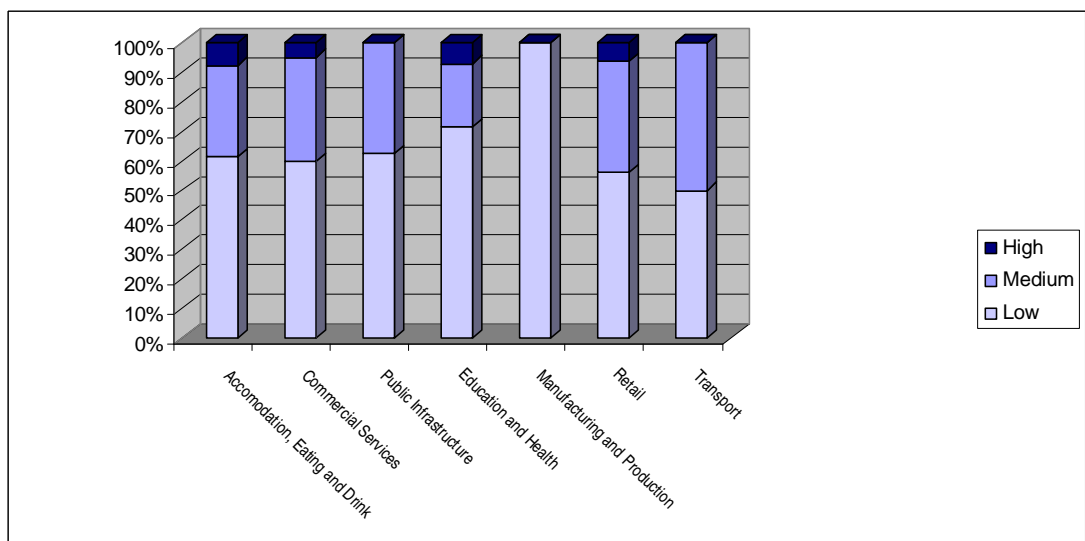


Figure 6-14: Westminster_SemiDetached_PI1_DescL1_3Class

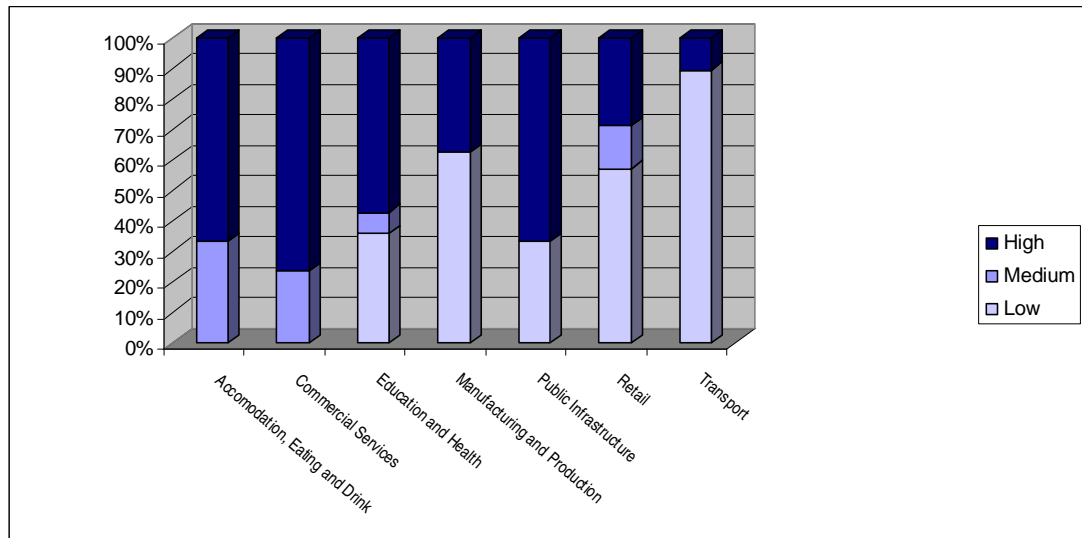


Figure 6-15: Westminster_Detached_P11_DescL1_3Class

Finally, detached houses (Figure 6-15) present quite a different picture. Proximity to services is positively priced, while in the majority of the rules proximity to transport is associated with low prices. Here, for the first time, proximity to Education and Health facilities in the majority of the rules contributes to high prices.

Kensington&Chelsea_PropertyLevel_P11_DescL1_3Class

The results here are similar to those of Westminster flats, with a mild increase in the lower prices associated with proximity to services. The bigger increase appears in the proximity to Manufacturing and Production and Transportation facilities. It is interesting to notice that the presence of Open Space in a rule contributes in the majority of the cases to medium and high prices, a pattern that also appears in the Westminster flats case.

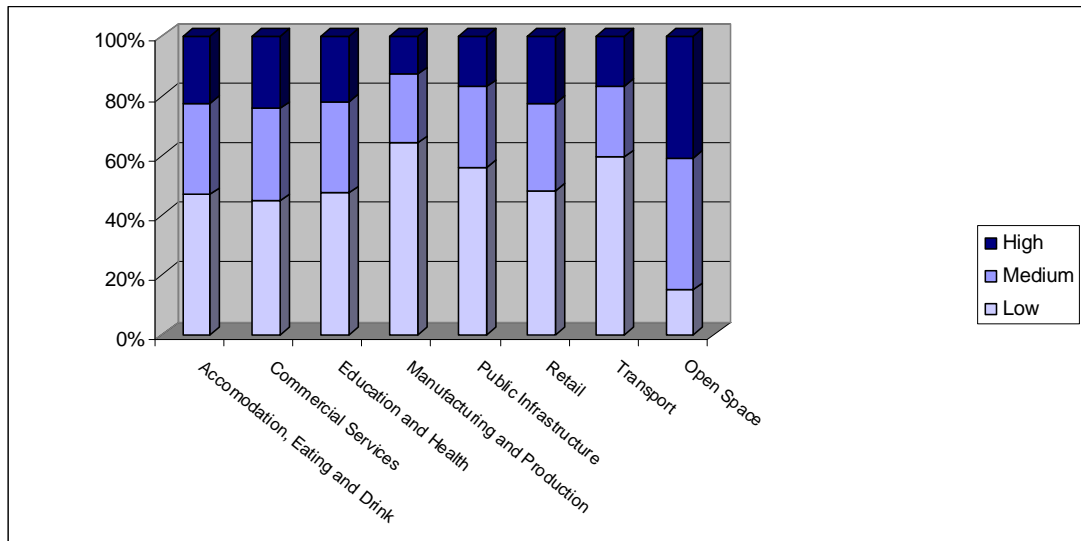


Figure 6-16: Kensington&Chelsea_Flat_PI1_DescL1_3Class

A similar pattern in the prices to that of Westminster can also be observed in Figure 6-17 that shows the results for the Kensington and Chelsea terraced houses. Again, there is an increase in the contribution to the lower prices but in the case of proximity to Open Space the contribution to higher and medium price bands have been increased.

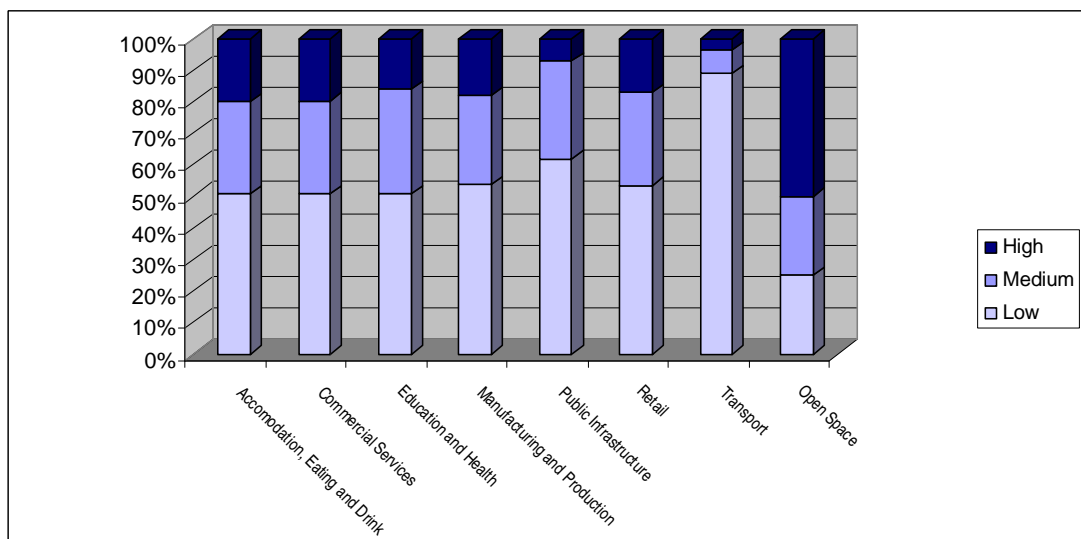


Figure 6-17: Kensington&Chelsea_Terrace_PI1_DescL1_3Class

Moving to the Semi Detached houses a similar behaviour to that of the flats in the same area can be observed. The difference here is that proximity to open space areas does not contribute to low prices at all.

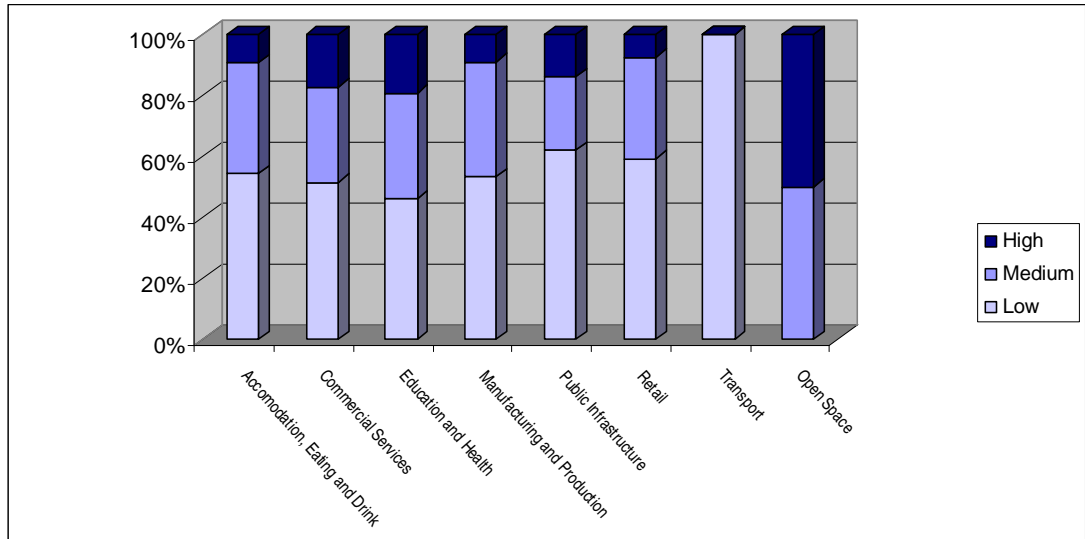


Figure 6-18: Kensington&Chelsea_SemiDetached_PI1_DescL1_3Class

In the final chart (Figure 6-19) the proportion of the contribution of proximity to non-residential landuses in the prices of detached houses is shown. Here, a different pattern can be observed. The contribution to high prices is extremely limited. Even in the case of proximity to Open Space where the contribution to low and medium prices is equal. A possible explanation could be the different spatial patterns of the detached houses compared for example to that of areas with blocks of flats. Detached houses are located in less dense areas and are always associated with private gardens. Therefore, it is possible that proximity to open spaces is less valued in these areas and is not considered an advantage against other properties of the same type.

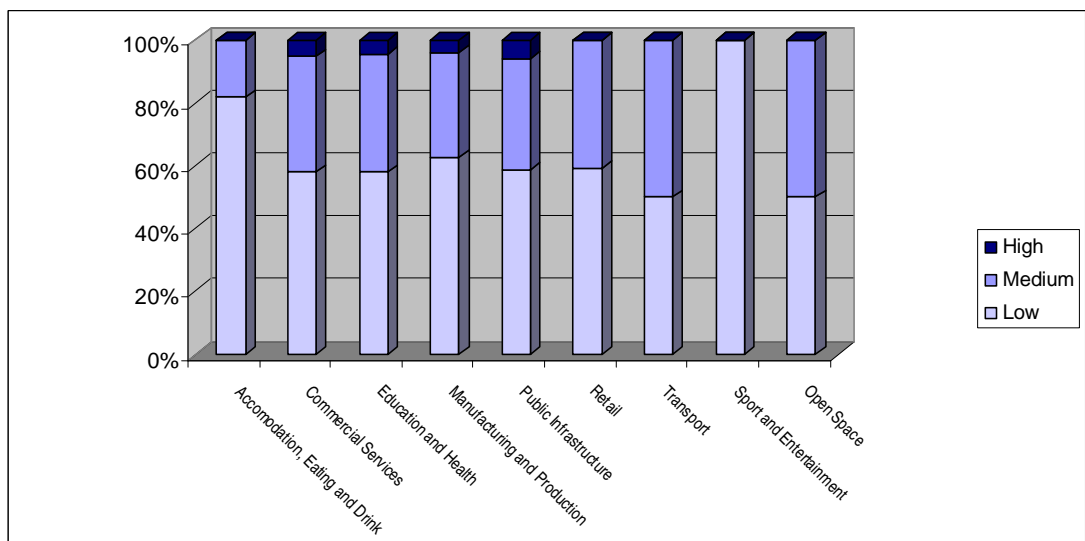


Figure 6-19: Kensington&Chelsea_Detached_PI1_DescL1_3Class

Hammersmith&Fulham_PropertyLevel_PI1_DescL1_3Class

The chart depicted in Figure 6-20 shows a similar pattern to that of the flats in Kensington and Chelsea area. Proximity to services almost equally contributes to low and medium-high price bands with a slight prevalence of the low price rules.

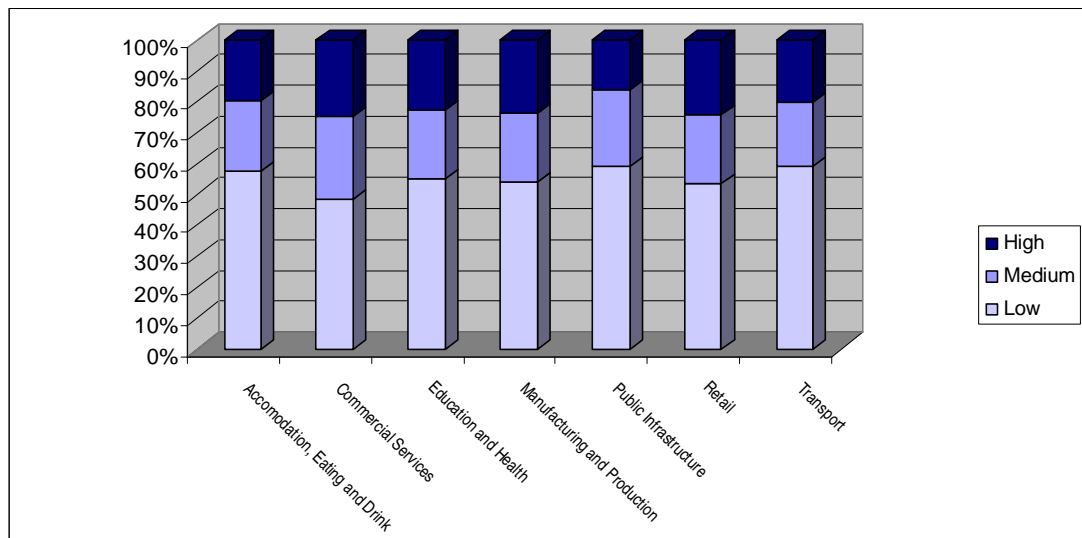


Figure 6-20: *Hammersmith&Fulham_Flat_PI1_DescL1_3Class*

In the case of the terraced houses, again the pattern is similar to that of the flats with the lowering of the contribution to the lower prices.

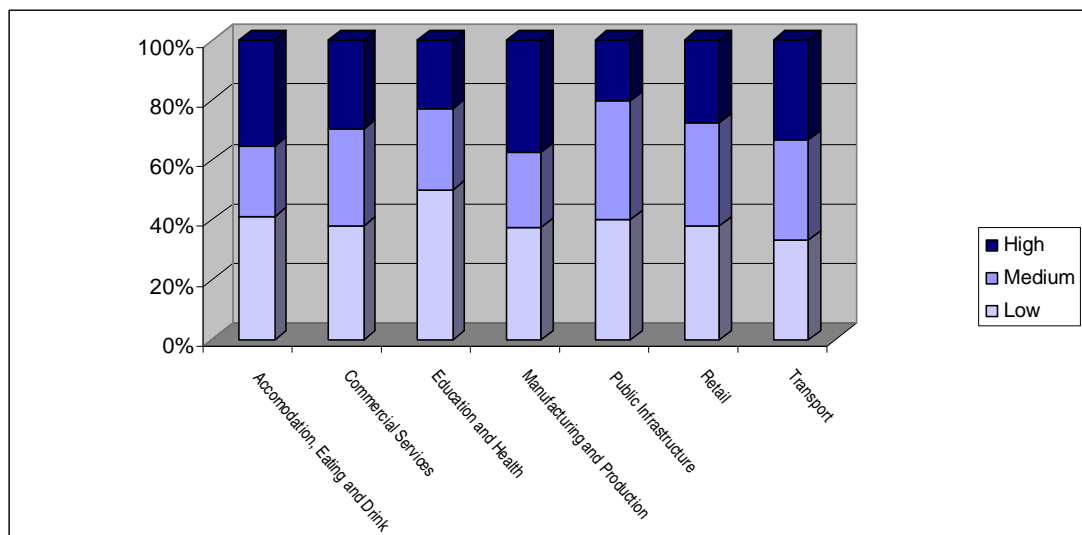


Figure 6-21: *Hammersmith&Fulham_Terrace_PI1_DescL1_3Class*

What is noticeable in both the flat and terraced houses charts, is the absence of the Open Space in the extracted rules. This can be explained by the possible low percentage of Open Spaces within areas that mainly consist from flats and terraced

houses. On the contrary, in both the charts that refer to the Detached and Semi-Detached housing stock Open space appears in the resulted rules.

In the case of the Semi-Detached house bands (Figure 6-22) proximity to services has extremely low contribution to the high price band. Also, proximity to Open Space areas entirely contributes to the middle price band.

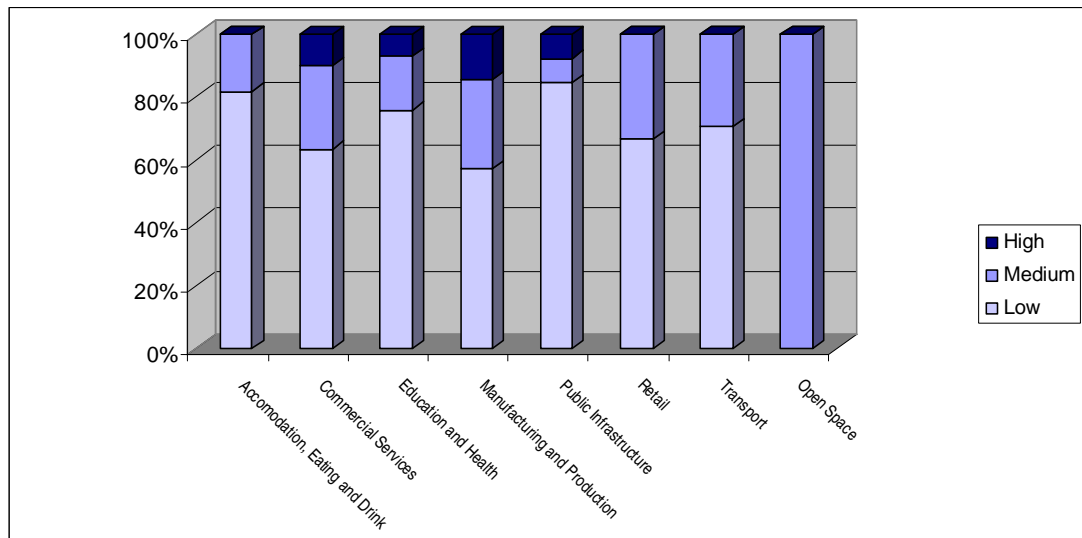


Figure 6-22: Hammersmith&Fulham_SemiDetached_PL1_DescL1_3Class

The pattern in the rules regarding the detached houses in Fulham and Hammersmith area (Figure 6-23) bare some similarities to those of the detached houses in the Westminster area. Unlike the semi-detached houses, here proximity to some services such as Education and Health, Public Infrastructure even Transportation facilities is associated in the majority of the cases with positive contribution (medium-high bands).

