EXPLORING THE CHARACTERIZATION OF UNCERTAINTY IN CENSUS AND BOREHOLE DATA USING ROUGH SETS

by

Gift Dumedah

BSc. University of Science & Tech. Ghana, 2002

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

In the Department of Geography

SIMON FRASER UNIVERSITY

Summer 2005

APPROVAL

Name:	**Gift Dumedah**
Degree:	**Master of Science**
Title of Thesis:	Exploring the Characterization of Uncertainty in Census and Borehole Data Using Rough Sets
Examining Committee:	
Chair:	**Dr. N.K. Blomley** Professor

Dr. N. Schuurman, Assistant Professor
Senior Supervisor
Department of Geography, Simon Fraser University

Dr. N. Hedley, Assistant Professor
Committee Member
Department of Geography, Simon Fraser University

Dr. J. Hodgson, Professor Emeritus
External Examiner
Department of Earth & Atmospheric Sciences,
University of Alberta

Date Approved: July 27, 2005

SIMON FRASER UNIVERSITY

PARTIAL COPYRIGHT LICENCE

ABSTRACT

This research introduces rough sets to better characterizing spatial relationships and uncertainty in two examples. First, scale issues in census data are addressed. Census data provide demographic and socio-economic information at specific area units. Hence, derived spatial information are scale-dependent leading to uncertainty when analyzing results at different scales. Rough sets mitigate scale distortions and provide scale-sensitivity measure during scale transition. It employs the metaphor of topology to illustrate the ability of rough sets to retain spatial relationships of adjacency and contiguity.

Second, rough sets and transition probability are used to characterize sediment distribution. The study simulates sediment state and transitions for low and high quality borehole data by providing better geological understanding. It also assesses Geological Survey of Canada standardization scheme for classifying borehole data. The utility of rough sets is demonstrated as a knowledge base tool for characterizing uncertainty irrespective of the data under study.

To Mum, Dad, brothers and sisters.

ACKNOWLEDGEMENTS

I am highly indebted to my senior supervisor, Nadine Schuurman for her practical criticisms and guidance that have brought scattered ideas into this coherent piece of work. Thank you for your stimulating questions and remarks by putting the reading audience first. You have inspired me to communicate complex ideas concisely while keeping the contents intact. My heartfelt gratitude goes to my supervisor, Nick Hedley; you have shown me the means to construct this research into a structure that is readily accessible and relevant for real world applications. You have asked the tough questions that have brought the structure of this thesis into an integrated view of human and physical GIScience applications. Special thanks to my external supervisor, Dr. M. John Hodgson, for being on my committee, despite such a short notice. Your contributions are valuable into making this research a success.

I am grateful for secondary funding from SSHRC and CIHR. I acknowledge Terrain Sciences Division of the Geological Survey of Canada (GSC) for providing the borehole data for this research without which this project would have been a dream. Thank you, Charles Logan for answering all my questions. My profound appreciation go to Bailey, Marcia, Jasper and all faculty and staff in the department of Geography for your support. You have made my experience at SFU a memorable one. Thanks to my colleagues in Geospatial Lab; Rob and Nathaniel and to all graduate students in the department of Geography.

I am grateful to my cousin; Worlanyo, and Smith, who is a friend and a brother. You guys have been there any time I need your help. Thank you, Tom & Edna, for your help and providing home. Thank you, Sara, Veronica and all members of Burnaby home group. Thank you, Cynthia and Jennifer for your encouragements. Finally, I thank those whose efforts have made my education a success, but have not been mentioned. Yours is the most valuable.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

FORMULAE

The four difference measures formulas:

- Mean Bias Error, MBE – describes the bias. The variability of (P – O) about the MBE is the variance of the distribution of differences.

$$MBE = N^{-1} * \sum_{i=1}^{N} (P_i - O_i)$$

- Root Mean Square Error, RMSE – defines the linear fit between model and observation and is an index of systematic error.

$$RMSE = \sqrt{N^{-1} * \sum_{i=1}^{N} (P_i - O_i)^2}$$

- Mean Absolute Error, MAE – is a weighted average of the absolute errors.

$$MAE = N^{-1} * \sum_{i=1}^{N} |P_i - O_i|$$

- Index of Agreement, d – measures the relative size of average difference or the nature of the differences comprising MAE or RMSE.

$$d = 1 - \left[\sum_{i=1}^{N} (P_i - O_i)^2 / \sum_{=1}^{N} \left(|P_i| + |O_i| \right)^2 \right]$$

where P and O denote predicted and observed values respectively.

GLOSSARY

Keywords:

- Aggregation: the grouping together of a selected set of like entities to form a single entity.
- Aquiclude: a body of impermeable or distinctly less permeable rock.
- Aquifer: a permeable geologic formation saturated with water and through which groundwater moves.
- Aquitard: low-permeability unit that can store groundwater and also transit it slowly from one aquifer to another.
- Background material: the most abundant geologic unit present in a borehole
- Bayesian: mathematical theory of probability which applies to the degree of plausibility of statements, or to the degree of belief of rational agents in the truth of statements
- Conditional probability: the probability of an event assuming another event.
- Correlation coefficientt: is a numeric measure of the strength of linear relationship between two random variables.
- Dempster-Shafer theory: a mathematical theory of evidence representing plausibilities.
- Drumlin: an elongated whale-shaped hill formed by glacial action.
- Facies: all characteristics of a geologic unit or a distinct kind of rock for a specific environment.
- Fuzzy sets: a set characterized by a membership-degree function.
- Golden spike: boreholes with continuous core recovery which provide high quality data
- Joint probability: is the probability of two events in conjunction
- Kriging: an interpolation technique using a regionalized variable.
- Lag: minimum spacing between and within sediments in a borehole

- Markov chains: a discrete-time stochastic process using transition probability of state and state transitions.
- Maximum entropy factor: ratio of the transition rate to the maximum entropy transition rate.
- Metadata: data about data and usage aspects of it.
- Moraine: glacial drift or till deposited chiefly by direct glacial action.
- Permeability: a rock's or sediment's capacity for transmitting a fluid.
- Porosity: a measure of the percent of void space in a rock or soil.
- Regression: a statistical method where the mean of one or more random variables is predicted conditioned on other random variables.
- Rough sets: a set with nonempty boundary when approximated by another.
- Specific retention: ratio of the volume of water a rock or sediment will retain against the pull of gravity to the total volume of rock or sediment.
- Specific yield: ratio of the volume of water a rock or soil will yield by gravity by gravity drainage to the total volume of rock or soil.
- Stratigraphy: the study of sequence and correlation of stratified rocks
- T-PROG simulation: sediment state and transition simulation in Markov chains using transition probability.
- Topology: the relative location of geographic phenomena independent of their exact position
- Transition probability: a conditional probability representing a system's state and transitions.

Acronyms:

- CSD: Census Sub-division
- CT: Census Tract
- d: Index of agreement
- DA: Dissemination Area
- DI : Deprivation Index
- GAI: Good Aquifer Index (Indicator)

- GSC: Geological Survey of Canada
- HDI : High Deprivation Index
- IDI: Identification Index
- LA: Lower approximation set
- LDI : Low Deprivation Index
- MAUP: Modifiable Areal Unit Problem
- MAE Mean Absolute Error
- MAI: Medium Aquifer Index (Indicator)
- MBE Mean Bias Error
- MDI : Medium Deprivation Index
- mef: maximum entropy factor
- MOEE Ministry of Environment and Energy
- NAI: Non-Aquifer Index (Indicator)
- NDI: Net Deprivation Index
- ORM: Oak Ridges Moraine
- PAI: Poor Aquifer Index (Indicator)
- pwd: private well drillers
- RIC: Recent Immigrant Concentration
- RIDI: Recent Immigrant Deprivation Index
- RMSE: Root Mean Square Error
- tpm: transition probability matrix
- UP: Upper approximation set
- VGAI: Very Good Aquifer Index (Indicator)
- VHDI : Very High Deprivation Index
- VLDI : Very Low Deprivation Index

CHAPTER 1:
INTRODUCTION

The representation of geographic phenomena is dependent on data collection which are, in turn, based on data acquisition concepts, tools and methods. Data acquisition techniques are designed to meet specific research requirements and to fully represent the phenomena under study. However, there is often a difference between the geographic model and the geographic reality which it represents. This disparity between geographic model and phenomena in reality may be termed uncertainty. Uncertainty in spatial data gained significance with the search for elements that define spatial data quality (Buttenfield and Beard 1994). Uncertainty has a direct influence on data quality because it distorts standards for data quality.

Spatial Data Transfer Standard (SDTS) is a set of standards established by the International Cartographic Association (ICA). SDTS aimed to provide detail information about data for users to assess the fitness of data for a particular use (Morrison 1995). The elements in SDTS include lineage, positional accuracy, attribute accuracy, completeness, logical consistency, semantic accuracy and temporal information (Morrison 1995). While the elements of spatial data quality attempt to enhance data validity for accurate and complete geographic inquiries, Buttenfield and Beard (1994) argued that SDTS should incorporate uncertainty of real world conditions. The elements of data quality standard, in other words, improved validity of data (uncertainty about database descriptions). Data quality standards have also focused narrowly on error in data rather than the wider consideration of uncertainty (Allan 2003). Buttenfield and Beard (1994) contended that the validity of geographic reality (uncertainty of real world conditions) be integrated into data quality records. Hence, standards for data quality are starting point for assessing geographic reality, but they need to address geographic validity by assessing uncertainty of real world

conditions. So, this study attempts to incorporate uncertainty into spatial database and analysis of geographic information.

The study investigates the effects of uncertainty in two data: census data and borehole data. These are seemingly different data and they represent different problems. Census data are used to characterize spatial aspects of human socio-economic and demographic information. Borehole data, on the other hand, are used in spatial characterization of the physical subsurface. They are, however, equally plagued with the problem of data uncertainty. Also a common analysis tool – rough sets technique is applied to minimize the effects of data uncertainty. Rough sets concept is an analytical theory for discovering hidden patterns in data in order to better describe phenomena about which data are collected.

In this study, uncertainty is described separately for each data while the methodology for specific uncertainties in each data is linked by the use of rough sets. For census data, uncertainty represents scale distortions from opposing analysis assumptions and poor spatial data integrity across scale changes. This uncertainty is often referred to as scale problem or scale issue. The scale problem can be restated as *the process where statistical results and relationships for aggregated data are different for the same set of data at different scales*. Borehole data uncertainty, on the other hand, includes geologic unit (or sediment) identification and description problems. These uncertainties are not the only uncertainties in the two data, but these are the uncertainties examined in this study. This research uses two data separately into two case studies in order to explore the nature of uncertainty using rough sets.

In the first case study, rough sets are applied to census data at different census scales for neighbourhood characterization using deprivation indices and recent immigrant population. Rough sets mitigate the scale issue by providing a scale sensitivity measure in order to map deprivation levels across multiple census scales. During transitions across multiple scales, data distribution retention is crucial in order to estimate scale translation indices. The rough sets technique employs the metaphor

of topology to illustrate its ability for retaining spatial relationships of adjacency and contiguity across multiple scales in attribute space.

For census data, spatial relationship outputs have shown that rough sets better represents spatial relationships than other spatial analysis techniques such as spatial regression. Specifically, rough sets may enhance spatial characterization and can replace other spatial analysis tools when characterizing geographic phenomena. The first case study illustrates this enhanced spatial characterization using rough sets by maintaining data distribution and providing scale sensitivity index for translating neighbourhood phenomena across different census scales. The rough sets outputs emphasize unique uncertainty levels at each census scale and provide thresholds within which results apply. The research also revealed limitations inherent in traditional spatial analysis tools, which make them poorly characterize spatial relations.

In the second case study, rough sets and transition probability are used to provide a means of characterizing the subsurface environment. These methods are used to enhance the use of borehole data of marginal quality (e.g. Ontario Ministry of Environment and Energy (MOEE) data) for accurate hydrogeological inquiry. The rough sets method also assesses the GSC (Geological Survey of Canada) standardization scheme and incorporates MOEE data variation when characterizing the subsurface.

Analysis tools for reducing the effects of erroneous sediment identification and description problem are mainly data classification. While this study recognizes the relevance of data standardization, it focuses on the assessment of data classification using rough sets. Also transition probability is used to characterize sediment variation in the subsurface. This translates accurate sediment distribution from quality borehole data to less accurate borehole data (i.e. MOEE data) in order to enhance poor quality data for reliable hydrogeological inquiry. The use of rough sets and transition probability compensate each other by assessing different effects of uncertainty due to erroneous sediment identification. This is illustrated in the borehole outputs especially when identifying high and less accurate sediments within boreholes.

In investigating the effects of uncertainty in these case studies, this research acknowledges traditional methods applied to the respective data for examining data uncertainties. It argues, however, that rough sets and transition probability are suitable analytical tools for exploring uncertainties of scale distortions in census data and sediment variation problems in borehole data. Hence, the study also underscores the utility and versatility of rough sets: its ability to enhance geographic inquiry irrespective of the data; and its flexibility to adapt to different effects of uncertainty in human and physical processes. Spatial characterization using census data represents a human phenomena characterization, while subsurface characterization signifies a physical process. This emphasizes that applications of rough set are broad because it adapts easily to different analysis problems. For example, scale problem in census data and sediment description problems in borehole data. This underscores the use of different data in this study: to illustrate the use of one analysis tool (i.e. rough sets) in both human (i.e. census socio-economic and demographic information) and physical (i.e. subsurface environment) phenomena analysis. Table 1.1 below shows the two different (but unified by data uncertainty and rough sets) case studies that constitute this research.

Table 1.1: Research description into two major case studies

Case study	Data	Effects of Uncertainty	Alternative & Traditional Techniques	Proposed Technique
One (1)	Census data	Scale problem: lack of spatial integrity across scales	Spatial regression, local and global statistical measures, etc	Rough sets
Two (2)	Borehole data	Sediment identification and description problems	Data standardization	Rough sets & transition probability

1.1 RESEARCH OBJECTIVES

The first case study is aimed at exploring the scale problem in census data: that is, the problem of lack of spatial integrity during transition across scales. The scale problem is the variation in analytical results when data for a particular set of spatial units are aggregated into smaller or larger spatial units for analysis (Openshaw 1984a).

The scale problem is closely linked to spatial data transition between two spatial units from small to high resolution or vice versa (Openshaw 1984a). This problem is examined with many spatial analysis tools but its effects have resisted analytical assumptions such as randomness and independency of data variables. This first empirical study uses rough sets to minimize the scale problem by examining the relationship between recent immigrants and deprivation indices. So, the research question is: during census characterization of spatial relationships how do we account for scale transitions at different census scales.

Spatial analysis techniques on spatially grouped data such as spatial regression and correlation examine global trends. They are inadequate in the prediction of future occurrence of local events and understanding spatial social structure at local levels (Lark 2000). They also have proven to poorly describe spatial relationships (Shaw and Wheeler 1994; Lark 2000). Methods are needed to describe spatial relationships at both global and local scales. The study applies a technique that utilizes high-quality large resolution census data for geospatial analysis while providing results that are relevant and appropriate to local spatial phenomena investigation. This is examined through rough sets ability to retain data variation and integrity across scales.

In sum, application of rough sets to spatially grouped data for multiple spatial units may allow us to quantify uncertainty associated with spatial transitions from small spatial unit to large units and vice versa. The metaphor of topology is used as a parallel to illustrate the ability of rough sets to retain data distribution and variation across multiple scales in attribute space. During spatial analysis, spatial integrity of adjacency and contiguity is maintained through topology. Topology in attribute space, however, becomes relevant when translating spatial phenomena across different scales. Rough sets analysis is used to retain data distribution and variation in attribute space, which is synonymous to spatial adjacency and contiguity in spatial space. So while spatial topology ensures spatial integrity during spatial analysis, rough sets technique enhances attribute integrity during scale transitions.

The second case study uses rough sets and transition probability on borehole data of marginal quality. A unique ability of rough sets is the discovery of hidden patterns in data and characterization of inter-data connections or relationships (Pawlak 1982). These patterns allow better understanding and description of phenomena about which data are collected. So, rough sets are used to better describe the subsurface environment using borehole data of marginal quality. Transition probability, on the other hand, is used in Markov chains to analyze complex systems using the concept of state and state transitions (Howard 1971). In this study, transition probability is used to simulate sediment variation for the borehole data. The research problem for this case study is to reduce the effects of sediment identification and description problems in borehole data of marginal quality (i.e. MOEE data). The properties of rough sets and transition probability are suitable uncertainty tools for subsurface data analysis in order to define borehole units using aquifer-supporting properties.

Borehole data of high accuracy (i.e. continuous core recovery borehole) acquired by the Geological Survey of Canada, GSC (referred to as golden spikes), and MOEE (Ontario Ministry of Environment and Energy) data mainly acquired by private well drillers are used for this case study. Rough sets and transition probability utilize golden spikes to enhance the MOEE data for reliable hydrogeological application. They enrich borehole data of marginal quality for geospatial analysis while at the same time accounting for the effects of uncertainty in subsurface environment.

Overall, rough sets and transition probability applications to uncertainty may allow borehole data of marginal quality to be incorporated into geographic inquiry – thereby improving characterization accuracy. Subsurface models developed using low quality data are validated with those built from high quality data (golden spikes from the GSC). Hence, the technique quantifies uncertainty inherent in the low quality data in order to enhance geological inquiry.

1.2 RESEARCH JUSTIFICATION

This research uses two case studies to address persistent geographical problems: scale problem in census data, and erroneous sediment identification and description problems in borehole data. First, the study characterized spatial relationship between recent immigrants and deprivation index in order to investigate the scale problem. Current literatures (Kassim and Laurel 2000; Kazemipur and Halli 1998) have identified strong correlations between poverty of census unit and the proportion of its population who are immigrants. The first case study examines the extent of this relationship by using rough sets.

In the first case study, census data provide valuable demographic and socio-economic information on people at a particular place and time. The application of these data in the management of resources and people has profound consequences. Resource management and distribution, for example, affect goods and services reaching a particular place and time. Census data form an integral component of market research databases for category management and planning, geo-demographic market segmentation and retail site selection and evaluation. The quality of these data has significant implications on subsequent applications and research. Hence, the study examines processes and implications of using census data for real world applications such as resource management and distribution, neighbourhood and retail site characterization.

Census data are often aggregated from small to coarse resolutions. Census data aggregation has important implications for protecting individual confidentiality and privacy. The inherent characteristics of census aggregate data are that they rarely retain original data distribution, leading to the creation of a different data with new data characteristics at each level of aggregation. Finally, these data are employed in model development and decisions are made for smaller spatial units, for example, neighbourhood characterization and retail site selection. In this study, rough sets aim to retain original data characteristics despite scale transitions.

In the second case study, borehole data remain an important data source for subsurface study. Borehole data are employed in subsurface mapping of geological settings and require accurate identification and description of the subsurface geology. They are also used in subsurface modelling to provide a means of understanding subsurface geologies which control the distribution and movement of fluids (e.g. groundwater). The efficiency of geological applications resulting from the use of borehole data is correspondingly dependent on the quality of borehole data. Consequently, the quality of these data is integral to the accuracy of any geological inquiry.

Additionally, borehole data may be employed in locating groundwater sources for economic and industrial activities. The environmental and economic benefits of aquifer locations cannot be overstated, as they are protected zones for water supply for domestic, industrial and agricultural purposes (Logan et al. 2001; Russell et al. 1998; Schuurman 2004). Groundwater represents a major proportion (about 0.6 per cent) of the earth's usable water resource and in some locations it is the only source of water supply (Price 1985). Groundwater can be developed when and where it is necessary or needed, thus they provide reliable water source with relatively good accessibility. However, aquifers – the earth's subsurface water repositories are not uniformly distributed throughout the earth's crust (Price 1985). They require geological settings that support adequate water movement and distribution, hence the need to protect and manage them effectively.

The importance and applications of well-log data are crucial. Well-log data allow subsurface geological investigations such as subsurface mapping for the representation of ground stability, aquifer (groundwater repositories) locations and mineral deposit sites and their assessment. Most geological applications are currently executed with little consideration to the quality of the initial data. Also there exist inaccuracies and assumptions built into analytical tools that are employed in model design. This causes the propagation of complex uncertainty that can result in

significant model departure from reality. Hence, this project seeks to design geographic models that retain input data variability using rough sets.

Overall, the use of different data in this study attempts to illustrate the utility of rough set for characterizing diverse uncertainties irrespective of the data under study. The study underscores the utility and versatility of rough sets: its ability to enhance geographic inquiry irrespective of the data; and its flexibility to adapt to different effects of uncertainty in human and physical processes. Hence, the applications of rough set are broad because it adapts easily to different analysis problems.

1.3 THESIS STRUCTURE

The preceding sections introduced the research focus and discussed the study objectives. This section outlines different components of the study which are organized into chapters. The study is organized into five (5) chapters including this introductory chapter.

Chapter two (2) introduces concepts and definitions of uncertainty and describes specific aspects of uncertainty in the two data: census data and borehole data. Chapter two also provides guidelines for identifying analytical tools in order to adequately accommodate specific effects of uncertainty in the data. This chapter describes the scale problem in census data and the problem of sediment identification and description in borehole data. Limitations which plague spatial analysis tools and impede methods for resolving the effect of scale distortions during spatial characterization are also emphasized. Finally, analysis tools for assessing uncertainty are described. Underlying assumptions and conditions which characterize these tools are outlined in order to assess their suitability for addressing specific uncertainties in the data.

Chapter three (3) constitutes the first case study. This chapter describes rough sets method, census data outputs and discussions. Rough sets method is developed in

order to mitigate the scale issue during scale transition. This chapter uses rough sets to explore neighbourhood characterization by assessing deprivation levels across multiple census scales. It also examines the extent of recent immigrant and deprivation level relationships. Scale transition estimates are approximated from large census units to small ones and vice-versa. The rough sets approach provides a scale sensitivity measure to enhance census estimates made from one census unit (say, DA[1]) using another (say, CT[2]).

Chapter four (4) constitutes the second case study. Chapter four describes rough sets and transition probability methods, borehole data outputs and discussions. It describes methods for limiting the effects sediment identification and description problem in borehole data. The section outlines data preparation and structuring and specific uncertainties examined separately by rough sets and transition probability. These two analysis tools enhance the use of low quality borehole data for characterizing the subsurface environment. Borehole data outputs are grouped into three categories. First, the GSC standardization scheme is assessed in order to estimate the extent of data variation in the MOEE data. Second, transition probability is used to simulate sediment transition sequence in the vertical direction. Finally, limitations of transition probability simulation are outlined and a simple illustration is used to estimate depth and spatial information for specific sediment states and transitions.

To conclude, chapter five (5) brings together research findings and contributions for both case studies. Chapter five integrates both case studies in order to demonstrate the utility of rough sets as a knowledge base tool for characterizing uncertainty irrespective of the data or area of application under study. This chapter also outlines limitations and recommendations for further research separately for census data and borehole data.

---

[1] DA denotes dissemination area

[2] CT denotes census tract

1.4 DATA SOURCES & DATA DESCRIPTIONS

This section outlines the two data: census data and borehole data. It also illustrates data sources and their descriptions. Data requirements for this study are categorized in two major groups: census data and borehole data, according to the two defining research focus. Census data acquired for the study is 2001 Canada census data for GVRD, a census metropolitan area (CMA). Three census units which are small constituents of the chosen CMA, (that is GVRD) are: census sub-division (CSD), census tract (CT) and dissemination area (DA). The 2001 census data are provided by the Statistics Canada through Simon Fraser University (SFU).

The borehole data were provided by Terrain Sciences Division of the Geological Survey of Canada (GSC). There are two categories of the borehole data: golden spikes and MOEE (Ministry of Environment and Energy) data. Golden spikes refer to boreholes with continuous core recovery and provide the highest quality data (Russell et al. 1996). Golden spikes were drilled by the GSC, OGS (Ontario Geological Survey) and IWA (International Water Association) (Russell et al. 1996). There are 32 boreholes that constitute the golden spike data scattered over the entire study area – southern Ontario. The MOEE data have limited application because they lack sediment sampling (Russell et al. 1996) and are provided by private well drillers (pwd). Sediment descriptions supplied by pwd are questionable because they often lack technical Geoscience training. Hence, resulting sediment identification and descriptions are often assigned multiple tags which limit the assessment of accuracy levels for the data. 'The sediment descriptions rely on washings brought to the surface during drilling and do not describe solid sediment core' (Russell et al. 1996, p196).

Boreholes from the MOEE data have low reliability (Russell et al. 1996) but constitute the most single abundant data available. There are about 62,325 boreholes available in the MOEE data. Hence, it provides a unique opportunity for characterizing the subsurface from high quality data (e.g. golden spikes) (Russell et al. 1996).

CHAPTER 2:
UNCERTAINTY & MODELS IN GEOGRAPHIC INFORMATION

The preceding chapter introduced the study and outlined key uncertainties for the both case studies. This chapter describes conceptual framework of uncertainty and common computational techniques for modelling uncertainty. The chapter also describes detail uncertainties in the two data: census data and borehole data.

2.1 SOME DEFINITIONS OF UNCERTAINTY

Uncertainty is a persistent and a common problem in most information systems. *Uncertainty* has many definitions. In the information sciences, for example, uncertainty 'relates to the truth or the conformity to reality of an information item' (Dubois and Prade 1988, 2). Uncertainty is assessed in relation to the degree of confidence in an information item (Dubois and Prade 1988). Confidence as a component of an information item is an index of reliability of an entity (Dubois and Prade 1988) which can be used to evaluate the uncertainty in information item.

In GIS, Dutton (1989, 126) defined spatial uncertainty as an inaccuracy that 'occurs when no model of ground truth exist or can be agreed upon in relation to a particular set of measurements'. Zhang and Goodchild (2002, 6) described uncertainty in relation to spatial databases as a 'measure of the difference between the actual contents of a database and the contents that a current user would have created by direct and perfect accurate observation of reality'. Dungan (2002, 26) described uncertainty as 'quantitative statement about the probability of error'. Allan (2003, 190) described uncertainty as a 'global term to encompass any facet of data, its collection, its storage, its manipulation or its presentation as information which may raise concern, doubt or scepticism in the mind of the user as to the nature or validity of the results intended message'. Pang (2001, 2) defined uncertainty as 'a multi-faceted characterization about data, whether from measurements and observations of some

phenomenon and predictions made from them'. It may include 'error, accuracy, precision, validity, quality, variability, noise, completeness, confidence and reliability' (Pang 2001, 2).

These descriptions of uncertainty show that certain aspects of uncertainty are exact and attainable (e.g. distance measurement errors), while others (e.g. indeterminate river banks) are not. The conformity or the simulation of reality in these descriptions of uncertainty with relation to geographic data is the basis for assessing uncertainty levels in this study.

In geography, uncertainty may be inherent in describing geographic phenomena, acquisition of geospatial data and manipulation processes. Continuous features such as mountains and rivers essentially exhibit vague boundary characteristics (Burrough and McDonnell 1998) such that their selection must accommodate some trade-offs in concepts used to describe them. Subsequent data collection techniques are dependent on how these features are conceptualized coupled with inaccuracies in the data acquisition tools and methods. Uncertainty can be deliberately introduced in geographic data (Worboys 1998) through information handling or mathematical operations applied on data. Data aggregation, in census data for example, conceals the original data variability and distribution. Analytical tools such as regression and correlation approaches to spatial data analysis may inadequately retain the input data variability during spatial analysis.

Uncertainty exists in the whole process of geographic data representation through data abstraction, collection, analysis and the use of data (Zhang and Goodchild 2002) partly due to the complex nature of geographic reality. Geographic reality however, must be simplified and represented in order to facilitate analysis and decision-making (Zhang and Goodchild 2002). These selection, generalization, symbolization or filtering processes are dependent on geospatial variations and heterogeneity in the real world (Lo and Yeung 2002). The use of these data acquisition tools for information gathering do not describe even the physical characteristics of the geographic environment because geographic reality cannot be reduced to models

without error (Duckham and Sharp 2004; Zhang and Goodchild 2002). The understanding and detection of a variety of uncertainties in geospatial data processes is fundamental to the modelling of uncertainty in geographic data. In the following section generalized factors which introduce uncertainty in geographic space characterization, are outlined.

2.2 UNCERTAINTY AND DATA QUALITY

The preceding section described uncertainty and its persistent occurrence in geographic phenomena. This section discusses specific elements of data quality through which uncertainty may emerge. The presence of uncertainty results in the deterioration data quality. Uncertainty has direct influence on data quality because it distorts standards for data quality. The components of Spatial Data Transfer Standard (SDTS) include lineage, positional accuracy, attribute accuracy, completeness, logical consistency, semantic accuracy and temporal information (Morrison 1995). Lineage describes original measurements, data acquisition and compilation methods, conversions, transformations, analyses and derivations that the data have been subjected to and the assumptions applied at any stage during data processing (Clarke and Clark 1995). Hence, lineage records the parentage of data by recording data changes in its nature, form and format. Positional accuracy refers to the nearness of the position of real world entity to the entity's true position in an appropriate coordinate system (Drummond 1995).