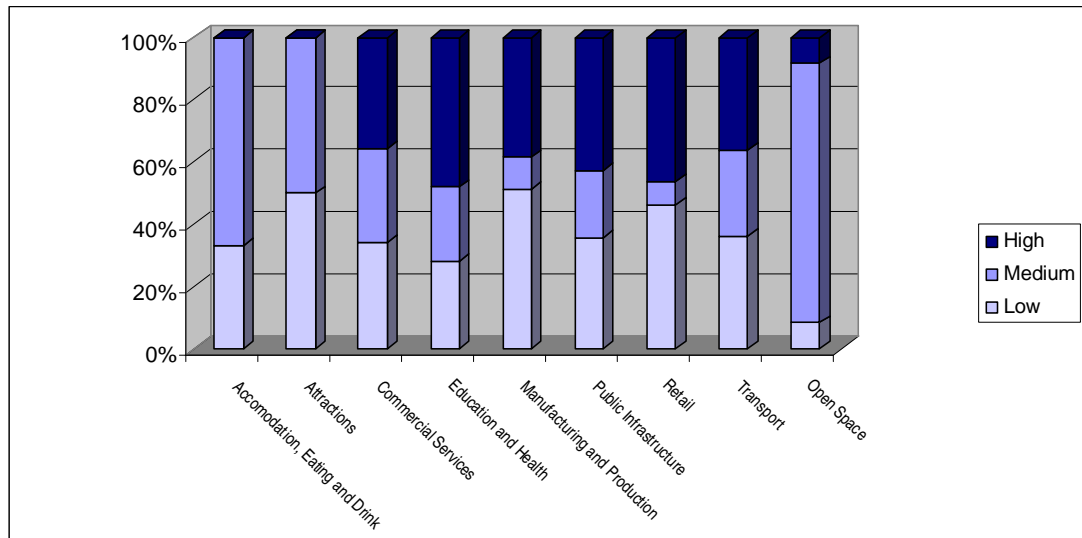


Figure 6-23: Hammersmith&Fulham_Detached_PL1_DescL1_3Class

All these tests were performed at a very coarse level of detail regarding the non-residential landuses. Also, a three-level classification scheme have been used that can be considered quite broad given the geographical areas covered and the price differences. Table 6-2 provides a comparison of these results for the four property types across the three boroughs. The different colours denote which class (low-medium-high) was supported by the majority of the rules for each landuse. Two and three colour circles denote that two or all the three classes were equally supported by the rules. The cells with the asterisk indicate the cases where all rules resulted in only one class.

		Landuses (Description Level 1)									
		Accommodation, Eating and Drinking	Attractions	Commercial Services	Education & Health	Manufacturing & Production	Public Infrastructure	Retail	Transport	Sport & Entertainment	Open Space
WESTMINSTER	Flat	●	* ●	●	●	●	●	●	●	●	●
	Terrace	●		●	●	●	●	●	●	* ●	●
	Semi-Detached	●		●	●	* ●		●	●		
	Detached	●		●	●	●	●	●	●		
KENSINGTON & CHELSEA	Flat	●		●	●	●	●	●	●		●
	Terrace	●		●	●	●	●	●	●		●
	Semi-Detached	●		●	●	●	●	●	* ●		●
	Detached	●		●	●	●	●	●		* ●	●
HAMMERSMITH & FULHAM	Flat	●		●	●	●	●	●	●		
	Terrace	●		●		●	●	●	●		
	Semi-Detached	●		●	●	●	●	●	●		* ●
	Detached	●	●	●	●	●	●	●	●		●

* 100% ● Low ● Medium ● High

Table 6-2: Comparative results (Landuse Description Level 1)

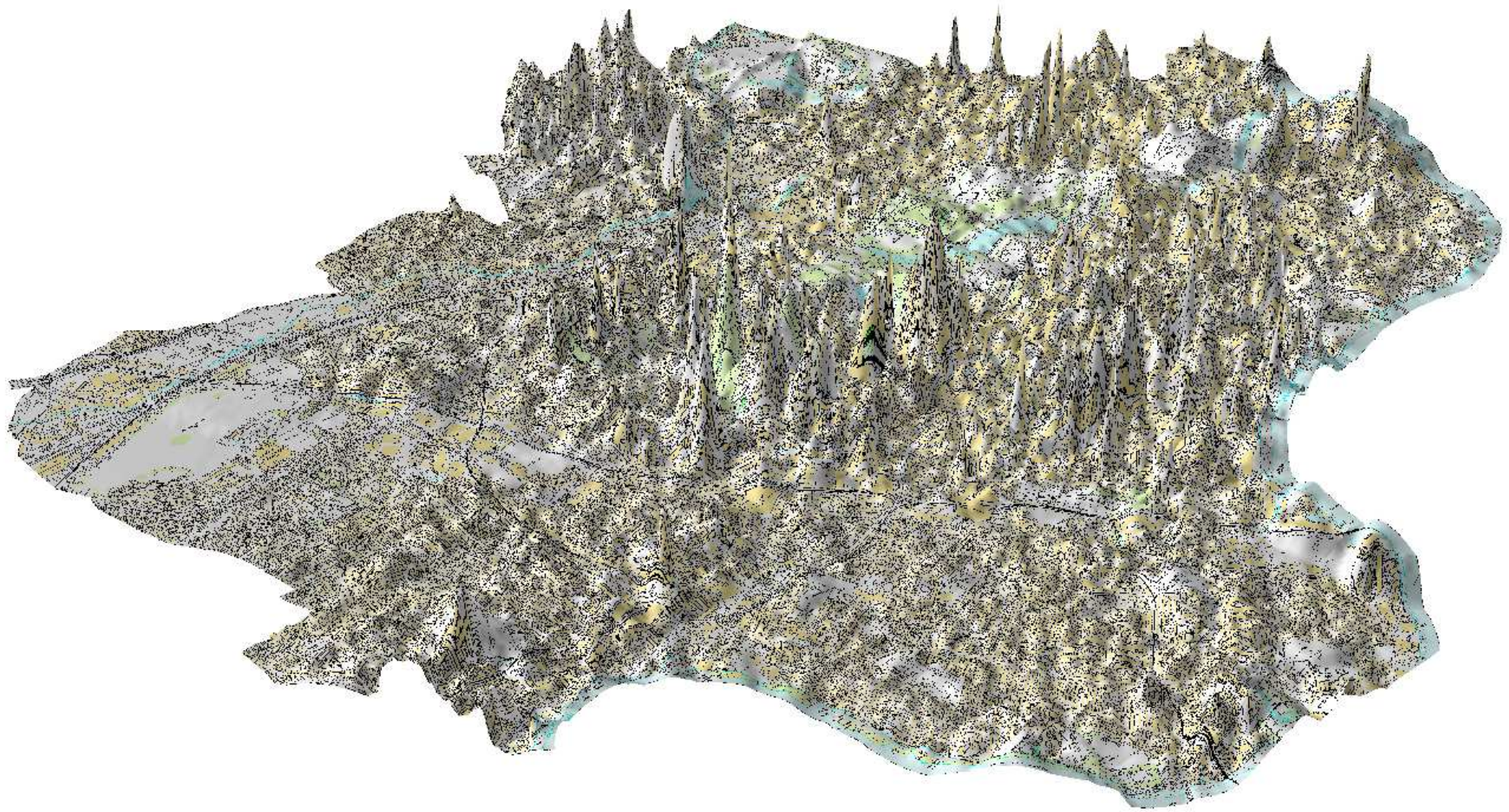


Figure 6-24: 3D value model

So far, the performed experiments used the landuse information at the highest level of detail. Despite this, some interesting results about the structure of the markets have still been achieved. For example, although variant across the different submarkets, it is apparent that proximity to Open Spaces is in most of the cases medium to highly valued. This is also depicted in Figure 6-24, where a 3D representation of the price variations across the study area is presented. As it is shown, major peaks are concentrated around major parks and other open spaces. As the higher level of classification of the non-residential landuses can be considered quite broad it would be interesting to examine whether any meaningful information can be retrieved at a lower level of classification but with still keeping the geographical area at a Borough level. Again the experiments have been performed per property type. The analysis charts can be found in Appendix D. What follows is a brief analysis of the results.

Borough_PropertyLevel_P11_DescL2_3Class

Table 6-3 summarises the results for this group of experiments. As the level of the locational information becomes more detailed a first observation that can be made is that a number of categories that existed at the first level of detail fail to reach the thresholds as they split into their sub-categories. One such example is Open Space where it does not participate in any of the extracted rules. That was expected given the way the algorithm performs in favour of the majority.

Nevertheless some results can still be extracted. There are some landuses that are associated with certain price bands irrespective the type of property and others that their influence varies upon property market. An example of the first type is proximity to Construction Services and Proximity to Multi-Item Retail facilities where in the majority of the cases in all boroughs result to low band prices. In the case of Transport, Storage and Delivery services when refer to Semi-Detached or Detached properties always contribute to lower price ranges. This is not the case when refer to Flats or Terraced properties where the contribution to lower band is proportionally very low.

		Landuses (Description Level 2)																			
		Accommodation	Construction Services	Consultancies	Eating & Drinking	Health Practitioners	Industrial Manufacturing & Production	Infrastructure & Facilities	IT, Advertising, Marketing & Media Services	Legal & Financial	Medical Establishments	Multi-Item Retail	Personal Consumer & Other Services	Primary, Secondary & Tertiary	Property & Development Services	Public Facilities	Recreational & Vocational Education	Research & Design	Road & Rail	Single Item Retail	Transport, Storage, Delivery
WESTMINSTER	Flat		*																		
	Terrace											*		*							
	Semi-Detached	*	*				*					*					*				*
	Detached											*									
KENSINGTON & CHELSEA	Flat											*					*				
	Terrace	*	*									*					*				
	Semi-Detached						*			*		*		*					*		*
	Detached					*	*											*			*
HAMMERSMITH & FULHAM	Flat									*	*	*					*	*			*
	Terrace			*											*						
	Semi-Detached			*			*									*					*
	Detached								*				*						*	*	

* 100% Low Medium High

Table 6-3: Comparative Results (Landuse Description Level 2)

		Landuses (Description Level 3)																						
		Accomodation	Bars	Beauty Salons	Bus Stops	Clinics & Surgeries	Clothing & Accessories	Construction Services	Consultancies	Electrical Features	Estate Agencies	Food & Drink	Health Practitioners	Household, Leisure & Garden	Industrial Products	IT, Advertising, Marketing & Media Services	Legal & Financial	Niche Goods	Personal Consumer & Other Services	Property & Development Services	Research & Design	Restaurants	Transport, Storage, Delivery	
WESTMINSTER	Flat																							
	Terrace																							
	Semi-Detached																							
	Detached																							
KENSINGTON & CHELSEA	Flat																							
	Terrace																							
	Semi-Detached																							
	Detached																							
HAMMERSMITH & FULHAM	Flat																							
	Terrace																							
	Semi-Detached																							
	Detached																							

* 100% Low Medium High

Table 6-4: Comparative Results (Landuse Description Level3)

Borough_PropertyLevel_P11_DescL3_3Class

Using even more detailed landuse classification per property type, the results in such a broad geographical area show a prevalence of the most commonly found landuses in an urban environment. Here, one can detect different behaviour of the market amongst the three boroughs (Table 6-4). In Westminster, in the case of flats for example, proximity to services that are most commonly located in the high street such as retail shops, bars, restaurants contributes positively in the price. On the contrary, in the same borough, in the case of semi-detached and detached houses such proximity results to low and medium price ranges. Looking at the Kensington Borough the percentage of lower price, due to proximity to high street, is noticeably higher even for flats and becomes prevalent as one moves to terraced, semi-detached and detached houses.

So far, the ability of the method to detect associations between the different locational features and price levels of prices has been demonstrated. One apparent limitation is that features that are most commonly found across the study area tend to overshadow sparse features. Landuses such as commercial use are spread over the study area while for example a more specialised use e.g. cemeteries is not so commonly found. As a result, the extracted rules consist of relationships to landuses that are most common.

This limitation is exaggerated by the fact that the study areas selected in the above experiments are big. As the percentage of the common landuses grows proportionally to the size of the area, the chances of less common landuses to reach the support and confidence thresholds and appear to the rules are reduced.

In order to investigate the affect of spatial relationships with the not so frequent landuses or landscape features, there is a need to lower the geographical level of the analysis. Therefore the same experiment was performed at a ward level. The following are some illustrative results of this. The selection of the ward was based on its distinctiveness in terms of landuses and geographical characteristics.

Ward_PropertyLevel_P11_DescL1_3Class

For this test, the Hyde Park (0BKGG) ward has been selected. In Figure 6-25, a similar situation to that shown in the Westminster general chart for flats is depicted. The additional information that is revealed here regards the relation of the proximity to Water with the property prices. Proximity to Running Water, in this case proximity to a canal, has in the majority of the cases a positive effect on the property prices.

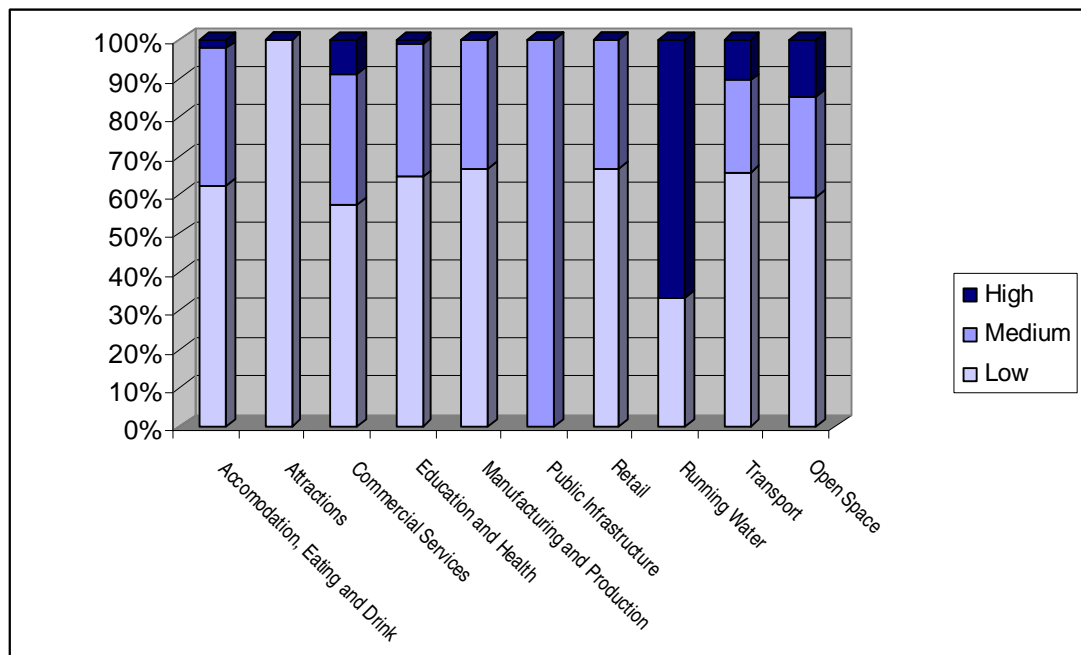


Figure 6-25: HydePark_Flat_P11_DescL1_3Class

In order to have a clearer idea about the combinations that result to high (or low) price bands within this study area the actual classification rules should be examined. Figure 6-26 shows an extract of the classification rules produced in this test. As it is shown, the combination of proximity to Open Space, Running Water, Transport and Commercial Services gives high range property prices. If we examine the whole list of the association rules we will see that even the generic rule Running Water= NEAR ==> ZPRICE_RANGE= (480000-3050000] appears with high support and confidence scores (support=20.4444, confidence=54.1176).

```

Rule 218: Open Space= NEAR Running Water= NEAR ==> PRICE_RANGE= (480000-3050000]
                                                    (support=13.3333, confidence=64.5161)
Rule 785: Open Space= NEAR Running Water= NEAR Transport= NEAR ==> PRICE_RANGE= (480000-3050000]
                                                    (support=13.3333, confidence=64.5161)
Rule 789: Commercial Services= NEAR Open Space= NEAR Running Water= NEAR ==>
PRICE_RANGE= (480000-3050000]
                                                    (support=13.3333, confidence=64.5161)
Rule 947: Commercial Service= NEAR Open Space= NEAR Running Water= NEAR Transport= NEAR ==>
PRICE_RANGE= (480000-3050000]
                                                    (support=13.3333, confidence=64.5161)
Rule 805: Open Space= NEAR Retail= NEAR Transport= NEAR ==> PRICE_RANGE= [90000-275000]
                                                    (support=10.8889, confidence=64.4737)
Rule 897: Education and Health= NEAR Open Space= NEAR Retail= NEAR Transport= NEAR ==>
PRICE_RANGE= [90000-275000]
                                                    (support=10.8889, confidence=64.4737)
Rule 1400: Accommodation, Eating and Drink= NEAR Education and Health= NEAR Open Space= NEAR Retail= NEAR
Transport= NEAR ==> PRICE_RANGE= [90000-275000]
                                                    (support=10.8889, confidence=64.4737)
Rule 1478: Accommodation, Eating and Drink= NEAR Commercial Services= NEAR Open Space= NEAR Retail= NEAR
Transport= NEAR ==> PRICE_RANGE= [90000-275000]
                                                    (support=10.8889, confidence=64.4737)
Rule 1727: Accommodation, Eating and Drink= NEAR Commercial Services= NEAR Education and Health= NEAR Open
Space= NEAR Retail= NEAR Transport= NEAR ==> PRICE_RANGE= [90000-275000]
                                                    (support=10.8889, confidence=64.4737)
Rule 1514: Commercial Services= NEAR Education and Health= NEAR Open Space= NEAR Retail= NEAR Transport=
NEAR ==> PRICE_RANGE= [90000-275000]
                                                    (support=10.8889, confidence=64.4737)
Rule 967: Commercial Services= NEAR Open Space= NEAR Retail= NEAR Transport= NEAR ==>
PRICE_RANGE= [90000-275000]
                                                    (support=10.8889, confidence=64.4737)
Rule 832: Accommodation, Eating and Drink= NEAR Open Space= NEAR Retail= NEAR Transport= NEAR ==>
PRICE_RANGE= [90000-275000]
                                                    (support=10.8889, confidence=64.4737)
.....
Rule 316: Education and Health= NEAR Public Infrastructure= NEAR ==> PRICE_RANGE= (275000-480000]
                                                    (support=10.2222, confidence=58.2278)
Rule 1909: Commercial Services= NEAR Education and Health= NEAR Public Infrastructure= NEAR ==>
PRICE_RANGE= (275000-480000]
                                                    (support=10.2222, confidence=58.2278)
.....

```

Figure 6-26: HydePark_Flat_PI1_DescL1_3Class Association Rule Extract

By further examining the case, by running the test for landuse classification of level 2 (see Figure 5-11) the rule Open Space= NEAR Running Water= NEAR continues to define the top range class in the form of Green Space = NEAR Canal = NEAR. This description coincides with the broader area of Little Venice which is one of the exclusive, fashionable and expensive residential areas.

Another interesting example can be found amongst the top rules that give high range classification at a more detailed landuse classification (level 3). In that we find the rule Bus Stops= NEAR Green Space= NEAR Rails= NEAR ==> ZPRICE_RANGE= (480000-3050000] (support=13.3333, confidence=64.5161). In this rule, the interesting part is the association of proximity to Railway with the high price band. Proximity to railway usually causes reduced prices due to aesthetic and other

reasons. The difference in this case is that the rule describes the locational situation around the Paddington station which is one of the most important railway and underground stations in London. Hence, this association reflects the real situation entirely.

Classification tests

These tests were performed to check whether it is possible to acquire a valid classifier and a satisfactory classification accuracy and with which parameters. The first group of tests were executed at a borough level. Although the geographical size of a borough is big and the price variations is not expected to be reflected in a small group of rules that form the classifier, it is interesting to explore with different parameterisation the behaviour of the algorithm. For the tests, 70% of the input dataset was used to build the associative classification model while the remaining 30% was used to test the classifier unless otherwise stated. The sample was automatically and randomly generated using the Oracle's Sample function.

Table 6-5 summarises the classification results performed at a Borough level for Kensington & Chelsea. For the cases of Flats and Maisonettes and Terraced houses, since they belong to the majority of the housing stock, the classification was performed by using also the year constraint. Detached and Semi-Detached houses did not form a big sample per year hence, the whole sample was used instead.

	Input Dataset			Build Data			Test Data			Classifier	
	Cases No	Cases No	Year	Cases No	Correct	Wrong	Unclassified	No of Rules	Accuracy		
Kensington&Chelsea_Flat_P11_DescL1_3Class	1527	1114	2000	413	132	281	0	3	31.96125908		
	1954	1376	2001	578	187	391	0	4	32.35294118		
	2058	1417	2002	641	219	422	0	9	34.16536661		
	1602	1108	2003	494	167	327	0	5	33.80566802		
	2112	1493	2004	619	228	391	0	4	36.83360258		
	1657	1131	2005	526	176	350	0	3	33.46007605		
	859	606	2006	253	84	169	0	5	33.20158103		
Kensington&Chelsea_Terrace_P11_DescL1_3Class	322	240	2000	82	23	59	0	4	28.04878049		
	388	284	2001	104	34	70	0	8	32.69230769		
	374	257	2002	117	38	79	0	5	32.47863248		
	311	229	2003	82	28	54	0	7	34.14634146		
	443	322	2004	121	50	71	0	6	41.32231405		
	369	253	2005	116	42	74	0	5	36.20689655		
	249	174	2006	75	18	57	0	5	24		
Kensington&Chelsea_SemiDetached_P11_DescL1_3Class	208	146	All	62	20	42	0	3	32.25806452		
Kensington&Chelsea_Detached_P11_DescL1_3Class	59	45	All	14	7	7	0	6	50		

Table 6-5: Borough_Level Classification Tests

As Table 6-5 shows, the average accuracy that is achieved is 34.18%. Although this level of accuracy can be considered as low, in this case it is an expected outcome. Given the size of the study area and the minimum structural description of the

properties, the price variation within the study area is wide. Additionally, the sample is not chosen on geographically based criteria, which means that it may not be uniformly distributed within the study area. This may lead to the non-representation of all the cases in the classification rules and therefore in the classifier, by resulting in unclassified or wrongly classified cases.

<p><i>HydePark_Flat_PL1_Desc1_3Class</i></p> <p>[Open Space=NEAR, Running Water=NEAR] -> [ZPRICE_RANGE=(510250-3000000)] [Confidence: 100.0% Support: 20.54%] [Transport=NEAR, Open Space=NEAR, Running Water=NEAR] -> [PRICE_RANGE=(510250-3000000)] [Confidence: 100.0% Support: 20.54%] [Commercial Services=NEAR, Running Water=NEAR, Open Space=NEAR] -> [PRICE_RANGE=(510250-3000000)] [Confidence: 100.0% Support: 20.54%] Default Class -> [(510250-3000000)] [Confidence: 0.0% Support: 0.0%]</p>
<p><i>HydePark_Flat_PL1_Desc2_3Class</i></p> <p>[Canal=NEAR, Green Space=NEAR] -> [ZPRICE_RANGE=(510250-3000000)] [Confidence: 100.0% Support: 20.72%] [Road and Rail=NEAR, Canal=NEAR, Green Space=NEAR] -> [PRICE_RANGE=(510250-3000000)] [Confidence: 100.0% Support: 20.72%] [Medical Establishments=NEAR, Personal, Consumer and Other S=NEAR, Repair and Servicing=NEAR] -> [PRICE_RANGE=[100000-326250]] [Confidence: 90.74% Support: 11.036%] [Medical Establishments=NEAR, Personal, Consumer and Other Services=NEAR, Accommodation=NEAR, Repair and Servicing=NEAR] -> [PRICE_RANGE=[100000-326250]] [Confidence: 90.74% Support: 11.04%] [Repair and Servicing=NEAR, Medical Establishments=NEAR, Personal, Consumer and Other S=NEAR, Legal and Financial=NEAR] -> [PRICE_RANGE=[100000-326250]] [Confidence: 90.74% Support: 11.036%] Default Class -> [(510250-3000000)] [Confidence: 0.0% Support: 0.0%]</p>
<p><i>HydePark_Flat_PL1_Desc3_3Class</i></p> <p>[Green Space=NEAR, Rails=NEAR] -> [ZPRICE_RANGE=(510250-3000000)] [Confidence: 100.0% Support: 20.40%] [Bus Stops=NEAR, Green Space=NEAR, Rails=NEAR] -> [PRICE_RANGE=(510250-3000000)] [Confidence: 100.0% Support: 20.40%] [Accommodation=NEAR, Cafe=NEAR, Estate Agencies=NEAR, Niche Goods=NEAR, Clothing and Accessories=NEAR] -> [PRICE_RANGE=[100000-326250]] [Confidence: 97.83% Support: 10.09%] [Cafe=NEAR, Clothing and Accessories=NEAR, Accommodation=NEAR, Restaurants=NEAR, Niche Goods=NEAR, Estate Agencies=NEAR] -> [PRICE_RANGE=[100000-326250]] [Confidence: 97.83% Support: 10.09%] [Property and Development Services=NEAR, Legal and Financial=NEAR, Chemists And Pharmacies=NEAR, Clinics And Surgeries=NEAR, Clothing and Accessories=NEAR] -> [PRICE_RANGE=[100000-326250]] [Confidence: 96.0% Support: 10.76%] [Chemists And Pharmacies=NEAR, Bus Stops=NEAR, Clothing and Accessories=NEAR] -> [PRICE_RANGE=[100000-326250]] [Confidence: 95.92% Support: 10.54%] [IT, Advertising, Marketing and=NEAR, Canal=NEAR] -> [PRICE_RANGE=(326250-510250)] [Confidence: 92.54% Support: 13.90%] Default Class -> [(510250-3000000)] [Confidence: 0.0% Support: 0.0%]</p>

Figure 6-27: Ward_Flat_PL1_DescL1,2,3_All

In order to investigate how the classification performs at lower geographic levels, the study area was reduced to the ward level. To demonstrate, the results acquired for the Hyde Park ward (Westminster Borough) that was also used for the locational influence tests are presented. A three level classification in the price range was kept and a 1% percent of the whole population was used for the creation of the test dataset. The sample was reduced, to make use of the majority of the known cases in the creation of the classifier. The test was performed for the three landuse levels of detail (see Figure 5-11). The resulted classifiers are shown in Figure 6-27. The accuracy obtained for levels of detail 1, 2, 3 were 28.57%, 50.0% and 75% respectively.

As we can see, by increasing the level of detail better results in terms of accuracy and in terms of classifier quality are acquired. Progressively all the classes are represented in the classifier improving in such way the result. This can be easily explained, as with specialisation different rules satisfy different classes while at the upper levels of detail there is an overlapping that result in the representation of the class by the most dominant rule.

Further tests at the Output Area level were omitted due to the limited number of transactions per Output Area. As illustrated by the examples, the method although works, it is extremely sensitive to the data. The absence of structural data limits the value of the outcome since it is bound to the wide price ranges.

6.4 Summary

A knowledge discovery process has been implemented for three London Boroughs. It consisted of all the typical stages of a knowledge discovery methodology that is: Application domain (Background Knowledge), Data selection, Cleaning/Pre-processing, Data transformation, Data mining and finally Interpretation and Evaluation.

In order to test the methodology, a case study has been carried out that consisted of two parts. The first part involved the investigation of the way location affects property prices. For this, a number of experiments have been performed. The first

group involved tests at Borough level and was performed for all the three levels of the landuse taxonomy. The sample was treated per property type. These tests resulted in the extraction of classification rules for the boroughs of Westminster, Kensington and Chelsea and Hammersmith and Fulham. The rules were further processed in order to investigate the positive or negative effect the different landuses have on property prices. Additional experiments have been performed at ward level to investigate the effect of less common landuses.

Despite data limitations, these experiments produced valid results and proved the ability of the method to identify the positive and negative relationships. They also highlighted the fact that locational effect varies upon the different submarkets.

The second part of the case study involved the testing of the association based classification. Similar to the locational influence tests, these were also performed at different geographical levels. The results demonstrated low accuracies at the higher geographical and detail levels. As the study area was reduced to the smaller units the accuracy and the quality of the classifier were improved.

7 Conclusions

This chapter provides a review of the research described in this thesis. Initially, a brief overview of the scope of the study is given followed by a reference to the importance and differentiation of the work. The first section concludes with a discussion about the strengths and limitations of the proposed methodology. In the next section, the research questions are reviewed and an account of the way they have been addressed is given. This is followed by a list of the research outcomes. Finally, recommendations about how this work could be expanded are provided.

7.1 Thesis Review

The research presented in this thesis involved the design of a knowledge discovery methodology and the implementation of a prototype platform that would accommodate such a process. The application area of this study was the area of property valuation and more specifically it involved the investigation of the way location affects the property prices and whether valuation could be based on such a method. An additional requirement was the complete lack of knowledge of the way the market behaves in the examined areas. That ensured an entirely data-driven approach.

This methodology is tightly associated with the application area that defines the nature of the input data. Hence, good data modelling is of great importance and strongly related to the success of the methodology. In this study, a graph theoretic approach was adopted for the data modelling that facilitates the method and also deals with the specialities of the different datasets employed in the study. This acts as

an input to the data mining system. The prototype system was designed to enable the uncovering of the desired and previously unknown information in an integrated manner.

This study was approached from two different perspectives. The first was to investigate the way the data mining technologies can be incorporated within the geographic domain and applied into real world problems. An additional consideration involved the viability of such approaches. Although knowledge discovery methodologies have been widely used in areas such as marketing, in geography it is considered as a relatively new area of research. The special nature of the data that associates with known geographical problems such as spatial autocorrelation played a major role in this delay in adaptation of such methodologies.

The second perspective emerged from the problem that the data mining process was called to solve. The fact that spatial arrangements can have a positive or negative effect on property prices has been widely discussed in the literature. The approach here was to investigate how far one can go by approaching the valuation process from an entirely spatial perspective.

Although a number of studies (see Sections 2.3 and 3.6) have been produced that face the challenges of both the above aspects, this research differentiates in a number of ways.

Despite of the recently increased interest in the adaptation and development of spatial data mining algorithms to deal with spatial data, studies have been mainly concentrated on the development aspect. Little focus has been placed on the viability of such approaches in real world problem solving within the geographical domain. This research investigates this, by shifting the focus in the application and in how the adopted methodology can produce optimal results. For this purpose, all the data that has been used in this study is real world data and reflects real situations.

In the majority of the previous property related studies the norm is to employ a statistical approach (see Section 3.6). In contrast, in this research an algorithmic approach was followed. An association rule mining technique, which is considered one of the core techniques of data mining, is used for the pattern uncovering while an association based classification method is used for the valuation. Furthermore, this

not only offers an integrated data mining system that accommodates the spatial data mining algorithm but also proposes a way to model and mine standard geographical datasets that is easily expandable.

In Chapter 3, the most popular property valuation methods and techniques have been reviewed. A special reference has been made on location aware techniques and how they incorporate location into their models. In Table 7-1, a comparison of these techniques and the one developed in this thesis is presented. The techniques are compared according to their output, ease of interpretation and general strengths and limitations.

In the comparative method, a number of similar properties to the one being valued are analysed in order to identify the different influences on the price that will lead to the accurate estimation of the value the property under consideration. To assess similarity, physical and legal characteristics of the properties are the most commonly used. To decide the degree of similarity between properties based on physical criteria is quite straightforward. On the other hand, accounting for locational differences is a complex task. In practice, there are two ways to deal with this when applying this methodology. The first is to take location into account by equating the external influences of the comparables by selecting them within a certain proximity to the property to be valued. The second way is to rely on the, often subjective, judgement of the valuer to quantify this. Both solutions have disadvantages. The success of the first solution depends entirely on the size of the cut off area since differences in prices can often occur within short distances especially in an urban environment. The latter requires knowledge of the mechanisms of the local market hence its success entirely depends on the experience of the valuer.

This methodology is a variation of the classic comparison method that attempts to enhance the way the influence of location is accounted. Instead of using a relatively small number of comparables to directly estimate the property value it uses a very large number of cases across wide geographical areas to identify frequent locational influence patterns. In this way, it is possible to identify the way locational factors affect the property value and assist the valuation process by identifying the pattern that reflects in the best way the locational situation of the property. This removes the reliance on a small number of comparables and any subjectivity issues.

<i>Approach</i>	<i>Model</i>	<i>Output</i>	<i>Ease of Interpretation</i>	<i>Strengths</i>	<i>Limitations</i>
Traditional	Comparative Method	Comparables	High	Simplicity Efficiency Ease of use Widely Accepted	Limited number of high quality comparables Indirect account for spatial influence Subjectivity
Hedonic models	Multiple regression	Parameter estimation	High	Conceptual soundness Benchmark Quantification of the effect caused by structural attributes or locational externalities	Issues related to the spatial nature of housing market Subjectivity High volume of data
ANN	Backpropagation or SOM	Classification Model or Clusters	Low	Patterns and trend detection Non-linearity handling Visualisation (SOM)	Lack of transparency Lack robustness Subjectivity High volume of data
Geospatial	Interpolation	Surface	High	Understanding of location effects through visualisation	Low predictive accuracy Subjectivity High volume of data
This work	Frequent Pattern Mining	Classification Rules	High	Frequent pattern detection Understandable results Transparency Objectivity	Numerical data handling High volume of data

Table 7-1: Comparison of approaches

Apart from the traditional comparative method there are also other approaches applied to the property valuation problem that have been extended to take into account location. Amongst the most popular are the hedonic modelling (Orford, 1999; Lake *et al.*, 2000) and the ANNs (Kauko, 2002). See also Table 3-5 and Table 3-6 for additional references in these areas. These techniques have been widely applied to the problem and presented interesting results.

Both techniques have advantages such as conceptual soundness and academic acceptance for the hedonic modelling and pattern recognition and non-linearity handling for ANNs. Weaknesses of the hedonic modelling technique include the multicollinearity, spatial autocorrelation and heteroscedasticity issues and also subjectivity in the identification of the variables. A number of methods that account for the spatial nature of the market have been proposed. Examples include the GWR, spatial expansion model and multi-level modelling (Orford, 1999) proposed to model spatial heterogeneity. Nevertheless, the subjectivity issue in the specification of the hedonic model remains. In the case of ANNs the subjectivity issue is also valid with the added problem of the lack of transparency. Although both techniques have similar track records of success the lack of explainability in the case of ANNs is considered a serious limitation for property valuation where justification and interpretation of results are of great importance. The approach developed in this thesis offers two advantages when compared to the other approaches. The first is clarity. The extracted classifiers are easy to understand and interpret by the end users. Hence, the valuation estimate is fully justified and supported. Moreover, this approach is objective since the classification rules are extracted following an inductive learning process based entirely on the dataset supplied. However, hedonic models and ANNs handle exact numeric values better.

An alternative approach to the one that extends the commonly used models to incorporate the location effect on property price is the use of geostatistical methods (Deddis *et al.*, 2002; Gallimore *et al.*, 1996). Interpolation techniques are used to quantify the location effects on property values in a combined manner. Justification of the valuation is achieved through 3D visualisation. Although there is an indication of the variation of prices upon location, there is no direct identification of the specific landuses or landscape features that cause this. In order to gain this information,

further examination with the use of supplementary information such as thematic maps is needed. This introduces subjectivity. In this approach, although the combined effect is taken into account there is also indication about the individual features in the output classification rules.

Overall, this methodology presents a number of strengths. It is based on standard data that is readily available, a fact that makes the application of this methodology easily applicable to other UK areas without requiring major adjustments.

As mentioned in Chapter 3 the most common way to take into account the location of the property in a valuation effort is the experience of the valuer. This introduces subjectivity in the result. In the case of automated location aware property valuation models or systems the design is such that the knowledge from the experts is appropriately captured. This approach faces problems related to subjectivity, knowledge elucidation and updating. Here, an inductive learning approach has been adopted instead. Unlike deductive learning, the learning process is based entirely on examples and hypotheses are generated based on similarities between them. In this case in particular, learning is achieved through associative classification rule induction that results to a classifier. This ensures objectivity since it does not require a priori assumptions about the relationship of the variables. It relies on the data for the uncovering of relationships, hence it is not biased. This is also supported by the way the system is designed that it does not require as an input the knowledge of the attributes that are included in the analysis.

The proposed data structure enables the easy expansion of the model. The graph model can be easily used to model other type of information without affecting the structure of the software. An example is the incorporation of other types of distances such as drive time and walking distance.

Due to the exploratory nature of this work, there are also some issues associated with it. One of the strengths of this methodology can potentially be seen as a weakness. Relying on data for the revealing of the possible associations makes the method vulnerable to data representation issues. That is, the results are entirely dependant on the level of detail. Additionally, the accuracy of the classification prediction is

entirely connected to the quality of the data. This means that dense and uniformly distributed sample points contribute to better results.

Another weakness is associated with the limited ability of the association rule mining algorithms to handle numeric data. This has a direct effect mainly in the classification, when this applied in large geographical areas (e.g. Borough Level), where fluctuations are big and the classifier fails to produce satisfactory results.

Finally, a limitation of the study relates to the limited information about the structure of the properties that constrain the analysis in the use of high level information such as flat, detached house etc.

The developed valuation system can be used in two ways. It can be used as a pattern recognition tool that aims to uncover patterns in housing markets. As demonstrated in the first part of the case study, the identification of most or least valued non-residential landuses or housing attributes is possible. This information can then be used to improve decision making in terms of land use planning. Furthermore, it can be used in conjunction with other valuation techniques such as hedonic modelling to assist in the model specification and improve objectivity.

The second use of the system is as an automated valuation system. Given that automated systems are commonly used in mass appraisals this approach can be used as such. Mass appraisals are useful when valuations must be conducted quickly and inexpensively. Data mining techniques are designed to deal efficiently with large amounts of data. Additionally, techniques that are traditionally associated with mass appraisal have weaknesses associated with the poor explainability and model complexity. It is more suitable for residential valuations since it is based on the principles of the comparative method where for commercial properties other methods are considered more suitable. Although there is a difference in the size of the target set between the mass and single valuations the main process remains the same. The system is designed so that can perform also single valuations. However, the use of price ranges instead of exact prices limits the applicability of this method in single valuations where the precision requirements are higher.

7.2 Research Questions Revisited

In Chapter 1 the research questions that set the framework for the current research were presented. In this section these questions are revisited and the way that they were addressed throughout the research is discussed.

What knowledge can be extracted from existing standard data sources. How could this be represented and stored into a spatial database?

A number of standard datasets, commonly used within the areas of geography have formed the geographical background that the mining process was based on. Based on extensive literature review in the areas of property valuation and knowledge discovery it was concluded that what was suited in this case was the extraction of knowledge in the form of association rules. This enabled the discovery of the combined way the location affects the property value. For the storage of the extracted association rules a relational database was used. This enabled the easy manipulation and examination of the rules since it is a dynamic form of storage as opposed to flat files.

How can location be modelled and successfully incorporated into a knowledge-based valuation model?

In accordance to the main requirement of the study that no a-priori assumption will be made in the way locational factors interact with the price, a data model that enables the investigation of the spatial arrangements was needed. To address this, a graph theoretic data model has been proposed and implemented. In this model spatial relationships are pre-calculated and stored as links of a weighted graph. By traversing the graph, spatial relationships can be extracted at several levels of depth. This enables the investigation of higher order spatial interactions without restricting in the immediate neighbourhood.

How does the spatial arrangement of landuses affect the property value based on real-world data?

Data from three central London boroughs has been collected resulting to a database of 50,000 known property transactions. A graph was constructed where the nodes

represented both the reference (known transactions) and task-relevant (non-residential landuses or landscape features) points. Three spatial relationships have been implemented and were represented as links in the graph. These were Adjacency, Containment and Proximity. A case study was performed at different geographical levels where association rules were extracted and further processed at different levels of abstraction.

Could such a location-driven methodology produce meaningful results? Does such an approach add value to the valuation process and how does this method compare to existing approaches?

Since the interest was also to investigate whether is possible to base a valuation process entirely on location, a second part of the case study was performed. In this, the extracted classification rules at different levels of abstraction and different geographical levels were used to form a classifier that assigned given test cases to the appropriate price range. In order to evaluate the accuracy of the method the sample at each case was split into build and test data. Results proved to be promising despite limitations of the input data, as discussed in Section 5.2.

7.3 Research Outcome

The main research outcomes that are the result of this work are listed below:

- An extensive review of the areas of knowledge discovery and geographical knowledge discovery. Identification and review of the most prevalent spatial data mining algorithms.
- An extensive review of the area of property valuation that covered all the aspects from current practices to the problem of incorporating location into valuation models.
- A knowledge based methodology that supports an entirely location aware approach to the property valuation problem.

- A novel approach for modelling location by using graph theory. This model was then used in the Apriori algorithm for frequent pattern mining and the CBA algorithm for classification data mining. These algorithms have been integrated into the property valuation system presented in this work.
- Design and development of a graph traversal algorithm that finds all paths comprising nearest neighbouring spatial objects with the property under investigation as the origin. This represents a novel approach for modelling location for the purpose of understanding the way it affects the value of the property under consideration.
- Design and development of a generic methodology for integrating key and yet easily accessible geographical datasets into a comprehensive landuse and landcover database on which property valuation decisions can be based.
- Use of that methodology for the implementation and population of an actual landuse and landcover database for the London boroughs of Westminster, Kensington and Chelsea and Hammersmith and Fulham. This contains a very large volume of data which represents 50,000 transactions leading to the construction of a spatial graph of approximately 80000 vertices and 3800000 edges. The algorithms were trained and tested on this very large dataset and led to conclusions for the case study that correspond to actual prices. This renders the proposed approach a viable and reliable tool for property valuation.
- Design and implementation of a database that combines detailed residential property data for the London boroughs of Westminster, Kensington and Chelsea and Hammersmith and Fulham.
- Implementation of the methodology and algorithms into a prototype software system that performs mining of spatial association rules and spatial associative classification on the dataset for the above mentioned boroughs of London.
- Design of a case study for the boroughs of Westminster, Kensington and Chelsea and Hammersmith and Fulham for the application of the system on

this, for the purposes of testing the methodology and algorithms. Further to this, the case study was used for the investigation of locational effects on property prices. This led to the understanding of the relationships between landuses and property location in these boroughs.

7.4 Future work recommendations

This research has investigated the application of a knowledge discovery methodology in the area of Geographic Information Science and in particular its use to assist in understanding the way location effects property prices. It addressed questions regarding the knowledge extraction and representation, data modelling, system design and the relationship between location and price. There are several directions that this research could be extended.

One aspect that can benefit from further investigation is that of the graph modelling. In the current research the graph comprises every spatial object in the database. This resulted to a very large number of edges that grew exponentially with the growth of the study area, leading to long computational times especially in path calculation. An efficient way to deal with such problems is the compression of the graph by performing a clustering technique that will result in a more compact and efficient data model. This entails research on suitable clustering techniques.

The conceptual design of the system (Figure 4-6) comprises an internet based architecture. The current prototype implementation however, includes only the algorithms and the database. This design can be extended to employ a Java Enterprise Edition application server such as JBoss, integration considerations with data sources and a web-based user interface. Direct interfaces to the data sources (perhaps employing web services) for the direct updating of the data in the model, especially the transactions are a technical extension that a commercial implementation could perhaps benefit.

The user interface is an area that justifies further research. As in any decision support system, visualisation is an important component that enables not only the better understanding of the results but also the input data and model as well. Such tools, can

therefore facilitate better understanding of methodologies, the problem and the proposed solution and result to a more precise decision making. The property valuation methodology and algorithms proposed in this work, will benefit from advanced visualisation tools to provide a better understanding of the association rules and more precision in the understanding of the effects of location in properties in different geographical areas.