Attribute accuracy refers to a fact about some location, set of locations or features on the surface of the earth (Goodchild 1995). Completeness shows whether each entity instance is present and whether all of its attributes are present, where the totality of entity instances is defined by the entities within an abstract universe (Brassel et al. 1995). Logical consistency refers to logical rules of structure and attribute rules for spatial data and describes the compatibility of a reference with other data in a dataset (Kainz 1995). Semantic accuracy is the quality of geographic object description in accordance with a selected model (Salge 1995). The quality of temporal information

describes the level of information adequacy (in terms of temporal, precision, frequency and process history) for describing geographic phenomena (Guptill 1995).

Elements of spatial data quality attempt to enhance data validity for accurate and complete geographic inquiries, but Buttenfield and Beard (1994) argue that SDTS should incorporate uncertainty of real world conditions. The elements of data quality standard seek to enhance validity of data (uncertainty about database descriptions). Data quality standards have also focused narrowly on error in data rather than the wider consideration of uncertainty (Allan 2003). So, Buttenfield and Beard (1994) contended that the validity of geographic reality (uncertainty of real world conditions) to be integrated into data quality records. Hence, while standards for data quality are starting point for assessing geographic reality, they need to address geographic validity by assessing uncertainty of real world conditions.

2.3 TYPES OF UNCERTAINTY

The above section described the influence of uncertainty on standards of data quality. This section outlines types of uncertainty. The assessment of uncertainty establishes the level of certainty for derived information from available (or known) data. This assessment involves application of a particular analysis process to certain data depending on whether the process is data-driven or method-driven. Cluster analysis for single variables, for example, is a data-driven process because data distribution remains unchanged and it determines the path of the analysis process. Weighting processes, on the other hand, are method-driven because data are subjected to analytical tool concepts and assumptions which determine output data distribution.

Data quality comprises several defining elements including: subjective aspects such as fitness-for-use; and objective measurables like deviation from observed or attainable true values (Worboys 1998; Lo and Yeung 2002). Restrictions on data quality resulting from imperfection can arise for a variety of reasons such as inherent and operational errors in data (Worboys 1998). These may be deliberately introduced (e.g. census data), inherent in the real-world objects that are under study, or during

data acquisition (Worboys 1998). Operational errors associated with uncertainty may occur during the process of collecting, managing and using geospatial data (Lo and Yeung 2002). Goodchild (1989) observed that errors inherent in geographic data describe the differences that exist between data model and the geographic truth that the model represents (Lo and Yeung 2002). Worboys (1998, 258) observed that "deficiencies in data quality, leading to various kinds of uncertainty may be the result of several factors:

- Inaccuracy and error: deviation from true value
- Vagueness: imprecision in concepts used to describe the information
- Incompleteness: lack of relevant information
- Inconsistency: conflicts arising from the use of information
- Imprecision: limitation on the granularity or resolution at which the observation is made or the information is represented".

Inaccuracy and error is the deviation from the truth or a value taken to be true (e.g. standardized value) with the assumption that the true value is achievable at least in theory (Worboys 1998; Zhang and Goodchild 2002). Vagueness (or inexactness) refers to the existence of indeterminate location or borderline cases or the lack of a clear boundary to define a set of values that fully characterizes an object (Bittner and Stell 2002; Worboys and Clementini 2001; Duckham et al. 2003; Dubois and Prade 1988).

Incompleteness arises due to the absence of information in which uncertainty can be assessed as the amount of information required for recovering the truth (Zhang and Goodchild 2002). Completeness describes the degree of replicability of reality through feature abstraction represented in databases. Data acquisition methods and standards employed in the creation of spatial databases are essential determinants of completeness (Veregin 1999). Incomplete data do not contain the relevant information required to fully describe a phenomena under study, partly because of research requirements underlying data acquisition and limitation of concepts for data collection methods. Incomplete data may be due to poor metadata information. Geographic data may be incomplete depending on which ministry or institution acquires the data

because geographic data collection is commonly selective based on the research goals and requirements. Hence, data can be identified or associated with a particular data collection agency and vice-versa (Schuurman 2004). A missing material description or lack of spatial location in a well-log description represents incompleteness for describing sediment distribution structure of that borehole.

Inconsistency exists as a result of lack of uniformity inherent in information and may be due to heterogeneous standards or lack of coherent classification rules applied to information (Smets 1997; Veregin 1999). Inconsistency may also result when models fail to render valid or reliable outcomes resulting in incoherent conclusions where variables function differently under similar conditions (Bosc and Prade 1997, 290; Smets 1997,229). Well-log data may exhibit inconsistency because there are multiple sediment tags and conflicts may persist for spatial and elevation values of borehole units. Absence of inconsistency is an indication of the level of internal reliability or validity, but its identification does not guarantee possible correction (Veregin 1999). Imprecision refers to lack of specificity (Worboys and Clementini 2001) in representation or lack of repeatability or the degree of spread of measurements. Imprecision and inconsistency relate to the substance or content of information item; information is imprecise because data are incompatible with reality; in the later, because no consistent pattern exist between reality and abstracted information (Smets 1997, 227). Hence, imprecision and consistency can be traced for particular geographic information and corrected as they are identified with an information item.

2.4 SOURCES OF UNCERTAINTY

In the preceding section, types of uncertainty have been outlined as broadly due to complexity in geographic phenomena and inaccuracies in concepts and tools employed for information extraction. This section focuses on sources of uncertainty. Uncertainty may arise from a variety of sources depending on the geographic

phenomena under study or it may be specific to the methods and tools employed in data acquisition process.

Allan (2003) described major sources of uncertainty as: intrinsic uncertainty, inherited uncertainty, operational uncertainty and uncertainty in use. Intrinsic uncertainty includes inaccuracies in observation, definition, generalization, natural variation and operator bias (Allan 2003). Intrinsic uncertainties are associated with primary data. For example, in spatial mapping, geometric error may occur as a result of measurements on the spherical surface of the earth and its corresponding projection onto plane surfaces (Hunsaker et al. 2001). Inherited uncertainties arise from the management of primary data for storage or for other applications. Inherited uncertainties are linked to secondary data and comprise errors due to: age, relevance, scale, format, coverage symbolization or semantics (Allan 2003; Worboys 1998). Operational uncertainties arise from inaccuracies inherent in data analysis tools and their conditions and assumptions of application. Uncertainty in use is associated with the use of data for decision-making. Beard (1989) observed that uncertainty in use arise from users' different perceptions or interpretations of the output information (Allan 2003).

A persistent difficulty remains because different disciplines conceptualize geographic space differently leading to semantic heterogeneity in spatial databases. Semantic heterogeneity is much researched (e.g. Kuhn 2001, 2003; Raubal 2001; Harvey et al. 1999) though methods for resolving attendant interoperability problems remain elusive. The rising need for data sharing requires integrating or reconciling different meaning or standards (Harvey et al. 1999). Harvey et al. (1999) suggested addressing semantic differences by constructing data sharing environments to develop cross-standard exchange mechanisms. This study is not focusing on interoperability, but it is worth acknowledging these challenges because they result into uncertain information. Plewe (2002) observed that generally, geographic complexity and other problems result in uncertainty via two processes: human conceptualization, involved

with the simplification of reality, and measurements from which formal representations are developed to form conceptual models.

Other sources of uncertainty may be due to physical changes of attribute information over space and time (Hunsaker et al. 2001). Most geographic databases are static snapshots of reality requiring temporal dimension of features to be considered as an integral component of geographic phenomena. For certain analysis such as well-log data, temporal change is large and can be ignored for short duration studies.

Overall, uncertainty manifests itself in diverse ways: through most stages of processes with data (e.g. data sampling, conversions, and transformations) to final geographical decisions. But, uncertainty may be detected, measured and characterized in order to assign a level of confidence to geographic information. Openshaw (1989) observed that what most applications need is not exact estimates of error but a level of confidence to protect validity of output information. The manifestations of uncertainty are diverse and may be specific to a particular data or area of application. The following sections, describe some manifestations, detection techniques and characterization of uncertainty in two data: census data and borehole data.

2.5 UNCERTAINTY IN CENSUS DATA

The preceding two sections outlined types and sources of uncertainty. This section describes specific aspects of uncertainty exclusive to census data. Census data acquisition and analysis, like any geographic data, are subject to inaccuracies in tools and methods employed in the data collection and reporting processes and also partly due to complexity of geographic reality. A fundamental characteristic of census data is that data are collected at the individual level and reported at area units (Schuurman 2004; Duckham et al. 2003). In essence, individual data are aggregated to spatial (i.e. area) units based on arbitrary subdivision of the area under study.

Openshaw (1984a) observed that spatial data aggregation is essential to generate relevant data and is a convenient way to report data. The collection of data requires

data aggregation to a larger area unit relative to the observed pattern of the phenomena under study. In population health studies, for instance, the spatial association of tuberculosis in relation to a particular ethnic group will require data aggregation to an area that includes at least one specified ethnic group and the tuberculosis case and vice-versa (Dragicevic et al. 2004). However, the spatial units at which data are aggregated are one out of many different ways of dividing non-overlapping area units for spatial analysis processes (Openshaw 1984a). This uncertainty of arbitrary subdivision of area units for spatial analysis purposes is often referred to as the Modifiable Areal Unit Problem (MAUP).

2.5.1 The Modifiable Areal Unit Problem (MAUP)

The MAUP is a fundamental spatial analysis problem inherent in all aggregated data where results are susceptible to the configuration (that is, shape and size) of spatial units at which data are analyzed (Openshaw 1984a). In other words, the MAUP is the process where different spatial unit configurations result into different and conflicting results (Fotheringham and Wong 1991; Reynolds 1998; Davis 2003; Klinkenberg 2003). This influences subsequent results of analysis made on such data, and also how these results are interpreted which may lead to the problem of ecological fallacy. Ecological fallacy describes the inaccuracy resulting from spatial analysis of area data applied to individual level or the application of aggregate data relationships to individual relationships (Tranmer and Steel 1998; Marceau 1999). Openshaw (1984a) provided a comprehensive study into the MAUP, and identified two subcomponents of the problem as the scale problem and the aggregation problem.

The scale problem is a variation in analytical results when data are increasingly aggregated into smaller or larger spatial units (Barber 1998). Aggregation effect results because the spatial units are modifiable (Openshaw 1984a). That is, there are many different ways of subdividing an area for aggregation at the same scale. The aggregation problem is the variation in analytical outcomes due to alternate or different aggregation of area units at the same or similar scales (Openshaw 1984a; Barber 1998).

Openshaw (1984a) and Horner and Murray (2002) observed that the aggregation effect occurs because of the uncertainty of how spatial configurations should be defined to generate a fixed number of area units.

The effects of MAUP has being tackled in many studies to minimize its effect on grouped data analysis. It effects seemed to have resisted many mathematical tools and methods. Openshaw (1984a), for example, suggested optimal zoning system to create homogeneous units and Barber (1998) recommended combining similar units during aggregation to preserve original data variability. The creation of spatial units with least variance to curtail the aggregation effect can only be optimized for a single variable at a time. In addition, spatial heterogeneity and variability inherent in geographic data will not permit uniform spatial units for multiple variables. Real-world scenarios will require single area definition for spatial analysis, and least-variance spatial unit definition may be varied depending on the variable under consideration. The aggregation problem investigation requires individual level data to create modifiable units for different aggregations at the same scale. The aggregation problem investigation is beyond the scope of this study due to lack of individual-level data.

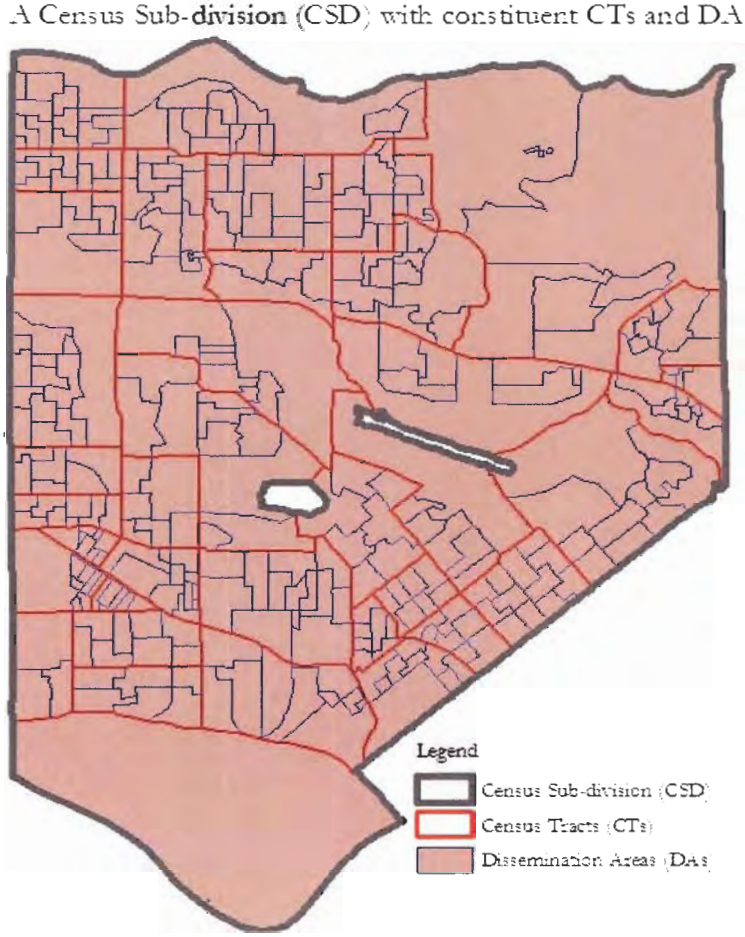
2.5.2 The Scale Problem

The scale problem is the variation in analytical results when data for a particular set of spatial units are aggregated into smaller or larger units for analysis (Openshaw 1984a). In the 2001 Canada census data, three common census units at which data are reported are: census sub-division (CSD), census tract (CT) and dissemination area (DA). The CSD represents the largest of these census units followed by CT, a medium size census unit and DA is the smallest census unit. Figure 2.1 illustrates these three census units (that is, CSD, CT and DA) and their relative sizes. A CSD represents a municipality or an area that is deemed to be equivalent to a municipality. A CT usually has a population of 2,500 to 8,000. They are located in large urban centres that have an urban core population of 50,000 or more. A DA

represents small areas consisting one or more neighbouring blocks, with a population of 400 to 700 persons.

So in the 2001 census data, for example, DA data aggregated to CT level represents a scale problem because a CT is a census unit which consists of two or more DA entities. This is a fundamental characteristic of all spatially grouped data for which census data are an example. Census data are gathered at fine spatial resolution and results are presented in aggregated form at coarser spatial resolution for privacy and other reasons (Schuurman 2004; Duckham et al. 2003). The scale problem can be considered analogous to scaling defined by Javis in Marceau (1999) as information transformation from one scale to another where upscaling is the information derivation from small to large scale and downscaling is information decomposition from one scale into its constituents at a smaller scale.

Figure 2.1: Three sample census units (CSD, CT and DA) and their relative sizes



A real challenge in the scaling process is the non-linearity between phenomena processes and the inherent heterogeneity of geographic properties that determines the rate of the processes (Jarvis 1995). Here, we are concerned about the effect of scale on statistical results and models when using aggregated data (Marceau 1999) and not the scaling in the natural sciences concerned with spatial patterns and associations at different scales. The scale problem considered with regard to aggregate census data can be restated as *the process where statistical results and relationships for aggregated data are different for the same set of data at different scales*. The scale issue is recognized as the uncertainty of the number of spatial units required for spatial analysis (Openshaw 1984a).

Figure 2.2: The Scale Problem - Map of Recent Immigrants at two scales (CT & DA)

The uncertainty specific to census data investigated in this study is the one associated with the scale problem. It involves the spatial transition of census data from DA to CT level in order to explore the spatial pattern of the scale effect. Figure 2.2 illustrates the spatial variability lost consequent to aggregation during the spatial transition from DA to CT and the increasing spatial heterogeneity within individual

census tracts (CT). Analysis of the spatial pattern due to the scale problem requires analytical tools that will retain the outputs at the original scales for accurate determination of disparities between different scales.

Hence, analysis tools which degrade input data distribution and introduce different spatial patterns between input variables are inadequate for this analysis. Many analytical methods have been applied on aggregated data to predict and minimize the effects of the MAUP. Among the analytical tools is multivariate statistical analysis by Fotheringham and Wong (1991, 1025) who applied multiple linear regression and multiple logit regression models in examining the sensitivity of estimates to variations in scale and spatial units. They concluded "The modifiable areal unit problem is shown to be essentially unpredictable in its intensity and effects in multivariate statistical analysis and is a much greater problem than in univariate or bivariate analysis. The results of this analysis are rather depressing in that they provide strong evidence of the unreliability of any multivariate analysis undertaken with data from areal units". Nakaya (2000) also examined the solution of optimal zoning system to MAUP and observed that the approach is suitable for spatial anomaly determination but presents a biased overall pattern. Openshaw (1984a) also investigated multiple dimensions of the MAUP and observed that the problem is best understood by empirical experiments and that MAUP should be considered a geographical problem rather than a statistical one and tools be developed as such to handle it. Further, he concluded that there are no convincing alternative techniques for managing spatially aggregated data in a statistically sound framework and suggested more radical non-statistical approaches to handling the problem. Following is a brief description of limitations identified with common analysis tools such as spatial regression and correlation, autocorrelation measures, multivariate statistical analysis, etc on spatially grouped data.

2.6 SOME LIMITATIONS OF COMMON SPATIAL ANALYSIS TOOLS

The preceding section discussed the relevance of data distribution as a precondition to resolving the scale issue and summarizes some early approaches to

limiting the effects of MAUP. This section outlines specific characteristics of spatial data which are difficult for most analysis tools to handle. The underlying problem is that analytical tool assumptions do not always accommodate characteristics of the data and resulting outputs are disjunct from reality because the method is tool-driven rather than data-driven. Below, is a description of common analysis tools with their inherent assumptions and conditions required from data.

Spatial analysis involving regression and correlation are valuable predictive and modelling tools allowing the creation of numerical terms to control one variable (dependent) from a single or multiple (independent) variables (Shaw and Wheeler 1994). These tools attempt to provide numerical prediction of geographic events based on specific assumptions between the predicted and predictor variable(s). An overview of some inherent inadequacies with regards to geographic information is described below.

The requirement of variable normality prohibits the use of percentages in regression models and requires a natural logarithm transformation which leads to complex conversions and loss of numerical clarity (Shaw and Wheeler 1994). Excessive disparity of intermediate output from the input data distribution due to data transformation can be considered loss of relative accuracy in the spatial model. Data normality requirements, however, are necessary for illustrating sound model-based inferences and determining statistics with variability that are less influenced by outliers (Griffith et al. 2003). Identification of outliers can easily be made for data acquired with repeated measurements mostly in the physical sciences. However, for geospatial data, spatial locations must be reconciled with attribute characteristics. This divide of analytical tool requirement and geographic data reality poses a conflict between derived model and data; resulting in data conformity to model requirements. But, the empirical data distribution (which conforms to the conceptual normal distribution theory) will yield reliable confidence intervals constructed for normally distributed error term assumptions (Griffith et al. 2003). So, the reliability and accuracy of these models are dependent on the input data characteristics, particularly data normality.

Next, the derivation of spatial relationships among data elements is a problem for unique data distributions. For example, in a complementary class such as gender groups of male and female have a pre-determined relationship such that as one increases the other decrease (Shaw and Wheeler 1994). These relationships may indicate strong or no spatial association but the variables may be spatially dependent on each other. Also the assumption that errors in regression models are statistically independent will often not be plausible due to spatial dependence in the sources of error (Lark 2000; Shaw and Wheeler 1994). Lesch et al. (1995) recognized this problem and suggested that the test of independence be applied to the residuals from regression models (Lark 2000). They recommended that regression only be used when the residuals appeared to be independent – a very restrictive condition for both single and multiple variable approaches (Lark 2000). Total errors in such models cannot be ascribed solely to the dependent variable alone because the independent variable may also be subject to error. The identification of independent and dependent variables may be difficult in cases where process–response relationship is not clear (Lark 2000; Shaw and Wheeler 1994).

In addition, statistical dependence may not necessarily imply valid geographical relationship between the variables (Lark 2000; Shaw and Wheeler 1994). A regression or correlation model may suggest a link or strong relationship between variables and at the same time the variables may just be responding simultaneously to different or unknown variable(s). Also there may be unique spatial considerations, such as spatial autocorrelation, geographic data dependencies, etc which may confound standard statistical approaches. An apparent result of these implications is the loss in variability of the predicted (or dependent) variable with regards to input (or independent) variable(s). These limit attempts to minimize the effects of MAUP in spatial analysis processes on aggregated data. Comparisons of various indices from spatial analysis processes at different spatial resolutions do not resolve or minimize the effects of MAUP because these analyses do not often retain the input data distribution.

Figure 2.3: Discernible spatial units –
Sample input variable

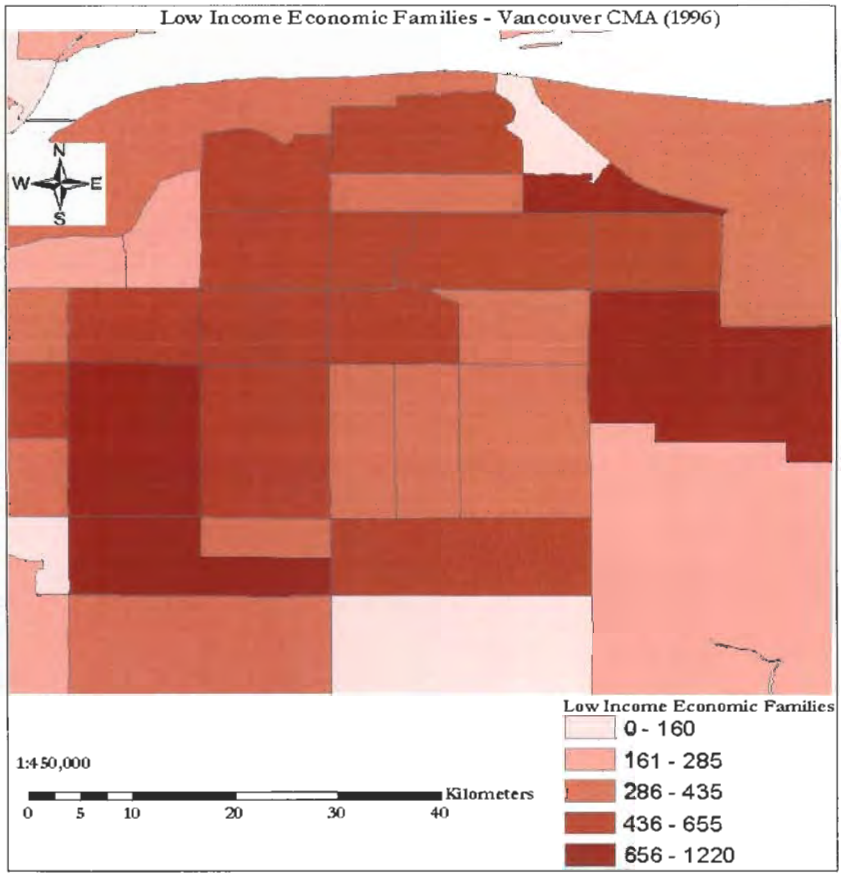
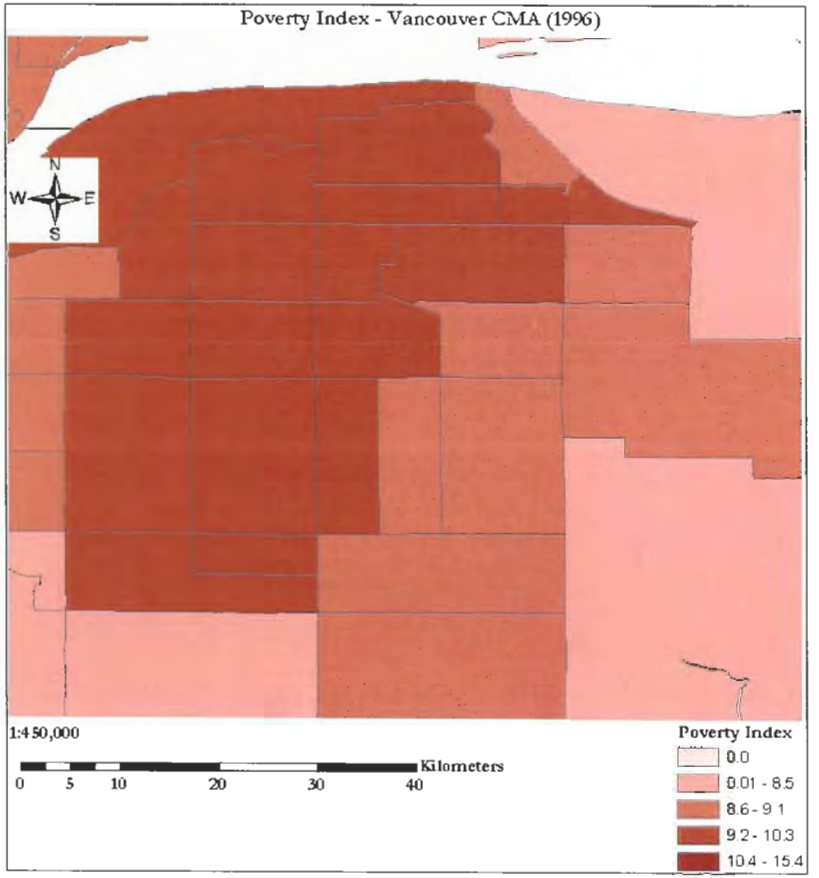


Figure 2.4: Indiscernible spatial units –
Loss of input data variability



Figures 2.3 and 2.4 are explicit examples of significant loss of input data variation and distribution inherent in spatial analysis processes. The maps are of the same spatial extents and census tracts. Figure 2.3 represents the input variable and Figure 2.4 is the spatial analysis output combining multiple variables. The spatial variation even at the same census scale is lost in this spatial analysis approach. This is not unique to spatial regression and correlation analyses alone but also to other spatial analysis tools such as factor and cluster analysis and spatial autocorrelation measures like Moran's I and Geary's Ratio.

These tools rely on global measures that must find best fit models based on some trade-offs in the input variable characteristics. But many approaches to improving model reliability and accuracy work by increasing input data volume or collecting data at small resolution or short temporal duration to satisfy central limit theory requirement. The central limit theorem which states that data normality can be achieved with sufficiently large data collected for a particular phenomenon does not apply to non-random data samples (Shaw and Wheeler 1994; Lark 2000). Also geospatial data are often dependent and heterogeneous requiring tools that can retain input data variation irrespective of spatial resolution or temporal dimension at which data are acquired.

2.7 UNCERTAINTY IN WELL-LOG DATA

The previous section outlined limitations associated with analysis tools in spatial data analysis. It illustrates how these inaccuracies propagate from known data into uncertainty during analysis. The section also suggested the need for non-statistical tools to retain data distribution during spatial analysis. This section identifies specific uncertainties inherent in well-log data.

Well-log data are geological samples mostly collected during groundwater investigation through borehole or water-well. A borehole or well is a vertical excavation constructed mainly for the purpose of groundwater extraction, subsurface exploration, artificial recharge and disposal of sewage or industrial waste (Tolman

1937; Tood 1964). These data may constitute key elements in determining the accuracy of subsequent applications using well-log data. The influence of data quality on resulting geographic representation cannot be over-emphasized as reliability of geological model resulting from data are determined by the quality of the original well-log data.

In Canada, borehole data are one major data source for subsurface study. Well-log data are employed in subsurface mapping of geological settings and requires accurate identification and description of the subsurface geology. Subsurface modelling provides a means of understanding subsurface geologies, which accounts for the distribution and movement of groundwater. However, because subsurface lithologies are hidden below the earth's surface, it is difficult and expensive to obtain samples (Schuurman 2004). In Canada, water well drillers are the primary source of well-log data (Schuurman 2002) involved in subsurface geologic material identification and description. Regrettably, due to the variability in experience and training of these private well drillers, the well-log data varies considerably in its level of detail, with many soil deposits and rock formations being misrepresented and several descriptions being given for one material unit. This has resulted in different lithological terms being used to describe the subsurface in British Columbia (Schuurman 2002) resulting in high variability in geological formation within a small region. Schuurman (2002) observed that this figure far exceeds the actual material distribution in the subsurface geology.

A snapshot of material descriptions of well-log data collected by the private well drillers is as shown in Figure 2.5. The data signify not only lack of experience in subsurface material identification and description on the part of the well drillers, but, also inherent complexity in the geological structure of boreholes. A borehole comprises continuous geological units which for purposes of geological analysis must be discretized and collected as distinct entities. The continuous characteristic of sediment units of a well-log data partly reflect the difficulty experienced by the well drillers in identifying the material limits and the struggle to differentiate between materials in assigning single tags to borehole units.

Figure 2.5: Sample borehole data material description by private well drillers

Description
Sand - medium to fine and brown clay
Blue clay some fine gravel
Coarse sand and fine gravel high silt content
Coarse sand and fine gravel
Coarse sand some medium sand
Coarse sand and fine gravel
Gravel medium to fine and coarse to medium sand
Sand - coarse to medium and fine gravel
Medium sand some coarse sand
Fine sand some medium sand high silt content
Brown silty clay some sand and fine gravel
Blue clay little gravel.
Gravel - medium to fine some coarse sand

Figure 2.6: A simple cross-section requiring no manual intervention – relatively homogeneous geological formation

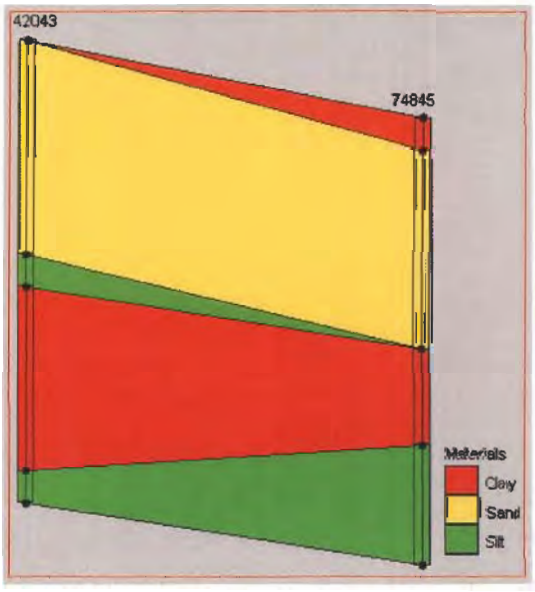
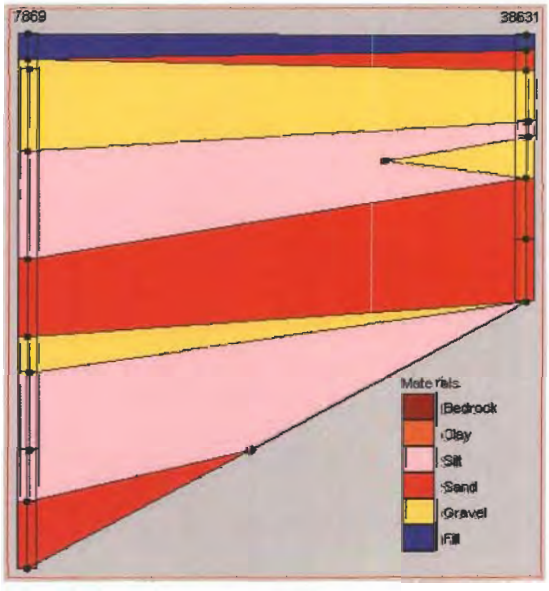


Figure 2.7: A simple cross-section requiring manual intervention – relatively heterogeneous geological formation



Also there exist gradual material transitions between different sediment types confusing the well drillers to assign multiple material tags. A single borehole unit is not given a distinct tag, suggesting that a material unit may belong to more than one

category whenever an attempt is made to place them in single group. This is an explicit illustration of inherent geological complexity and the uncertainty experienced in classifying geological space. This uncertainty is what is investigated in this study with regards to well-log data.

Immediate and apparent consequences of this well-log data collection process is the high geological variability and distribution within boreholes and between boreholes, which many applications find challenging and demanding to support. Figures 2.6 and 2.7 illustrate this high geological material heterogeneity within and between boreholes from such data.

2.8 WHY CONSIDER GEOGRAPHIC DATA UNCERTAINTY

The preceding sections (sections 2.3 and 2.5) have identified specific uncertainties in the two data used in this study. This section discusses the relevance of integrating uncertainty into geospatial analysis. Uncertainty exist in all forms of geographic realities, mandating uncertainty integration into all geographic information systems (Smets 1997).

Models resulting from geographic information systems are always imperfect (Zhang and Goodchild 2002; Smets 1997; Hunsaker et al, 2001). This may partly be due to inherent complexity in the real-world and inaccuracies in tools and methods employed in geographic model development. For instance, most geographic spaces are continuous, while observations must be discretized to allow data analysis and facilitate decision-making (Goodchild et al. 1994; Zhang and Goodchild 2002). Optimal and accurate representation of geographic complexity in simple and complex models may be impractical, but the assessment of model accuracy is important for the use of data to certain applications. This was clearly recognized by Openshaw (1989, 265) who observed that "... what many applications seem to need is not precise estimates of error but some confidence that the error and uncertainty levels are not so high as to render in doubt the validity of the results in a particular data specific situation".

Openshaw (1989, 265) identified two approaches to minimizing uncertainty: "... develop adequate means of representing and modelling uncertainty and error characteristics of spatial data; and secondly, to develop GIS related methods and techniques that can explicitly take error into account during operations with spatial data". The first approach requires conceptual model and methods that can fully accommodate spatial data properties; it can be considered a general broad-based approach for most geographic analysis. The second technique is what is considered in this study: applying analytical tools which minimize uncertainty in spatial analysis processes to safeguard final decisions. The method is not aimed at total eradication of uncertainty in geographic analysis processes, but the application of analytical tools that utilize available information to minimize or lessen quantifiable inadequacies which render spatial decision-making sour.

Additionally, digital geographic information from computers are trusted to be of high accuracy and precision (Keukelaar 1999; Duckham and Sharp 2004; Motro 1997). The integrity of geographic information is related, however, to geographic reality. Geographic inquiry dependent on digital databases must be comparable to reality in order to maintain spatial data reliability and protect integrity of geographic decisions.

2.9 ANALYTICAL TOOLS - UNCERTAINTY MODELS

The preceding sections have outlined specific aspects of uncertainty in the two data: census data and well-log data. The sections below will identify and describe analytical tools for characterizing uncertainty, their assumptions and conditions of application.

There are a variety of analytical tools and methods employed in minimizing the effects of uncertainty in geographic information. Depending on what is identified to describe uncertainty, certain methods are applied in controlling its occurrence and propagation. Dominant and common in many approaches is the conventional error propagation inherent in many analytical processes using probability distribution.

Error models may be formulated using stochastic processes capable of simulating a variety of known error possibilities or inaccuracies identified with analytical processes (Abbaspour et al. 2003). These error models aimed at minimizing residuals in the modelling process in order to bring the model close to input data. Uncertainty assessment, however, goes beyond reducing error during modelling processes but attempts to replicate reality or data distribution. This chapter outlines different aspects of uncertainty modelling by using available information to develop models, which closely characterize reality. Four prominent tools for uncertainty modelling are probability and stochastic models, fuzzy set theory and rough set theory. Following is a brief outline of their respective approaches to uncertainty modelling.

2.10 PROBABILITY AND STOCHASTIC MODELS

The theory of probability is an advanced mathematical tool with clear and standard concepts. The probability of an event is a number expressing the degree of chance or belief that an event will occur or a proposition is true (Henri et al. 1997). Probability values range from zero (0), signifying a false proposition or an event will not occur to one (1), indicating occurrence of an event or a true proposition.

Probability theory has enriched and strengthened the analysis of geographic data and their applications in scientific research have an advantage of well understood concepts (Zhang and Goodchild 2002; Duckham and Sharp 2004). Probability concepts have been employed in diverse disciplines of geography with certain degrees of success, for example, Manslow and Nixon (2002) evaluate the ambiguity in a sensor's point spread function (PSF) on the information it acquires; Dowd and Pardo-Iguzquiza (2002) investigated the model of uncertainty in geostatistical analysis for geological data using stochastic method; Diggle and Ribeiro (2002) utilized Bayesian approach to spatial interpolation smoothing in order to accommodate uncertainty associated with unknown value determination for model parameters; Brunsdon 2001 also employed Bayesian models for determining catchment zones for schools. The list continues, but there are conditions and assumptions, which must be satisfied for satisfactory results.

Probability and stochastic models are dependent on random variation and independent data. The random requirement does not accommodate spatial dependencies, which are highly associated with geographic data. The independency condition is inflexible to allow gradual transition of spatial phenomena that are characterized by vagueness. The conditions for probability and stochastic models are complex and are difficult to maintain (Duckham and Sharp 2004). For example, we apply probability if and only if residuals appear to be independent. Some analytical processes, however, are carried out with less regard to these conditions, probably ignoring their consequence or hoping their effects are minimal. Errors from spatial data are characteristic of the data under study and Zhang and Goodchild (2002, 87) observed that "all spatially distributed data demonstrate spatial dependency to a certain degree and so do spatial errors". In other words, uncertainties in geographic data must reflect trends underlying the data. Hence, for probabilistic and stochastic approaches to yield satisfactory results, both data and error characteristics must satisfy prevailing conditions.

2.11 STOCHASTIC SIMULATION - MARKOV CHAIN MODEL

The section above discussed probabilistic and stochastic restrictions on data in order to yield satisfactory outcomes. This section describes one stochastic model – Markov chains – and its approach to characterizing uncertainty. The Markov chain is a dynamic probabilistic model for analyzing complex systems using the concept of state and state transitions (Howard 1971). A system's state represents all descriptive values, which characterize the system at any instant. The dynamic behaviour experienced by the system from one state to another is called state transition or simply transition. Some processes whose states and transitions are finite and whose probabilistic character is random possess a Markov chain. So, if the probability of a process's state is dependent on only the present state for a given transition period then the process is called a Markov chain (Elfeki and Dekking 2001; Howard 1971; Carle and Fogg 1997). The statistical description of a system's process using Markov chains requires the

specification of conditional probabilities of all individual states in the system. The specification of these probabilities can in itself be a problem and its implementation could be complex (Howard 1971).

Markovian assumption is employed to simplify both the complex behaviour of the system and the problem of specifying the process. The Markov dependence assumption is: given the present, the future does not depend on the past, in other words, only the present state characteristics of the process are relevant in determining its future behaviour (Elfeki and Dekking 2001; Howard 1971; Carle and Fogg 1997). In other words, the probability of future transitions of each state in the system depend only on the present state occupied. For instance, consider a borehole with which equal depth intervals (say 0.5 metres) are marked from the top down to a specified depth. By using this equal depth interval as borehole state transitions, different material states could be identified (say sand, gravel, silt, etc). Hence, borehole material transition from one state (say, silt) to another, given the present conditions, is not dependent on the material's past occurrence. That is, Markov model of spatial variability assumes that local occurrence of a category depends entirely upon the nearest presence of another category and independent of more distant occurrences (Carle and Fogg 1997).

The validity of this assumption is the concern of many complex system analysts because the compatibility of this assumption to practical observations is integral to model accuracy. It is however, analogous to first law of geography which states that all things are related but near things are more related than far things (Tobler 1970). While, no experimental results can fully support the Markov assumption, there are neither practical processes that are entirely non-Markovian (Howard 1971).

2.11.1 Transition Probability Matrix

The preceding section discussed the Markov chain concept and assumptions for the treatment of uncertainty. This segment describes a key ingredient – transition probability for implementing the Markov chain. The conditional probability specification of a system's state and transition processes forms an integral component

to the implementation of the Markov model. So, a Markov process is defined by specifying for each state and transition time, the probability of making the next transition to each other state given the present conditions. The transition probability, P_{ij} is the probability that a process presently in state i will transit to state j after its next transition (Howard 1971). A transitional probability is thus, obtained by dividing the frequency of transitions by the frequency of the state in question. It represents a conditional probability, the probability to move to state j, given that the subject is in state i. The transitional probability is not the same as joint probability, which reflects the overall probability of observing a certain transition (it is computed by dividing by the total number of observations). The transition probability, P_{ij} satisfies the probability unit scale requirement that is:

$0 \leq Pij \leq 1$ for $1 \leq i, j \leq n$

where n is the number of transitions.

For a finite number of transitions, n where the process must occupy one of its n states after each transition, the sum of all transition probabilities must be one (1.0).

$$\sum_{j=1}^{n} P_{ij} = 1.0 \text{ For all i = 1, 2, 3, ..., n}$$

The transition probability that describes a Markov process is represented by a square (n×n order) matrix called transition probability matrix (P_t) with P_{ij} elements.

$$P_t = P_{ij} = \begin{bmatrix} p_{11} & p_{12} \cdots & p_{1n} \\ p_{21} & p_{22} \cdots & p_{2n} \\ \vdots & & \\ p_{n1} & p_{n2} \cdots & p_{nn} \end{bmatrix}$$

This transition probability matrix, P_t is a stochastic matrix, that is a matrix whose entries cannot lie outside a unit scale (0, 1) and whose row's sum is one (Howard 1971). The transition probability describes categorical data dynamics by revealing information about the underlying structure of the data sequence (Lemay 1999). The

transition probability matrix forms a fundamental framework and remains the first step into the Markov process modelling.

2.11.2 Multistep Transition Probability

The assignment of probability values to future states for n number of transitions given the present condition is one practical problem most complex system analyst's face. In other words, given the present borehole condition what is the borehole state after n number of transitions? This raises the question of how does the present condition and the number of transitions influence the condition of future states? The quantity $\phi_{ij}(n)$ is called the n-step transition probability of the Markov process from state i to state j (Howard 1971). The multistep transition probability $\phi_{ij}(n)$ is related to the transition probability p_{ij}. The multistep transition probability can be written into an $n \times n$ matrix as:

$$\Phi(n) = \{\phi_{ij}(n)\} = \begin{bmatrix} \phi_{11}(n) & \phi_{12}(n) & \dots & \phi_{1n}(n) \\ \phi_{21}(n) & \phi_{22}(n) & \dots & \phi_{2n}(n) \\ \vdots & & & \\ \phi_{n1}(n) & \phi_{n2}(n) & \dots & \phi_{nn}(n) \end{bmatrix} \qquad n = 0, 1, 2, \dots$$

For a process starting at state i to another state j after $(n+1)$ transitions, the multistep transition probability matrix $\Phi(n)$ and the transition probability P are related by the equation:

$$\Phi(n+1) = \Phi(n)P \qquad \text{for } n = 0, 1, 2, \dots$$

$$\Phi(0) = I \qquad \text{where I is an identity matrix}$$

For successive n transitions, the multistep transition probabilities are:

$$\Phi(0) = I$$

$$\Phi(1) = \Phi(0)P = IP = P$$

$$\Phi(2) = \Phi(1)P = P^2$$

$$\Phi(3) = \Phi(2)P = P^3$$

Thus, in general, $\Phi(n) = P^n$ for $n = 0, 1, 2, \dots$

The multistep transition probability equation is a fundamental relation with each row specifying the probability distribution that will exist over the states of the process after n transitions for each possible starting point (Howard 1971).

The multistep transition probability equation raises a question of validity for extremely large numbers, that is, the response of the matrix elements as n increases. It is apparent that when n increases then the difference between rows of n and (n+1) decreases and will eventually approach zero. For example, consider the following multistep transition probability matrix for increasing power of n:

$$\Phi(0) = I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \Phi(1) = I = \begin{bmatrix} 0.6 & 0.4 \\ 0.5 & 0.5 \end{bmatrix}$$

$$\Phi(2) = I = \begin{bmatrix} 0.56 & 0.44 \\ 0.55 & 0.45 \end{bmatrix} \qquad \Phi(3) = I = \begin{bmatrix} 0.556 & 0.444 \\ 0.555 & 0.445 \end{bmatrix}$$

$$\Phi(4) = I = \begin{bmatrix} 0.556 & 0.444 \\ 0.556 & 0.445 \end{bmatrix} \qquad \Phi(5) = I = \begin{bmatrix} 0.556 & 0.444 \\ 0.556 & 0.444 \end{bmatrix}$$

Hence, for extremely large values of n, the multistep transition probability becomes:

$$\Phi(\infty) = I = \begin{bmatrix} 0.6 & 0.4 \\ 0.6 & 0.4 \end{bmatrix}$$

So, for very large values of n: $\Phi = \Phi(\infty) = P^{\infty}$ is called the limiting multistep transition probability (Howard 1971) because the deviation of subsequent rows for closed states approaches a limiting value of zero (0). This mathematical proof validates the Markov assumption, because there is a lost of dependency of future states on past states and it is only the current state that reflects what happens next. In sum, the transition probability provides a means of simulating hidden patterns in complex systems.

2.12 FUZZY SET THEORY

The section above discussed one stochastic simulation, that is, Markov chains and its basic framework for implementation in minimizing the effects of uncertainty.

This section describes another tool – fuzzy set theory – for characterizing uncertainty. Fuzzy sets concept was introduced by Zadeh in the 1960s to deal with uncertainty in complex systems (Dubois and Prade 1988).

In classical logic, sets are constructed with definite set characteristics enforcing strict outcomes; either an element belongs to or it does not belong to a set. For instance, an observed area is classified as water or land with sharp grades of set membership which does not recognize the interaction that may occur for gradual transition of one geographic phenomenon to another. While conventional set theory can be employed to certain applications (crisp and independent physical features such as roads, houses, etc) with high degrees of success, this approach needed modification for quantitative analysis of geographic information.

Geographic space is a continuum with particles interacting at different levels to form features abstracted even at one scale of observation coupled with complexity of geographic reality. To account for this complexity, fuzzy set theory was introduced to measure the degree of membership of an element to a set (Dubois and Prade 1988; Jiang 1998). In essence, fuzzy set theory was designed to accommodate partial set memberships, vague boundaries and allow gradual transition of one phenomenon to another. In fuzzy sets, each element is identified with a real number within the unit interval (0, 1) which describes the degree of membership or belongingness of that element to a set (Duckham and Sharp 2004). Zero (0) signifies no membership while one (1) indicates full membership, so the closer the result of fuzzy membership value is to 1 the higher the degree of its membership.

Interest in the application of fuzzy set theory to uncertainty in geographic information has advanced over the past two decades (Duckham and Sharp 2004). Fuzzy sets have been applied in a variety of areas within geography such as remote sensing, environmental and ecological systems, etc. Lagacherie et al. (1996), for instance, used fuzzy sets to handle uncertainty in delineating soil boundaries; Fortin and Edwards (2001) also employed fuzzy sets to demarcate vegetation boundaries in remote sensing; Allen et al. (2002) utilized fuzzy set theory to model and visualize

three dimensional structure of an aquifer; Carranza and Hale (2001) mapped mineral distribution using fuzzy set theory to redirect surficial exploration processes to potential sites; Dragicevic et al. (2001) modelled urban growth dynamic using fuzzy logic. There are numerous and diverse applications of fuzzy set in geographic information.

Fuzzy membership values are appropriate for accommodating different levels of complexity in geographic data which are characterized with vague or fuzzy characteristics. However, the assignment of fuzzy membership values is difficult because values are often subjective, a characteristic usually regarded as a weakness in the application of fuzzy set theory (Duckham and Sharp 2004). Fuzzy membership values are integral components of fuzzy sets (Lin 2001), thus, the level of accuracy of fuzzy set theory is highly dependent on set membership function formulation. Lin (2001) observed that since probability is the most common concept of uncertainty, fuzzy membership functions can be easily confused as probability. The unit interval for both probability and fuzzy membership values can also be misinterpreted to apply these numerical values interchangeably. Probability values are determined specifically with reference to probability space as a measure of event occurrence based on some observed population. While fuzzy membership values can refer to probability values, Lin (2001), pointed out that the identification or association of probability space is required for satisfactory application. Fuzzy membership values have primarily been assigned using expert knowledge, probability and possibility distributions. The section below describes probability and possibility distributions for assigning fuzzy membership values.

2.12.1 Probability and Possibility Distributions

In the preceding section, fuzzy sets theory is described. The above showed the problem of assigning membership function values, which may be subjective. This section discusses two ways of assigning membership values; excluding expert knowledge.

When dealing with uncertainty in geographic information, certain truth-values must be assigned to data elements of the real world they represent in order to assess the degree of confidence in data. This truth-value assignment may be difficult due to uncertainty caused by geographic complexity. Approximate methods are used in determining possible values based on available information (Bonissone 1997). Commonly used approximate techniques are probability and possibility techniques.

Probability concepts are well developed and theoretically advanced uncertainty model. The probability of an event is a number expressing the degree of chance or belief that an event will occur or a proposition is true (Henri et al. 1997). Probability values range from zero (0), signifying a false proposition or an event will not occur to one (1), indicating occurrence of an event or a true proposition. Henri et al. (1997, 256) described three probability perspectives namely; propensity view, frequency view and subjective probability.

In propensity perspective, probability is a physical characteristic of a device (e.g. a fair coin), the tendency of a particular coin to show up heads, for example. The occurrence of different events is dependent on the physical characteristic of the device, and devices could be partially constructed. In frequency view, probability is the convergence limit of relative frequencies for repeated random events or the characteristic of a population of like events (Smets 1997; Henri et al. 1997), the occurrence of a particular geological material (e.g. clay) in a specified borehole, for example. Smets (1997) observed that this probabilistic view is the most widely accepted, however, the observation of convergence limits is impossible. The assumption that past event occurrence pattern will be the same for future occurrences is not possible for single events. Threshold specifications are required when convergence limits are reached for certain phenomenon. Geographic data, for example, exhibit spatial dependency limits called 'sill' used in kriging interpolation technique that indicates the variance value at which spatial dependency ceases among data variables.

Subjective probability is a numerical value indicating a person's confidence or degree of belief in a proposition or the occurrence of an event using the person's knowledge about the phenomena under study (Henri et al. 1997). This probability perspective is open to many criticisms such as personal subjectivity and as Henri et al. (1997) pointed out, it is contrary to propensity and frequency views as subject to a particular observer's perspective. Generally, probability values are determined based on some evidence that may be subjective or objective based on available information. However, optimal decisions arise from these numerical values with no degree of confidence on the real validity or reliability of an event's occurrence or the truth of a proposition.

In addition, geospatial data are often spatially dependent, a characteristic contrary to randomness condition of probability distribution. Spatial variability and geographic heterogeneity characteristics do not allow the application of convergence limits, for example, the application of probability value of clay in one borehole to another. Probability applications assume symmetrical pattern of event outcomes (Smets 1997); spatial data, however, may exhibit disparate or unequal classes of outcomes. The number of geological units (materials) in different boreholes is not bound to be uniform in order to enforce symmetrical outputs. Probability values for a particular geological material will be different because there are different sediment populations in different boreholes resulting in failure of the symmetric condition.

Possibility theory, on the other hand, was introduced by L. A. Zadeh in 1978 in connection with fuzzy set theory to deal with uncertainty that accounts for an element's association with one or more classes when one attempts to place them in specific category (Rokos et al. 2004; Dubois and Prade 1988). The possibility approach was introduced with the identification of non-probabilistic uncertainties in information systems. Spatial autocorrelation and geographic data dependency, for example, do not assume probabilistic conditions of randomness and independency of event occurrences. Lagacherie et al. (1996) observed that in fuzzy set theory, the grade of membership does not necessarily exhibit random characteristics but does exhibit a

possibility character. Possibility concept assesses the degree of event occurrences or to what extent the occurrence of events are possible and the certainty of event occurrence without prior knowledge of probability of its occurrence (Rokos et al. 2004; Dubois and Prade 1988). Possibility theory also has descriptive interval values ranging from zero (0) signifying impossible events, to one (1) indicating complete possible events.

Fuzziness is different from randomness which deals with the probability of an element's membership to distinct sets, while fuzziness entails the uncertainty of belonging to a fuzzily defined set (Piatetsky-Shapiro 1997). Probability values are based on evidence from well-defined sets and set definition whose constituents describes set population play an integral role in probability value determination. However, in most information systems, these set characterizations are not uniquely defined and imposition of defined (that is, crisp) sets over fuzzy set will result in conflicting membership values. Possibility distribution is designed to allow the description of entities which conform to fuzzy constraints or belong to ill-defined sets (Smets 1997). A set of highly permeable geologic materials in a borehole based on multiple material tags may be vaguely defined for subsurface material classification because constituent materials may belong to different sets. Smets (1997) observed that numerical values of a possibility distribution do not matter; it is the ordinal system that imposes an order on the elements of the domain. He further emphasized that possibility and probability values do not necessarily correlate to imply high possibility mean high probability and vice-versa, but if an event is impossible it is also bound to be improbable. This paradigm does not only approximately quantify uncertainty but also evaluates the real validity or reliability of an event occurrence.

2.12.2 Rough sets Theory and other set (Boolean & Fuzzy sets) Concepts

The preceding section discussed the assignment of fuzzy membership values using probability and possibility distributions. This section discusses rough sets and other set concepts: Boolean and fuzzy sets. It is important to differentiate rough sets theory from other set concepts and why these concepts are not suitable for the kind of

uncertainty examined in this study. The following distinguishes rough sets from Boolean sets (or classical sets) and fuzzy sets.

In classical logic, sets are constructed with definite set characteristics enforcing strict outcomes, either an element belong to or it does not belong to a set. For instance, a river bank may be classified as water or land with sharp grades of set membership that do not recognize and account for interactions which may occur for gradual transition of one geographic phenomenon to another. Classical logic is based on Aristotelian logic where:

- everything is what it is – law of identity,
- something and its negation or inverse cannot both be true – law of non-contradiction and
- every statement is either true or false – principle of excluded middle (Burrough 1996).

So in Boolean sets, an element is assigned one (1) or zero (0) (or 'Yes' and 'No') to categorize the element as member of a set or not a member respectively. In other words, Boolean sets do not accommodate varying set memberships; that is, an element is either part of a class or it is not. Boolean sets can be used for some applications (e.g. crisp and independent physical features such as roads or houses) with high degrees of success. But this concept needs modification for quantitative analysis of geographic information.

In fuzzy sets theory, on the other hand, membership functions enable elements to exhibit partial class memberships of different and overlapping sets in order to account for multiple states of an entity. Confusion sets, however, may result in cases where zones of different fuzzy sets intersect (Burrough and McDonnell 1998). This may arise where an element is a partial member of three or more fuzzy sets which may generate two or more intersecting fuzzy zones. For instance, when we select a land area based on three characteristics; slope, vegetation cover and soil type. It is likely, at

least in practice, that some land areas may exhibit overlapping characteristics which may result into confusion sets.

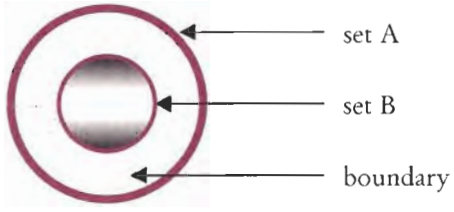
Rough sets accommodate confusion sets by applying set rules to derive lower and upper approximation sets, and allow the possibility of partial set memberships. So considering the case where land areas are selected based on three conditions: slope, vegetation and soil type. When confusion zones arise, then we develop set rules to characterize elements that definitely belong and those with partial memberships to construct lower and upper approximation sets respectively. The design of set rules is dependent on specific data patterns; the two case studies in chapters 3 and 4 demonstrate sample derivation of rough sets rules. The following section describes set characterization using the concept of rough sets.

2.13 ROUGH SET THEORY

The preceding section discussed rough sets and other set concepts: Boolean and fuzzy sets. This section describes a generalized concept of rough sets in order to establish a background upon which to implement rough sets in spatial analysis: for neighbourhood characterization in census data; and for subsurface characterization using aquifer properties.

A rough set is a set (that is, a classical set extension) that has a nonempty boundary when approximated by another set (Pawlak et al. 1995). In other words, a rough set is a conventional set whose boundary is too rough to be approximated by a crisp set.

Figure 2.8: Crisp set characterization into a rough set scenario



In Figure 2.8, the set A may be described as having elements in both shaded regions. While at the same time it may strictly be described to contain only elements in the inner set B. Based on different or the same criteria of observation, set A may be characterized as having either elements in set B only or both set B elements and those outside (boundary elements). Elements outside set B may exhibit slightly different properties based on the spatial or measurement resolution. To handle this, a threshold could be set to define set characteristics of set A. But, some elements may still exhibit properties that do not necessarily exclude them from the set criteria. Also some elements may exhibit different characteristics at different times (today it belongs to set A and tomorrow it does not).