The data mining technique employed in the mining process was that of the association rule mining. Although interesting results acquired, one of the problems encountered was that of the prevalence of the frequent landuses over those that occurred only rarely. For example, Commercial use could have overshadowed the effect of the proximity to a major park simply because commercial use is widely spread. An interesting exploration would be the use of weighted association rules that enable the application of weights that also contribute to the final result. In order to comply with the main requirement of not assuming the effect of different landuses, the weights could be applied not based on the type of the use but on the geographical area they cover. In that case, the problem described above could have been avoided.

For the valuation, the classification method that was used was based on an associative classification technique. The sorting scheme used in the development of the classifier was proposed in the CBA algorithm that bases the classification on the best rule. An interesting study would be the use of other classification schemes e.g. all rules, in order to investigate their effect on the classification accuracy.

Finally, for the realisation of the case study presented in Chapter 6 a sample parameterisation was used in order to demonstrate the use of the system and also to produce some initial results for the evaluation of the method. This could be extended in a more complete exploration that would lead in comparisons between the different parameterisations.

7.5 Conclusion

Analysing complex systems that involve human decisions and try to understand the reasoning behind their behaviour is not a trivial task. Relying on a computerised knowledge discovery methodology to reveal previously unknown knowledge is one of the approaches that attracted a lot of interest within the academic community. Although a number of success stories underline the importance and usefulness of such practices still the human role as a guide remains irreplaceable.

“We suppose ourselves to possess unqualified scientific knowledge of a thing, as opposed to knowing it in the accidental way in which the sophist knows, when we think that we know the cause on which the fact depends, as the cause of that fact and of no other, and, further, that the fact could not be other than it is.”

Aristotle, Posterior Analytics

References

- Adair, A.S., Berry, J.N., McGreal, W.S., 1996, Hedonic modelling, housing submarkets and residential valuation, *Journal of Property Research*, 13(1), p.67-83.
- Adamo, J-M, 2001, *Data Mining for Association Rules and Sequential Patterns*. Springer-Verlag, New York.
- Agrawal, R., Srikant, R., 1995, Mining Sequential Patterns. In: *Proceeding of the 11th International Conference on Data Engineering*, Taipei, Taiwan, p.3-14.
- Agrawal, R., Srikant, R., 1994, Fast algorithms for mining association rules. In: *Proceedings of the 20th International Conference on Very Large Databases*, p.487-499.
- Agrawal, R., Imielinski, T., Swami, A., 1993, Mining association rules between sets of items in large databases. In: *Proceedings of the ACM SIGMOD International Conference on Management of Data*, p.207-216.
- Agrawal, R., Gehrke, J., Gunopoulos, D. and Raghavan, P., 1998, Automatic subspace clustering of high dimensional data for data mining applications. In: *Proceedings of the 1998 ACM-SIGMOD International Conference on Management of Data*, p.49-60.
- Almond, N., Lewis, O.M., Jenkins, D.H., Gronow, S.A., Ware, J.A., 1997, Intelligent systems for the valuation of residential property. In: *Proceedings of the 1997 RICS-Cutting Edge Conference*. Available from: <http://www.glam.ac.uk/sot/doms/research/ai.php> [Accessed 25th November 2005]

- Alonso, W., 1965, *Location and Land Use – Toward a General Theory of Land Rent*. Harvard University Press, Cambridge.
- Alonso, W., 1960, A Theory of the Urban Land Market, Papers and Proceedings, Regional Science Association, 6, p.149-157. Cited in: Lloyd, P.E., Dicken, P., 1972, *Location in space: a theoretical approach to economic geography*. Harper & Row LTD.
- Ankerst, M., Breunig, M. M., Kriegel, H-P., Sander, J., 1999, OPTICS: Ordering Points To Identify the Clustering Structure. In: *Proceedings the 1999 ACM SIGMOD International Conference on Management of Data*, Philadelphia, p.49-60.
- Anon, 1996, Surveyor Wallow in Doom and Gloom. Daily Telegraph. Cited in: Almond, N., Lewis, O.M., Jenkins, D.H., Gronow, S.A., Ware, J.A., 1997, Intelligent systems for the valuation of residential property. In: *Proceedings of the 1997 RICS- Cutting Edge Conference*. Available from: <http://www.glam.ac.uk/sot/doms/research/ai.php> [Accessed 25th November 2005]
- Anselin, L., 1998, GIS Research Infrastructure for Spatial Analysis of Real Estate Markets. *Journal of Housing Research*, 9(1), p.113-133.
- Anselin, L., 1988, *Spatial Econometrics: Methods and Models*. Kluwer Academic Publishers, Netherlands.
- Appraisal Institute, 1992, The appraisal of Real Estate. Appraisal Institute, Chicago. Cited in: Wyatt, P., 1995, Using a Geographical Information System for Property Valuation, *Journal of Property Valuation and Investment*, 14(1), p.67-79.
- Asuncion, A & Newman, D.J. (2007). UCI Machine Learning Repository <http://www.ics.uci.edu/~mllearn/MLRepository.html>. Irvine, CA: University of California, Department of Information and Computer Science.

- Barbara, D., DuMouchel, W., Faloutsos, C., Haas, P.J., Hellerstein, J.H, Ioannidis, Y., Jagadish, H.V., Johnson, T., Ng, R., Poosala, V., Ross, K.A., Servcik, K.C., 1997, The New Jersey data reduction report. Bulletin of the Technical Committee on Data Engineering, 20(4). Cited in: Miller, H.J., Han, J., 2001, Geographic data mining and knowledge discovery-An Overview. In: Miller, H.J., Han, J.(eds), 2001, *Geographic Data Mining and Knowledge Discovery*. Taylor and Francis, p.3-32.
- Basu, S., Thibodeau, T.G., 1998, Analysis of Spatial Autocorrelation in House Prices, *The Journal of Real Estate Finance and Economics*, 17(1), p.61-85.
- Beckmann, N., Kriegel, H.P., Schneider, R., Seeger, B., 1990, The R*-tree: An efficient and robust access method for points and rectangles. In: *Proceedings of the 1990 ACM SIGMOD International Conference on Management of Data*, Atlantic City, USA, p.322-331.
- Benson, E.D., Hanson, J.L., Schwartz, A.L, and Smersh, G.T., 1998, Pricing Residential Amenities: The Value of a View, *Journal of Real Estate Finance and Economics*, 16 (1), p.55-73.
- Borgatti, S.P., M.G. Everett, and L.C. Freeman, 1999, *UCINET 5.0 Version 1.00*. Natick: Analytic Technologies.
- Borgelt, C., Kruse, R., 2002, Induction of association rules: Apriori implementation. In: *Proceedings of the 15th Conference on Computational Statistics*, p. 395–400.
- Britton, W., Davies, K., Johnson, T., 1989, Modern methods of valuation of land. London, Estates Gazette. Cited in: Wyatt, P., 1995, Using a Geographical Information System for Property Valuation, *Journal of Property Valuation and Investment*, 14(1), p.67-79.
- Butenfield, B., Gahegan, M., Miller, H.J., and Yuan, M., 2001, Geospatial Data Mining and Knowledge Discovery [Online]. *University Consortium for Geographic Information Science Research White Paper*. Available from: http://www.ucgis.org/priorities/research/research_white/2000%20Papers/emerging/gkd.pdf [Accessed 25th November]

- Carlson, E., 2002, Kohonen Map, GIS and the Analysis of Real Estate Sales. In: *Proceedings of FIG 2002, The XXII FIG International Congress*, The International Federation of Surveyors (FIG), Washington.
- Casetti, E., 1972, Generating Models by the Expansion Method: Applications to Geographical Research, *Geographical Analysis*, 4, p.81-91.
- Ceci, M., Appice, A., 2006, Spatial Associative Classification: Propositional vs Structural approach, *Journal of Intelligent Information Systems*, 27(3), p.191-213.
- Ceci, M., Appice, A., Malerba, D., 2004, Spatial Associative Classification at Different Levels of Granularity: A Probabilistic Approach. In: *Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases*, p.99-111.
- Chatfield, C., 1995, Model uncertainty, data mining and statistical inference (with discussion), *Journal of Royal Statistical Society*, 158, p.419-466
- Chawla, S., Shekhar, S., Wu, W.L., Tan, X., 2000, *Spatial Data Mining: An emerging tool for policy makers* [online]. CURA REPORTER. Available from: <http://www.cura.umn.edu/reporter/00-Sep/article2.pdf> [Accessed 25th November 2005]
- Chawla, S., Shekhar, S., Wu, W., Ozesmi, U., 2001, Modelling Spatial Dependencies for Mining Geospatial Data. In: Miller, H.J., Han, J. (eds), *Geographic Data Mining and Knowledge Discovery*, Taylor and Francis, p.131-159.
- Chawla, S., Shekhar, S., Wu, W., 2000, Modelling Spatial Dependencies for Mining Geospatial Data: A Statistical Approach, Technical Report.
- Cheshire, P., Sheppard, S., 1995, On the Price of Land and the Value of Amenities, *Economica*, 62, p.247-267.
- Clementini, E., Di Felice, P., Koperski, K., 2000, Mining Multiple-Level Spatial Association Rules for Objects with a Broad Boundary, *Data Knowledge and Engineering*, 34, p.251-270.

- Cliff, A.D., Ord, J.K., 1975, Model Building and the Analysis of Spatial Patterns in Human Geography. *Journal of the Royal Statistical Society*, 37, p.297-348.
- Cliff, A.D., Hagget, P., Ord, K., 1979, Graph Theory and Geography. In: Wilson, R.J, Beineke, L.W. (eds), 1979, Applications of Graph Theory, Academic Press, p. 293-326.
- Coenen, F., Leng, P., 2004, An Evaluation of Approaches to Classification Rule Selection. In: *Proceedings of the 4th IEEE International Conference on Data Mining (ICDM-04)*, Brighton, UK, p.359-362.
- Dale, P.F., McLaughlin, J.D., 1989, *Land Information Management-An Introduction with special reference to cadastral problems in Third World countries*, Oxford University Press.
- Deddis, W.G., Lamont, I., McCluskey, W.J., 2002, The Application of Spatially Derived Location Factors Within A GIS Environment. In: *Proceedings of the 2002 Pacific RIM Real Estate Society Conference, New Zealand*.
- Din, A., Hoesli, M., Bender, A., 2001, Environmental Variables and Real Estate Prices, *Urban Studies*, 38(11), p.1989-2000.
- Do, Q., Grudnitski, G., 1992, A neural network approach to residential property appraisal. *Real Estate Appraiser*, p.38-45. Cited in: Worzala, E., Lenk, M., Silva, A., 1995, An exploration of Neural Networks and its applications to Real Estate valuation. *The journal of Real Estate research*, 10(2), p.185-201.
- Dubin, R.A., 1998, Spatial Autocorrelation: A Primer, *Journal of HousingEconomics*, 7, p.304-327.
- Dubin, R.A., Pace, K.R., Thibodeau, T.G., 1999, Spatial autoregression techniques for real estate data. *Journal of Real Estate Literature*, 7(1), p.79-95. Cited in: Kauko, T., 2003, On current neural network applications involving spatial modelling of property prices. *Journal of Housing and the Built Environment*, 18(2), p.159-181.

- Dubin, R.A., Sung, C-H., 1990, Specification of Hedonic Regressions: Non-nested Tests on Measures of Neighbourhood Quality, *Journal of Urban Economics*, 27, p.97-110.
- Dubin, R.A., Sung, C-H., 1987, Spatial Variation in the Price of Housing: Rent Gradients in Non-Monocentric Cities, *Urban Studies*, 24, p.193-204.
- Dunn, E.S., 1954, The location of agricultural production. University of Florida Press, Gainesville. Cited in: Johnston, R.J., Gregory, D., Pratt, G., Watts, M., 2000, *The Dictionary of Human Geography*. WileyBlackwell.
- Dzeroski, S., 2003, Multi-Relational Data Mining: An Introduction, *SIGKDD Explorations*, 5(1), p.1-16.
- Eckert, J.K., 1990, Property appraisal and assessment administration. Chicago, IAAO. Cited in: RICS, 1998, *The Price is Right? Using computer-based mass appraisal techniques to value residential property*. The Royal Institution of Chartered Surveyors, London.
- Egenhofer, M., 1991, Reasoning about Binary Topological Relations, *Lecture Notes in Computer Science*, 525, Springer-Verlag, p. 143-160.
- Egenhofer, M., Franzosa, R., 1991, Point-Set Topological Spatial Relations, *International Journal of Geographical Information Systems*, 5 (2), p.161-174.
- Ester, M., Kriegel, H-P., Sander, J., 2001, Algorithms and Applications for Spatial Data Mining. In: Miller, H.J., Han, J.(eds), *Geographic Data Mining and Knowledge Discovery*. Taylor and Francis, p.131-159.
- Ester, M., Frommelt, A., Kriegel, H-P., Sander, J., 1999, Spatial Data Mining: Database Primitives, Algorithms and Efficient DBMS Support, *Data Mining and Knowledge Discovery an International Journal*.
- Ester, M., Kriegel, H-P., Sander, J., 1997, Spatial Data Mining: A Database Approach. In: *Proceedings of the Fifth International Symposium on Large Spatial Databases*, p.47-66.

- Ester, M., Kriegel, H-P., Sander, J., Xu, X., 1996, A density-based algorithm for discovering clusters in large spatial databases with noise. Cited in: Han, J., Kamber, M., Tung, A.K., 2001, Spatial clustering methods in data mining. In: Miller, H.J., Han, J. (eds), *Geographic Data Mining and Knowledge Discovery*. Taylor and Francis, p.188-217.
- Evans, A., 2004, *Economics, Real Estate and the Survey of land*, Wiley-Blackwell, p. 61-62.
- Evans, A.W., 1973, *The Economics of Residential Location*. Macmillan, London, p.281.
- Evans, A., James, H., Collins, A., 1991, Artificial neural networks: an application to residential valuations in the UK. *Journal of Property Valuation and Investment*, 11(1), 195-204. Cited in: Worzala, E., Lenk, M., Silva, A., 1995, An exploration of Neural Networks and its applications to Real Estate valuation. *The journal of Real Estate research*, 10(2), p.185-201.
- Evans, J.R., Minieka, E., 1992, *Optimization algorithms for networks and graphs*. CRC Press.
- Fayyad, U., 1997, Data Mining and Knowledge Discovery in Databases: Implications for Scientific Databases. In: *Proceedings of the 9th International Conference on Scientific and Statistical Database Management*, 11(13), p.2-11.
- Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., 1996A, From data mining to knowledge discovery: An Overview. In: Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (eds), 1996, *Advances in Knowledge Discovery and Data Mining*. AAAI Press, p.1-30.
- Fayyad, U.M, Piatetsky-Shapiro, G., Smyth, P., 1996B, From data mining to knowledge discovery in databases. *AI Magazine*, 17(3), p.37-54.
- Follain, J.R., Jimenez, E., 1985, Estimating the demand for housing characteristics: A critique and survey, *Regional Science and Urban Economics*, 15, p.77-107.

- Fotheringham, A. S., Charlton, M. E., Brundson, C. F., 1998, Geographical weighted regression: A natural evolution of the expansion method for spatial data analysis, *Environment and Planning A*, 30, p.1905-1927.
- Frank, A., 1996, Qualitative spatial reasoning: cardinal directions as an example, *International Journal of Geographical Information Science*, 10(3), p.269-290.
- Frank, A., 1991, Properties of geographic data: Requirements for spatial access methods. In: Proceedings of the Second International Symposium on Large Spatial Databases. Cited in: Guting, R.H, 1994, An introduction to spatial database systems. *The International Journal on Very Large Data Bases*, 3(4), p.357-399.
- Frans C, Paul L, 2004, An Evaluation of Approaches to Classification Rule Selection. In: *Proceedings of the 4th IEEE International Conference on Data Mining (ICDM'04)*. pp. 359-362
- Fraser, W.D., 1993, *Principles of property investment and pricing*. 2nd edition, Macmillan.
- Gahegan, M., 2001, *Data mining and knowledge discovery in the geographical domain* [online]. A National Academies white paper. Available from: <http://www7.national-academies.org/cstb/publications.html> [Accessed 25th November 2005]
- Gahegan, M., 2000, On the application of inductive machine learning tools to geographical analysis. *Geographical Analysis*, 32. Cited in: Miller, H.J., Han, J., 2001, Geographic data mining and knowledge discovery-An Overview. In: Miller, H.J., Han, J.(eds), 2001, *Geographic Data Mining and Knowledge Discovery*. Taylor and Francis, p.3-32.
- Gallimore, P., Fletcher, M., Carter, M., 1996, Modelling the influence of location on value, *Journal of Property Valuation and Investment*, 14(1), p.6-19.

- Gelfand, A.E., Ecker, M.D., Knight, J.R., Sirmans, C.F., 2004, The Dynamics of Location in Home Price. *Journal of Real Estate Finance and Economics*, 29(2), p.149-166.
- GeoMiner, 2004, GeoMiner: A knowledge discovery system for spatial databases and geographic information systems [Online]. Available from: <http://db.cs.sfu.ca/GeoMiner/> [Accessed 25th November 2005]
- Goethals, B., 2003, *Survey on frequent pattern mining*. Technical report, Helsinki Institute for Information Technology.
- Goodall, B., 1977, The economics of urban areas. Oxford, Pergamon Press. Cited in: RICS, 1999, *Geographical analysis in property valuation*. London, The Royal Institution of Chartered Surveyors.
- Gopal, S., Liu, W., Woodcock, C., 2001, Visualisation Based on the Fuzzy ARTMAP Neural Network for Mining Remotely Sensed Data. In: Miller, H.J., Han, J. (eds), *Geographic Data Mining and Knowledge Discovery*. Taylor and Francis, p.315-336.
- Gould, P., 1970, Is Statistix Inferens the Geographical Name for a Wild Goose?. *Economic Geography*, 46, p.439-448. Cited in: Haggett, P., A. D. Cliff and A. E. Frey., 1977, *Locational Analysis in Human Geography*, Second Edition. Arnold, London.
- Granger, C., 1969, Spatial data and time series analysis. *Studies in Regional Science*, p.1-24. Cited in: Haggett, P., A. D. Cliff and A. E. Frey., 1977, *Locational Analysis in Human Geography*, Second Edition. Arnold, London.
- Griffith, D., 1999, Statistical and mathematical sources of regional science theory: Map pattern analysis as an example. *Papers in Regional Science*, 78(1), p.21-45.
- Gronow, S., Scott, I., 1987, Information Technology and Building Society Valuations. *The Valuer*, p.58. Cited in: Almond, N., Lewis, O.M., Jenkins, D.H., Gronow, S.A., Ware, J.A., 1997, *Intelligent systems for the valuation of residential property* [online]. Available from:

- <http://www.glam.ac.uk/sot/doms/research/ai.php> [Accessed 25th November 2005]
- Grossman, R., Kasif, S., Moore, R., Rocke, D., Ullman, J., 1999, A Report on three NSF Workshops on Mining Large, Massive, and Distributed Data.....
- Guenther, O., Buchmann, A., 1990, Research issues in spatial databases. In: Proceedings of *ACMSIGMOD*. Cited in: Guting, R.H, 1994, An introduction to spatial database systems. *The International Journal on Very Large Data Bases*, 3(4), p.357-399.
- Guha, S., Rastogi, R., Shim, K., 1998, CURE: An Efficient Clustering Algorithm for Large Databases. In: *Proceedings the 1998 ACM-SIGMOD International Conference on Management of Data*, p.73-84.
- Guting, R.H, 1994, An introduction to spatial database systems. *The International Journal on Very Large Data Bases*, 3(4), p.357-399.
- Haggett, P., A. D. Cliff and A. E. Frey., 1977, *Locational Analysis in Human Geography*, Second Edition. Arnold, London.
- Han, J and Fu, Y., 1999, Mining multiple-level association rules in large databases. *IEEE Transactions on Knowledge and Data Engineering*, 11(5), p.798-805.
- Han, J. and Fu, Y.,1996, Attribute-oriented induction in data mining. In: Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P. and Uthurusamy, R. (eds), *Advances in Knowledge Discovery and Data Mining*. MA: MIT Press, Cambridge, p.399-421.
- Han, J. and Fu, Y., 1995, Discovery of Multiple-Level Association Rules from Large Databases. In: *Proceedings of the 21st International Conference on Very Large Databases*, p.420-431.
- Han, J., Kamber, M., Tung, A.K., 2001A, Spatial clustering methods in data mining. In: Miller, H.J., Han, J. (eds), *Geographic Data Mining and Knowledge Discovery*. Taylor and Francis, p.188-217.