In addition, observations can be made at different spatial resolution or scales. At a fine resolution, an element may be found to belong to a set, but found excluded at a coarse resolution. So the question is what scale or resolution should we use to define or discern observation or measurements made close to boundary elements? At what resolution will boundary elements exhibit different characteristics? In other words, what is the spatial resolution do we classify elements to belong to one set for having indiscernible properties? Rough sets analysis of elementary granules for a geographic data can be employed to develop indiscernible relations in categorizing geographic space.

The primary concept of rough set is the indiscernibility relation (Pawlak et al. 1995) that is developed from elementary sets. Information subsets of entities characterized by the same or similar information that are indiscernible are called elementary sets (Zhang and Goodchild 2002; Pawlak et al. 1995). Borehole units with multiple distinct tags represent indistinguishable elements with each tag characterizing the material differently into different categories. Also a census tract with constituent distinct dissemination areas (that is, elementary sets) represents a rough set. A rough set may be definable if there is a finite union of all elementary sets (Pawlak et al. 1995).

The concept of rough sets is based on approximation space to effectively categorize an information space using patterns inherent in available information

represented in information (or decision) tables. Rough sets have three approximated regions as identified in Figure 2.8 as: lower approximation set, upper approximation set and the boundary region. Lower approximation set is a consistent union of all elementary sets that are definitely members of the rough set (Zhang and Goodchild 2002; Duckham and Sharp 2004). Upper approximation set is the union of all elementary sets that have nonempty intersection with the entire set; in essence, upper approximation set characterizes set items that possibly belong to the set (Zhang and Goodchild 2002; Duckham and Sharp 2004). The boundary region comprises elementary set elements that result because of the nonempty characteristic of the upper approximation set, in other words, the failure of congruency between the upper and lower approximation sets. Zhang and Goodchild (2002, 181) and Pawlak et al. (1995, 91) described the boundary region as the "disparity between the upper and lower approximation sets" or the collection of "elementary set elements that are members of the upper approximation set but not members of the lower approximation set".

Hence, there are two necessary conditions for the implementation of rough set to multiple or single variables:

- the variable must be categorized to have nonempty boundary when approximated by another variable (e.g. sub variable(s))
- the sub variable(s) which constitute the rough set variable must be crisp sets (that is, elementary sets)

So, a single variable can only be treated as a rough set if and only if it has nonempty boundary when approximated by a crisp set. For example, an orange is a crisp set (or elementary set) when it is classified into a family of citrus fruits. It is however, a rough set when it is approximated by the peel (that is, orange skin) and the pulp (that is, inside tissue). Hence, a variable is a rough set; other words it is a crisp set with uniform constituent elements (Pawlak et al. 1995).

Consequently, the elementary set characterization and the definition and quality of lower and upper approximation sets play an integral role in the effectiveness

of rough sets assessment of uncertainty. The quality of set approximation, indiscernibility and attribute dependency are concepts which indicate how accurate we can predict outputs with a particular set of data (Katzberg and Ziarko 1994). Lower approximation set constituents are used to design certain rules because their characteristic can be definitely derived from the available information with certainty. Similarly, the upper approximation set properties are used to develop possible rules since there is some possibility that their characteristics have certain truth values. Pawlak et al. (1995) compared the quality of the lower and upper approximation sets as belief and plausibility functions respectively which are commonly used in Dempster-Shafer evidence theory (a generalization of the Bayesian concept of subjective probability). Pawlak et al. (1995) also defined quality of lower approximation as the ratio of the number of all elements in the lower approximation set to the total number of elements; and quality of the upper approximation is the ratio of the total number of elements in the upper approximation set to the total number of elements. For any approximation, the accuracy of set element estimation can also be determined to assess which approximation process adequately represents the original set characteristics.

CHAPTER 3:
FIRST CASE STUDY - CENSUS DATA METHODS, RESULTS & DISCUSSIONS

In the preceding chapter, uncertainties associated with spatial data were described. Also common analytical tools for handling uncertainty and their assumptions and conditions of application were outlined. This chapter constitutes the first case study and it recognizes the numerous uncertainties in spatial data but focuses on only the scale problem. The chapter illustrates the use of rough sets to mitigate the scale issue in census data. In order to examine this scale issue, the case study investigates spatial relationship between recent immigrants and deprivation indices in the Greater Vancouver Regional District (GVRD). A number of socio-economic factors are considered in relation to recent immigrant settlement in Greater Vancouver Regional District (GVRD) at census tract (CT) and dissemination area (DA) levels. So, rough sets will assess recent immigrant and deprivation index patterns at multiple census units (that is, CSD, CT and DA).

Current literatures have identified strong correlations between poverty of CT and the proportion of its population who are immigrants (Kassim and Laurel 2000). Also rising poverty in Canada raises the possibility that the misery is absorbed by certain segments of the population, particularly, immigrants of certain ethnic origins (Kazemipur and Halli 1998). This study examines the extent of this relationship in order to verify whether the overall index does provide information on whether or not the observed trend is applicable to all recent immigrants.

Census data like other aggregate data pose a persistent challenge that renders resulting patterns from this data subject to the census resolution used. Policies are implemented with little regard to census units utilized; that is policies may not be in conformity to spatial resolution of the census data used to inform policy development. Since one fundamental focus of policy development is to ensure that services reach the

people who need them, spatial data analysis must preserve unique data characteristics in model development. Policies may have profound impact on services reaching people at a particular place and time, but data analysis methods that aim to replicate reality should play an integral role in the development of these policies. But, assumptions and conditions of analysis techniques must conform to data characteristics. Hence, methods used in this project are geared towards retaining data distribution and variation rather than analytical tool characteristics.

The implementation of census data for spatial model design is aimed at the scale problem investigation. The scale issue can only be examined by considering two or more spatial resolutions to explore the pattern of each spatial resolution relative to each data distribution. A prerequisite for such model development is the preservation of data characteristics to permit optimal approximation of model discrepancy due to scale transition. A unique technique is rough sets analysis of spatially grouped data. While the assumption of variable dependency is required for assessing relationship between two or more variables in many analytical methods, it is not necessary in rough sets technique. This assumption of variable dependency is required in selecting dependent and independent variables, for instance, in most statistical analysis like regression and correlation. The spatial dependency assumption among variables has fundamental model implications because this may not be valid hypothesis and it is predetermining unknown spatial relationship.

Rough set implementation for spatial data analyzes variables independently with no assumption of spatial dependency. For example, a deprivation index for socio-economic variables is developed from various variables interacting with different characteristics. The following section shows an explicit implementation of sample variables in a rough set fashion.

3.1 ROUGH SET MODEL FOR DEPRIVATION INDICES - CT & DA

This section describes the use of rough sets in a spatial analysis case study which mitigates the scale dimension of MAUP. The scale issue is the result of data

characteristic changes across scales. A precondition to accommodating scale changes is to maintain data distribution from one census resolution to another. Data characteristics can be described as the inherent relationships between the data elements, which are analogous to spatial relationships in topology. Topology is central to spatial analysis functions and is considered most suited for complex spatial analysis (e.g. neighbourhood search) (Theobald 2001; Zlatanova, Abdul Rahman, and Shi 2004). Rough sets are used to facilitate appropriate grouping of multiple data elements in order to characterize attribute space for neighbourhood definition using deprivation levels in the following.

In urban poverty studies, socio-economic factors and ethno-cultural features tied to immigration include: immigrant concentrations in census units, country of origin, ability to use official language, period of arrival. Ley and Smith (1997) used these attributes to describe the extent of deprivation within a census unit. Their methodology was to map poverty levels as a chloropleth map. These socio-economic variables such as level of unemployment and education and dependency upon government transfer payments were identified as representing a significant contribution to spatial variation of urban poverty (Ley and Smith 1997). In this study, these poverty indicators are extracted from census data as deprivation indices and examined using rough sets. Deprivation, therefore, is a neighbourhood characterization process where different sub-variables categorize an area unit differently. The focus however, is not how these sub-variables are chosen but an appropriate representation of each sub-variable distribution in the final output.

There are several socio-economic variable indicators reported in the census data, but these individual data elements have little relevance and application for direct and appropriate development of relationships between deprivation index and recent immigrants. Recent immigrant information is directly available, but deprivation indicators such as standard of education, income level status are available for entire census units. So deprivation indices are not specific to recent immigrants, but to entire population within a census unit. Major deprivation indices considered are broadly

classified into categories as education, employment, housing and income levels. The individual variables which constitute these categories are outlined in Appendix B1.

The abstraction of similar deprivation variables for census units (CSD, CT, DA) has a limitation on the number of individual indices that forms one major category. For example, disparities in data at two census units: CT and DA are reflected in the constituent elements for the housing category (see Appendix B1). A diagrammatic illustration of the deprivation index derivation procedure is shown in Appendix B2 with no spatial dependency assumed. This derivation process is hierarchical and it allows independent analysis of individual variables. Intermediate results from the derivation process can be used to explain individual variable patterns, for example, to select variables that constitute major deprivation categories. An illustration of this pattern is evident in the choice between 'Incidence of low income' and 'Low income' (see Appendices B3 and B4). These are different variables assessing different aspects of income, but they have consistent patterns of data distribution and variation. The frequency distributions which are shown in charts and tables (see Appendices B3 to B6) illustrate a uniform trend between the two variables.

The deprivation variables (e.g. income, education, etc) are combined with their individual normal distribution values, which retained their unique data characteristics. All the normal distribution values are re-scaled into unit intervals from zero (0.00) to one (1.00) to allow appropriate and mathematically sound assessment of the deprivation indicators. It is worth noting that each data element is homogenized from the census data. So the re-scaling of variable intervals into unit interval scale does not change original data distribution.

A sample illustration of deprivation index values is shown in Table 3.1 where each deprivation index is considered as a unique category. The rough sets assessment of these indicators is to define each index as a unique category. For example, high deprivation index category is a rough set because distinct census units have constituent deprivation values belonging to multiple categories (e.g. education, income, etc) (see Table 3.1). That is, a census tract (say CTUID: 0187.05) has constituent indices (each of

which is treated as a rough set) belonging to multiple deprivation categories. Set criteria for the approximation of lower and upper sets are defined in Table 3.2.

Hence, it is evident that individual census tracts do not have consistent values across the various deprivation indicators, and set constituents cannot define each category because elements have multiple set memberships.

Table 3.1: Sample deprivation indicator values for set approximations

CTUID[3]	Education	Employment	Housing	Income
0187.05	0.781	0.471	0.201	0.501
0056.01	0.101	0.784	0.600	0.360
0186.01	0.451	0.120	0.520	0.602
0184.06	0.811	0.123	0.536	0.254
0191.02	0.521	0.435	0.491	0.289

Table 3.2: Set category key and criteria for set approximations

Set Categories	Unit Interval	Set Approximation Criteria
Very low deprivation indicator (VLDI)	0.00 – 0.20	Lower approximation: three to four set inclusion
Low deprivation indicator (LDI)	0.21 – 0.40	
Medium deprivation indicator (MDI)	0.41 – 0.60	Upper approximation: at least one set inclusion
High deprivation indicator (HDI)	0.61 – 0.80	
Very high deprivation indicator (VHDI)	0.81 – 1.00	Modified upper approximation: at least two set inclusion
Note: If an entity belongs to all four upper approximation sets then determine the average of its constituent values and place the entity into the upper approximation set of the set range it falls. Also apply same rule if an entity belongs to two upper approximation sets.		

In Table 3.1, the set constituents for medium deprivation category (MDI) can be defined as below:

- For Education: Census tracts, CTUID: 0186.01, 0191.02
- For Employment: Census tracts, CTUID: 0187.05, 0191.02
- For Housing: Census tracts, CTUID: 0056.01, 086.01, 0191.02, 0184.06
- For Income: Census tracts, CTUID: 0186.01, 0187.05

[3] CTUID is a unique identifier for a CT.

Applying the set approximation criteria, the lower and upper approximation sets for this deprivation category - medium deprivation indicator, MDI are:

MDI_{lower} = 0186.01, 0191.02

MDI_{upper} = 0187.05, 0191.02, 0186.01, 0184.06, 0056.01

Modified MDI_{upper} = 0187.05, 0191.02, 0186.01

In addition, an entity may belong to all four upper approximation sets, for example, CTUID: 0184.06. In an effort to reduce set memberships to a minimum for manageable assessment of entities, the last criterion stated in Table 3.2 is applied to specify entities to distinct sets. The average of the constituent values for CTUID: 0184.06 is (0.811 + 0.123 + 0.536 + 0.254) / 4 = 0.431 (see Table 3.1). Since 0.431 is within the MDI range (that is 0.41 - 0.60), the entity: CTUID = 0184.06 is therefore placed into the upper approximation set of MDI.

In sum, input variables are analyzed based on their distinct property at an instant, so entities do not necessarily assume their group characteristics. The technique consequently, recognizes the common variance shared by all the input variables and the unique variance that distinctively identifies and separates each data characteristic from another. Common variances, which exist among variables, situate these variables into upper approximation sets, while unique variance isolates specific variables into lower approximation sets. Set property is the result of individual data distribution and variation, which may be similar to other data characteristics (e.g. see Tables 3.1 and 3.2).

3.1.1 Approximation of Deprivation Index Spaces at Dissemination Areas into Census Tracts

The preceding section illustrated the use of rough sets to group disparate data characteristics irrespective of the spatial unit (that is, census unit) in question. This section employs rough sets to characterize a large census unit (e.g. a CT) using a small one (e.g. a DA). A CT is a census unit, which comprise two or more DA entities, and a

census sub-division (CSD) is a census entity, which comprises multiple CT. Deprivation index models are developed for the same group of variables at CSD, CT and DA. Distinct DA which constitute a CT may belong to different deprivation categories and approximating these component values for a particular CT requires aggregation. This grouping of constituent DA values for a CT is necessary to determine the deprivation index discrepancy due to the scale transition from DA to CT. The resulting model of this technique is two independent deprivation index models for all CT. The first model is a deprivation index model with direct variables from CT resolution, while the second is a deprivation index model of DA grouped into CT.

The approximation process for individual DA values which constitute different CT is illustrated in Appendix B7. Each set value is grouped independently, for example, all DA 'Education' constituent values for a distinct CT are aggregated. This result into a similar information table developed in Table 3.1 for which set approximation criteria can be applied. The set elements for the medium deprivation indicator, MDI in the various major deprivation categories are shown below:

- For Education: Census tracts, CTUID: 0133.02
- For Employment: Census tracts, CTUID: 0132.00, 0250.02
- For Housing: Census tracts, CTUID: 0133.02
- For Income: Census tracts, CTUID: 0133.02, 0132.00

Again, applying the set approximation criteria in Table 3.2, using MDI category the following constituent units for the lower and upper approximation sets are:

MDI_{lower} = 0133.02

MDI_{upper} = 0133.02, 0132.00, 0250.02

Modified MDI_{upper} = 0133.02, 0132.00

This illustrates the use of small census units as elementary sets for large census units, which are considered rough sets. For example, when a CT is treated as rough sets, then its elementary sets are the constituent DA.

3.1.2 Recent Immigrant Deprivation Indicator (RIDI) Deduction

The foregoing sections have demonstrated the derivation of the deprivation index irrespective of the census unit under study and with no dependency on any specific variable(s). Next, with reference to Appendix B2, this section illustrates the derivation of deprivation index for the target population group (that is, recent immigrants) to establish a spatial association. The spatial association we attempt to examine is the spatial relationship between the deprivation index (DI) and recent immigrant concentrations (RIC) within specific census units. This spatial association is denoted RIDI (that is, recent immigrant deprivation index).

Table 3.3: Recent Immigrant Deprivation Indicator estimation

Recent Immigrant Population	Census Unit Deprivation Index	Resulting Recent Immigrant Deprivation Index	Recent Immigrant Population	Census Unit Deprivation Index	Resulting Recent Immigrant Deprivation Index
Very Low Recent Immigrants 0.00 – 0.20	0.00 – 0.20	1.0	Low Recent Immigrants 0.21 – 0.40	0.00 – 0.20	1.0
	0.21 – 0.40	1.0		0.21 – 0.40	2.0
	0.41 – 0.60	1.0		0.41 – 0.60	2.0
	0.61 – 0.80	1.0		0.61 – 0.80	2.0
	0.81 – 1.00	1.0		0.81 – 1.00	2.0
Recent Immigrant Population	Census Unit Deprivation Index	Resulting Recent Immigrant Deprivation Index	Recent Immigrant Population	Census Unit Deprivation Index	Resulting Recent Immigrant Deprivation Index
Moderate Recent Immigrants 0.41 – 0.60	0.00 – 0.20	1.0	High Recent mmigrants 0.61 – 0.80	0.00 – 0.20	1.0
	0.21 – 0.40	2.0		0.21 – 0.40	2.0
	0.41 – 0.60	3.0		0.41 – 0.60	3.0
	0.61 – 0.80	3.0		0.61 – 0.80	4.0
	0.81 – 1.00	3.0		0.81 – 1.00	4.0
Recent Immigrant Population	Census Unit Deprivation Index	Resulting Recent Immigrant Deprivation Index	Key	Meaning	
Very High Recent Immigrants 0.81 – 1.00	0.00 – 0.20	1.0	1.0	Very Low Deprivation Index	
	0.21 – 0.40	2.0	2.0	Low Deprivation Index	
	0.41 – 0.60	3.0	3.0	Moderate Deprivation Index	
	0.61 – 0.80	4.0	4.0	High Deprivation Index	
	0.81 – 1.00	5.0	5.0	Very High Deprivation Index	

First, DI is estimated for all census units as shown in section 3.1. So each census unit is assigned a DI value (within 0.00 – 1.00 interval). To allow rough sets operation, RIC values for the census resolutions (that is, DA, CT and CSD) are also transformed into a unit scale (that is, 0.00 to 1.00). These RIC values are grouped into intervals, which conform to the class interval developed for the deprivation index. Few examples are: very low RIC interval (0.00 – 0.20), low RIC interval (0.21 – 0.400), medium RIC interval (0.41 – 0.60), etc (see Table 3.3 for details). The final step generates values for RIDI (also within 0.00 – 1.00 interval) in each census unit. The recent immigrant deprived census units are those census entities for which recent immigrant values (that is, RIC) match a designated deprivation index (that is, DI). For instance, low (that is, 0.21 – 0.40) RIC value with very low, low, medium, high and very high DI values is assigned very low, low, low, low and low RIDI values respectively. So, to assign a RIDI to a census unit, we examine its RIC and DI values. The RIDI value is the RIC value or less because RIDI cannot be assigned to census unit with DI value which does not match RIC and vice versa. The set rules for computing RIDI values are illustrated in Table 3.3.

This spatial association (that is, RIDI) is characterized with model magnitude and strength. The model magnitude is computed directly by examining the RIDI values. But the strength of the spatial association is analogous to the correlation coefficient, which is deduced as the ratio of the number of match-value-categories to the total number of census units.

$$Correlatio\ n\ coefficien\ t,\ r = \frac{number\ \ of\ \ matched\ \ -\ value\ -\ categories}{total\ \ number\ \ of\ \ census\ \ \ units}$$

The number of matched-value-categories is the number of census units which have RI and DI values belonging to the same group, say, high RI and high DI are classified into high RIDI set. The computation of the correlation coefficient is illustrated in section 3.6 for the three census units (that is, DA, CT and CSD).

3.2 INCLUSION OF CENSUS SUBDIVISION DATA

The inclusion of a third census resolution has become necessary to validate the same information at two different spatial resolutions with a different census resolution data. The immediate census resolution after census tract is census subdivision (CSD). All the data extracted for CT and DA are also collected for CSD in the Greater Vancouver Regional District (GVRD). Similar analytical procedures are also taken to derive values into a unit scale from zero (0.0) to unity (1.0). Census subdivision unit characteristics are approximated from CT and DA by calculating both the average and the median of each constituent CSD entity. Hence, all CSD have their calculated deprivation indices coupled with two approximated values each of average and median values from CT and DA. This permits the investigation of deprivation model discrepancy for approximating CSD from CT and DA. Resulting discrepancy can be applied to estimated values as net error to be incurred in scale transition between similar census data resolutions.

3.3 CENSUS DATA RESULTS AND DISCUSSION

The choice of tools and methods applied to spatial data in this study is focused on retaining individual data characteristics in resulting outputs. For instance, sub-variables (e.g. income, education, etc) used to derive deprivation index are made to assume their unique characteristics rather than their generalized pattern. Spatial data characteristic preservation has both analytical and practical importance because geographic models should not obscure the data from which they are developed. Data property manipulation during analytical implementation can result in geographic model failure such as inadequate resource distribution or inaccurate targeting of consumer groups. Data distribution retention during analytical process is paramount, particularly in spatial data manipulation where each data element is characteristic of a spatial location. Also the representation of each data characteristic in outputs becomes crucial when we identify extreme data elements. Extreme data elements identified

based on any attribute must be reconciled with its spatial location to fully classify such data entity as an outlier.

Figure 3.1: Analytical tools (Spatial Regression and Rough sets) perspective of census data for the city of Burnaby at census tract resolution

This is important because statistical values identified as outliers may be different in a spatial framework since statistical calculation may neglect the spatiality of the data. So, the first output, Figure 3.1, attempt to illustrate the preservation of data characteristic during spatial data analysis using rough sets. Spatial data distribution and variability underlies the accuracy of many analytical processes because relative accuracy between multiple variables is reliant on individual data characteristics. In the rough sets process, the steps for deriving recent immigrant deprivation index (RIDI) census units are based on designated recent immigrant concentration (RIC) values, which match similar deprivation index (DI) values. It is apparent that RIDI cannot be

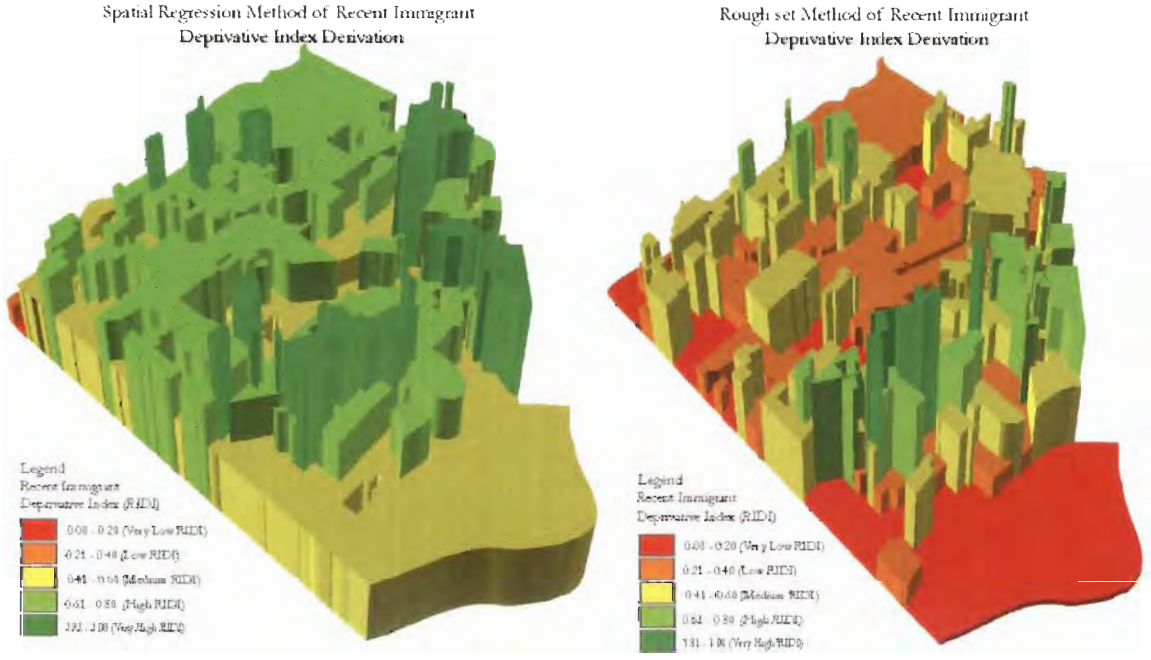
assigned to census units with DI, which does not match RIC, and vice-versa. That is, a census unit of low (0.21 - 0.40) DI category and corresponding very low (0.00 - 0.20) recent immigrant concentration (RIC) cannot be designated a moderate (that is, 0.41 - 0.60) RIDI category. This is the erroneous pattern observed from the regression output; due to the initial assumption that DI is dependent on RIC. All census units with high RIC are assigned high RIDI irrespective of the original DI in the census entities. So, certain census tracts are designated higher RIDI than their initial DI using the regression method. But, the rough sets approach has indicated a strong parallel between the two input variables (that is, RIC and DI): retaining and effectively integrating two data distribution and their spatial variation.

The rough sets method is comparable to a regression technique for which all the assumptions of dependency, independency of observations, linearity, no missing values and outliers are exhibited by the data under consideration. Fundamentally, the rough sets technique will yield similar outputs to a regression method for exactly two variables, which exhibit the above regression assumptions. But conflicting results arise when multiple variables are combined hierarchically such that each variable exhibits inconsistent data distribution towards the dependent variable. An explicit example is shown in the regression output where RIC obscures the DI distribution (see Figure 3.1). This data characteristic pattern is beyond the control of many analytical tools. So values are constrained to assume general data pattern in order to determine results for points that do not comply with the overall data distribution pattern. This results in loss of unique data characteristics and their subsequent loss of representation in the output model.

Also the rough sets categorization can be considered analogous to data clustering techniques whose classification is dependent on the global measures rather than individual data characteristic used in rough sets. Global patterns are valuable quantitative measures, which give accurate descriptive information for uniformly distributed data. But, for heterogeneous data where extreme variables must be treated equally and not allowed to assume the general data pattern, global measures

inadequately describe the data distribution. The problem becomes compounded for multiple variables for which extreme variables in one data must be considered in relation to other variables. While the approximation of data distribution and variation is inescapable during analytical processes, the rough sets method represents all input data characteristics at optimum thresholds. This guarantees consistency between the spatial model and data from which it is generated. This is what has been achieved in this first output (see Figure 3.1) and this characteristic underlies subsequent outputs. Hence, the rough sets concept assumes no dependency in multivariate analysis, indicating that one attribute clustering does not necessitate clustering in geographic space or another attribute characteristic.

Figure 3.2: Sample Recent Immigrant Deprivation Index estimation using Spatial Regression & Rough sets



The disparity in the model development using spatial regression and rough sets approach is reinforced in Figure 3.2 showing significant differences between the two models. Figure 3.2 is a rough sets and regression output for recent immigrant deprivation index for the first output shown in Figure 3.1. The green colour indicates more deprived areas with exaggerated z (height) values while the red colour shows less

deprived neighbourhoods with relatively small z values. There is significant difference between the two outputs; difference in both distribution and quantity of neighbourhoods assigned into a particular RIDI category. The rough sets output shows a more diverse distribution (high variance) and small number of census units with high RIDI. This pattern reflects the data characteristics because RIDI value of each census unit is comparable to the input variables. The regression output shows the contrary, small variability for RIDI values and large number of census units with high RIDI values. The foregoing has illustrated the preservation of multiple data characteristics in derived models at specified census unit. The scale issue however, is the problem of translating data characteristics across different scales.

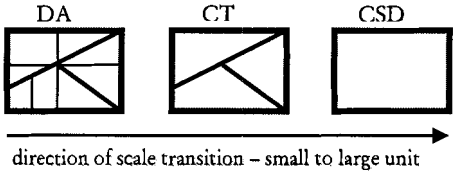
3.4 SCALE TRANSITION & ASSESSMENT FOR CSD FROM CT & DA

The preceding section showed the utility of rough sets in accommodating unique data characteristics during spatial analysis. The rough sets outputs were compared to the results from the regression method, which does not retain single or multiple data distribution at specific spatial unit. This section examines the scale transition over multiple census units (CDS, CT and DA) from small to large units in order to estimate scale translation parameters. This is a necessary step in order to reverse the scale transition from large units to small ones.

The scale transition for census sub-division, CSD involves the approximation of deprivation index constituents of CT and DA for a particular CSD. Due to different characteristics of each census unit both the average and the median values of these component values are determined to unearth any distinct pattern. In rough sets context, a CSD is a rough set while both the DA and CT within a particular CSD are elementary sets. The scale assessment entails determining estimated values of one census resolution using another, say approximate CSD using DA. Figure 3.3 illustrates the direction of the scale transition and shows which census units to designate as rough sets and elementary sets for a specified scale transition. For scale transition from DA to CT, for example, constituent DA that are subsets of a particular CT are designated

elementary sets, while the specified CT becomes a rough set. This estimation is dependent on the appropriate pattern, which emerge from these three census resolutions.

Figure 3.3: A CSD represented at different spatial resolutions



The first scale transition output (see Table 3.4) is that estimated for CSD using both DA and CT separately. The sample deprivation index values for CSD shown in Table 3.4 shows that approximated values from CT and DA are generally less than the calculated values using CSD resolution data. While the average estimated values from CT indicated a smaller discrepancy, corresponding average values from DA indicated the opposite. The overall spread and variation however, shown by the standard deviation and the data interval for CT are higher than those of DA. Despite the small total deviation and mean deviation values of CT, the descriptive measures (e.g. sum, maximum, mean, etc) indicated for these two census resolutions; CT and DA have demonstrated more uniform data distribution for DA than for CT. Increased accuracy indicators for DA are also apparent in the difference measures (e.g. RMSE, MAE, d, etc) with a smaller error and high index of agreement than for CT, indicating a more coherent outcome.

Secondly, scale transition from DA to CT is also examined because it is a spatial transition from small to large unit. The DA approximation to CT (see Table 3.5) has indicated a comparable deviation of the estimated values from the known ones. The DA approximation suggests an increased accuracy irrespective of the increased number of census units involved (that is, CT is 424 units and DA is 3369 units). Increased accuracy does not guarantee comparable pattern in the approximation processes, but the results shown demonstrate coherent pattern with respect to size of the census resolution. That is, decrease in size of census unit results in increased accuracy of

estimating its data distribution and variation in order to approximate constituent census entities.

Table 3.4: Sample deprivation index approximation for CSD from CT and DA

Census Subdivision	CSDUID[4]	NDI Value	NDI Value–CT	NDI Value–DA	Difference of NDI–CT	Difference of NDI-DA
Langley	5915001	0.5091	0.4431	0.4467	0.0660	0.0625
Langley	5915002	0.6777	0.6170	0.5634	0.0607	0.1142
Surrey	5915004	0.6426	0.5677	0.5290	0.0749	0.1136
White Rock	5915007	0.4707	0.4158	0.4060	0.0549	0.0647
Delta	5915011	0.5094	0.4365	0.4468	0.0729	0.0627
Richmond	5915015	0.5889	0.4920	0.4565	0.0969	0.1324
Greater Vancouver A	5915020	0.4059	0.2795	0.3299	0.1264	0.0761
Vancouver	5915022	0.5635	0.5108	0.4720	0.0527	0.0914
Burnaby	5915025	0.6058	0.5252	0.4742	0.0806	0.1316
New Westminster	5915029	0.6357	0.6008	0.5297	0.0349	0.1060
Coquitlam	5915034	0.5698	0.4781	0.4757	0.0917	0.0941
Port Coquitlam	5915039	0.3555	0.5376	0.3342	-0.1822	0.0212
Port Moody	5915043	0.3555	0.4679	0.3342	-0.1124	0.0212
North Vancouver	5915046	0.2388	0.2782	0.2157	-0.0394	0.0231
North Vancouver	5915051	0.2388	0.4806	0.2157	-0.2418	0.0231
West Vancouver	5915055	0.5814	0.1435	0.5063	0.4379	0.0751
Bowen Island	5915062	0.6785	0.2406	0.6088	0.4379	0.0697
Pitt Meadows	5915070	0.3467	0.5283	0.3514	-0.1816	-0.0047
Maple Ridge	5915075	0.2570	0.5311	0.2133	-0.2741	0.0436
Musqueam 2	5915803	0.7403	0.1735	0.6070	0.5668	0.1333
				Sum	1.2237	1.4549
Key	Meaning			Minimum	-0.2741	-0.0047
NDI	Net Deprivation Index			Maximum	0.5668	0.1333
CT	Census Tracts			Mean	0.0612	0.0727
DA	Dissemination Areas			Std. Deviation	0.2189	0.0419

CT Difference Measures	Value	DA Difference Measures	Value
Root Mean Square Error (RMSE)	0.2219	Root Mean Square Error (RMSE)	0.0834
Mean Absolute Error (MAE)	0.1643	Mean Absolute Error (MAE)	0.0732
Index of agreement (d)	0.9461	Index of agreement (d)	0.9925

[4] CSDUID is a unique identifier of a CSD

This reduction in size of census unit, for example, from CSD to CT has indicated a consistent data characteristic resulting in the increased accuracy of estimating CT deprivation index using DA.

Table 3.5: Descriptive and difference measures of DA approximation to CT

Descriptive Measures	CT	DA	Residuals	Difference Measures
Sum of net deprivation index	207.1302	195.1527	11.9775	Root Mean Square Error (RMSE) = 0.0056
Minimum	0.0884	0.1692	-0.1478	
Maximum	0.8603	0.713	0.2315	Mean Absolute Error (MAE) = 0.0605
Mean	0.4885	0.4603	0.0282	
Standard deviation	0.1771	0.1173	0.0696	Index of agreement (d) = 0.9943
Number of census units	424	3369	424	

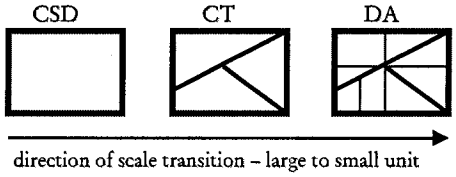
The rise in model accuracy is also reflected in the difference measures of the DA approximation to CT, resulting in a decline in both root mean square and mean absolute errors of 0.0778 and 0.0127 respectively. There is also a surge in the index of agreement of 0.0018; all from previous DA approximation to CSD compared to DA estimation of CT values (see Tables 3.4 and 3.5). Hence, the three scale transition examined from small to large spatial unit suggests that size and number of census units used in the transition process influence data characteristics across scales. The persistent problem however, during scale transition is not the transition from small to large units but the converse of this transition; the derivation of data characteristics for small area units using large units.

3.5 SCALE TRANSTION FROM LARGE TO SMALL CENSUS UNITS

Census data have been approximated for large areas using small census units in the above section. The scale transition results and their accuracy assessment values in the preceding section are used here. This will help to reverse the census data approximation of large units using small entities in order to determine small entities using large census units. The importance of this approximation is an integral concern for many grouped data applications where decisions are made for small areas from large area data. Small area data estimation is desirable, for example, optimum

distribution of goods and services, site characterization for retail location, etc. Small area data approximation, however, is difficult for many analytical tools because data distribution and variation must be controlled at discrete level rather than using general pattern to describe distinct data characteristics.

Figure 3.4: A CSD represented at different spatial resolutions



In addition, it involves data pattern estimation, which may not only be nonlinear but also heterogeneous. Figure 3.4 shows the direction of the scale transition process from large units to small entities. This estimates scale transition across scales at CSD, CT and DA, and evaluates the estimated values against known values in the following.

The mean bias error (MBE) and the mean absolute error (MAE) both measure the effective error due to rough sets approximation. The MBE assesses the variance of the residual distribution, while the MAE measures the weighted average of the absolute errors. The application of both the MBE and MAE are to correct for the error due to the approximation. The error incurred however, is characterized by both quantity and its distribution. The error distribution is measured using the standard deviation of the census residuals. So, the application of MBE/MAE and the standard deviation (STD) of the census residuals to any census unit should describe the predicted census distribution and variation. Specific to the census resolutions considered; three census approximations were examined; estimation of CT using CSD and determining DA using both CT and CSD. In the determination of CT, the formula below is applied to component CSD as:

$$CT = CSD_{value} + MBE \pm STD$$

A CSD consists of multiple CT and census error distributions are assumed as related to the number of its constituent census units. This assumption is apparent in

section 3.1.1 where there are disparities among constituent census entities, which are contained in a larger census unit (e.g. DA within CT). The rationale behind this assumption is to distribute computed values of constituent CT of a CSD using the number of CT contained in a CSD. This is because it is unlikely, at least in practice, for all CT values to exhibit no variation and distribution. A CSD exhibits variable smaller units say CT, hence, for a CSD with n number of CT then:

$$CT_i = CSD_{value} + MBE \pm STD \times i/n$$

where i starts from 1 to n and MBE and STD are from CT to CSD approximation.

The number of constituent census units in a larger census resolution can be even or odd, affecting the error distribution for a particular approximation. For even number of census units, the standard deviation, STD is applied as positive to half the number of census units and to the others as negative. In cases where there are odd number of census units, the middle census unit is applied zero STD and the same even number of census unit procedure is repeated for the rest of the census entities. Below are excerpts for two numbers of census units; four (4) and five (5), which are even and odd respectively.

Even number of census units e.g. four (4):

$$CT_1 = CSD_{value} + MBE + STD \times 1/1$$

$$CT_2 = CSD_{value} + MBE + STD \times 1/2$$

$$CT_3 = CSD_{value} + MBE - STD \times 1/1$$

$$CT_4 = CSD_{value} + MBE - STD \times 1/2$$

Odd number of census units e.g. five (5):

$$CT_1 = CSD_{value} + MBE + STD \times 1/1$$

$$CT_2 = CSD_{value} + MBE + STD \times 1/2$$

$$CT_3 = CSD_{value} + MBE$$

$$CT_4 = CSD_{value} + MBE - STD \times 1/1$$

$$CT_5 = CSD_{value} + MBE - STD \times 1/2$$

Similar census approximation is repeated for DA using both CSD and CT separately for corresponding difference and descriptive measures. This summarizes the approximation of small census units from large ones and subsequent redistribution of estimated values for constituent census units. For instance, DA values are estimated using a CT, and these estimated values are redistributed to the constituent DA in specified CT. The succeeding section discusses the scale transition results for small census units derived from large units.

3.5.1 Census Estimation Results Summary

The census approximation results discussed here are in threefold: CT results using CSD, and DA results using both CT and CSD separately. First, a total of 39 CSD are used to estimate deprivation index values for 424 CT by reversing the error quantity and redistributing the error variation for the approximation of CSD from CT. The CT outputs from CSD shown in Table 3.6 have produced results that are consistent with predicted values. The descriptive statistics for the known and predicted values have indicated minimum deviations, particularly, the minimum, maximum and the standard deviation of the deprivation index. The predicted model has optimum approximation to an absolute error margin of 0.0011 (approx. 0.1%) and to a worst absolute error of 0.6512 (approx. 65%).

Table 3.6: CT Deprivation Index estimation using CSD – Results summary

Descriptive statistics	Deprivation Index Estimates		Residuals	Difference Measures
	CT	CSD	CT – CSD	CT
Sum	207.1302	265.7081	-58.5779	MAE = 0.1790
Minimum	0.0884	0.0811	0.0011	
Maximum	0.8603	0.9578	0.6512	RMSE = 0.0494
Average	0.4885	0.6267	-0.1382	
Standard deviation	0.1771	0.1067	0.1743	d = 0.9557
No. of census units	424	39	424	

Despite this large error margin, the MAE value is as small as 0.1790, which is the average absolute error with error distribution of ± 0.1743. This results in extreme

error quantity and distribution to the interval: -0.1382 ± 0.1743 (that is between 0.0415 and -0.3071 using MBE). The model degree of agreement also has high accuracy value of 0.9557 (approx. 96%). These values indicate high accuracy approximation process for deprivation index values at the CT level. It is partly dependent on the earlier approximation of CSD from CT from which the error quantity and its distribution were computed. It is worth noting that the predicted values have less distribution (0.1067) than the known (0.1771) demonstrating that the smaller the census unit the more diverse its data characteristics. The census data variation caused by this aspect of the scale transition cannot possibly be described from a larger census resolution data. This data characteristic pattern is also apparent in subsequent results for DA discussed in the following.

The deprivation index estimation at the DA level using both CT and CSD uses 424 and 39 census units respectively to approximate 3369 DA. The DA results shown in Table 3.7 also have high accuracy values and are consistent with earlier CT results. The descriptive measures such as the minimum, maximum, mean and the standard deviation values have minimum deviations from the known values. The minimum and maximum absolute errors from CT are 0.0000 and 0.6104, and for CSD they are 0.0000 and 0.5946 respectively. This is remarkable because in estimating deprivation index for some DA there will be zero error (that is, absolutely no error). Also the average absolute error for both CT and CSD is 0.1020 and 0.1758 respectively. The approximation process has shown an increased accuracy level demonstrated from the descriptive and the difference measures. This increased accuracy is partly due to the reverse approximation involving DA, CT and CSD to determine the error quantity and distribution values used in this current estimation.

The extreme error magnitude and distribution from the CT approximation is described by the MBE (that is, -0.0642) and the standard deviation (that is, ±0.1138) of the residuals. This results in an interval of -0.0642 ± 0.1138 (that is, from 0.0496 to -0.1780) within which errors will be distributed. Likewise, for approximation values from CSD, the error magnitude and distribution interval is -0.1642 ± 0.1328 (that is,

from -0.0314 to -0.2990). A more descriptive model assessment is the degree of agreement value, indicating 0.9828 and 0.9594 for CT and CSD respectively. The increased accuracy illustrated by estimated values from CT can be described in relation to the similarity of census data distribution characterized by its size. That is, CT data distribution is more similar to DA data distribution because their sizes are also comparable.

Table 3.7: DA Deprivation Index estimation using CT & CSD – Results summary

Descriptive statistics	Deprivation Index Estimates			Residuals		Difference Measures	
	DA	CT	CSD	DA -CT	DA - CSD	CT	CSD
Sum	1796.06	2132.83	1579.69	-216.364	-553.1362	MAE = 0.102	MAE = 0.176
Minimum	0.047	0.073	0.058	0.0000	0.0000		
Maximum	0.958	0.855	0.839	0.6104	0.5946	RMSE = 0.017	RMSE = 0.045
Average	0.533	0.633	0.469	-0.0642	-0.1642		
Standard deviation	0.176	0.086	0.142	0.1138	0.1328	d = 0.9828	d = 0.9594
No. of census units	3369	424	39				

Also the predicted deprivation index values are always less distributed than the known values. The scale transition from small to large resolution exhibits data distribution, which is highly comparable to known data because smaller resolution data have richer data characteristic, which competes with what may be observed. The converse is not true because patterns derived from data cannot describe explicitly distinct data distribution and variation. This is the challenge, which geospatial model development tools encounter with regards to scale transition because it is more than replicating data distribution and variation for different spatial resolutions. Observing scale transition pattern from multiple resolutions can approximate this data characteristic pattern.

This section concludes the scale transition investigation across multiple census resolutions while maintaining data characteristics during the estimation process. Accuracy indices for the estimation are also illustrated to validate the scale transition

process. The error magnitudes defined by MBE/MAE and standard deviation provides a scale sensitivity measure for the scale transition. Following, we examine spatial associations derived from multiple variables across different census scales.

3.6 RECENT IMMIGRANT AND DEPRIVATION INDEX RELATIONSHIP

We have examined the effect of varying scales on data characteristics. It is shown that the size of census unit is a major defining characteristic of data distribution. Census units with relatively smaller sizes (e.g. DA) exhibit more diverse data distribution. So, DA data are most distributed compared to census data at CT and CSD levels. But, what is the effect of data characteristics across different scales on spatial relationships derived at these varying scales? In other words, how does the size (scale) of a census unit and its corresponding data characteristics affect the derived spatial relationships?

The immediate concern for which spatial resolution of census data poses a challenge in geographic model development is the accuracy of spatial relationships and associations generated from these data to inform social and physical policies. Spatial relationships developed at different census units are often different. The understanding of accuracy limitation introduced by data resolution on derived spatial models is crucial to utilize these results as generalized patterns rather than distinct occurrences in the real world. So, how do policy makers incorporate spatial resolution standards into their decisions? Are there any accuracy indicators that inform them of the spatial thresholds, which limit their decisions?

The essence of spatial resolution underlines the purpose of social and physical applications of census data to distribute resources and services to where they are needed. The derivation of these spatial relationships is characteristic of the analytical tool from which they are developed. So, integration of data characteristics into spatial models generated from the data is the first accuracy index achieved (see section 3.3) which enforces harmony between data and the spatial model. It is also indicated that

the data and model conflict may degenerate into erroneous conclusions from derived spatial associations between variables. The spatial relationship we examine here is the spatial association between recent immigrants and deprivation index. So, the relationship between recent immigrant and deprivation indices are developed from outputs, which are consistent with the data from which they are developed, and at various spatial resolutions whose patterns have been observed carefully.

Figure 3.5: Deprivation Index for the entire census unit population – CT

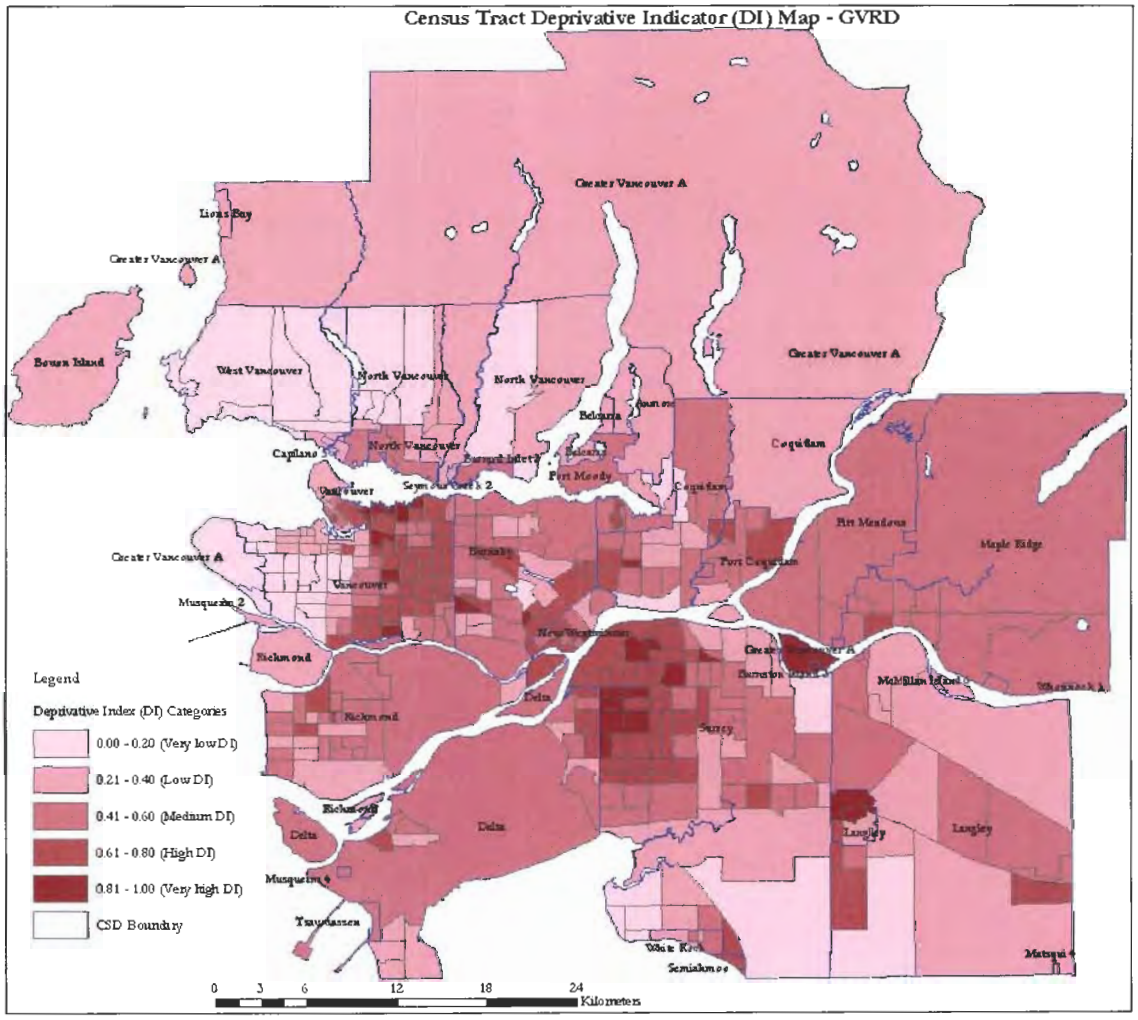


Figure 3.6: Recent Immigrant concentrations – CT

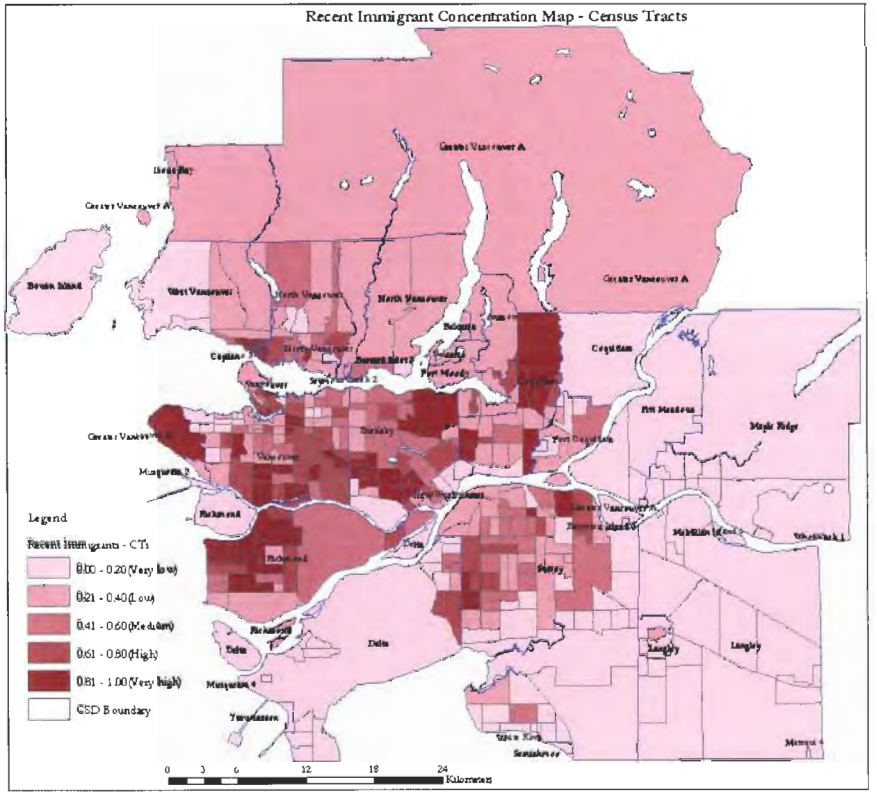
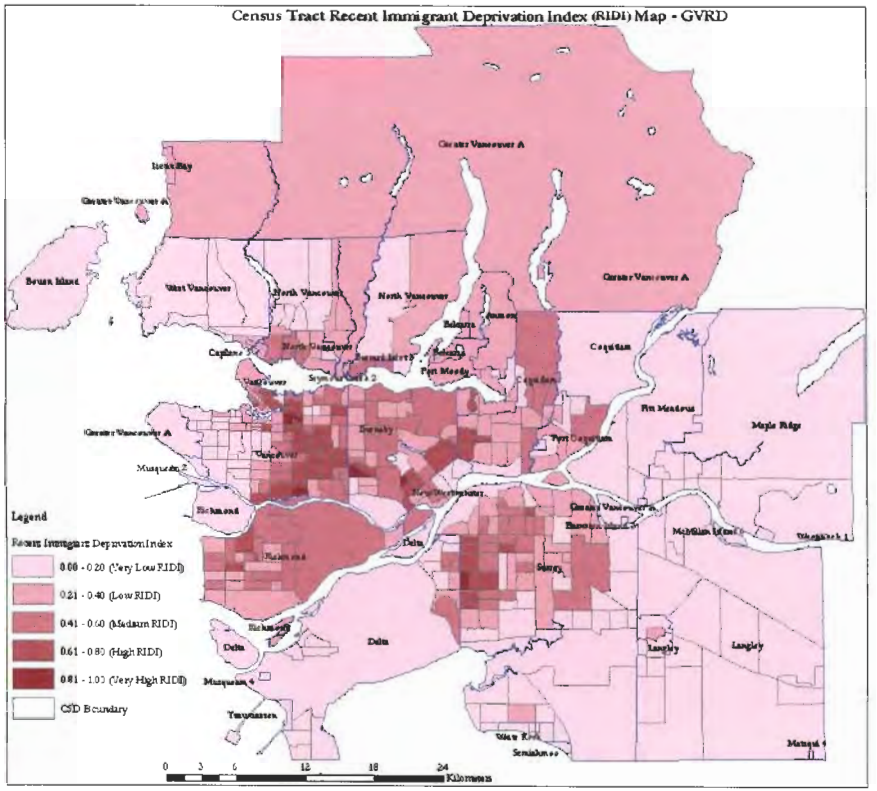


Figure 3.7: Recent Immigrant Deprivation index – CT



Spatial relationships are developed at specific census resolutions; CSD, CT and DA. The problem becomes: are there indices that quantify and translate values from one resolution to another? While the relative generalized pattern observed for the three census units are consistent suggesting unbiased data reporting for the census resolutions, the individual data pattern are incoherent. For example, recent immigrant concentration in one census unit, say, CT can be traced to a similar proportion of recent immigrants at DA level. But, the variance among specific census units (e.g. DA, CT) and nonlinearity of the multiple variables are non-transferable between different census units. This individual data disparity discourages the derivation of consistent spatial relationships among multiple variables. Figures 3.5 and 3.6 illustrate the individual data disparity, which may exist among multiple variables, and how the spatial model generated in Figure 3.7 harmonizes the two different data characteristics.

As shown in Figure 3.5, the CT indicates the overall deprivation index due to the entire population in respective CT units with no dependency on any distinct population group. Next, the target population group, recent immigrants are shown in Figure 3.6. The resulting model is a deprivation index due to recent immigrants indicated in Figure 3.7 at census tract level. This model characteristic has no data distribution conflict but reflects optimum integration of the two data distributions in Figures 3.5 and 3.6. The model demonstrates the strength and magnitude of the spatial relationship between recent immigrants and deprivation index at CT level. The measure of this spatial association between recent immigrants (RI) and the deprivation index (DI) that is analogous to the correlation coefficient (r) is determined as ratio of the number of matched-value-categories to the total number of census units.

$$\text{Correlation coefficient, } r = \frac{\text{number of matched - value - categories}}{\text{total number of census units}}$$

The number of matched-value-categories is the number of census units which have RI and DI values belonging to the same group interval, say; high RI and high DI are classified into high RIDI (recent immigrant deprivation index) set. The correlation coefficient at CSD, CT and DA levels are 0.600, 0.549 and 0.468 respectively. The CSD

correlation coefficient indicates that 60% of the total deprivation index is linked to recent immigrants while it is approximately 55% at CT level. The coefficient of correlation at the DA level has declined significantly illustrating the scale transition effect even for data manipulation in which data characteristics are retained from one census resolution to another. The variation in the correlation coefficient value of the census resolutions shows the inaccurate conclusions, which could be reached, based on any specific census data resolution. The steady decline in the correlation coefficient value also demonstrates that census household or individual level data will have a smaller correlation index. It is worth noting that the CSD and the CT models should be used for reconnaissance study in order to identify potential spatial locations to be further examined. While the DA model cannot be used to uniquely describe reality, the individual level model is not far from it. So, it is essential to investigate the scale transition over multiple census resolutions to approximate what could describe the household level model.

From the foregoing, it is necessary to investigate further at smaller extent the recent immigrant and deprivation index relationship. From patterns observed at CSD and CT levels, the following CSD units: Burnaby, Vancouver and North Surrey have shown high RIDI correlations and their compact DA settings will be suitable for manipulation. Hence, one of these CSD units: Burnaby is examined further with no characteristic preference.

3.6.1 Household Census Data Approximation using Large Resolution Data

The results for the census resolutions involving CSD, CT and DA have significant patterns that describe the estimated value characteristics with respect to the size of census unit. In other words, data characteristic is a function of size of census unit. Table 3.8 shows the summary pattern defined by various descriptive and difference measures.

Table 3.8: Summary of descriptive patterns observed from multiple census resolutions data

Small To Large Census Resolution (e.g. DA to CSD)	Large To Small Census Resolution (e.g. CSD to DA)
Predicted values defined by the mean of the census residual (MBE) are smaller	Predicted values defined by the mean of the census residual (MBE) are higher
Predicted values defined by the absolute mean of the census residual (MAE) are smaller	Predicted values defined by the absolute mean of the census residual (MAE) are higher
The error distribution characterized by the standard deviation of the residuals are smaller	The error distribution characterized by the standard deviation of the residuals are higher
Estimated deprivation values indicate a richer and more varied distribution than the known values	Approximated deprivation values are less distributed than known values suggesting a less varied prediction
Model index of agreement are generally higher	Model index of agreement are smaller
The strength of derived spatial associations and relationships declined significantly with decrease in size of census resolution	

The essence of these patterns is to describe a generalized model for derived spatial relationship between recent immigrants and deprivation index at a smaller scale (e.g. household or individual level). Zhang and Goodchild (2002) observed that error distribution cannot be different from the data characteristics from which these errors are generated. The census data at CSD, CT and DA resolutions have shown this property of each error distribution representing their respective data characteristics. For instance, error distribution for CSD data are less varied while residuals from DA data showed a more diverse distribution. These error distributions are characteristic of their respective data traits. The deviation of an estimated value from its known value however, is characterized by the error magnitude and its distribution. The patterns of census resolutions and their error characteristics due to the scale transition, and the approximation process describe the error magnitude and its distribution for the census household model derivation from larger resolution data.