- Han, J., Tung, A.K., He, J., 2001B, SPARC: Spatial Association Rule-Based Classification. In: Grossman, R.L., Kamath, C., Kegelmeyer, P., Kumar, V., Namburu, R. (eds), *Data Mining for Scientific and Engineering Applications*. Kluwer Academic Publishers, Netherlands, p.461-486.
- Han, J., Pei, J., Yin, Y., 2000, Mining Frequent Patterns without Candidate Generation. In: *Proceedings SIGMOD, Dallas*.
- Han, J., Yin, Y., Mao, R., (2004), Mining frequent patterns without candidate generation: A frequent-pattern tree. *Data Mining and Knowledge Discovery*, 8, Kluwer Academic Publishers, p.53-87.
- Harvey, D., 1972, *Society, the city and the space-economy of urbanism*. Washington DC, Association of American Geography. Cited in: Orford, S., 1999, *Valuing the Build Environment: GIS and house price analysis*. Ashgate Publishing Ltd.
- Haykin, S., 1998, *Neural Networks: A Comprehensive Foundation*. Prentice Hall.
- Jenkins, D.H., Lewis, O.M., Almond, N., Gronow, S.A., Ware, J.A., 1998, Towards an intelligent residential appraisal model. *Journal of Property Research*, 16(1), p.67-90.
- Johnston, R.J., Gregory, D., Pratt, G., Watts, M., 2000, *The Dictionary of Human Geography*. WileyBlackwell.
- Isaak, D., Steley, T., 2000, *Property Valuation Techniques*. Palgrave, New York.
- Isard, W., 1956, *Location and Space Economy*. John Wiley and Sons, New York. Cited in: Lloyd, P.E., Dicken, P., 1972, *Location in space: a theoretical approach to economic geography*. Harper & Row LTD.
- Karypis, G., Han, E-H., Kumar, V., 1999, CHAMELEON: A Hierarchical Clustering Algorithm Using Dynamic Modelling. *IEEE Computer*, 32(8), p.68-75.

- Kaufman and Rousseeuw, 1990, Finding groups in data, An introduction to cluster analysis. Belgium, John Wiley & Sons. Cited in: Han, J., Kamber, M., Tung, A.K., 2001, Spatial clustering methods in data mining. In: Miller, H.J., Han, J. (eds), *Geographic Data Mining and Knowledge Discovery*. Taylor and Francis, p.188-217.
- Kauko, T., 2002, Modelling the locational determinants of house prices: neural network and value tree approaches. PhD Thesis.
- Kauko, T., 2003, On current neural network applications involving spatial modelling of property prices. *Journal of Housing and the Built Environment*, 18(2), p.159-181.
- Klemettinen, M., Mannila, H., Ronkainen, P., Toivonen, H., Verkamo, A.I., 1994, Finding Interesting Rules from Large Sets of Discovered Association Rules. In: *Proceedings the 3rd International Conference on Information and Knowledge Management*, p.401-407.
- Klosgen, W., Zytrow, J.M., 1996, Knowledge discovery in databases terminology. In: Fayyad, U.M., Piatetski-Shapiro, G., Smyth, P., Uthurusamy, R.(eds) *Advances in knowledge discovery and data mining*. Cambridge, MA:MIT Press, p.573-592. Cited in: Miller, H.J., Han, J., 2001, Geographic data mining and knowledge discovery-An Overview. In: Miller, H.J., Han, J.(eds), 2001, *Geographic Data Mining and Knowledge Discovery*. Taylor and Francis, p.3-32.
- Knobbe, A.J., Blockeel, H., Siebes, A., Van der Wallen, D.M.G., 1999, Multi-Relational Data Mining. In: *Proceedings of Benelearn '99*, Leuven, Belgium.
- Knorr, E., Ng, R.T., 1996, Finding Aggregate Proximity Relationships and Commonalities in Spatial Data Mining, *IEEE Transactions on Knowledge and Data Engineering*, 8(6), p.884-897.
- Knox, P.L., 1995, Urban Social Geography: An Introduction. Routledge. Cited in: Orford, S., 1999, *Valuing the Build Environment: GIS and house price analysis*. Ashgate Publishing Ltd.

- Kohonen, T., 1982, Self-organised formation of topologically correct feature maps, *Biological Cybernetics*, 43, p.59-69.
- Kolari, P., Joshi, A., 2004, Web Mining – Research and Practice, *IEEE Computing in Science and Engineering*, 6(4), p.49-53.
- Koperski, K., Han, J., 1995, Discovery of Spatial Association Rules in Geographic Information Databases. In: *Proceedings of the 4th International Symposium of Advances in Spatial Databases*, 951, p.47-66.
- Koperski, K., Han, J., Adhikary, J., 1998A, Mining Knowledge in Geographical Data [online]. *IEEE Computer*. Available from: <http://www.db.cs.sfu.ca/sections/publication/kdd/kdd.html> [Accessed 25th November 2005]
- Koperski, K., Han, J., Adhikary, J., 1996, Spatial Data Mining: Progress and Challenges survey. In: *Proceedings of the ACM SIGMOD Workshop on research issues on data mining and knowledge discovery*. Montreal, Canada.
- Koperski, K., Han, J., Stefanovic, N., 1998B, An efficient two-step method for classification of spatial data. In: *Proceedings of the 1998 International Symposium on Spatial Data Handling*, Vancouver, Canada.
- Kothuri, R., Godfrind, A., Beinat, E., 2004, *Pro Oracle Spatial*. Apress.
- Kriegel, H.P., Borgwardt, K.M., Kroger, P., Pryakhin, A., Schubert, M., Zimek, A., 2007, Future trends in data mining, *Data Mining and Knowledge Discovery*, 15, p.87-97.
- Kuba, P., 2001, *Data structures for spatial data mining* [online]. Faculty of Informatics, Masaryk University, Report series. Available from: <http://www.fi.muni.cz/veda/reports/files/2001/> [Accessed 25th November 2005]
- Lafore, R., 2003, *Data Structures and Algorithms in Java*. Sams.

- Lake, I.R., Lovett, A.A., Bateman, I.J., Day, B., 2000, Using GIS and large-scale digital data to implement hedoning pricing studies. *International Journal of Geographical Information Science*, 14(6), p. 521-541.
- Lake, I.R., Lovett, A.A., Bateman, I.J., Langford, I.H., 1998, Modelling Environmental Influences on Property Prices in an Urban Environment. *Computers, Environment and Urban Systems*, 22(1), p. 121-136.
- Land Registry, 2004, Latest Property Prices [Online]. Available from: <http://www.landreg.gov.uk/propertyprice/interactive> [Accessed 25th November 2005]
- Land Registry, 2007, Quarterly Reports [Online]. Available from: <http://www.landreg.gov.uk/propertyprice/interactive/> [Accessed 19th December 2007]
- Lavrac, N., Motoda, H., Fawcett, T., Holte, R., Langley, P., Adriaans, P., 2004, Introduction: Lessons Learned from Data Mining Applications and Collaborative Problem Solving. *Machine Learning*, 57, p.13-34.
- Lawrance, D.M., Rees, W.H., Britton, W., 1971, *Modern Methods of Valuation of Land, Houses and Buildings*. The Estates Gazette Limited.
- Lee, K.J., Williams, P.D., Cheon S., 2008, Data Mining in Genomics. *Clinics in Laboratory Medicine*, 28(1), p.145-166.
- Lee, P. and Nevin B., Using GIS to Research low and Changing Demand for housing. In: Kidner, D., Higgs, G., White, S., (eds), 2003, *Socio-Economic Applications of Geographic Information Science*, Taylor and Francis, London, p. 119-132.
- Li, S., 1995, Marcov random field modelling. Computer Vision. Cited in: Shekhar, S., Schrater, P.R., Vatsavai, R.R., Wu, W., Chawla, S., 2002, Spatial Contextual Classification and Prediction Model for Mining Geospatial Data. *IEEE Transactions on Multimedia*, 4(2), p.174-187.

- Li, W., Han, J., Pei, J., 2001, CMAR: Accurate and Efficient Classification Based on Multiple Class-Association Rules. In: *IEEE International Conference on Data Mining*, p.369-376.
- Li, D., Di, K., Li, D., 2000, Land Use Classification of Remote Sensing Image with GIS Data based on Spatial Data Mining Techniques. In: *International Archives of Photogrammetry and Remote Sensing*. XXXIII (B3), Amsterdam, p.238-245.
- Li, M.M., Brown, H.J., 1980, Micro-Neighborhood Externalities and Hedonic Housing Prices, *Land Economics*, 56(2), p.125-141.
- Lisi, F., Malerba, D., 2004, Inducing multi-level association rules from multiple relations. *Machine Learning*, 55, p.175–210.
- Liu, G., Lu, H., Lou, W., Xu, Y., Yu, J.X., 2004, Efficient Mining of Frequent Patterns Using Ascending Frequency Ordered Prefix-Tree. *Data Mining and Knowledge Discovery*, 9(3), p.249-274.
- Liu, B., Hsu, W., & Ma, Y., 1998, Integrating classification and association rule mining. In: *Proceedings of 4th International Conference on Knowledge Discovery and Data Mining*, New York, USA, p. 27-31.
- Liu, B., Hsu, W., Mun, L-F., Lee, H-Y., 1996, Finding interesting patterns using user expectations. *IEEE Transactions on Knowledge Data Engineering*, 11, p.817-832.
- Local Economy, 2006, borough profile [online]. Available from: http://www.lbhf.gov.uk/Images/4.%20FINAL%20proof%20-%20Local%20economy_tcm21-56260.pdf [Accessed 20th April 2007]
- Lloyd, P.E., Dicken, P., 1972, *Location in space : a theoretical approach to economic geography*. Harper & Row LTD.
- Longley P., Higgs G., Martin D., 1996, The rates revisited? A geographical reassignment of property valuations and local tax burdens under the council tax, *Environment and Planning C*, 14, p.101–120.

- Longley, P., Higgs, G., Martin, D., 1994, The predictive use of GIS to model property valuations, *International Journal of Geographical Information Systems*, Vol. 8(2), p.217-235.
- Longley, P., Higgs, G., Martin, D., 1993, A GIS-based appraisal of Council Tax valuations, *Journal of Property Valuation and Investment*, 11(4), p.375-383.
- Lovell, M.C., 1983, Data Mining, *The Review of Economics and Statistics*, 65(1), p. 1-12.
- Lu, W., Han, J., Ooi, B.C., 1993, Discovery of General Knowledge in Large Spatial Databases. In: *Proceedings of the Far East Workshop on GIS*, Singapore, p.275-289.
- Mackmin, D., 1994, *The Valuation and Sale of residential Property*. Routledge.
- Malerba, D., Esposito, F., Lisi, F.A., Appice, A., 2002, Mining Spatial Association Rules in Census Data. *Research in Official Statistics*, 5(1), p.19--44.
- Malerba, D., Lisi, F.A., An ILP method for spatial association rule mining. *Working notes of the First Workshop on Multi-Relational Data Mining*, Freiburg, Germany, p.18--29.
- Mannila, H., Smyth, P., 2001, *Principles of Data Mining*. The MIT Press.
- Mannila, H., Toivonen, H., Verkamo, A.I., 1997, Discovery of Frequent Episodes in Event Sequences. In: *Proceedings of the 1st Conference on Knowledge Discovery and Data Mining*, Montreal, Quebec, p.210-215.
- Marshall, A., 1890, Principles of Economics [Online]. Macmillan and Co., Ltd., London. Available from: <http://www.econlib.org/library/Marshall/marP.html> [Accessed 19th December 2005]
- Marx, K., 1867, Capital: A Critique of Political Economy, Vol. I. The Process of Capitalist Production [Online]. Charles H. Kerr and Co., Chicago. Available from: <http://www.econlib.org/library/YPDBooks/Marx/mrxCpA.html> [Accessed 19th December 2005]

- Matheus, C.J., Chan, P.K., Piatetsky-Shapiro, G., 1993, *Systems for Knowledge Discovery in Databases*. IEEE Transactions on Knowledge and Data Engineering, 5(6), p.903-913.
- Matheus, C., Piatetsky-Shapiro, G., McNeil, D., 1994, An Application of KEFIR to the Analysis of Healthcare Information. In: *Proceedings of the 11th International Conference on Artificial Intelligence AAAI-94, Workshop on Knowledge Discovery in Databases*, p.25–36.
- May, M., Savinov, A., 2003, SPIN! An Enterprise Architecture for Spatial Data Mining [online]. Available from: <http://www.ais.fraunhofer.de/savinov/publicat/> [Accessed 25th November 2005]
- McCluskey, W., Anand, S., 1999, The application of intelligent hybrid techniques for the mass appraisal of residential properties. *Journal of property Investment and Finance*, 17(3), p.218-238.
- McCluskey, W., Deddis, W., Lamont, I., Borst, R., 2000, The application of surface generated interpolation models for the prediction of residential property values. *Journal of Property Investment and Finance*, 18(2), p.162-176.
- McCluskey, W., Deddis, W., Mannis, A., McBurney, D., Borst, R., 1997, Interactive Application of Computer Assisted Mass Appraisal and Geographic Information Systems. *Journal of Property Valuation and Investment*, 15(5), p.448-465.
- McLeod, P.B., 1984, The demand for local amenity: an hedonic price analysis, *Environment and Planning A*, 16(3), p.389-400.
- Meen, G., 1996, Spatial autoregression, spatial dependence and predictability in the UK housing market. *Housing studies*, 11(3), p.345-372. Cited in: Kauko, T., 2003, On current neural network applications involving spatial modelling of property prices. *Journal of Housing and the Built Environment*, 18(2), p.159-181.

- Meen, G, 2001, *Modelling spatial housing markets - Theory, Analysis and Policy*. Kluwer Academic Publishers.
- Michalski, R., 1980, Pattern recognition as rule-guided induction inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2, p.349-361. Cited in: Liu, B., Hsu, W., & Ma, Y., 1998, Integrating classification and association rule mining. In: *Proceedings of 4th International Conference on Knowledge Discovery and Data Mining*, New York, USA, p. 27-31.
- Miller, H.J, 2004, Geographic Data Mining and Knowledge Discovery. In: Wilson, J.P., Fotheringham, A.S. (eds), in press, *Handbook of Geographic Information Science* [online]. Blackwell. Available from: <http://www.geog.utah.edu/%7Ehmler/research.html> [Accessed 25th November 2005]
- Miller, H.J., Han, J., 2001, Geographic data mining and knowledge discovery-An Overview. In: Miller, H.J., Han, J.(eds), 2001, *Geographic Data Mining and Knowledge Discovery*. Taylor and Francis, p.3-32.
- Millington, A.F., 2001, *An Introduction to Property Valuation*. Estates Gazette.
- Militino, A.F., Ugarte, M.D., Garcia-Reinaldos, L., 2004, Alternative Models for Describing Spatial Dependence among Dwelling Selling Prices. *Journal of Real Estate Finance and Economics*, 29(2), p.193-209.
- Morimoto, Y., 2001, Mining Frequent Neighbouring Class Sets in Spatial Databases. In: *Proceedings of the International Conference on Knowledge Discovery and Data Mining ACM SIGKDD*, p.353-358.
- Muggleton, S., De Raedt, L., 1994, Inductive Logic Programming: Theory and Methods. *Journal of Logic Programming*, 19, p.629-679.
- National Research Council, 2003, *IT RoadMap to Geospatial Future* [online]. Available from: http://www.books.nap.edu/html/geospatial_future/ [Accessed 25th November 2005]

- Netz, A., Chaudhuri, S., Bernhardt, J., Fayyad, U., 2000, Integration of Data Mining and Relational Databases. In: *Proceedings of the 26th International Conference on Very Large Databases*, Cairo, Egypt, Morgan Kaufmann, p.719-722.
- Ng, R.T, 2001, Detecting Outliers from Large Datasets. In: Miller, H.J., Han, J. (eds), *Geographic Data Mining and Knowledge Discovery*. Taylor and Francis, p.188-217.
- Openshaw, S., 1999, Geographical data mining: key design issues [Online]. In: *Proceedings of the 1999 GeoComputation International Conference*, USA. Available from:
http://www.geovista.psu.edu/sites/geocomp99/Gc99/051/gc_051.htm [Accessed 25th November 2005]
- Openshaw, S., 1984, *The modifiable areal unit problem, Concepts and Techniques in Modern Geography*. Geo Books, Norwich.
- Openshaw, S., Turner, A., Turton, I., Macgill, J., 1999, Testing space-time and more complex hyperspace geographical analysis tools [Online]. In: *Proceedings of the 1999 GISRUUK Conference*. UK. Available from:
<http://www.ccg.leeds.ac.uk/smart/hyper.html> [Accessed 25th November 2005]
- Ordnance Survey, 2004, OS MasterMap [Online]. Available from:
<http://www.ordsvy.gov.uk/oswebsite/products/osmastermap/> [Accessed 25th November 2005]
- Orford, S., 1999, *Valuing the Build Environment: GIS and house price analysis*. Ashgate Publishing Ltd, England.
- O'Sullivan, D., Unwin, D.J., 2002, *Geographic Information Analysis*. John Wiley & Sons, New Jersey.
- Pace, R.K., Barry, R., Gilley, O.W., Sirmans, C.F., 2000, A method for spatial-temporal forecasting with an application to real estate prices, *International Journal of Forecasting*, 16(2), p.229-246.

- Page, D., Craven, M., 2003, Biological applications of multi-relational data mining. *ACM SIGKDD Explorations Newsletter*, 5(1), p.69-79.
- Pagourtzi, E and Assimakopoulos, V., Hatzichristos, T., French, N., 2003, Real estate appraisal: a review of valuation methods. *Journal of Property Investment and Finance*, 21(4), p.383-401.
- Piatetsky-Shapiro, G., 2007, Data mining and knowledge discovery 1996 to 2005: overcoming the hype and moving from “university” to “business” and “analytics”. *Data Mining and Knowledge Discovery*, 15, p.99-105.
- Piatetsky-Shapiro, G., 1991, Knowledge discovery in real databases: A report on the UCAI-89 Workshop. *AI Magazine*, 11(5), p.68-70.
- Piatetsky-Shapiro, G., 1991b, Discovery, Analysis and Presentation of Strong Rules. In: Piatetski-Shapiro, G., Frawley, W.J.(eds), *Knowledge Discovery in Databases*. AAAI Press/ The MIT Press, p.229-248.
- Pompe, J.J., Rinehart, J.R., 1995, Beach Quality and the Enhancement of Recreational Property Values, *Journal of Leisure Research*, 27, p.143-154.
- Preparata, F.P., Shamos, M.I., 1988, Computational Geometry: An Introduction. Springer-Verlag, New York. Cited in: Ng, R.T, 2001, Detecting Outliers from Large Datasets. In: Miller, H.J., Han, J. (eds), *Geographic Data Mining and Knowledge Discovery*. Taylor and Francis, p.188-217.
- Pyle, D.,1999, *Data preparation for data mining*. Morgan Kaufmann, San Francisco.
- Quinlan, J.R., 1986, Induction of decision trees. Machine learning. Cited in: Ester, M., Kriegel, H-P., Sander, J., 1997, Spatial Data Mining: A Database Approach. In: *Proceedings of the Fifth International Symposium on Large Spatial Databases*, p.47-66.
- Quinlan, J.R., 1993, *C4.5: Programs for machine learning*. Morgan Kaufmann Publishers Inc. San Francisco, USA. Cited in: Li, D., Di, K., Li, D., 2000, Land Use Classification of Remote Sensing Image with GIS Data based on

- Spatial Data Mining Techniques. In: *International Archives of Photogrammetry and Remote Sensing*. XXXIII (B3), Amsterdam, p.238-245.
- Quinlan, J.R. and Cameron-Jones, R.M., 1993, FOIL: A Midterm Report. In: *Proceedings of 1993 European Conference on Machine Learning*, Vienna, Austria, p.3-20.
- Reinartz, T., 1999, Focusing solutions for data mining. Lecture Notes in Artificial Intelligence 1623. Cited in: Miller, H.J., Han, J., 2001, Geographic data mining and knowledge discovery-An Overview. In: Miller, H.J., Han, J.(eds), 2001, *Geographic Data Mining and Knowledge Discovery*. Taylor and Francis, p.3-32.
- Ricardo, D., 1817, *On the Principles of Political Economy and Taxation* [Online]. John Murray, London. Available from: <http://www.econlib.org/LIBRARY/Ricardo/ricP.html> [Accessed 19th December 2005]
- RICS (2003), *RICS Appraisal and Valuation Manual (Red Book)*. Royal Institution of Chartered Surveyors, London.
- RICS, 1999, *Geographical analysis in property valuation*. The Royal Institution of Chartered Surveyors, London.
- RICS, 1998, *The Price is Right? Using computer-based mass appraisal techniques to value residential property*. The Royal Institution of Chartered Surveyors, London.
- Rogers A., 1969, Quadrat analysis of urban dispersion: 1. Theoretical techniques. *Environment and Planning*, 1(1), p.47 – 80. Cited in: Haggett, P., A. D. Cliff and A. E. Frey., 1977, *Locational Analysis in Human Geography*, Second Edition. Arnold, London.
- Rosen, S., 1974, Hedonic prices and implicit markets: product differentiation in pure competition. *The journal of political economy*, 82(1), p.34-55.

- Rumbaugh, J., Jacobson, I., Booch, G., 2005, *The Unified Modelling Language Reference Manual*, Second Edition, Pearson Education.
- Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986, Learning internal representations by error propagation. In: Rumelhart, D.E., McClelland, J.L., PDP Group, *Explorations in the microstructure of cognition*, Cambridge, MA:MIT Press.
- Savinov, A., 2003, Mining spatial rules by finding empty intervals in data. In: *Proceedings of the 2003 International Conference on Knowledge-Based Intelligent Information & Engineering Systems*, Oxford, University of Oxford, p.1058-1063.
- Scarrett, D., 1991, *Property Valuation – The Five Methods*. E & FN Spon, London.
- Sheikholeslami, G., Chatterjee, S., Zhang, A., 1998, WaveCluster: a multi-resolution clustering approach for very large spatial databases. In: *Proceedings of the 1998 International Conference on Very Large Data Bases*, New York City, p. 428-439.
- Shekhar, S., Chawla, S., 2003, *Spatial Databases: A Tour*. Prentice Hall, p.182-226.
- Shekhar, S., Huang, Y., Wu, W., Lu, C.T., Chawla, S., 2001, What's spatial about spatial data mining: three case studies. In: Kumar, V., Grossman, R., Kamath, C., Nambuku, R. (eds), *Data mining for scientific and engineering applications*. Kluwer Academic Publications, p.1-28.
- Shekhar, S., Lu, C.T., Zhang, P., 2003, A Unified Approach to Detecting Spatial Outliers. *GeoInformatica*, 7(2), p.139-166.
- Shekhar, S., Schrater, P.R., Vatsavai, R.R., Wu, W., Chawla, S., 2002, Spatial Contextual Classification and Prediction Model for Mining Geospatial Data. *IEEE Transactions on Multimedia*, 4(2), p.174-187.
- Silberschatz, A. and Tuzhilin, A., 1996, What makes patterns interesting in knowledge discovery systems. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 8(6), p.970-974.

- Smith, A., 1776, *An Inquiry into the Nature and Causes of the Wealth of Nations* [online]. Methuen and Co., Ltd., London. Available from: [URL:http://www.econlib.org/LIBRARY/Smith/smWN.html](http://www.econlib.org/LIBRARY/Smith/smWN.html) [Accessed 19th December 2005]
- SPIN!, 2004, *SPIN! Spatial mining for data of public interest* [Online]. Available from: <http://www.ccg.leeds.ac.uk/spin/overview.html> [Accessed 25th November 2005]
- Srikant, R., Agrawal, R., 1996, Mining Quantitative Association Rules in Large Relational Tables. In: *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data*, Montreal, Canada, p.1-12.
- Srikant, R., Agrawal, R., 1995, Mining Generalised Association Rules. In: *Proceedings of the 21st International Conference on Very Large Databases*, p.407-419.
- Stapleton, T., 1989, Property research (The Estates Gazette Professional Guides). England, Estates Gazette.
- Tay, D.P.H., Ho, D.K.K., 1992, Artificial intelligence and the mass appraisal of residential apartments. *Journal of property valuation and investment*, 10(2), p.525-540. Cited in: Worzala, E., Lenk, M., Silva, A., 1995, An exploration of Neural Networks and its applications to Real Estate valuation. *The journal of Real Estate research*, 10(2), p.185-201.
- Thabtah, F., Cowling, P., Peng, Y., 2005, A Study of Predictive Accuracy for Four Associative Classifiers. *Journal of Digital Information Management*, 3(3), p.209-212.
- Thériault M., Rosiers F.D., Villeneuve P., Kestens Y., 2003, Modelling interactions of location with specific value of housing attributes, *Property Management*, 21(1), p.25-62.
- Thrall, G., 2002, *Business Geography and New Real Estate Market Analysis*. Oxford University Press, London and New York.

- Thrall, G., 2001, Data Resources for Real Estate and Business Geography Analysis: A Comprehensive Structured Annotated Bibliography. *Journal of Real Estate Literature*, 9(2), p.175-225.
- Tobler, W.R., 1970, A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46(2), p.234-240.
- Tukey, J.W., 1975, Mathematics and picturing data. In: Proceedings of International Congress of Mathematics, Vancouver, 2, p.523-531. Cited in: Ng, R.T, 2001, Detecting Outliers from Large Datasets. In: Miller, H.J., Han, J. (eds), *Geographic Data Mining and Knowledge Discovery*. Taylor and Francis, p.188-217.
- Tukey, J.W., 1977, Exploratory Data Analysis. Addison-Wesley, Reading. Cited in: Ng, R.T, 2001, Detecting Outliers from Large Datasets. In: Miller, H.J., Han, J. (eds), *Geographic Data Mining and Knowledge Discovery*. Taylor and Francis, p.188-217.
- Turner A.G.D., 2002, State of the Art Geographical Data Mining [Online]. SPIN!-project *Work Package 5 Report, CCG Working Paper*. Available from: <http://www.geog.leeds.ac.uk/people/a.turner/publications/archive/Turner2002.html> [Accessed 25th November 2005]
- Turner, D.M., 1977, *An approach to Land Values*. Geographical Publications Limited, United Kingdom.
- Valuation Office, 2004, *Council tax homepage* [Online]. Available from: http://www.voa.gov.uk/council_tax/index.htm [Accessed 25th November 2005]
- Vickers, T., 2003, *Mapping UK Property Values Today* [Online]. Available from: www.landvaluescape.org/papers/Mapping_UK_Property.pdf [Accessed 25th November 2005]
- Vickers, T., Thurstain-Goodwin, M., 2002, Visualising Landvaluescape without a Cadastre. In: *Proceedings of FIG XXII International Congress*, Washington.

- Wabe, J.S., 1971, A Study of House Prices as a means of Establishing the Value of Journey Time, the Rate of Time Preference and the Valuation of some Aspects of Environment in the London Metropolitan Region, *Applied Economics*, 3(4), p.247-255.
- Wang, F, 2006, *Quantitative methods and applications in GIS*. CRC Press.
- Wang, W., Yang, J., Muntz, R., STING: A statistical information grid approach to spatial data mining. In: Proceedings of the 1997 International Conference on Very Large Data Bases. Cited in: Han, J., Kamber, M., Tung, A.K., 2001, Spatial clustering methods in data mining. In: Miller, H.J., Han, J. (eds), *Geographic Data Mining and Knowledge Discovery*. Taylor and Francis. p.188-217.
- Washio, T., 2007, Applications eligible for data mining. *Advanced Engineering Informatics*, 21(3), p.241-242.
- Webb, G., 2000, Efficient Search for Association Rules. In: *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p.99-107.
- West, D.B., 2000, *Introduction to Graph Theory*. Prentice Hall.
- Wilkinson R.K., 1973, House Prices and the Measurement of Externalities, *The Economic Journal*, 83 (329), p.72-86.
- Wilson, A.G., 2000, *Complex Spatial Systems: The Modelling Foundations of Urban and Regional Analysis*. Pearson Education.
- Wolverton, M., Diaz, J., 1996, Investigation into Price Knowledge Induced Comparable Selection Bias. Cited in: Almond, N., Lewis, O.M., Jenkins, D.H., Gronow, S.A., Ware, J.A., 1997, Intelligent systems for the valuation of residential property. In: *Proceedings of the 1997 RICS-Cutting Edge Conference*. Available from: <http://www.glam.ac.uk/sot/doms/research/ai.php> [Accessed 25th November 2005]
- Worboys, M., Duckham, M., 2004, *GIS A Computing Perspective*, CRC Press.

- Worzala, E., Lenk, M., Silva, A., 1995, An exploration of Neural Networks and its applications to Real Estate valuation. *The journal of Real Estate research*, 10(2), p.185-201.
- Wu X.D., Kumar V., Quinlan J.R., Ghosh J., Yang Q., Motoda H., McLachlan G.J., Ng A., Liu B., Yu P.S., Zhou Z.H., Steinbach M., Hand D.J., Steinberg D., 2008, Top Ten Algorithms in Data Mining. *Knowledge and Information Systems*, 14(1), p. 1-37.
- Wyatt, P., Ralphs, M., 2003, *GIS in Land and Property Management*. Spon Press, London.
- Wyatt, P., 1997, The development of a GIS-based property information system for real estate valuation. *International Journal of Geographical Information Science*, 11(5), p.435-450.
- Wyatt, P., 1996, Using a Geographical Information System for Property Valuation. *Journal of Property Valuation and Investment*, 14(1), p.67-79.
- Yin, X., Han, J., 2003, CPAR: Classification based on predictive association rules. In: *Proceedings of 2003 International Conference on Data Mining (SDM'03)*, Siam, .
- Yoo, J-S., Shekhar, S., 2006, A Joinless Approach for Mining Spatial Colocation Patterns. *IEEE Transactions on Knowledge and Data Engineering*, 18(10), p.1323-1337
- Yoo, J-S, Shekhar, S., 2004, A Partial Join Approach for Mining Co-location Patterns. In: *Proceedings. of the 12th ACM International Symposium on Advances in Geographic Information Systems(ACM-GIS)*, Washington D.C., USA.
- Yuan, M., Battenfield, B., Gahegan, M., Miller, H., 2001, *Geospatial Data Mining and knowledge discovery* [online]. A UCGIS White Paper on Emergent Research Themes. Available from: http://www.ucgis.org/priorities/research/research_white/2000%20Papers/emerging/ [Accessed 25th November 2005]

- Zhang, T., Ramakrishnan, R., Livny, M., 1996, BIRCH: An Efficient Data Clustering Method for Very Large Databases. In: *Proceedings of the 1996 ACM-SIGMOD International Conference on Management of Data*, Montreal, Canada, p.103-114.
- Zheng, Z., Kohani, R., Mason, L., 2001, Real World Performance of Association Rule Algorithms. In: *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, California, p.401-406.

Appendix A- I2I Landuse

<i>NLUD CODE</i>	<i>DESCRIPTION</i>
1.2 Ploughed field	
1.3 Fallow land	
4.2 Standing water	
4.3 Running water	
4.5 Salt marsh	
7.1 Indoor recreation	Amusement and show places Libraries, museums and galleries Sports facilities Holiday camps and parks
7.2 Outdoor Recreation	Outdoor Amenity & Open Spaces Sports Facilities Holiday camps and Parks
7.3 Allotments	Allotment and City Farms
8.1 Roads	Transport tracks and ways Transport terminal and interchanges
8.2 Car parks	Public car parks
8.3 Railways	Transport tracks and ways Transport terminal and interchanges Vehicle storage Goods and freight terminals
8.5 Docks	
9.1 Residential	Dwellings
9.2 Institutional & Communal/Accommodation	Hotels, boarding and guest houses Residential Institutions
10.1 Institutional Buildings	Medical and health care services Community Services
10.2 Educational buildings	Education
10.3 Religious buildings	Places of worship
11.1 Industry	Manufacturing
11.2 Office	Offices Financial and Professional Services
11.3 Retailing	Shops Restaurants and Cafes Public houses and bars
11.4 Storage & warehousing	Storage Wholesale distribution Energy production and distribution
11.5 Utilities	Energy production and distribution Water supply and treatment Cemeteries and Crematoria Post and telecommunications
12.1 Vacant land previously developed	Vacant
13 Defense land & Buildings	Defense

Appendix B - Output .txt file

Chelsea&Kensington_Detached_P11_Lall1_3Class_2000.txt

Computer Aided Property Valuation: [version 0.0.0.10-alpha]
Copyright (C) 2007 Aikaterini Christopoulou

DATA TRANSFORMATION

Data preparation steps

Create binned data view...
Prepare_Build_Data is started, please wait. Prepare_Build_Data is successful.
Create test table...
Create build data view...
Create build data view...
Execute column format transformation...
Execute column format transformation...
Execute pivot query...
Execute pivot query...

BUILD MODEL

Configuration of the Association Rule Mining algorithm

arBuildTask_jdm is started, please wait. arBuildTask_jdm is successful.
BuildSettings Details from the arSettings_jdm table:
Table : arSettings_jdm
SETTING_NAME SETTING_VALUE
ALGO_NAME ALGO_APRIORI_ASSOCIATION_RULES
ASSO_MAX_RULE_LENGTH 10
ASSO_MIN_CONFIDENCE 0.1
ASSO_MIN_SUPPORT 0.1
JDMS_FUNCTION_TYPE ASSOCIATION
BuildSettings Details from the arSettings_jdm model build settings object:
Algorithm Name: aprioriAssociationRules
Function Name: association
Max Number of Rules: 10
Min Confidence: 10
Min Support: 10
Model Name: ARMODEL_JDM

DISPLAY CLASSIFICATION RULES

Extracted Classification Rules

Rule 5: Commercial Services= NEAR ==> PRICE_RANGE= [1610000-5800000] (support=100, confidence=100)
Rule 9: Manufacturing and Production= NEAR ==> PRICE_RANGE= [1610000-5800000] (support=100, confidence=100)
Rule 11: Open Space= NEAR ==> PRICE_RANGE= [1610000-5800000] (support=100, confidence=100)
Rule 16: Commercial Services= NEAR Manufacturing and Production= NEAR ==> PRICE_RANGE= [1610000-5800000] (support=100, confidence=100)
Rule 19: Commercial Services= NEAR Open Space= NEAR ==> PRICE_RANGE= [1610000-5800000] (support=100, confidence=100)
Rule 22: Manufacturing and Production= NEAR Open Space= NEAR ==> PRICE_RANGE= [1610000-5800000] (support=100, confidence=100)
Rule 25: Commercial Services= NEAR Manufacturing and Production= NEAR Open Space= NEAR ==> PRICE_RANGE= [1610000-5800000] (support=100, confidence=100)

SORT CLASSIFICATION RULES

CBA based sorting of the Classification Rules

[Pos in List: 0]RuleId:11: [Open Space=NEAR] -> [PRICE_RANGE=[1610000-5800000]](confidence: 100.0, support: 100.0) (CRule = true Strong cRule= true)
[Pos in List: 1]RuleId:9: [Manufacturing and Production=NEAR] -> [PRICE_RANGE=[1610000-5800000]](confidence: 100.0, support: 100.0) (CRule = false Strong cRule= false)
[Pos in List: 2]RuleId:5: [Commercial Services=NEAR] -> [PRICE_RANGE=[1610000-5800000]](confidence: 100.0, support: 100.0) (CRule = false Strong cRule= false)
[Pos in List: 3]RuleId:22: [Manufacturing and Production=NEAR, Open Space=NEAR] -> [PRICE_RANGE=[1610000-5800000]](confidence: 100.0, support: 100.0) (CRule = false Strong cRule= false)
[Pos in List: 4]RuleId:19: [Commercial Services=NEAR, Open Space=NEAR] -> [PRICE_RANGE=[1610000-5800000]](confidence: 100.0, support: 100.0) (CRule = false Strong cRule= false)
[Pos in List: 5]RuleId:16: [Manufacturing and Production=NEAR, Commercial Services=NEAR] -> [PRICE_RANGE=[1610000-5800000]](confidence: 100.0, support: 100.0) (CRule = false Strong cRule= false)
[Pos in List: 6]RuleId:25: [Commercial Services=NEAR, Manufacturing and Production=NEAR, Open Space=NEAR] -> [PRICE_RANGE=[1610000-5800000]](confidence: 100.0, support: 100.0) (CRule = false Strong cRule= false)

PRINT CLASSIFIER

Classifier: Classification Rules and the Default Class

[Open Space=NEAR] -> [PRICE_RANGE=[1610000-5800000]] [Confidence: 100.0% Support: 100.0%]
Default Class -> [[1610000-5800000]] [Confidence: 0.0% Support: 0.0%]

CLASSIFY

Test Case Classification Result

Transaction id = 48011 Class label = [1610000-5800000]

DISPLAY ACCURACY

Classification Accuracy Information

correctClassCounter = 1
unclassifiedCounter = 0
wrongClassCounter = 0
Num case = 1
accuracy = 100.0
Accuracy = 100.0

Appendix C- Procedures

geoReference*

Procedure 1

```
CREATE OR REPLACE PROCEDURE geoReference is
BEGIN
    UPDATE transactions A
    SET reftotopoarea =
    (SELECT referencetotopographicarea FROM addresspoint B
    WHERE B.postaladdress_buildingnumber = A.buildingno AND
    B.postaladdpostalcodepostalcode = A.postcode AND rownum=1
    OR A.postaladdress_buildingname = B.buildingname AND
    A.postaladdpostalcodepostalcode = B.postcode AND rownum=1);
    COMMIT;
    UPDATE transactions A
    SET polygon = SELECT polygon FROM topographicarea B
    WHERE B.TOID = A.reftotopoarea;
    COMMIT;
END geoReference;
/
```

**Procedure that adds polygonal reference to transactions*

getRelations*

Procedure 2

```
CREATE OR REPLACE PROCEDURE getRelations AS
BEGIN
    INSERT /*+APPEND*/ INTO relation
    SELECT A.landuse_id, B.landuse_id, 'ADJACENCY'
    FROM landuse A, landuse B WHERE
    SDO_RELATE(A.geometry, B.geometry, 'MASK=TOUCH')= 'TRUE'
    COMMIT;
    INSERT /*+APPEND*/ INTO relation
    SELECT A.LANDUSE_ID, B.LANDUSE_ID, 'EQUALS' FROM landuse A,
    landuse B WHERE
    SDO_RELATE(A.geometry, B.geometry, 'MASK=EQUAL')= 'TRUE' AND
    A.landuse_id != B.landuse_id
    COMMIT;
```

```

INSERT /*+APPEND*/ INTO relation
SELECT A.landuse_id, B.landuse_id, 'INSIDE'
FROM landuse A, landuse B WHERE
SDO_RELATE(A.geometry, B.geometry, 'MASK = CONTAINS') = 'TRUE'
COMMIT;
INSERT /*+APPEND*/ INTO relation
SELECT A.landuse_id, B.landuse_id, 'NEAR' FROM
landuse A, landuse B
WHERE
SDO_WITHIN_DISTANCE(A.geometry, B.geometry,
'DISTANCE = 80M') = 'TRUE' AND SDO_TOUCH(A.geometry,
B.geometry, 'MASK = TOUCH + CONTAINS +EQUAL') != 'TRUE'
ORDER BY A.landuse_id
COMMIT;
INSERT /*+APPEND*/ INTO relation
SELECT B.unique_reference_number, A.landuse_id, 'ADJACENCY'
FROM landuse A, trans_poly B
WHERE SDO_TOUCH(A.geometry, B.geoloc)= 'TRUE'
COMMIT;
INSERT /*+APPEND*/ INTO relation
SELECT B.unique_reference_number, A.LANDUSE_ID, 'EQUALS'
FROM landuse A, trans_poly B
WHERE SDO_EQUAL(B.geoloc,A.geometry)= 'TRUE'
COMMIT;
INSERT /*+APPEND*/ INTO relation
SELECT A.unique_reference_number, B.landuse_id, 'NEAR'
FROM trans_poly A, landuse B
WHERE
SDO_WITHIN_DISTANCE(A.geoloc, B.geometry, 'DISTANCE = 80M') =
'TRUE' AND SDO_TOUCH(A.geoloc, B.geometry) != 'TRUE'
AND SDO_CONTAINS(A.geoloc, B.geometry) != 'TRUE' AND
SDO_INSIDE(A.geoloc, B.geometry) != 'TRUE'
COMMIT;
END getRelations;
/

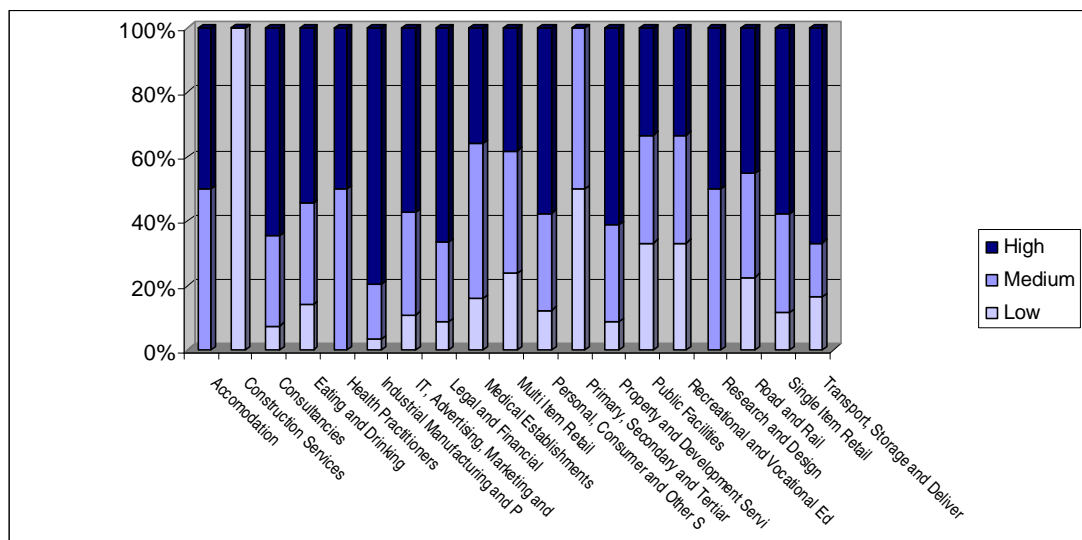
```