From the descriptive and the difference measures indicated for the census approximation residuals and the deviation patterns observed, the census household (CH) level data derivation from CSD, CT and DA are shown as follows:

CSD census household data approximation:

$$CH = CSD_{value} - MBE \pm STD$$

$$CH = CSD_{value} - 0.0915 \pm 0.1328$$

CT census household data approximation:

$$CH = CT_{value} - MBE \pm STD$$

$$CH = CT_{value} - 0.0360 \pm 0.1138$$

DA census household data approximation:

$$CH = DA_{value} - MBE \pm STD$$

$$CH = DA_{value} - 0.0282 \pm 0.0696$$

These approximations follow the same estimation procedure used in section 3.4 separately for odd and even numbers of constituent census units. The approximations have shown a declining trend in both the error magnitude and its distribution from CSD through CT to DA. These decreasing values (that is, MBE and STD values) show the increased approximation accuracy from using small resolution census data (e.g. DA) and its accompanying diverse distribution.

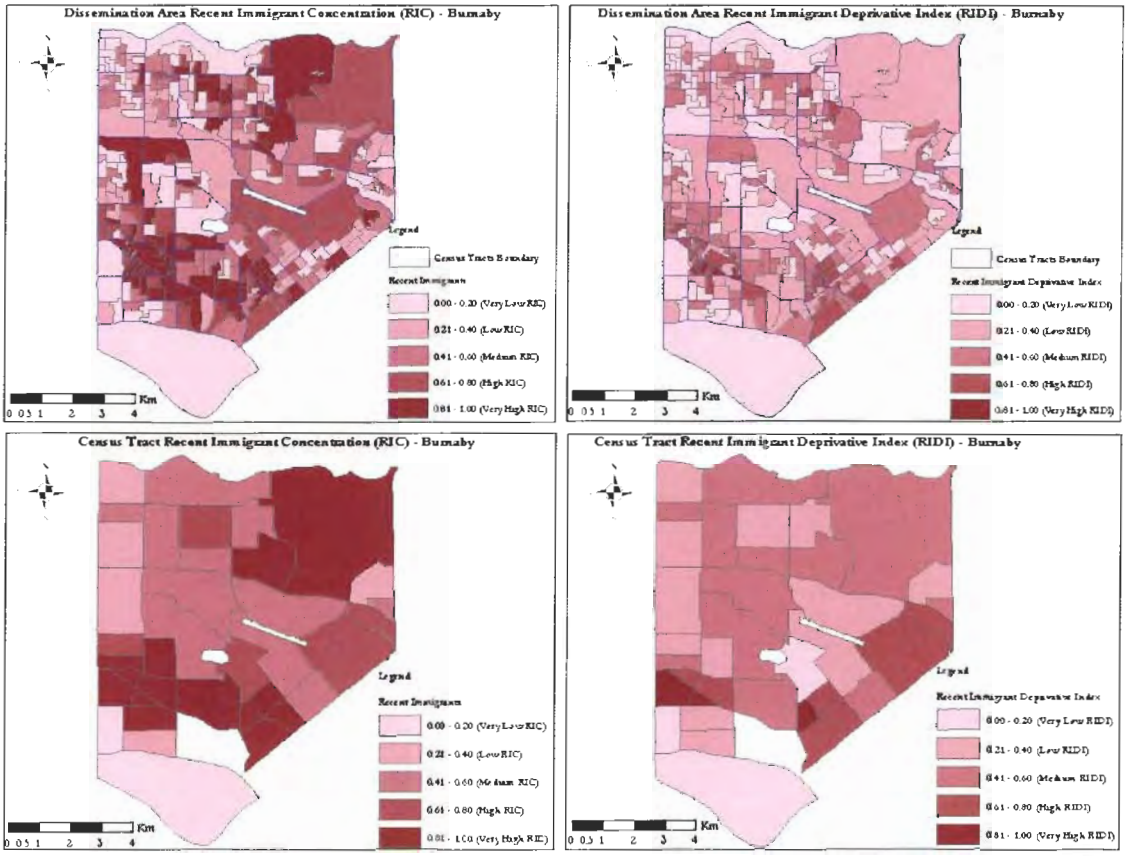
The validation of these estimates is essential to evaluate the reliability of the method used in the census approximation process. Census data elements reported are sampled from a particular area, ensuring that samples statistically represent the study area under consideration (Schuurman 2002). For other spatial analysis methods, the number of census households used in the sampling process will affect the accuracy of the approximation. The rough sets census approximation adopted however, is independent of the number of constituent units used. This is because irrespective of the number of census entities used, the error and its distribution values ensure that the average value is retained. So, while the individual census values remained unchanged, the error magnitude and its variation are the key descriptors for the approximation process.

3.6.2 RIDI Relationship for Selected CSD - Burnaby

The derivation of spatial relationships from different census resolutions has been described for the entire Greater Vancouver Regional District (GVRD) at CSD and CT levels. A more detailed consideration of the effects of the scale transition and the size of census resolution on derived spatial associations is addressed in relation to this selected CSD unit: Burnaby. Figure 3.8 shows recent immigrant concentrations and their derived spatial relationship with deprivation index at DA and CT levels. The correlation coefficients for derived recent immigrant and deprivation index relationship at CT and DA levels are 0.5405 (that is, 20/37) and 0.5590 (that is, 180/322) respectively. The proximity of these two values is remarkable because this shows that multiple census resolutions can result in very close and similar outputs of derived spatial relationships from areas characterized with homogeneity and randomness.

Also the closeness of the two correlation coefficients indicates the versatility and efficacy of the rough sets process in retaining data characteristics. Appropriate treatment of conflict within input data variables and convenient representation of multiple data distributions were recognized as key merits in rough sets analysis. While the rough sets tool does not improve nor create new data, its worth is in reasoning with data to develop models, which replicate single or multiple data characteristics. From the derived spatial relationship, it is evident that regarding the choice of census resolution for particular policy implementation, there is a limiting threshold within which derived models from these multiple census resolution could yield very close results. This approach can be employed in census resolution specification in data collection and model development for certain applications in order to specify size for optimal census resolution. This also reduces the uncertainty of selecting census scales to describe deprivation levels. This is because if small resolution data are expensive to collect and manage then it becomes inefficient to develop spatial models from small resolution data if the same model can be generated from large census resolution with comparable accuracy.

Figure 3.8: Recent Immigrant Concentration and their corresponding Deprivation Index at DA & CT



It is worth noting however, that while there is no alternative to indicating spatial specification for areas smaller than the census resolution used in model development: they can give indications of likely spatial locations within which certain occurrences are possible. This pattern is apparent in Figure 3.8 where certain deprivation occurrences at DA level can be traced from a larger area at CT level.

CHAPTER 4:
SECOND CASE STUDY - BOREHOLE DATA METHODS, RESULTS & DISCUSSIONS

The preceding chapter discussed census data methods and results, and marked the end of the first case study. This chapter describes borehole data methods and outputs. The chapter constitutes the second case study. The second case study uses rough sets and transition probability to minimize the effects of uncertainty due to erroneous sediment identification and description problems. The technique is to identify hydrogeologic properties, specifically, hydrostratigraphic units of subsurface materials. These are used to approximate geological spaces with the aim of classifying borehole units for conceptual model development using aquifer-supporting criteria.

The definition of hydrostratigraphic units from geological information, for example, well-log data are paramount for many aquifer flow investigations. Hydrostratigraphic units comprise geologic units of similar hydrogeologic characteristics (Anderson and Woessner 1992). The development of a conceptual model for groundwater flow system, for example, requires accurate hydrostratigraphic description. Such hydrogeologic unit information underlies the overall performance of derived numerical models. The modeling of regional flow systems, aquifers and confining beds are suitably described, for example, using the concept of hydrostratigraphic unit (Anderson and Woessner 1992). The reconstruction of material deposition is also reliant on stratigraphic information to unfold depositional history. Understanding of depositional account in a study region can be helpful in discovering the occurrence of sediment types when geologic information is sparse (Anderson and Woessner 1992).

Generally, hydrostratigraphic information is developed from detailed site-specific information on stratigraphy and hydraulic conductivity (Anderson and Woessner 1992; Dolgoff 1996). While hydrostratigraphic information may be most

suitable for regional simulation of geologic systems, at small scales stratigraphic and hydrogeologic information are required. Site-specific information becomes necessary because facies model which are idealized representations of environments of deposition do not represent the characteristics of any one site (Anderson and Woessner 1992). A facies is a unit of material with similar physical characteristics that are deposited in the same geological setting (Dolgoff 1996; Anderson and Woessner 1992). Metamorphic facies, for example, are formed by material assemblage under the same set of temperature-pressure conditions regardless of their original compositions (Dolgoff 1996). Thus, facies models describe the expected distribution of predicted geologic units (Anderson and Woessner 1992) and such information can be used to define hydrostratigraphic units.

4.1 HYDROLOGIC CHARACTERISTICS OF SUBSURFACE MATERIALS

To investigate the occurrence of groundwater necessitates a clear understanding of the geological settings that support its existence, distribution and movement. Subsurface environments are not homogenous, but highly heterogeneous with varied hydrologic characteristics which control the quantity and distribution of groundwater (Tolman 1937; Tood 1964). Geological settings, formations (or structures) that are sufficiently porous to store water and permeable enough to transmit water in adequate and economic quantities are called aquifers (Price 1985; Tolman 1937). The defining characteristics of these subsurface water repositories – aquifers are discussed below.

The principal hydrologic characteristics of rocks are porosity, effective porosity or specific yield, specific retention, permeability and the direction of maximum ease of percolation (Tolman 1937; Price 1985; Tood 1964). These hydrological properties are dependent on porosity, size of openings or voids (or interstices) and shape, arrangement, interconnection and continuity (Tolman 1937). Porosity is the ratio of the volume of voids (that is, openings or pores in rock) in the rock to the total volume of the rock (Price 1985; Tood 1964). Porosity controls the entrance of water into aquifers by assessing the rock's capacity to hold water. Tolman (1937, 111-112)

investigated the pattern of voids in relation to porosity and the direction of ease of water percolation and identified the following:

- "percentage of void space does not increase with the size of material
- from the previous, rock heterogeneity reduces pore space
- the size of the finest void material which occurs in sufficient amount to surround the coarser grain materials controls the velocity of percolation in heterogeneous material
- the larger the proportion of large grains enclosed in fine material, the greater the reduction in average porosity of the formation".

Appendix A1 shows the porosity of selected geological materials.

Permeability controls the combined effect of material void size and their interconnectedness to enable appreciable passage of water through them. Simply, permeability is the measure of the ease with which water flows though rock pores. Tolman (1937) observed that permeability varies with the degree of material assortment or the percentage of fine material and arrangement of coarse grains with fine material (that is, sedimentary structures). Water permeability is called hydraulic conductivity which is the volume of water that flows though a unit cross-sectional area of a geological formation in unit time under unit hydraulic gradient at a particular temperature (Brassington 1988). Permeability is, thus, measured by assessing the hydraulic conductivity of rocks. Specific yield also called effective porosity is a measure of the water moved under gravity influence or the volume of water that is drained from a rock or soil material under gravity effect when initially saturated (Tolman 1937; Price 1985). Specific yield increases with grain size and assortment. Specific retention is the measure of water that is not drained from the pores when a saturated rock or soil material is drained under gravity (Price 1985). Apparently, specific retention decreases with grain size and assortment.

Specific yield and specific retention contributes to the water-holding ability described as porosity (Price 1985; Tolman 1937). Specific yield and permeability are

also broadly related. In general, geological formations with high specific yield tend to be more permeable and vice-versa (Brassington 1988). Appendix A2 indicates the specific yield in percent and Appendices A3 and A4 show the permeability in terms of hydraulic conductivity of selected geological materials. Tolman (1937) observed that decreasing grain size and increasing fineness and proportion of void material can gradually alter the geological settings from an aquifer to aquiclude. An aquiclude is a geological formation which although porous and capable of absorbing water, will not transmit it fast enough to furnish an adequate supply of water. Essential aquicludes are silt and clay and their extent and structure formation control groundwater distribution and movement in aquifers.

4.2 STUDY SITE AND HYDROGEOLOGIC CONSIDERATIONS

Major hydrologic properties which support the existence, flow and distribution of groundwater are outlined in the preceding section. This section describes hydrogeologic characteristics of the study area (ORM) and its major depositional information. Depositional information provides a baseline for evaluating the subsurface environment. Depositional history may be reconstructed from stratigraphic information. The ORM deposition environment is a moraine. A moraine is a general term for debris of all sorts originally transported by glaciers or ice sheets that have since melted away (Wicander and Monroe 1995; Skinner and Porter 1989). That is, a moraine is accumulation of glacial sediments (drift) deposited directly by glaciers (Levin 1981). Moraines are characterized with sediments (e.g. sand, silt, gravel, etc) and unconformity. The ORM, as an example, is built on high relief, erosion surface (unconformity) and a network of tunnel valleys (Barnett et al. 1998). So, geological deposits are predominantly sediments as evident from the golden spikes, MOEE data and depositional information. This limits the geologic units (that is, gravel, silt, etc) to be considered in the modelling process. However, the key problem is excessive complexity in sediment distribution due to varied extents of subsurface deposits. These are revealed in aquifers and aquitards having varied extents and geometry (Sharp et al.

1996). This problem requires modelling tools to accommodate local geologic property. The section following outlines varying sediment types present in the ORM.

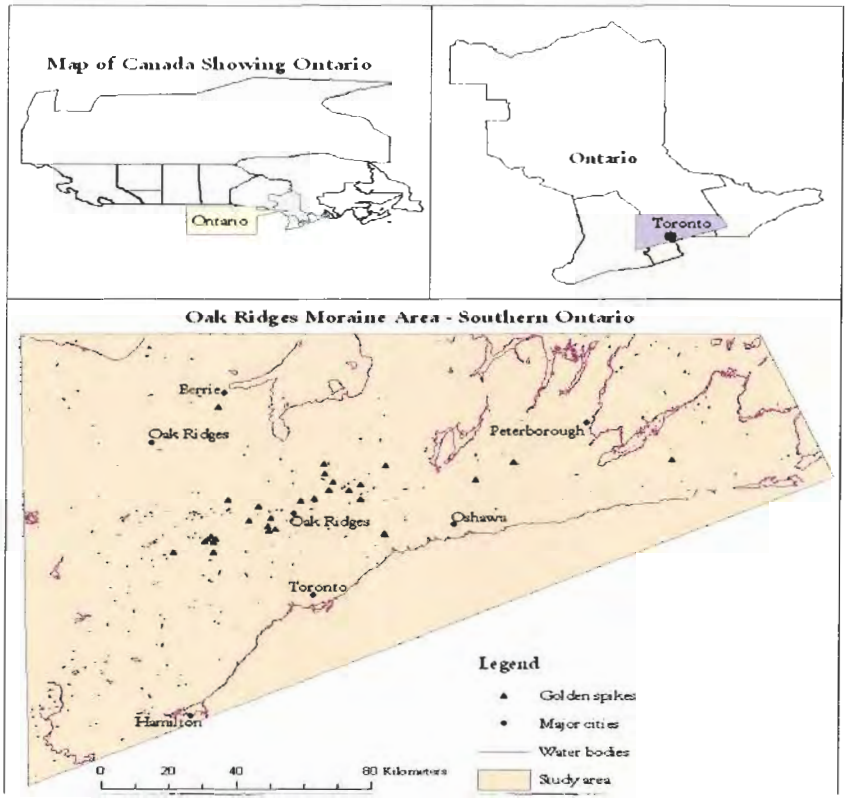
4.2.1 Oak Ridge Moraine (ORM) - Southern Ontario

Oak Ridge Moraine (ORM) is located in southern Ontario adjacent Lake Ontario. The ORM (Figure 4.1) subsurface environment is an aquifer complex which provides large amounts of potable water within the Greater Toronto Area (GTA) (Sharp et al. 1996). Groundwater potential of the ORM has attracted many researches into understanding its origin, nature and architecture. The Geological Survey of Canada (GSC) in 1993 initiated a three year regional hydrogeological study of ORM principally for aquifer delineation (Sharp et al. 1996). Among many research objectives, the following problems constitute core ORM challenges:

- weak geological framework for addressing hydrogeological and planning related problems,
- excessive complexity due to complex subsurface glacial deposits because aquifers and aquitards have varied regional extents and geometry,
- groundwater flow-paths are difficult to trace and
- 3-D geological mapping is necessary to identify geological controls on groundwater flow (Natural Resources, Canada 2003).

In their paper, 'On the origin of the Oak Ridges Moraine' Barnett et al. (1998) outlined past and present geological model of major sediments and their distribution in the ORM. ORM forms a drainage divide of high sandy ground between Lake Ontario and Georgian Bay, extending from the Niagara Escarpment to beyond Rice Lake (Barnett et al. 1998). Barnett et al. (1998) and Sharp et al. (1999) observed that the Peterborough drumlin field occurs to the north of ORM (Chapman and Putnam 1943; 1951; 1984) that forms a regional NE-SW-oriented surface underlain by thick, deposits of Newmarket Till.

Figure 4.1: Study area location for golden spikes and MOEE data



The drumlin field is cut by a complex NE-SW-oriented, network of deep valleys. The valleys have deep sides, a branching pattern, inset eskers and large bedform. These features have being suggested as tunnel channels by high energy subglacial meltwater flow (e.g. Barnett 1989; 1990; Shaw and Gorrell 1991; Brennand and Shaw 1994). In their paper 'Regional geological mapping of the Oak Ridges Moraine, Greater Toronto Area southern Ontario' Sharp et al. (1999), described seven physiographic areas of the ORM as the following:

- the Niagara Escarpment is an elevated landform that affect melt water flow across the area (Barnett et al. 1998)
- drumlinized uplands of the Peterborough drumlin field occur north and south of the ORM and they underlie it (Barnett et al. 1998)

- large flat-floored valleys are eroded into the drumlin upland north of the ORM; some continue south of the moraine (Sharpe and Barnett 1997; Kenny 1997)
- ORM forms a drainage divide of high sandy ground between Lake Ontario and Georgian Bay, extending from the Niagara Escarpment to beyond Rice Lake (Barnett et al. 1998)
- broad, gently sloping plains border the south-western margin of the ORM (Barnett et al. 1991)
- Lake Iroquois shoreline truncates this plain
- river valleys dissect the area rising in drumlinized uplands or in the ORM.

Depositional information of the ORM is crucial to understand sediment distribution in the subsurface environment. Sharp et al. (1999) identifies sediment origin and thickness for the ORM subsurface environment. Major sediment units are:

- Halton till: are drifts occurring as surface tills and lake sediments. It comprise clayey silt to silt till with interbedded sand and silt (Sharpe 1996).
- Oak Ridge Moraine: constitutes extensive surface deposit, 160km long and 2 to 11km wide but may be more extensive beneath the Halton drift (Sharpe 1996). Interbedded fine sands and silts constitute major sediments, but coarse sands and gravel are prominent locally (Sharpe 1996).
- Newmarket till: have drumlins and erosion elements and occurs at the surface north of ORM. It comprise a thick gravel, silty sand to sandy diamicton separated by sandy interbeds (Sharpe 1999; Sharpe 1996).
- Unconformity: is regional erosions surface marked with channels and drumlins (Barnett et al. 1998). Coarse grained drifts form part of the erosion surface.
- Lower deposits: lies between the bedrock at the bottom and Newmarket till at the top. It comprises mainly sand, silt, clay and till. White (1975) observed that outcrops of this formation occur north of Lake Ontario shoreline (Sharpe 1996).

- Channel fill: comprise dense buried drifts with 10 to 25m thick of gravel sequence. This unit has 10 to 75m thick of sandy drifts that fine upwards to silt and clay (Barnett et al. 1998). But surface sediments contain fine sand, silt and organic material. Gwyn and Dilabio (1973); Sharpe et al. (1994) observed that sandy and stony till extend beneath the ORM.

In sum, the ORM is built on regional unconformity comprising irregular drumlins of Newmarket Till in the broad upland areas and the base of the deep, wide, inter-upland valleys (Barnett et al. 1998). The section below describes ways to accommodate varied sediment distribution for characterizing the subsurface environment.

4.3 APPROXIMATING GEOLOGICAL SPACES USING BOREHOLE UNITS

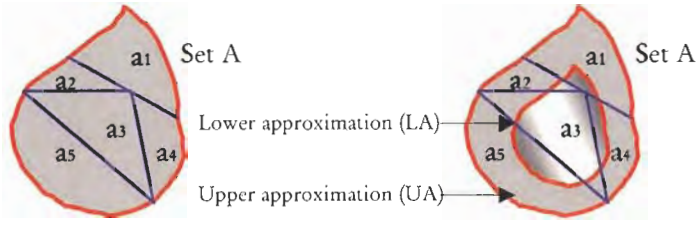
Depositional information which describes the nature of geologic units that make-up the ORM is outlined in the preceding section. This section discusses the means of combining aquifer characteristics in order to characterize the subsurface environment. Geological spaces are inherently heterogeneous with increasing material variability exhibited by borehole units. Approximation of these borehole units that give descriptive properties of the geological formation is essential for geographic analysis and decision-making. Hydrological properties of water-bearing rock materials identified from the preceding section are employed in geological space approximation. However, this approximation process is only accurate as the available data, so the accuracy of the modelled or the resulting geological space must retain the original data distribution and inherent variability. A rough sets approach is employed here as the approximation tool in classifying or categorizing these borehole units.

Borehole units referred to as subsurface materials are identified with aquifer characteristics as: porosity, permeability, specific retention, specific yield and grain size. The approach categorizes well-log units using these hydrologic properties. The rough sets method aims at retaining individual variability of input variables

irrespective of spatial granularity of observed variables. This characteristic will enable accurate assessment of disparities in outputs for different uncertainty levels of data. For example, to define whether an area is classified as an aquifer, we start by imposing elementary sets using borehole units. Borehole units considered as elementary sets are characterized by discernible information, that is, well-log units (e.g. sand, clay, etc) are elementary granules of an aquifer formation.

For example, in Figure 4.2, we consider the area in set A as a borehole and then use borehole unit labels; a1, a2, a3, a4, a5 as elementary sets or partitions of set A. These borehole units are used to define the lower and upper approximations of set A as LA and UA respectively. The elementary sets are analyzed for set criteria using the hydrologic properties, and LA and UA are derived accordingly, for example, as shown in Figure 4.2. The set A in Figure 4.2 is an individual borehole and the elementary sets are the various layers of subsurface material for a specified borehole (set A). LA and UA specified for the individual sets will be identified with heights.

Figure 4.2: Rough set characterization into approximation sets (LA & UA) using elementary sets



By this way, boundaries could be defined in three dimensions because each LA and UP have both spatial and height information. Table 4.1 which comprise porosity, permeability, specific yield and grain size have been designed into five categories for the borehole material group approximations. Increasingly, the moderate/medium categories in the various hydrologic characteristics except for permeability have been identified consistently as ideal geological setting for optimal groundwater conduction and storage. Table 4.2 shows the key for the various material groupings which specify their levels of aquifer-supporting characteristics.

Table 4.1: Summary list of hydrologic characteristics for borehole material (sediments) category approximation

POROSITY (per cent)				
Very Low	Low	Moderate/Medium	High	Very High
0 – 20	21 – 33	34 – 45	46 – 53	54 – 95
Glacial till	Coarse gravel	Fine gravel	Silt	Peat
	Medium gravel	Coarse sand		Organic clay
	Till (sand)	Medium/fine sand		
	Till (silt)	Clay, Dune sand		
PERMEABILITY (Hydraulic conductivity in metres per day)				
Very Low	Low	Moderate/Medium	High	Very High
$10^{-5} - 10^{-4}$	$10^{-3} - 10^{-2}$	$10^{-1} - 1$	$10 - 10^{2}$	$10^{3} - 10^{4}$
Clay,	Silt, Till	Fine sand	Clean sand	Gravel
	Till (sand, clay & silt)	Silty sand	Sand	
	Sandy silts, clayey sand			
SPECIFIC YIELD (per cent average)				
Very Low	Low	Moderate/Medium	High	Very High
0 – 12	13 – 21	22 – 23	25	26 – 27
Clay	Coarse gravel	Medium gravel	Fine sand	Coarse sand
Till (sand, gravel, silt)	Medium & fine gravel	Coarse gravel	Gravely sand	Medium sand
	Fine sand, Silt	Dune sand		
GRAIN SIZE (mm)				
Very Small	Small	Moderate/Medium	Large	Very Large
63μ – 374μ	375μ – 500μ	501μ – 1999μ	0.002 – 16	17 – 256
Silt	Medium sand	Coarse sand	Granules	Cobble
Clay		Fine gravel	Gravel	Boulders
Fine sand			Medium pebbles	Large pebbles

Developed as a summary table from Appendices A1, A2, A3 & A4 and also from (Brassington 1988, appendix III).
Modified and categorized to suit research application.

Table 4.2: Descriptive key for material category approximations

KEY			
	Very good Aquifer Indicators (VGAI)		
	Good Aquifer Indicators (GAI)		Poor Aquifer indicators (PAI)
	Moderate Aquifer Indicators (MAI)		Non – Aquifer Indicators (NAI)

These grouping constituents do not form consistent material categories for the various hydrologic properties. Some materials belong to more than one category and even within one group, category material constituent are not consistent. These material distribution and category characteristic are evident in the real world. The material group characterization illustrated by these hydrologic properties describes what Pawlak et al. (1995) defined as a set that is indefinable by given attributes called a rough set. The set of good aquifer indicators, for example, has material constituents that are inconsistent for each hydrologic property. This inherent data characteristic can be appropriately handled using the rough set for set approximation. In rough set fashion, each material category (e.g. low, medium, etc) can be said to comprise lower and upper approximation sets, with each material considered as elementary entity constituting a particular category. It is worth noting that the approximation sets are not uniquely defined for probabilistic applications which require exact and known population of sample spaces

4.3.1 Approximation and Set Derivation from Well-log Material Characteristics

The preceding section outlined geologic material grouping using aquifer properties for subsurface materials. This section employs rough set rules to categorize geologic materials into different levels of aquifer index. As apparent from Table 4.1 (see section 4.3), well-log materials exhibit multiple categories for the various material characteristics. Distinct material classification is not possible using the individual borehole material properties, because of inconsistency of constituent well-log materials which characterize the material properties.

Table 4.3: Approximation set derivation from borehole material properties

Set Category	Upper Approximation set	Modified Upper Approximation set	Lower Approximation set	Set categorization criteria
Very Good Aquifer Indicators (VGAI)	Fine/medium gravel	Dune sand	Fine gravel	Upper Approximation set = at least one set inclusion
	Coarse/clean sand		Clean/coarse sand	
	Medium & fine sand			
	Clay, Dune sand			
Good Aquifer Indicators (GAI)	Coarse/medium gravel	Medium sand	Coarse/medium gravel	Modified Upper Approximation set = at least two set inclusion & with higher indicator presence
	Fine gravel, gravel		Gravel	
	Till (sand), Till (silt)			
	Fine sand, silt			
Moderate Aquifer Indicators (MAI)	Glacial till	Glacial till	Fine sand	Lower Approximation set = at least three set inclusion & with prior aquifer indicator presence
	Fine sand	Till(sand, silt)		
	Silty sand	Granules		
	Coarse/medium sand	Medium pebble		
Poor Aquifer Indicators (PAI)	Silt, sandy silts, clayey sand	Till	Silt	Note: Prior aquifer indicator presence signifies significant possibility of being classified into a higher aquifer indicator category, hence such materials have higher weights in the next set group
	Till, Till(sand, clay, silt)	Cobbles, Boulders	Sandy clay	
	Fine sand, gravely sand	Large pebbles	Gravely sand	
	Cobbles, Boulders	Till (sand, gravel, silt)		
	Large pebbles			
Non– Aquifer Indicators (NAI)	Peat, Organic clay	Peat	Clay	
	Clay	Organic clay		
	Till(sand, gravel, silt)			
	Fine sand			

However, borehole materials for different properties can be classified into various levels for aquifer indicator suitability. This allows the possibility of applying set rules using the four aquifer properties to approximate well-log materials into set categories. The result of this approximation for the material properties from Table 4.1 is shown in Table 4.3.

In fuzzy set theory, membership functions enable elements to exhibit partial class memberships of different and overlapping sets. Confusion sets, however, may result in cases where zones of different fuzzy sets intersect (Burrough and McDonnell 1998). This may arise where an element is a partial member of three or more fuzzy sets to generate two or more intersecting fuzzy zones. As indicated in Table 4.1, silt for example, is a partial member of three different sets (that is, poor aquifer indicator, non-aquifer indicator and good aquifer indicator). This and similar situations where an element is characterized by multiple set memberships require tools that categorize individual elements while retaining their varied class characteristics and fuzzy memberships. It worth noting that no particular tool can account for this data characteristic, so the approach is to implement these analytical tools (that is, rough set theory and fuzzy set theory) to accommodate different aspects of the data. Consequently, the rough set process for borehole material categorization does not account for the inherent geologic unit transition within boreholes, but allows the classification of any geologic unit into a single group.

4.3.2 Application of Geologic unit Categories to MOEE Data

In the above section, a generalized grouping of unconsolidated subsurface materials is developed by characterizing sediments using major aquifer properties such as porosity, permeability, etc. This section applies this grouping to MOEE data which are standardized by the Geological Survey of Canada (GSC) with predefined sets of materials. Appendix A5 lists major subsurface sediments present in MOEE data after standardization by the GSC while Appendix A6 shows geologic material groupings for the MOEE data using rough sets. In reference to Table 4.3, which illustrates geologic

materials into aquifer index categories, Appendix A6 shows GSC material details which constitute different aquifer indicator groups. These materials are further detailed into Table 4.4 with 'fill' regrouped into 'others'. These material clusters are used in subsequent subsurface characterization process such as Markov chains.

Table 4.4: List of geologic material details grouped into categories

Material Tag	Material	Material Details	Material Tag	Material	Material Details
Others	fill	fill	Gravel	gravel	gravel
	organic	organic	Sand	sand	sand
	covered	covered, previously bored		sand_diamicton	sand, diamicton
	bedrock	bedrock		sand_diamicton	gravel,diamicton
	limestone	limestone	Silt	silt	silt
	shale	shale		silt	sand, clay
	granite	granite		silt_diamicton	silt, sand
	dolomite	dolomite		silt_diamicton	gravel, clay
	pot_bedrock	potential bedrock	Clay	clay	clay
	sandstone	sandstone		clay_diamicton	clay, silt
	limestone_shale_inter	limestone, shale			
	unknown	*unknown*			

4.4 ASSESSING GSC STANDARDIZATION SCHEME

In the preceding section, different sediment types have been categorized using aquifer supporting properties. But these groupings are standardized geologic units from MOEE water well data which are characterized with diverse geologic units and terms. The accuracy and extent of the standardization process however is not known. So, this section evaluates the representation of original data in the standardized MOEE data.

MOEE water well data for ORM is mainly collected by private well drillers (pwd). The quality of MOEE data undermines its application into research and subsurface studies. Lack of training for pwd have been identified as one major factor limiting geologic accuracy of the MOEE data (Russell et al. 1998; Schuurman 2002). A rational approach to resolving this problem is the use of standardization schemes to homogenize the data onto a common platform. This is a necessary step to both

prepare the data for modelling processes and derive relevant geologic information. Standardization, however, should retain original data variability. The variability of output data from the standardization process is crucial for at least two reasons.

First, variability output data from standardization with respect to the original data evaluates the accuracy and the extent of the classification system. This is important because different classification systems have different goals. This goal may be data reduction and filtering while others may aim at replicating original data distribution. The measure of accuracy is the degree of how standardized outputs represent original data. The accuracy assessment describes not only explicit representation of terms such as gravel, silt, etc but also descriptive information carried by those terms. For example, 'sand and clay' classified as silt has 100% accuracy. The extent of classification is the degree by which standardization rules reduce original data into standard terms. For example, all instances of 'gravel and clay' are converted to 'silt_diamicton'.

Second, standardization scheme assessment could be used to design training programs for well-log data collectors (e.g. pwds). Hence, the assessment result should identify geologic materials with excessive high degree of error and vice versa. There are however, problems in assessing data variability, particularly, categorical data. The section below outlines few of these problems.

4.4.1 Challenges in Assessing Variability of Categorical Data

The above section discussed the need to assess classification systems and incorporate assessment outputs into validating rule-based standardization processes. This section, however, identifies some problems associated with categorical data classification using standardization schemes. Data variability assessment is a measure of variance between original data and the output data from standardization process. It includes both the magnitude and the direction of the variance. Variance assessment for numerical data is relatively easy because prior understanding and domain knowledge are not necessary preconditions to determine variance. Numerical data, though, not

independent of the parameter under study (e.g. income levels, elevation, etc) have properties, and are understood by their magnitude. Population counts, elevation, income level, etc are examples of numerical data. Data variability can be assessed without underlying information about the data or consulting domain experts.

Conversely, variability of categorical data requires domain knowledge. Variance of categorical data is not just the disparities that exist between two or more data elements but the unique properties inherent in them. For groundwater considerations for instance, fill, overburden and topsoil are considered one geologic unit since their hydrogeologic property is similar - hence no (or zero) variance. So, expert knowledge is a precondition to establishing variance in categorical data.

A common approach to handling poor quality categorical data is to classify its terms to standard terms so relevant information are embedded into specific terms. This enhances information retrieval (Russell et al. 1998) and accuracy parameters can be assessed easily. Validation of such classification systems do not fall into mainstream standardization approaches. But, not until we have validated rule-based schemes which impose strict grouping of data elements (e.g. geologic units), do we actually begin relating standardization scheme to the real world. Original data collected by pwd express the variability inherent in the real world. Hence, the assessment of the standardization should enhance ways to incorporate original data variation into standardized outputs. The derivation of this variability is illustrated in the following section.

4.4.2 Determination of Variability Index

Problems encountered during categorical data assessment are discussed in the above section. This section outlines the approach adopted for computing the variability index for categorical borehole data. The variance computation employs set rules to derive three identification categories. First, if the original material is fully represented in the standardized output then identification index (IDI) of one (1) is assigned to that borehole unit (for example, if silt is standardized as silt then its IDI is

1). Second, a borehole unit is assigned half (0.5) if the original data element is partially represented in the output. For example, if 'sand and gravel' is classified as 'gravel' then IDI is 0.5. Third, a unit is given zero (0) IDI if there is no relationship between the original data and the standardized output. The set rule essentially compares the sum of all identifiable materials to the standardized material for each well-log unit within a particular borehole. All borehole units with IDI values less than one (1) are extracted in order to identify geologic units in error and also to compute percent error for material identification within each borehole. A summary table is constructed to outline the minimum, maximum, mean, standard deviation and coefficient of variation for the percent error of all boreholes. The summary table also includes the percent accuracy for identifying a particular geologic unit, say sand, gravel, etc.

4.5 ACCOMMODATING GRADUAL TRANSITION BETWEEN BOREHOLE UNITS

The preceding section illustrated methods for assessing the GSC standardization scheme. This section describes techniques for accommodating gradual transition between geologic units. The approximation of borehole units developed towards categorizing geological space must also accommodate gradual transition between these geologic materials which are continuous entities. The transport of geologic materials from their source origins to various deposition sites through agents such as streams, glaciers, winds, etc accumulate sediments into layers whose transition from one to the other vary (Dolgoff 1996). While other geological investigations and deposition history of sediment accumulation sites can be employed to model the pattern of various geologic units, the geologic unit transition from one material to the other can hardly be estimated. Also such information is often lacking from mainstream well-log information. In cases where knowledge of geologic unit pattern may be sufficient for certain applications, an account of this added information on transition zones of various materials can enhance the performance of derived numerical models.

The essence and model performance capability of this information for environmental and geologic applications are not farfetched. Control of contaminant transport in groundwater systems may be endangered resulting in failure of numerical models that are used for the location and design of waste disposal sites. Road and dam settlements may also occur where there are significant transition zones between major geologic units that are ignored. The definition of the transition zone in fuzzy objects is related to the limits at which the object indicates differing characteristic which is significantly dependent on the precision of measuring the phenomena under consideration (Burrough and McDonnell 1998, 271). For data measured at certain locations (or points) the width of the transition zone could reflect the known accuracy of the measurement technique; for interpolated grid data using Kriging, the width of the transition zone could be given by the Kriging standard error (Burrough and McDonnell 1998, 271). For diffuse geographic boundaries, width of the transition zone of the membership functions related to geographic boundaries could be defined using expert knowledge from the terrain (Burrough and McDonnell 1998).

To accommodate indeterminacy in geologic unit transitions, relevant data on transition zone of geologic unit is essential coupled with application of an appropriate analysis tool that fully represents the fuzzy phenomena. Geophysical borehole logging such as resistivity and radiometric logs are potential means of estimating the width of transition boundaries between subsurface units. Lagacherie, Andrieux and Bouzigues (1996) identified the collection of soil indeterminacy information and the selection of a theoretical framework to model soil indeterminacy. The collection of geologic unit transition zone information in well-log data are required in the simulation of gradual transition between geologic units. The estimation of transition zone width may be highly inaccurate from geologic unit characteristic because transition boundary length between two geologic units is a property of the boundary rather than the geologic unit characteristics in consideration. That is, the boundary width of a transition zone may be defined separately for geologic unit and study area in question. It is worth noting, however, that the depositional history of a particular geological setting may influence

the boundary width and the arrangement of constituent sediments. Hence, fuzzy set theory is not implemented in this study due to lack of boundary information for different geologic units.

4.6 STOCHASTIC SIMULATION USING MARKOV CHAIN MODEL

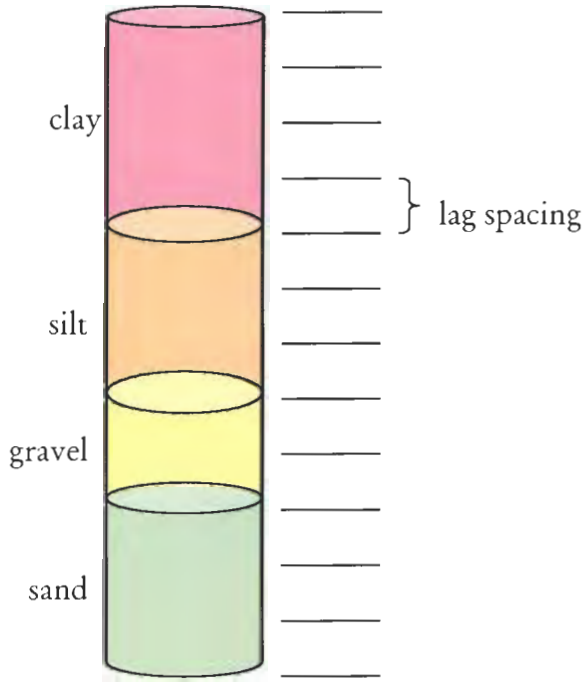
The above section discussed information requirement for applying fuzzy sets to borehole data for characterizing the subsurface environment. This section describes stochastic approach for simulating material transition sequence in the subsurface. The modelling of spatial surfaces using various interpolation techniques, for example, Kriging is employed to approximate possible estimates at unsampled locations (Burrough and McDonnell 1998). Variograms and the variance of estimated regionalized variable in Kriging do not necessarily exhibit original data characteristics, but ensures minimum error in the interpolation process. These interpolation techniques strive for the accuracy of the approximation process with little regards to geographic reality.

Geographic simulation, on the other hand, attempts to provide infinite number of realizations (or renditions) that replicate the distribution and variability of original data (Shibli 2003). Hence, simulation attempts to reproduce reality by considering the uncertainty involved in data characteristics. Stochastic simulation is an alternate approximate probability simulation (Bardossy 2003) employing the use of random number sequence to generate multiple representations of reality. Stochastic simulation by Markov chains generates geographic representations of random field by employing simulated annealing, a generalization of a random sampling from probability distributions for examining a system's varying states in finite transitions (Bardossy 2003; Bevington and Robinson 1992). Simulations are very useful for possible representations of reality that are not final outputs of analysis (Bardossy 2003).

Markov chain models are applied in geology for categorical data (e.g. lithologies or geologic units) modelling to provide random patterns of spatial variability and also for relatively structured patterns with asymmetry and cyclical trends (Elfeki and

Dekking 2001; Carle and Fogg 1997). Increasingly, spatial heterogeneity (e.g. lens length variation) common in geological material units requires analytical tools to simulate the geological distribution characterized with random geological states. Carle and Fogg (1997) contended for the appropriateness of the use of transition probability in Markov chains for accommodating asymmetric geological patterns. Most indicator geological models assume geological symmetry to quantify geographic variability, for example, cross-variogram or indicator models.

Figure 4.3: Sample illustration of transition probability estimation using borehole geologic units



Markov chain models were developed in Groundwater Modeling System (GMS) by Brigham Young University into T-PROGS interface in the Borehole Module of GMS. T-PROGS application performs a transition probability geostatistics to generate multiple realizations of aquifer heterogeneity which are conditioned to a well-log data (Jones 2003). The Markov chain model in geological applications starts by defining n number of possible geologic material states e.g. S_1, S_2, S_3, ..., S_n. The probability P_{ij} of material transition from state S_i to state S_j is estimated. Stationary and

transition probabilities are then generated from the borehole material to develop the Markov chain model. Figure 4.3 illustrates the superimposition of a vertical line of equidistant points along a borehole at a particular interval.

The transition frequencies between material states are determined as the ratio of the number of times a given state S_i is followed by itself or the other states S_j in the process to the total number of transitions (Elfeki and Dekking 2001). The transition probability of material j to k, t_{jk} (h) is defined by the conditional probability as:

$$t_{jk}(h) = \Pr(j \text{ occurs at } x+h \mid k \text{ occurs at } h)$$

where x is the spatial location, h is the lag spacing and j, k are the material categories (Jones 2003; Carle and Fogg 1997). A curve of transition probability against the lag spacing represents the Markov chain. Multiple material sets are generated during the simulation phase by fitting Markov chain curves to measured transition probability curves (Jones 2003). Markov chain model applied to one-dimensional categorical data in a direction ϕ assumes a matrix exponential form:

$$T(h_\phi) = \exp(R_\phi h_\phi) \text{ and}$$

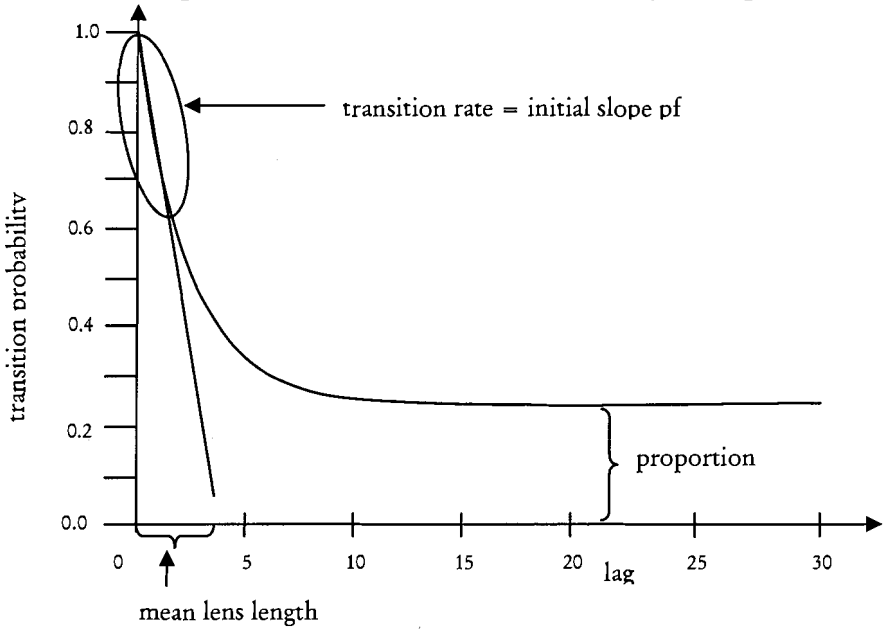
$$R_\phi = \begin{bmatrix} r_{11,\phi} & \cdots & r_{1k,\phi} \\ \vdots & & \vdots \\ r_{k1,\phi} & \cdots & r_{kk,\phi} \end{bmatrix}$$

where ϕ is a lag in the direction ϕ, R_ϕ represents the transition rate matrix and $r_{jk,\phi}$ denotes the conditional rates of change from material category j to k per unit length in the direction ϕ (Jones 2003; Carle and Fogg 1997). Transition rates ensure an optimum fit between Markov chain model and the observed transition probability data.

The Markov chain plot is characterized by three key descriptive features; material proportions, lens length and transition rates as illustrated in Figure 4.4. In the sample Markov chain shown in Figure 4.4, the transition probability corresponding to the flat portion of the curve represents the mean proportion for the material

considered. The mean lens length corresponds to the lag separation on the horizontal where the tangent drawn against the initial portion of the curve intersects the horizontal axis. The initial gradient of the curve represents the transition rate.

Figure 4.4: Sample transition curve demonstrating key descriptive features



The preceding computations for the Markov chain is considered for only the vertical direction, horizontal transition probability matrix need to be estimated to generate the horizontal Markov chain. Borehole data are generally rich in vertical direction but not sufficiently dense in the horizontal direction. Walther's law is employed to approximate the horizontal Markov chain model from results generated from the vertical direction calculations. Walther's law states that "any juxtapositional tendencies observed in the vertical direction will also hold true in the horizontal directions" or in other words, the vertical successions of deposited facies represent the lateral successions of environments of deposition (Jones 2003). The application of Walter's law assumes uniform material proportions in all directions to allow the derivation of horizontal transition rates from the vertical transition rates.

4.7 BOREHOLE DATA RESULTS AND DISCUSSION

The preceding sections outlined methods for implementing the borehole data in order to enhance the means of characterizing the subsurface environment. Outputs from the borehole data are discussed in the following sections and are categorized into three groups. First, the GSC standardization system is assessed to estimate the extent of data variability reduction. It also identifies specific sediment frequencies and the accuracy of geologic material identification level for all water wells in the MOEE data. Second, the transition probability matrix (tpm) for sediments grouped by rough set process is determined in order to simulate sediment transition sequences in the vertical direction. T-PROG simulations which describe sediment transitions are defined for golden spike clusters and sample MOEE data. Conflicting and similar sediment transition patterns are identified to describe relative correlation between both golden spikes and MOEE data. Third, limitations of T-PROG simulation are outlined in the subsurface characterization process. While the tpm outputs have both theoretical and practical relevance, its drawbacks are also carefully noted and outlined. Finally, a simple illustration is used to estimate depth and spatial information where specific sediments and sediment transitions are likely to occur.

4.8 STANDARDIZATION ASSESSMENT RESULT

The standardization assessment result is summarized in two tables: Tables 4.5 and 4.6. In Table 4.5, original geologic descriptions are standardized into specific terms defined by the GSC. All possible geologic unit descriptions are identified from the original description and labelled as 'mat1', 'mat2', 'mat3', etc. These geologic units approximate the variability inherent in the initial description and are compared to the GSC standardized terms. Identification index, IDI is assigned to each description (that is, borehole unit) and for each borehole, percentage accuracy (ratio of the sum of IDI values for each geologic unit to the total number of geologic units that constitute the borehole expressed as a percentage) is computed.

Table 4.5: Sample output for GSC standardization scheme assessment – MOEE data

WTN	Unit Number	Description	GSC_stdmat	mat1	mat2	mat3	Material	ID Index	Percent Accuracy	Inconsistent Material
17-128	1	previously bored	Covered				covered	1.0		
17-128	2	medium sand - muck	Sand	sand	organic		sand	0.5	50	sand, organic
17-128	3	hardpan - -	Silt_Diamicton	diamicton			clay, gravel	0.0		diamicton
17-102	1	clay - stones -	Silt_Diamicton	clay	gravel		clay, gravel	1.0	100	
17-102	2	limestone - -	Limestone	limestone			limestone	1.0		
17-1005	1	topsoil –clay -medium sand	Fill	fill	clay	sand	fill	0.5		fill, clay, sand
17-1005	2	gravel - -	Gravel	gravel			gravel	1.0	75	
17-1005	3	clay - -	Clay	clay			clay	1.0		
17-1005	4	medium sand - stones	Gravel	sand	gravel		gravel	0.5		sand, gravel
68-9481	1	sand - gravel -	Gravel	sand	gravel		gravel	0.5		sand, gravel
68-9481	2	limestone - -	Gravel	limestone			gravel	0.0	16.667	limestone
68-9481	3	- -	Unknown							

Material Exceptions			
Original material	Standardized material	Original material	Standardized material
sand_diamicton	sand, diamicton	silt_diamicton	silt, sand
sand_diamicton	gravel, diamicton	silt_diamicton	gravel, clay
diamicton	hardpan	clay_diamicton	clay, silt
covered	previously bored	limestone_shale_inter	limestone, shale
silt	sand, clay		

If IDI is less than one (1), then geologic units corresponding to the original description are identified as inconsistent materials.

Table 4.6: Geologic material characteristics from MOEE data and summary statistic measures

Material	Material details	Number of Overall Occurrences From Golden Spike data (32 wells)	Number of Overall Occurrences	Number of Occurrences in Error	Percent Error
Gravel	gravel	62	75844	28705	37.847
Sand	sand, sand_diamicton	998	92404	35626	38.555
Silt	silt, silt_diamicton	615	12370	10152	82.070
Clay	clay, clay_diamicton	447	120977	25573	21.139
Others	bedrock, pot_bedrock, sandstone, shale, limestone_shale_inter, limestone dolomite, covered, fill, organic, unknown	459	113456	17826	15.712
Total		2581	415051	117882	195.445
Summary Statistics for Percent Accuracy of geologic units within Boreholes					
Sum	5620220.56	Number of borehole units		262,650	
Minimum	16.67	Number of boreholes		62,325	
Maximum	100.00	Standard deviation		5.2504	
Mean	90.18	Coefficient of variation		0.0583	

In Table 4.6, the number of instances a geologic unit occur both in original description and as inconsistent material are used to compute the percentage error (ratio of the number of occurrence in error to the number of overall occurrence expressed as a percentage). The geologic units considered are gravel, sand, silt, clay and all other subsurface materials are labelled 'others'. This material grouping is in reference to the rough set approximation. The geologic material details are shown in Table 4.6. The grouping illustrates the insignificant presence of other geologic units with respect to gravel, sand, silt and clay. Silt has the highest percentage error and 'others' records the least percentage error. The significant percentage error for silt raises many questions. Silt has the least occurrences both in error and in the original description; but for every ten (10) occurrence, at least eight (8) of these occurrences will be in error. This represents a considerable error and limits the accuracy of the classification system.

Clay has the highest overall occurrence but registers a small presence in error. This suggests most occurrences of clay have consistent classification and without 'others' geologic units, clay represents the most consistent sediment classified. Hence, the GSC rule has well represented (or targeted) clay in the classification process. Sand and gravel have approximately the same percentage error, though sand has higher occurrences both in original description and as inconsistent material.

From golden spike data, the number of occurrence of these geologic units is also computed (see Table 4.6). Sand has the largest number of occurrences while gravel has the least. Comparing overall occurrence of these sediment groups, sand is well represented in both golden spike data and MOEE data. Clay is over-represented in the MOEE data; hence it is reduced significantly (78.9%) in the classification process. This supports what Russell et al. (1998) observed for two boreholes in the Humber River watershed that less than 2% of clay was over-represented to about 40% in the MOEE data (Russell et al. 1998-E). The significant percentage error for silt may be because it is the least (17.9%) reduced material unit despite its predominance (approx. 24%) in the golden spike data. Significant reduction targeted on clay should also be directed onto silt in order to reduce its presence in error.

Table 4.6 also shows summary statistic values for the accuracy of all geologic units that constitute each borehole. The accuracy for each borehole is computed as the sum of IDI values for each geologic unit to the total number of geologic units that constitute the borehole expressed as a percentage. For 62,325 boreholes in the MOEE data; there are 262,650 geologic units, the minimum and maximum accuracies are 16.67% and 100% respectively. There is, however, only one occurrence of 16.67% and the average accuracy is 90.18% with a standard deviation of 5.25. Hence, for one standard deviation from the average, the accuracy range is between 95.43% and 84.93%. This represents a high accuracy measure for the classification system.

4.9 CHARACTERIZING SEDIMENT VARIABILITY - ORM SUBSURFACE

The above section assessed the GSC classification system which standardizes diverse geologic units into specific terms. It also illustrated major sediment types present in the ORM and their relative frequency of occurrence. This section however, employs golden spike data to characterize the subsurface in a vertical direction using transition probability matrix from Markov chains.

4.9.1 Group Selection for Golden spikes

The spatial distribution of golden spike data requires the grouping of golden spikes into small clusters based on proximity and similar sediment types. The accuracy of determining sediment transition sequence is a function of a deposition environment that exhibit similar sediment distribution. The grouping is very crucial to the accuracy of determining sediment proportions, lens lengths and the measure of transition of one sediment to another. Figure 4.5 illustrates sample borehole grouping.

The boreholes which constitute group 6 (constituent clusters are groups 6a, 6b and 6c) are related by proximity and can be categorized as one cluster. But, the sediment distribution patterns observed for this group of boreholes have directions to it. The depositional pattern has north-west (NW) to south-east (SE) direction where sediments of larger sizes are towards NW side and small sized sediments are to the SE. Hence, sediments are generally transported from NW to SE. Employing these properties of proximity and sediment distribution, golden spikes are grouped into small clusters shown in Table 4.7.

Figure 4.5: Sample borehole grouping prior to subsurface characterization

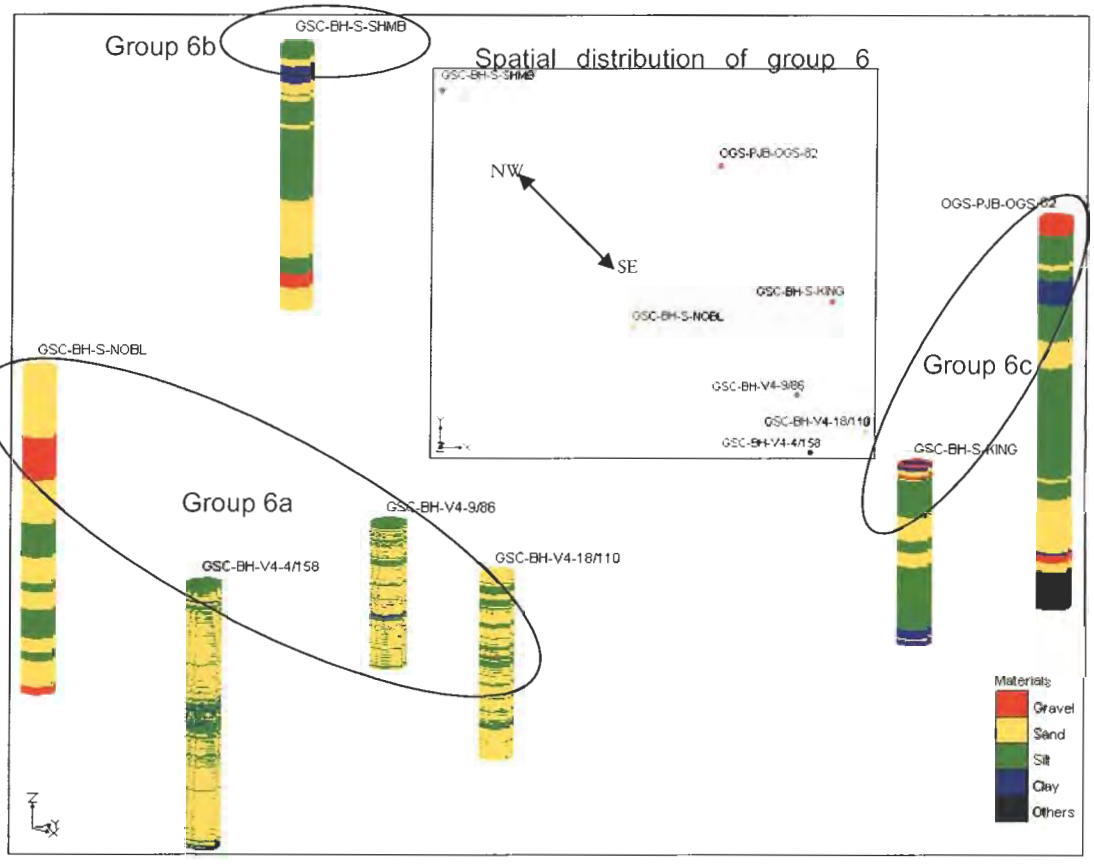


Table 4.7: Golden spike clusters for T-PROG simulation

Group Tag	Golden Spikes	Group Tag	Golden Spikes
Group 1	GSC-BH-S-RICE	Group 6a	GSC-BH-S-NOBL
Group 2a	GSC-BH-S-PON		GSC-BH-S-VGHN
Group 2b	GSC-BH-S-GHR		GSC-BH-V4-18/110
Group 3	RMD-UX-01		GSC-BH-V4-4/158
Group 4	GSC-BH-VSR		GSC-BH-V4-9/86
	OGS-PJB-14	Group 6b	GSC-BH-S-SHMB
	OGS-PJB-15	Group 6c	GSC-BH-S-KING
	OGS-PJB-16		OGS-PJB-OGS-82
	OGS-PJB-17	Group 7a	GSC-BH-C34B-14
	OGS-PJB-18		GSC-BH-C34B-17A
	OGS-PJB-19		GSC-BH-C34B-21
	GSC-BH-S-AUR		GSC-BH-C34B-28A
	GSC-BH-S-BAL		GSC-BH-C34B-29
	GSC-BH-S-MSR		GSC-BH-C48-4A
Group 5	GSC-BH-EE11-1/1	Group 7b	GSC-BH-S-CVC
	GSC-BH-EE11-9/1	Group 8	OGS-PJB-10

4.9.2 Transition Probability Outputs

The preceding section described factors considered for grouping golden spikes in order to enhance sediment transition determination. This section employs these golden spike clusters into simulating sediment distribution. These golden spike clusters are used in Groundwater Modeling System (GMS) employing T-PROG simulation to characterize sediment distribution in the vertical direction. For each simulation, the predominant sediment type is chosen as the background material and lag is specified. The T-PROG vertical simulation output includes; sediment proportions, lens lengths, transition rates, embedded transition probabilities and frequencies, maximum entropy factors and Markov chain graphs. These outputs for all the golden spike groups are shown in Appendices C1 to C24.