**Procedure that calculates and stores the topological and metric relationships to the graph structure.*

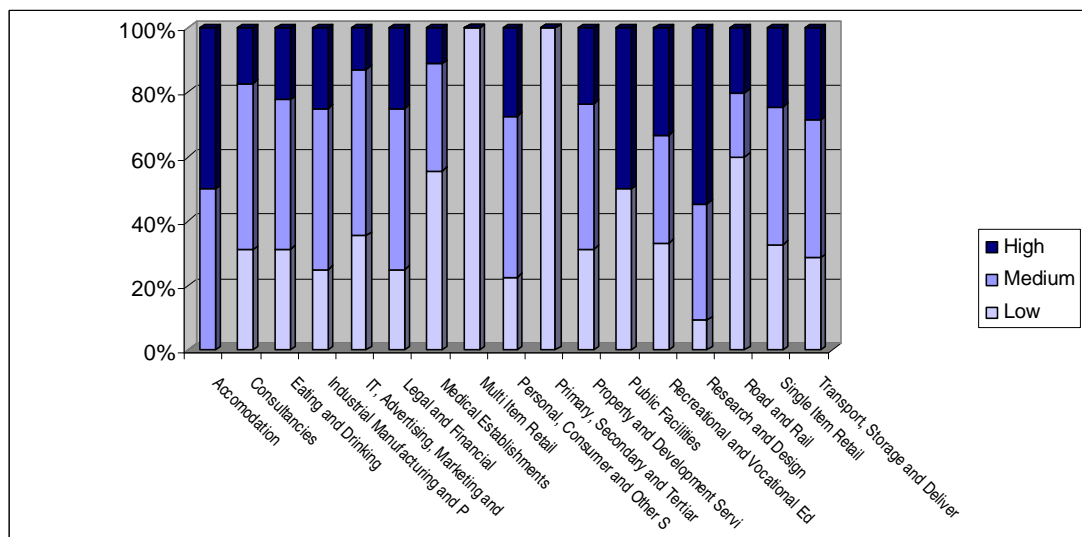
Appendix D - Results

Borough_PropertyLevel_P11_DescL2_3Class

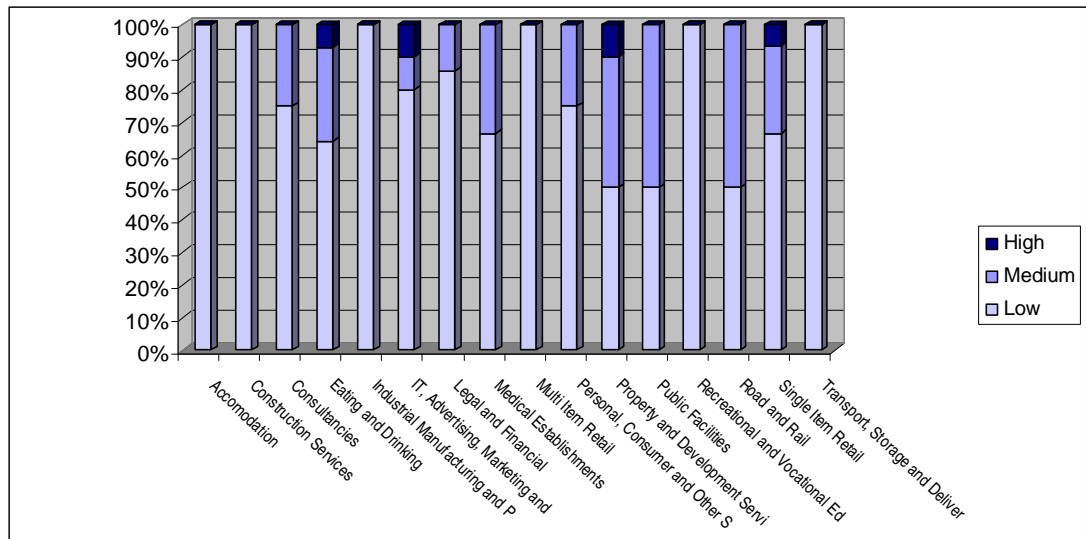
Westminster_Flat_P11_DescL2_All_3Class



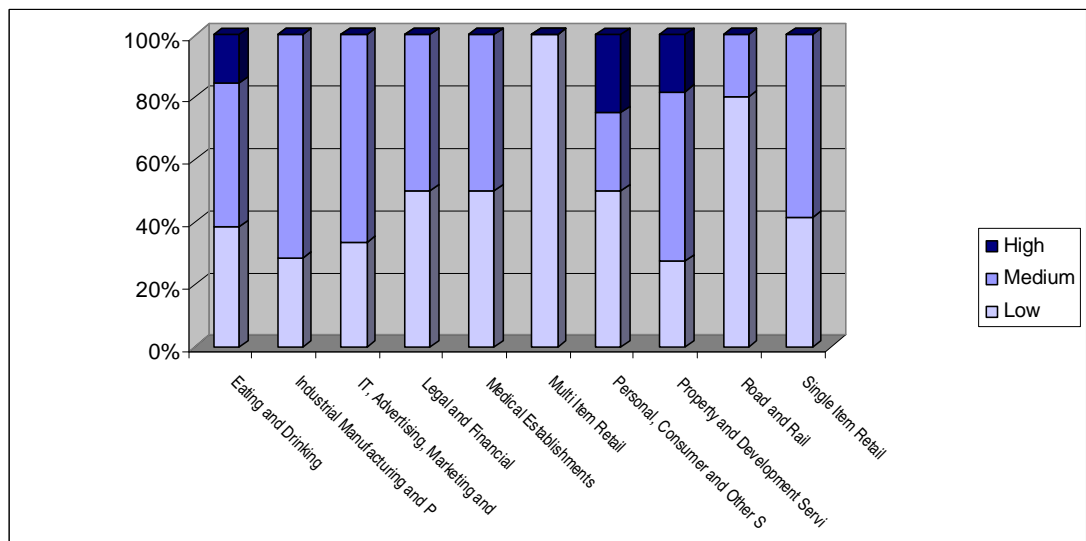
Westminster_Terrace_P11_DescL2_All_3Class



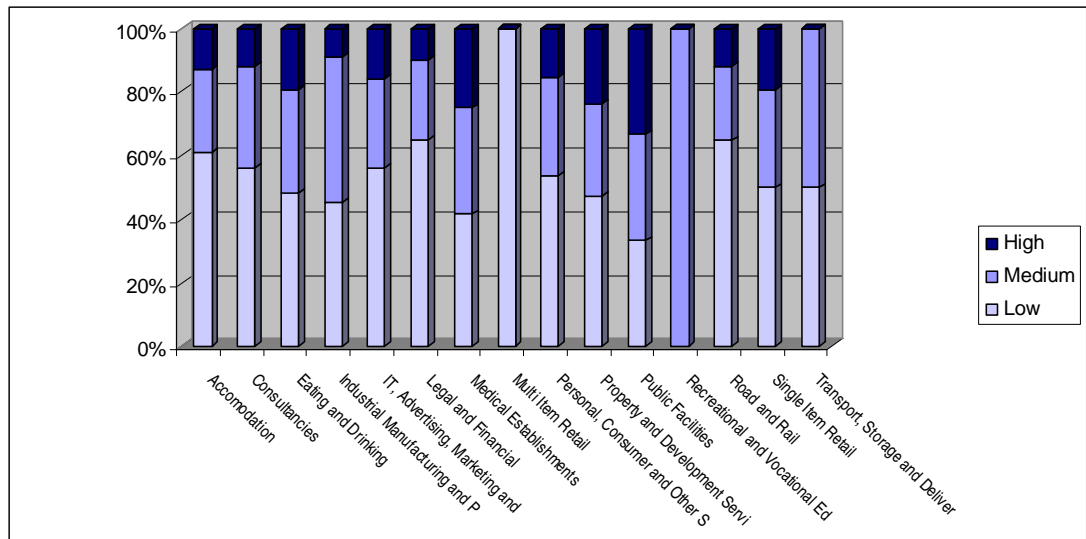
Westminster_SemiDetached_Pl1_DescL2_All_3Class



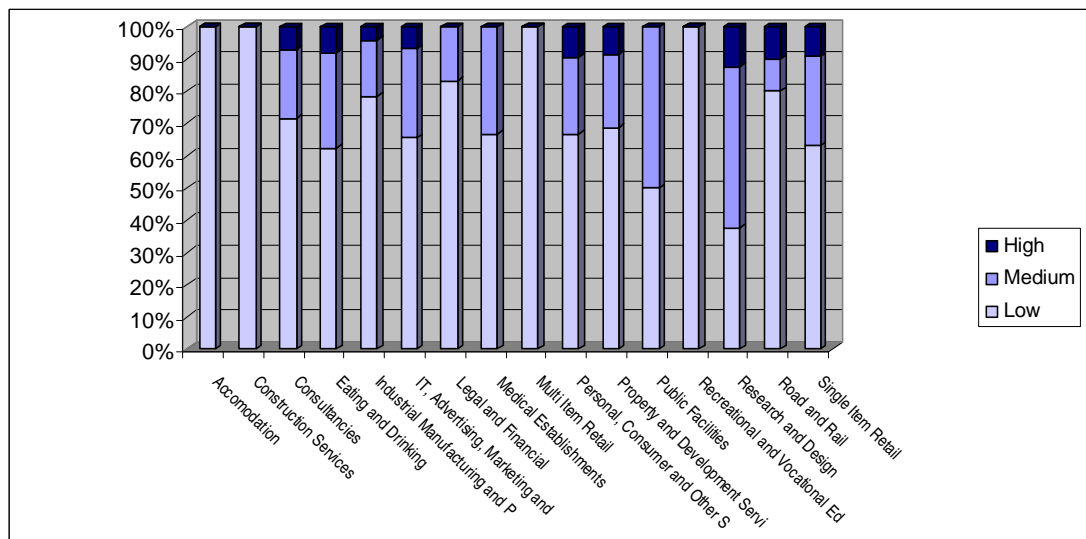
Westminster_Detached_Pl1_DescL2_All_3Class



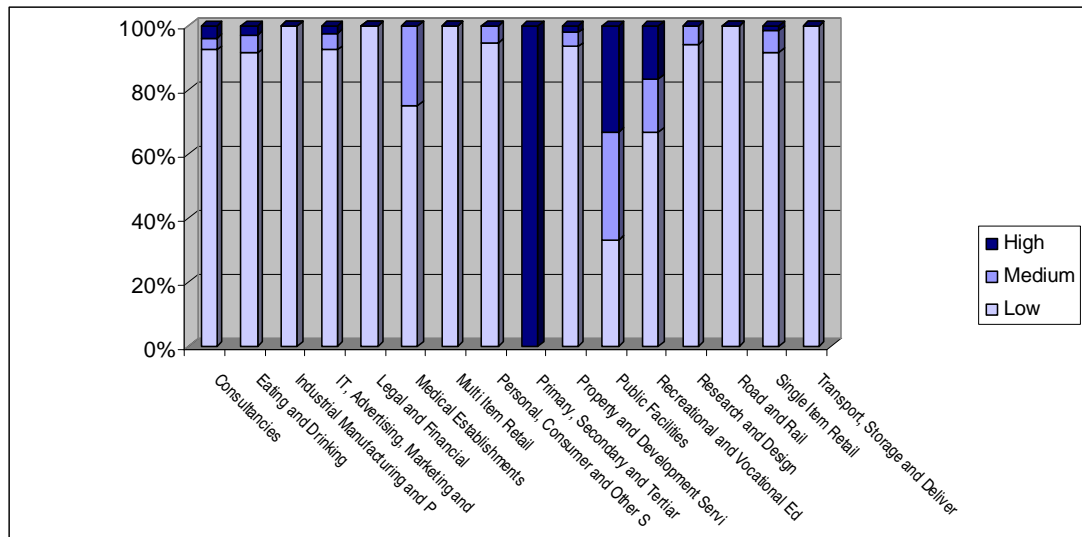
Kensington&Chelsea_Flat_P11_DescL2_All_3Class



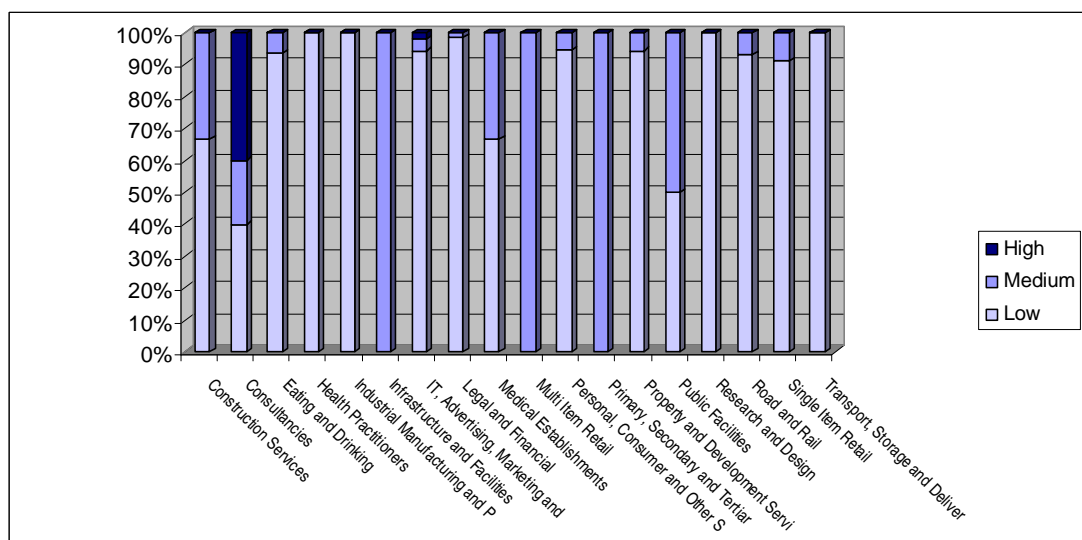
Kensington&Chelsea_Terrace_P11_DescL2_All_3Class



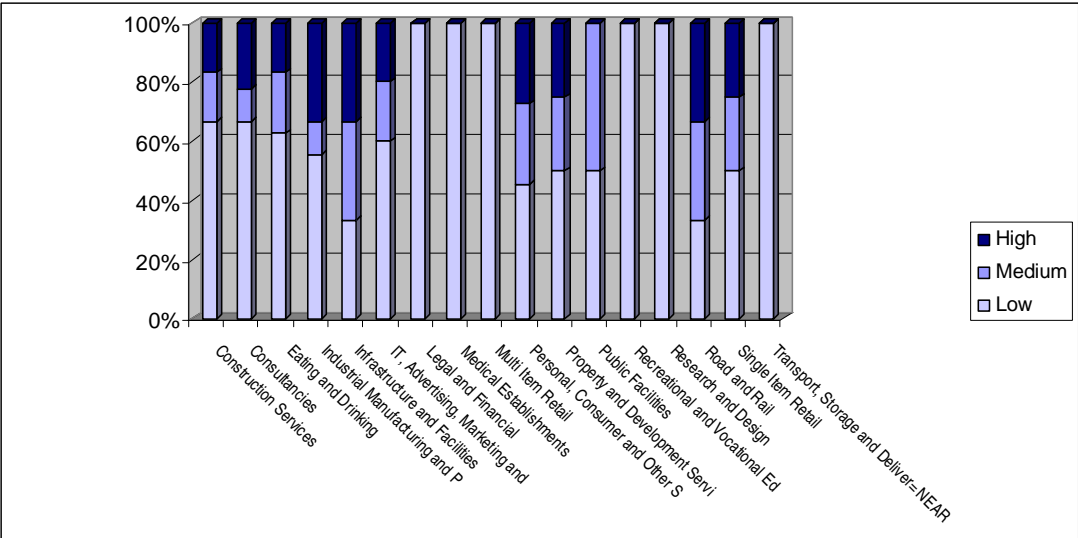
Kensington&Chelsea_SemiDetached_P11_DescL2_All_3Class



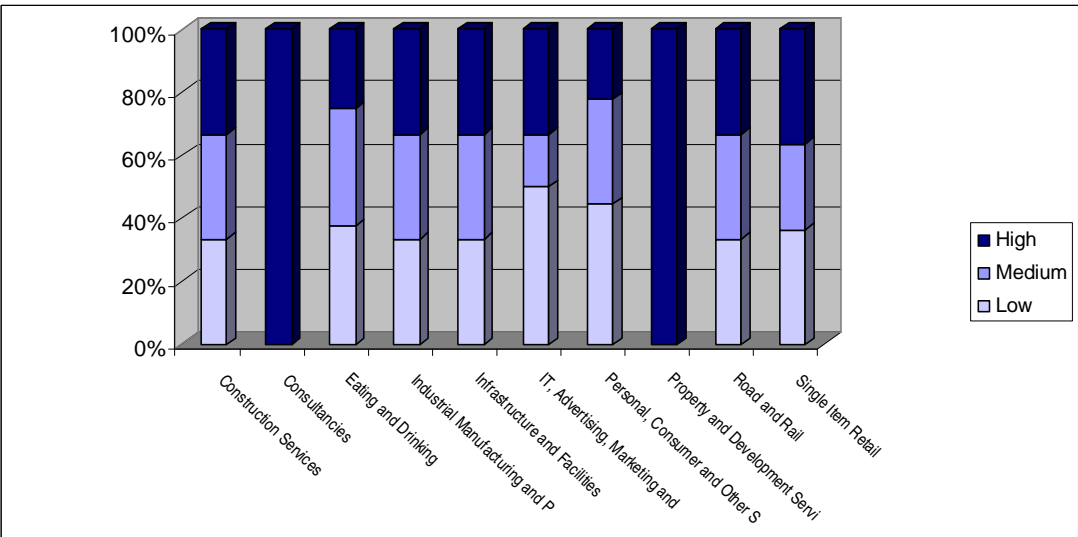
Kensington&Chelsea_Detached_P11_DescL2_All_3Class



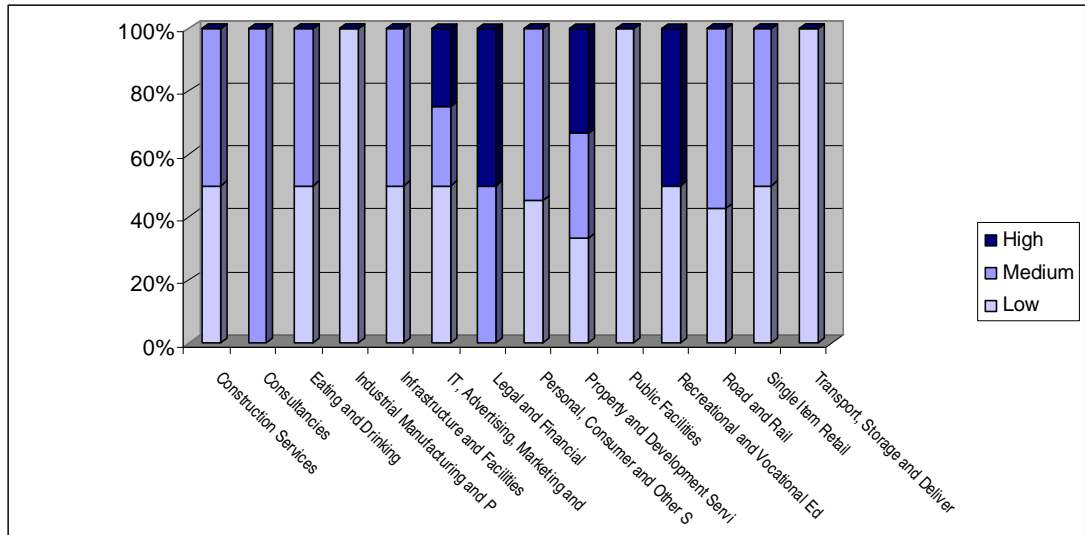
Hammersmith&Fulham_Flat_P11_DescL2_All_3Class



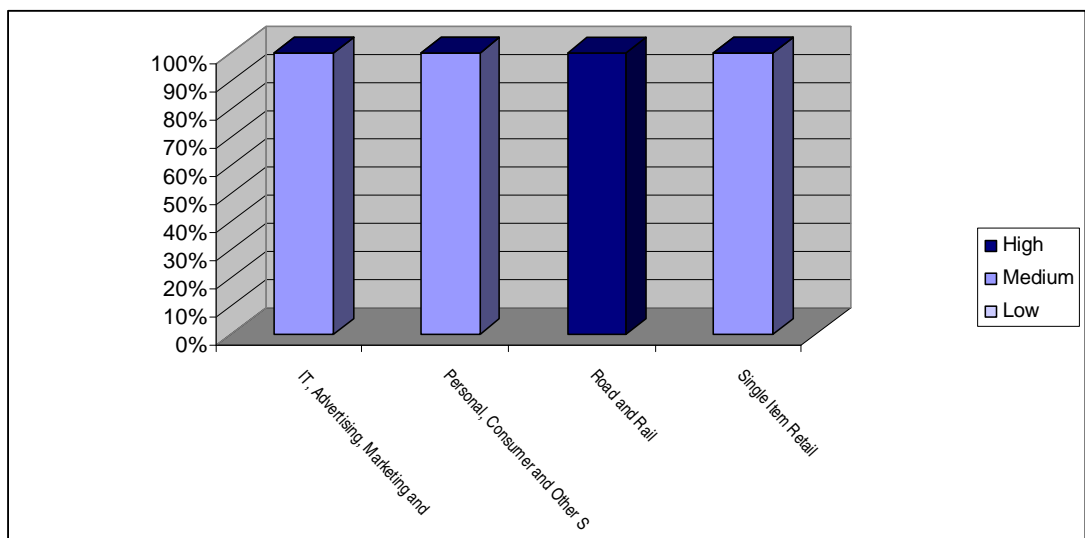
Hammersmith&Fulham_Terrace_P11_DescL2_All_3Class



Hammersmith&Fulham_SemiDetached_P11_DescL2_All_3Class

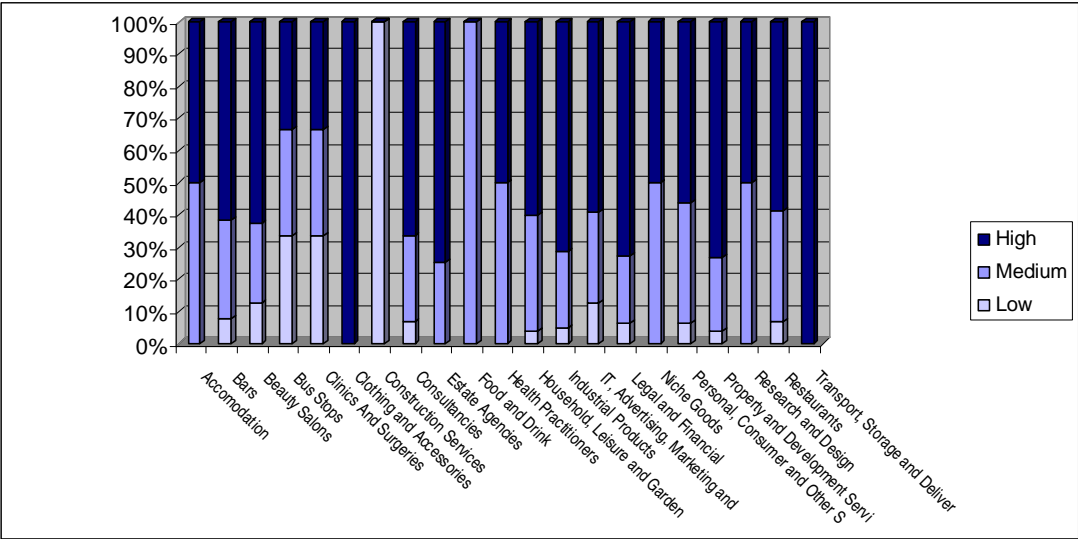


Hammersmith&Fulham_Detached_P11_DescL2_All_3Class

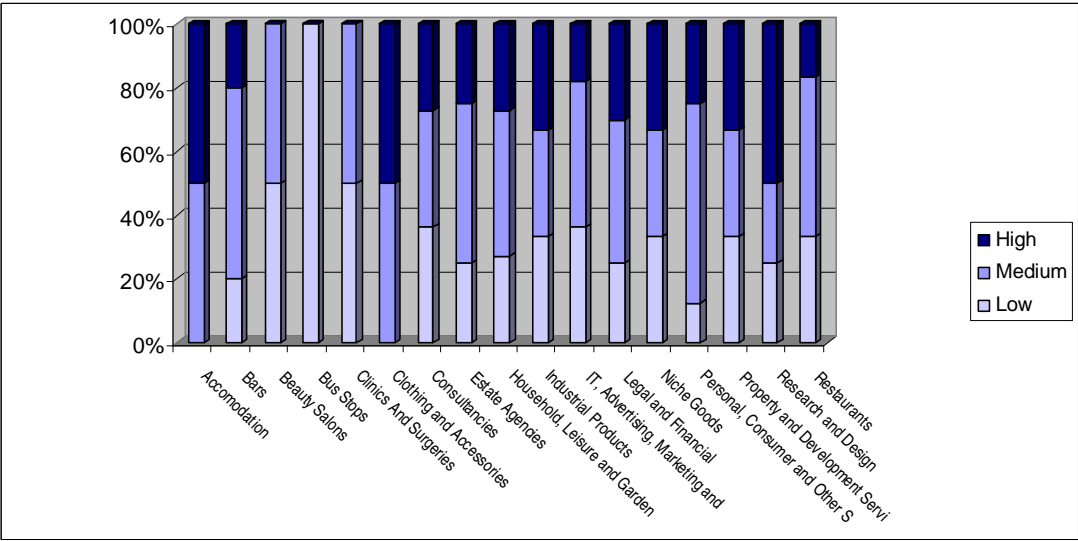


Borough_PropertyLevel_PL1_DescL3_3Class

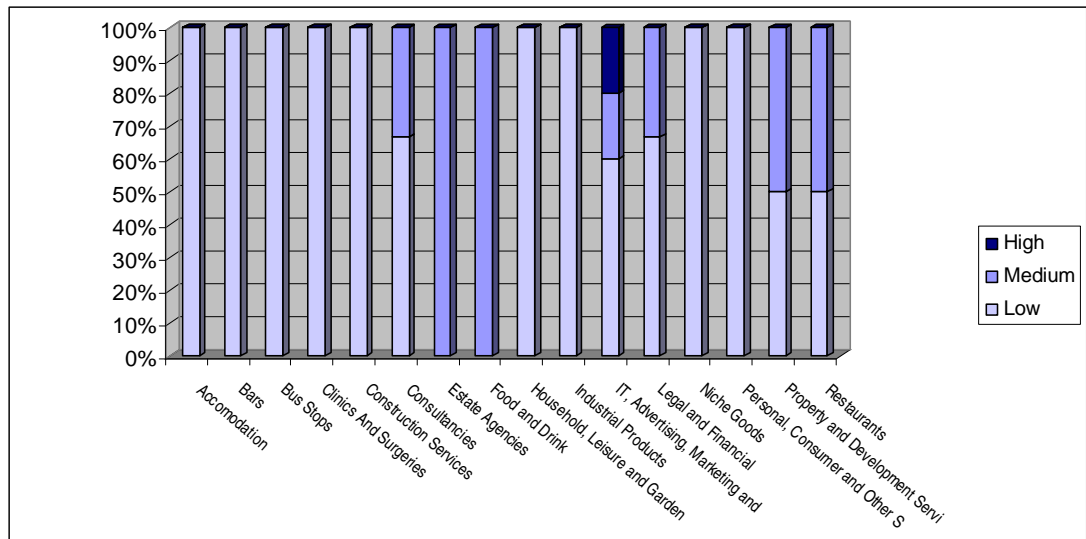
Westminster_Flat_PL1_DescL3_All_3Class



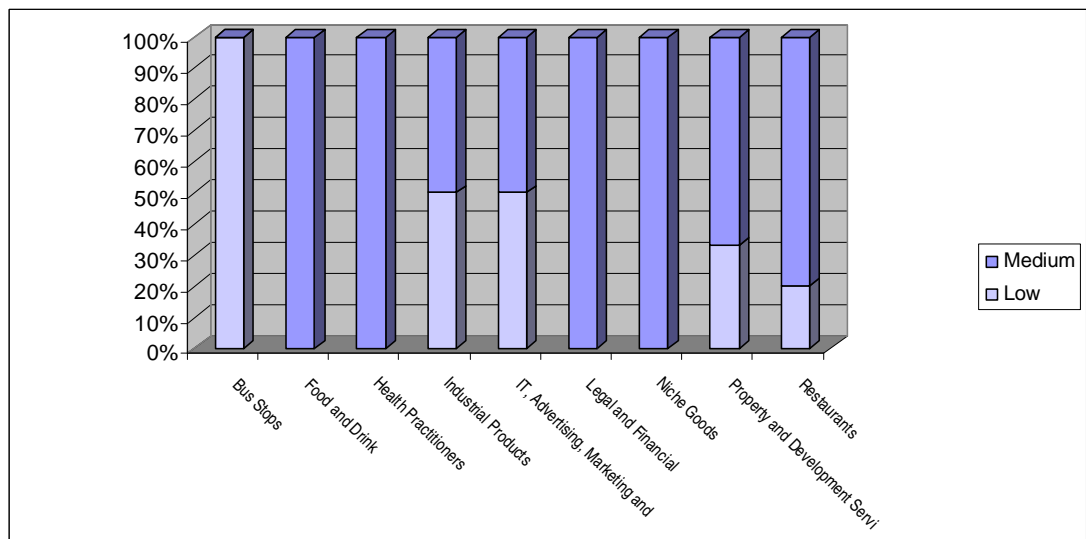
Westminster_Terrace_PL1_DescL3_All_3Class



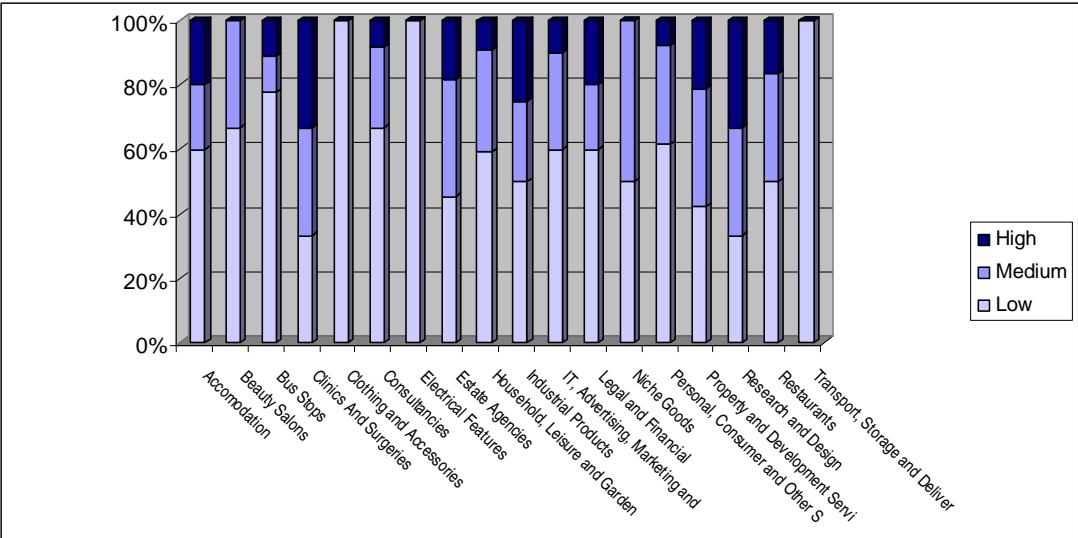
Westminster_SemiDetached_P11_DescL3_All_3Class



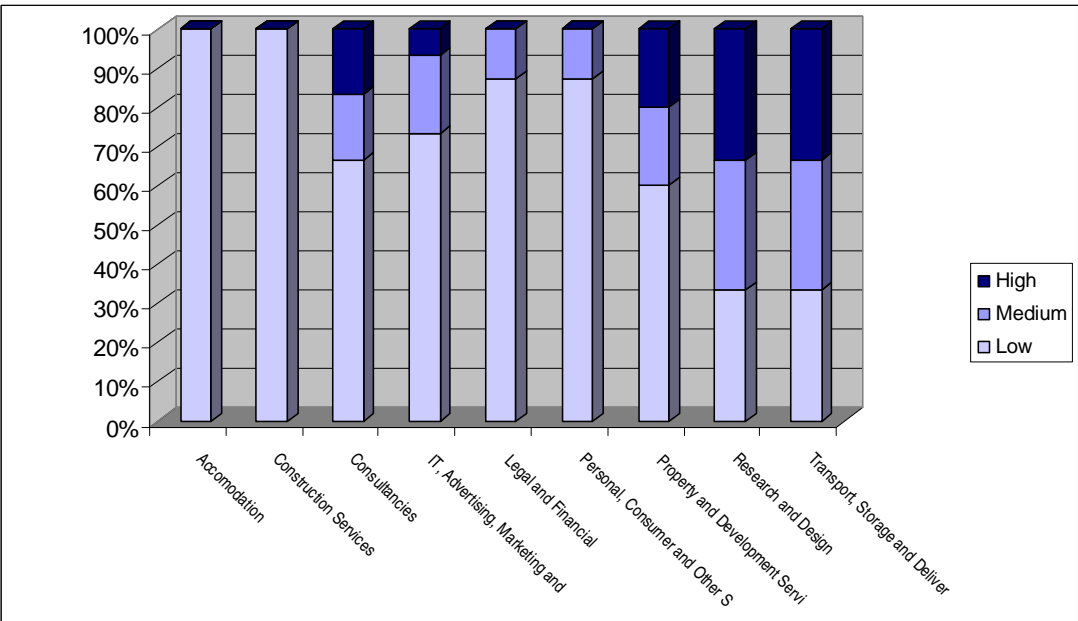
Westminster_Detached_P11_DescL3_All_3Class



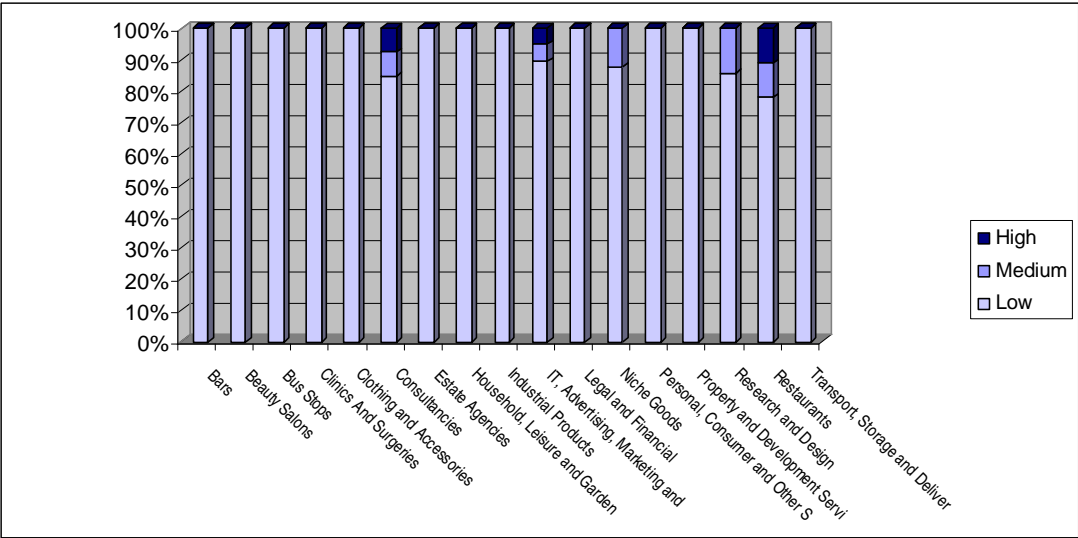
Kensington&Chelsea_Flat_P11_DescL3_All_3Class



Kensington&Chelsea_Terrace_P11_DescL3_All_3Class



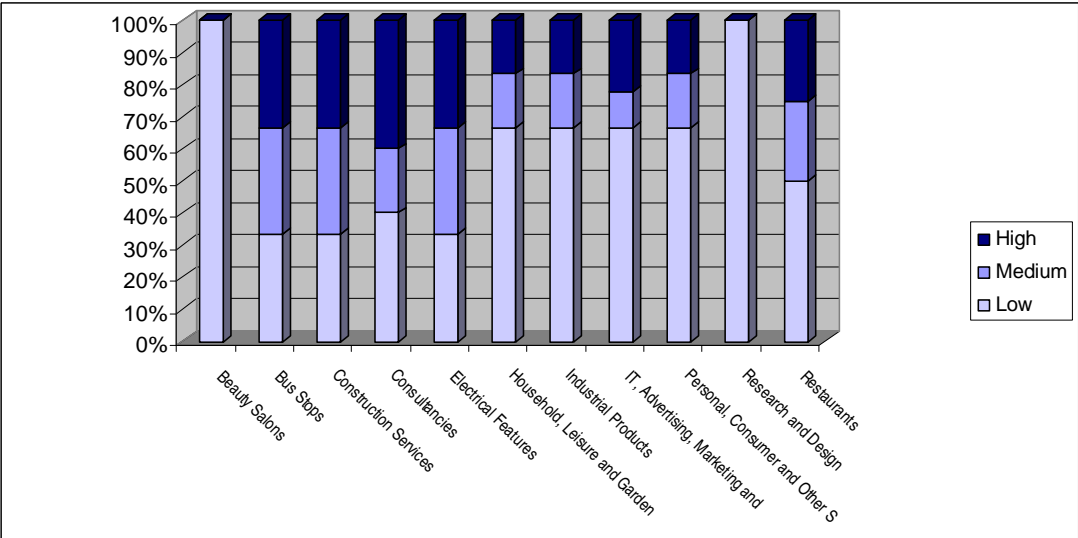
Kensington&Chelsea_SemiDetached_P11_DescL3_All_3Class



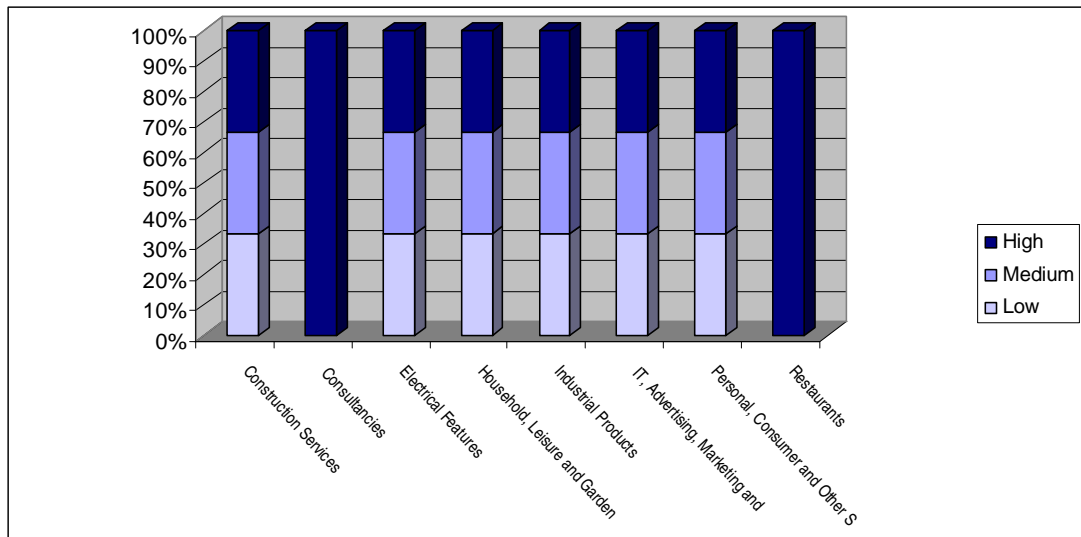
Kensington&Chelsea_SemiDetached_P11_DescL3_All_3Class

Failed to produce results

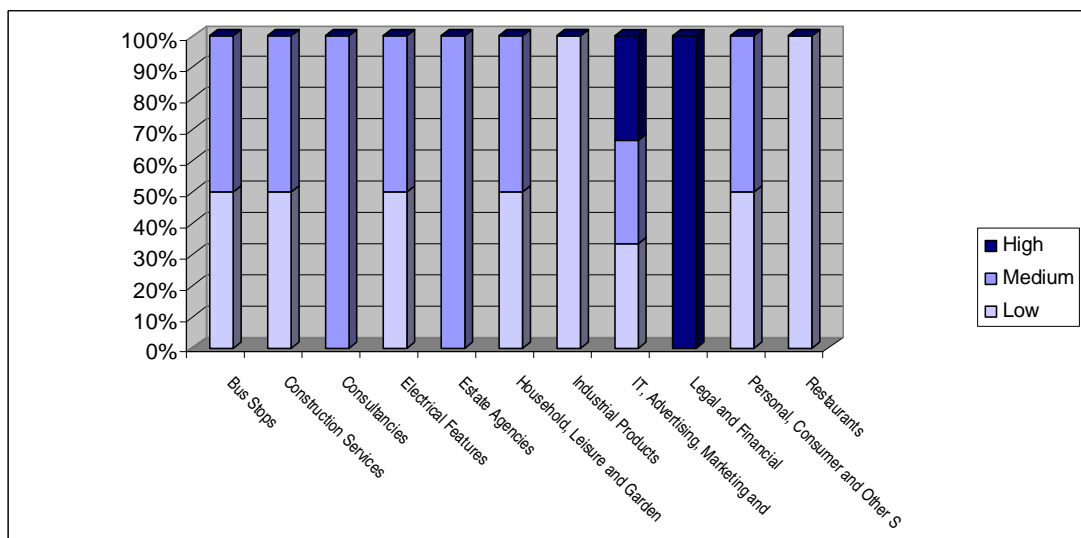
Hammersmith&Fulham_Flat_P11_DescL3_All_3Class



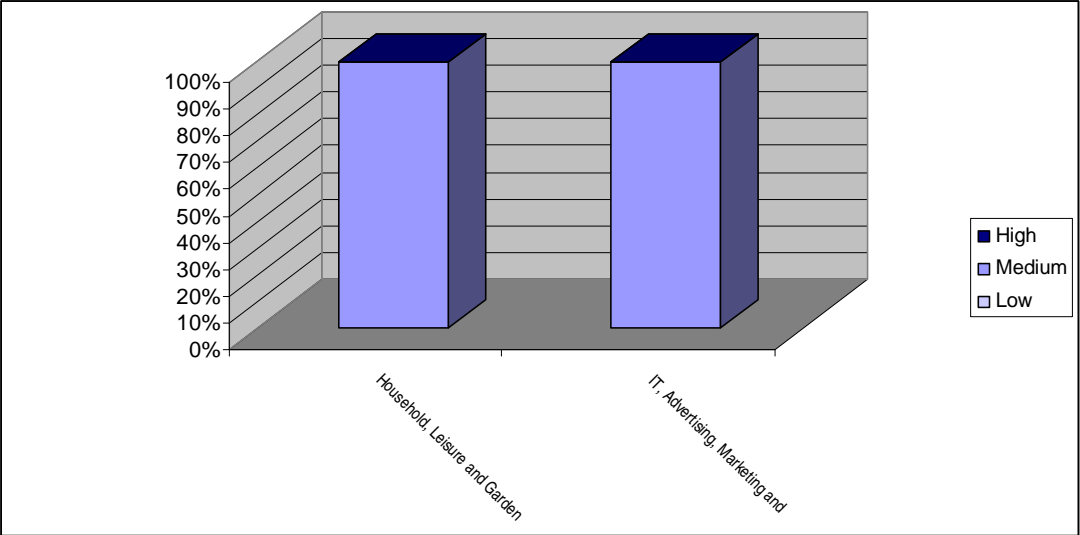
Hammersmith&Fulham_Terrace_P11_DescL3_All_3Class



Hammersmith&Fulham_SemiDetached_P11_DescL3_All_3Class



Hammersmith&Fulham_Detached_PL1_DescL3_All_3Class



Appendix E – CD Contents

On the enclosed CD-ROM, the original output files of the experiments presented in Chapter 6 are included. They are organised in three directories: Initial_Tests, Locational_Tests and Classification_Tests and they correspond to the results presented in pages 198, 211-224 and 225-226 respectively.