In Table 4.8 and Figure 4.6 is a sample output for group 4 (see Figure 4.6) golden spikes. Background material is sand (most predominant sediment) and lag is 0.3m. The transition rates corresponds to the slope of the transition probability curve at the initial lag (that is lag = 0). Diagonal values of the transition rate matrix are negative because for the same sediment type, transition rate decreases with an increase in lag (see Figure 4.7). For the same sediment type, consistent transition occurs when transition probability varies uniformly with lag. Sand transits most uniformly than any other sediments and have a transition rate of -0.084, while gravel and clay exhibit a relatively rapid transition. Transition rates for different sediments (off-diagonal terms) are positive; transition probability increase together with lag. Clay to sand (0.181) and gravel to sand (0.169) transitions have the highest rates. The converse of this is not true because sand to clay (0.028) and sand to gravel (0.029) transition have smaller rates. Hence, it is more likely to transit from clay to sand and from gravel to sand than the opposite of these transitions.

Embedded transition probabilities (or transition frequencies) are also computed so that they are conditioned to the sediment lens lengths. So, diagonal entries for the tpm correspond to the lens length values. Off-diagonal values are transition

probabilities for different sediment transitions. Clay to sand (0.727) and gravel to sand (0.571) transitions again, indicate the highest probabilities.

Table 4.8: Vertical T-PROG simulation output for group 4 golden spikes

		Transition Rates				Embedded Transition Probabilities			
Material	Proportion	Gravel	Sand	Silt	Clay	Gravel	Sand	Silt	Clay
Gravel	0.068	-0.312	0.169	0.092	0.051	3.210	0.571	0.238	0.190
Sand	0.538	0.029	-0.084	0.028	0.028	0.346	12.762	0.362	0.292
Silt	0.296	0.016	0.055	-0.086	0.015	0.217	0.522	11.609	0.261
Clay	0.098	0.008	0.181	0.045	-0.234	0.045	0.727	0.227	4.265
		Embedded Transition Frequencies				Maximum Entropy Factors			
Material	Lens Length	Gravel	Sand	Silt	Clay	Gravel	Sand	Silt	Clay
Gravel	3.210	3.210	0.106	0.046	0.036	3.210	0.889	1.256	0.797
Sand	12.762	0.133	12.503	0.135	0.113	1.092	12.503	0.914	0.874
Silt	11.609	0.046	0.126	11.609	0.055	0.962	0.578	11.609	0.833
Clay	4.265	0.009	0.150	0.046	4.265	0.190	1.090	0.795	4.265
Background material: sand					lag: 0.3m				
Golden spikes: GSC-BH-VSR, OGS-PJB-14, OGS-PJB-15, OGS-PJB-16, OGS-PJB-17, OGS-PJB-18, OGS-PJB-19, GSC-BH-S-AUR, GSC-BH-S-BAL, GSC-BH-S-MSR.									

Figure 4.6: Spatial and sediment type distribution for group 4 golden spikes

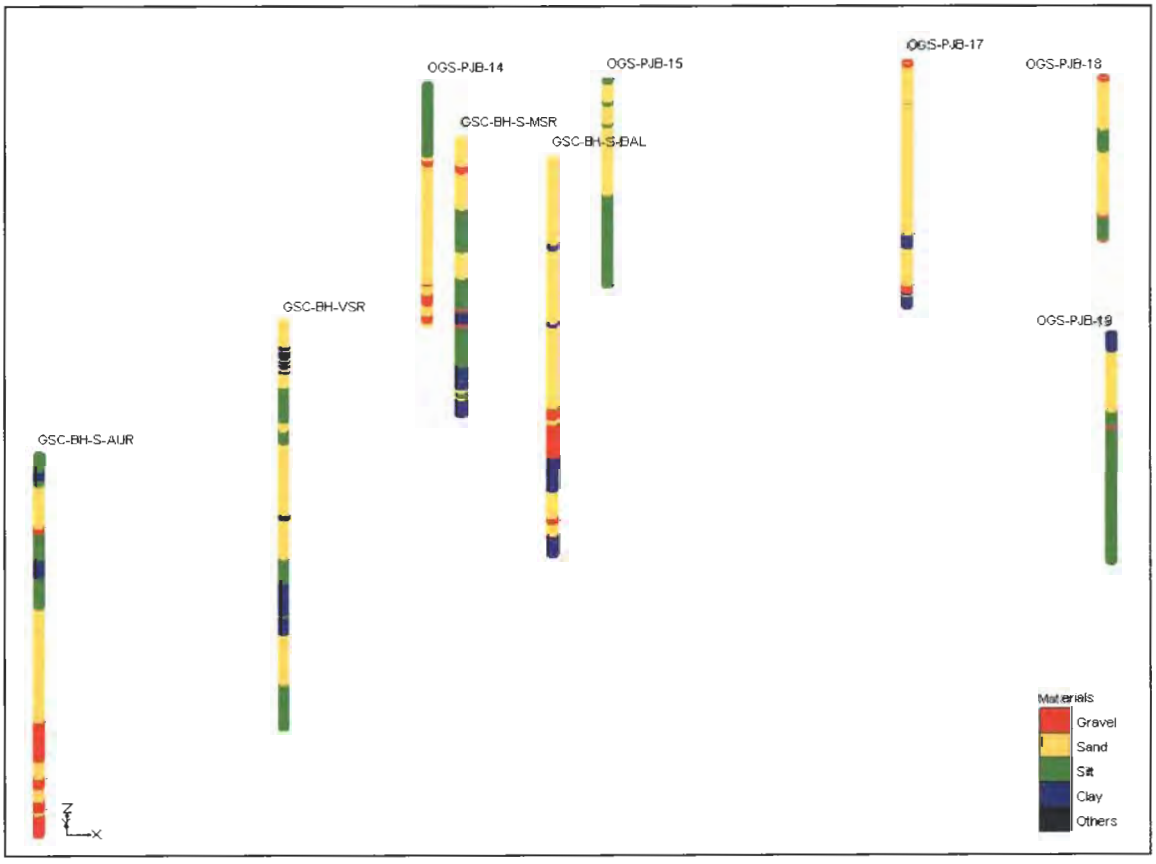
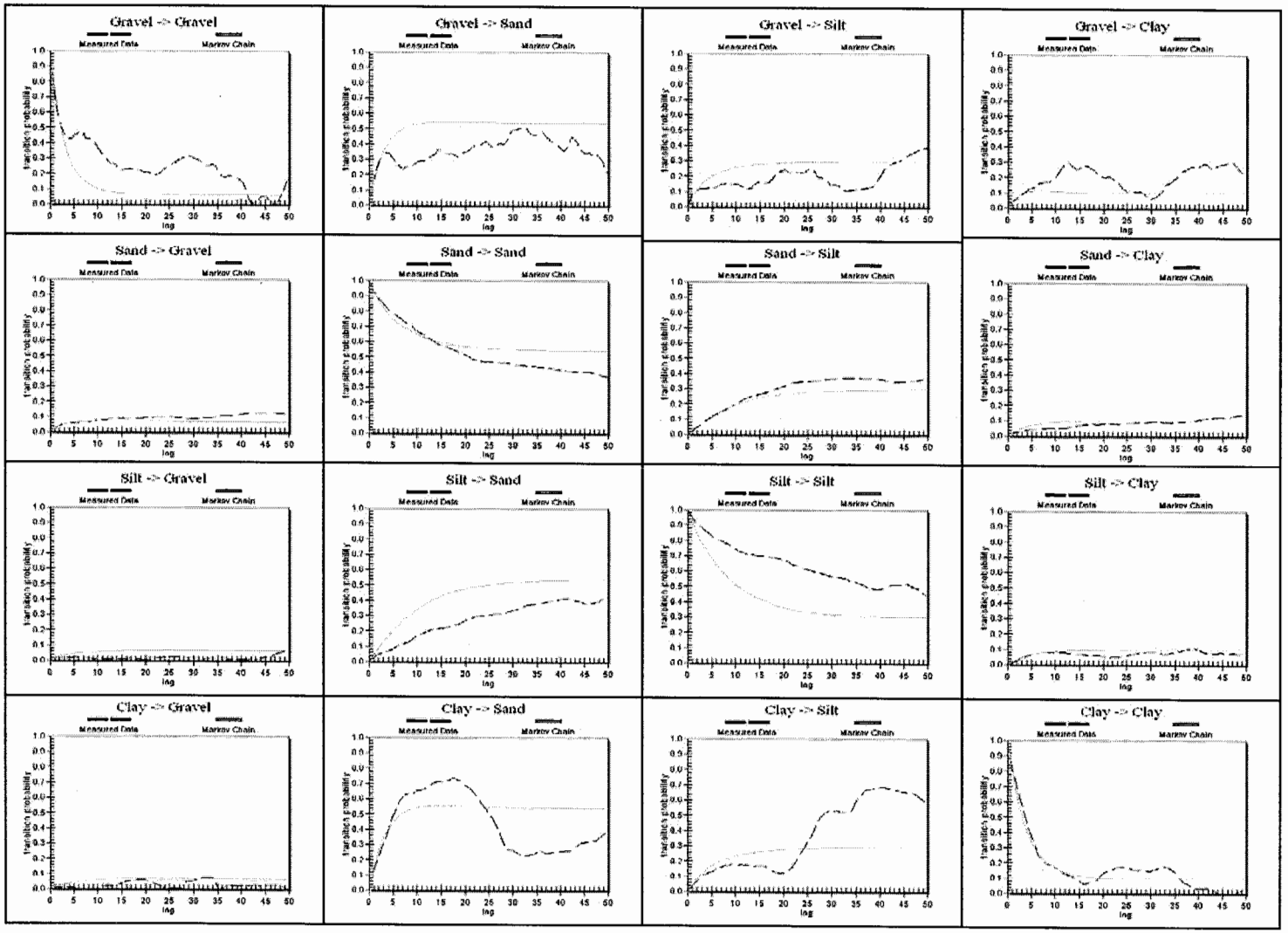


Figure 4.7: Markov chain graphs for group 4 golden spikes



Finally, maximum entropy factors (mef) represent the ratio of the transition rate to the maximum entropy transition rate. A mef of 1.0 represents maximum disorder in depositional tendencies (Jones 2003; Environmental Modeling Systems 2004). A mef greater than 1.0 indicates that two sediments tend to occur next to each other and when the rate is less than unity then the opposite occurs (Environmental Modeling Systems 2004). In other words, for a mef of 1.0, the transition probability of single sediment to another is consistent with random distribution of the sediments. So, transition rate is dependent only on sediment proportions of these two materials (Jones 2003). This is an intuitive method of generating Markov chains because it enables logical incorporation of anisotropy into the model with the maximum entropy factors (Environmental Modeling Systems 2004; Jones 2003). In Table 4.8, high maximum entropy factors close to or greater than unity are; gravel to silt (1.256), sand to gravel (1.092), sand to silt (0.914), silt to gravel (0.962) and clay to sand (1.090) transitions. These transitions have high juxtaposition tendencies while the least transition is clay to gravel (0.190).

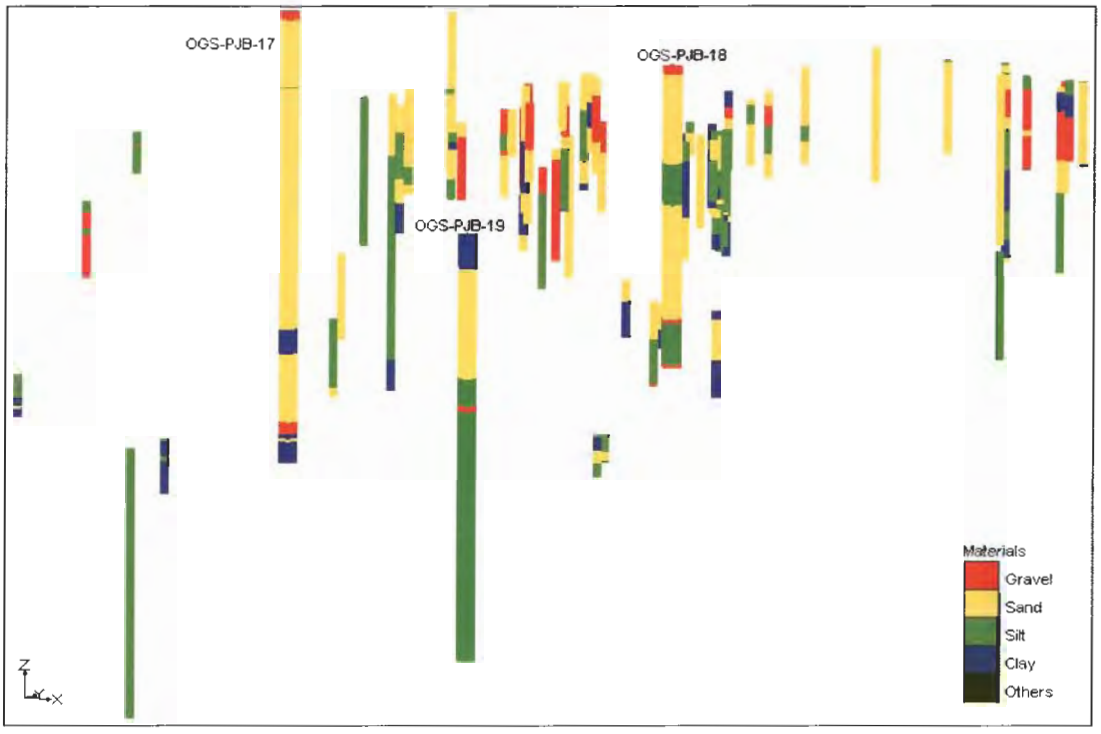
4.9.3 Sample Transition Probability Outputs for Golden spikes and MOEE Data

In the section above, golden spikes data are used to characterize the ORM subsurface in the vertical direction using T-PROG by generating transition probability simulations. This section outlines a common spatial extent for which both golden spikes and MOEE data coexist and are used to generate transition probability simulations. This is necessary in order to examine sediment transitions for the two data in order to identify comparable and dissimilar patterns.

A subset of group 4 golden spikes used in the above section (that is, section 4.9.2) is selected for this exercise. This sub-selection becomes necessary due to lack of spatial coordinates for MOEE data covering the entire group 4 golden spikes area. A combination of golden spikes and MOEE data chosen is shown in Figure 4.8. The widths of golden spikes are exaggerated relative to the MOEE data, but this does not affect sediment states and transitions. T-PROG simulations were run for golden spikes

and MOEE data separately. The outputs for transition probability results are shown in Tables 4.9 and 4.10 while Markov chain graphs are displayed in Appendices C25 and C26 respectively.

Figure 4.8: Spatial and sediment distribution of sample golden spikes and MOEE data



The tpm for the golden spikes are very distinct, that is, specific sediment transitions such as silt to clay, clay to sand, etc are consistently measured with unique values that separate them from other transitions. Silt to clay transition measured using transition rate, embedded transition probability and frequency and maximum entropy factor all indicate zero (0.000) (see Table 4.9). The converse of this transition is true; that is clay to silt transition exhibit the same properties. Hence, possible sediment transitions and juxtaposition tendencies are non-existent for silt to clay and clay to silt transitions. The opposite of this transition property is clay to sand. Clay to sand transition measured using transition rate, embedded transition probability and frequency and maximum entropy factor indicate 0.211, 1.000, 0.172 and 1.844

respectively (see Table 4.9). These values show that wherever clay occurs in this environment, the only sediment type to transit to will be sand. The clay to sand transition has 100% probability and shows the most prevalent juxtaposition trend. So clay to sand transition properties are unique based on measured tpm values. The other transitions exhibiting transition properties close to clay to sand are gravel to silt and silt to gravel. These observed transitions (that is, clay to silt or vice versa and clay to sand) show two extreme transition patterns which characterize vertical sediment distribution for this environment.

Table 4.9: Vertical T-PROG simulation output for sample Golden spikes

		Transition Rates				Embedded Transition Probabilities			
Material	Proportion	Gravel	Sand	Silt	Clay	Gravel	Sand	Silt	Clay
Gravel	0.028	-0.792	0.251	0.452	0.090	1.262	0.333	0.500	0.167
Sand	0.499	0.027	-0.072	0.018	0.027	0.285	15.647	0.338	0.377
Silt	0.399	0.022	0.033	-0.055	0.000	0.600	0.400	18.242	0.000
Clay	0.074	0.000	0.211	0.000	-0.211	0.000	1.000	0.000	4.730
		Embedded Transition Frequencies				Maximum Entropy Factors			
Material	Lens Length	Gravel	Sand	Silt	Clay	Gravel	Sand	Silt	Clay
Gravel	1.262	1.262	0.068	0.130	0.043	1.262	0.522	1.650	0.511
Sand	15.647	0.112	15.638	0.108	0.128	0.499	15.638	1.401	1.010
Silt	18.242	0.130	0.108	18.242	0.000	1.123	0.445	18.242	0.000
Clay	4.730	0.000	0.172	0.000	4.730	0.000	1.844	0.000	4.730
Background material: sand					lag: 0.3m				
Golden spikes: OGS-PJB-17, OGS-PJB-18, OGS-PJB-19.									

Table 4.10: Vertical T-PROG simulation output for sample MOEE data

		Transition Rates				Embedded Transition Probabilities			
Material	Proportion	Gravel	Sand	Silt	Clay	Gravel	Sand	Silt	Clay
Gravel	0.135	-0.088	0.059	0.018	0.011	11.420	0.600	0.250	0.150
Sand	0.423	0.018	-0.090	0.052	0.019	0.172	13.657	0.638	0.189
Silt	0.320	0.008	0.050	-0.083	0.024	0.140	0.488	12.117	0.372
Clay	0.123	0.012	0.114	0.016	-0.142	0.158	0.632	0.211	7.024
		Embedded Transition Frequencies				Maximum Entropy Factors			
Material	Lens Length	Gravel	Sand	Silt	Clay	Gravel	Sand	Silt	Clay
Gravel	11.420	11.420	0.073	0.040	0.024	11.420	0.880	0.709	0.743
Sand	13.657	0.065	14.013	0.235	0.051	0.661	14.013	0.523	0.755
Silt	12.117	0.048	0.131	12.117	0.128	0.804	0.555	12.117	1.463
Clay	7.024	0.024	0.147	0.032	7.024	0.759	0.568	0.370	7.024
Background material: sand					lag: 0.3m				

For the MOEE data, while there are similar transition patterns, important sediment transition conflicts persist. First, clay to sand transition is consistent with its observed pattern in golden spikes. The similarity in this transition is only limited to sediment sequences but not in juxtaposition trend. That is, transition rate, embedded transition probability and frequency have high (see Table 4.10) values (0.114, 0.632 and 0.147 respectively) enforcing a similar clay to sand transition sequence in golden spikes. But, entropy factor (0.568) is considerably below unity (less juxtaposition tendency) indicating that it is less likely to transit from clay to sand in this environment. This juxtaposition pattern is opposite to that observed for golden spikes.

Second, the only sediment transition which exhibits juxtaposition pattern is silt to clay transition with mef of 1.463. But, this transition (silt to clay) constantly indicate a zero (0.00) value for all tpm measures (that is, transition rate, embedded transition probability and frequency and mef) for golden spike simulation. Hence, silt to clay transition sequence in MOEE data in this environment is the most conflicting sediment transition. So to adjust sediment transition sequence, silt to clay transition should be the first to correct in order to simulate or replicate sediment transition pattern from golden spikes into MOEE data.

4.10 SEDIMENT DISPARITIES FOR GOLDEN SPIKES AND MOEE DATA

The preceding section characterized sediment variability in the vertical direction using state transitions of geologic materials. T-PROG simulations were used to describe material distribution for sample golden spikes and MOEE data and disparate sediment transition patterns are identified. Geologic material arrangement patterns described by T-PROG simulation in the subsurface environment are not the only characteristics hydrogeologist need. For example, while sediment transition sequence becomes vital input to determine depth limits for water wells during drilling processes, the exact depth and spatial information may remain unknown.

Also when adjusting sediment transition sequence in MOEE data to conform to observed pattern in golden spikes, there is no depth information to indicate where

specific transitions and sediments are likely to occur. For example, if sand to gravel transition has predominant juxtaposition trend in golden spikes then to replicate this pattern in MOEE data, one of these sediments (that is, either sand or gravel) must occur in the MOEE data. This becomes a necessary precondition to determine discrepancy patterns for different sediments in both data and such conditions may be difficult to accomplish. So, this section illustrates a means of identifying conflicting sediment types in golden spikes and MOEE data.

Figure 4.9: Spatial and sediment distribution for selected golden spikes (horizons between sediment contacts are in metres)

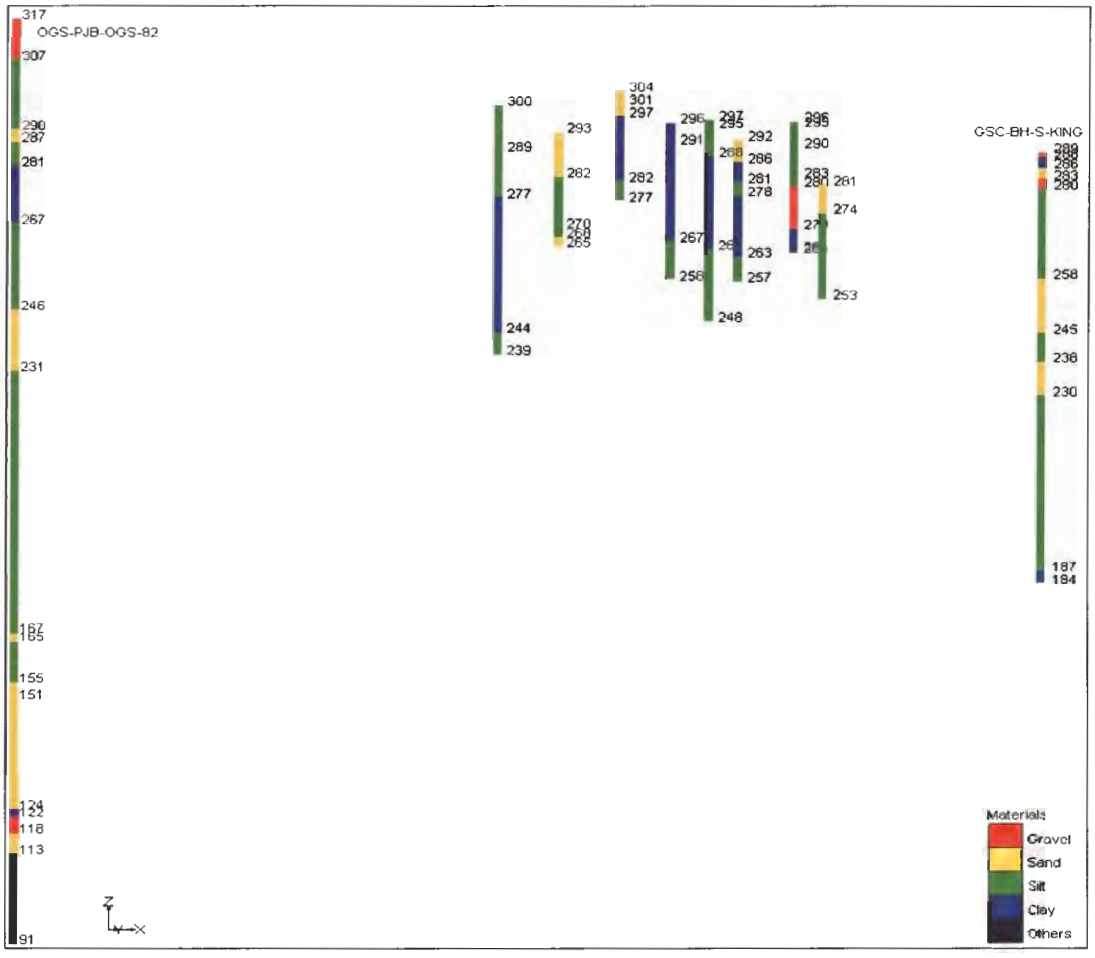
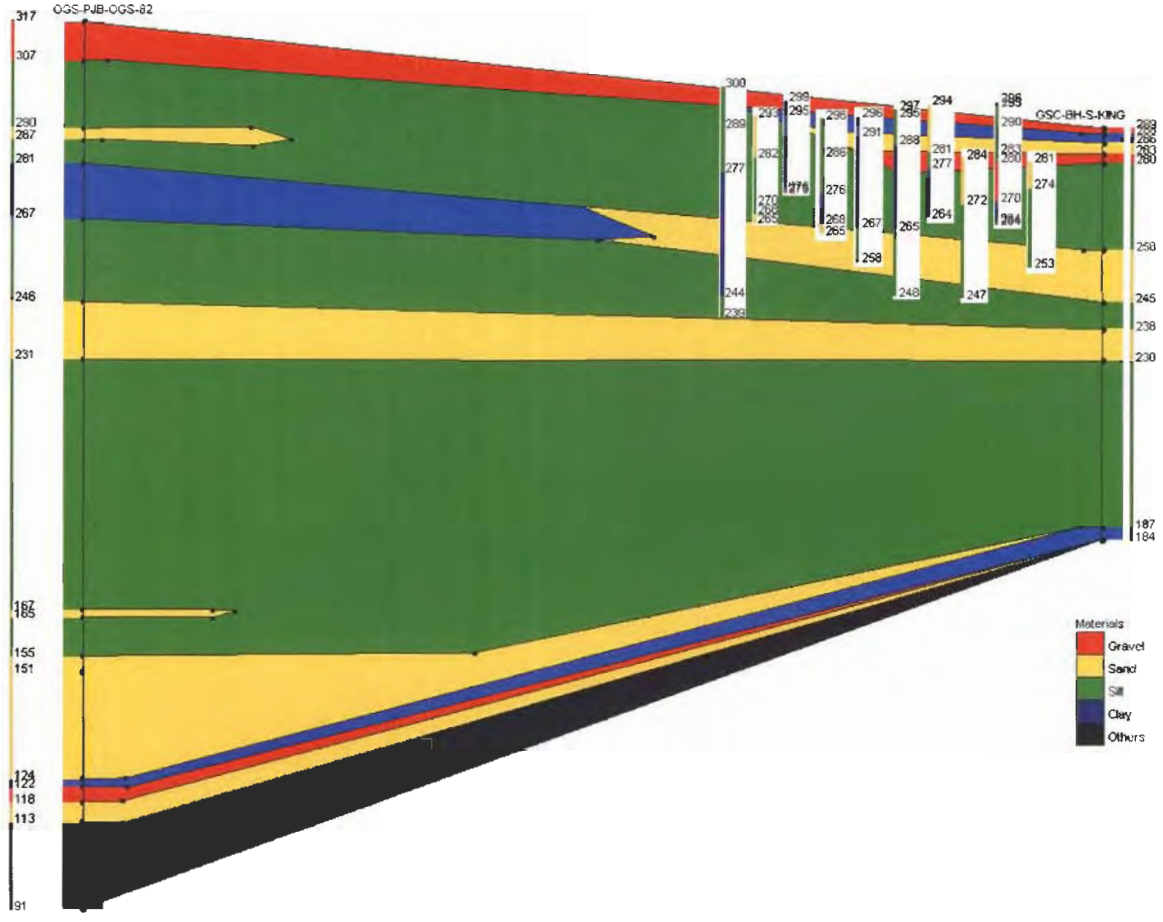


Figure 4.9 shows spatial and sediment distribution for sample golden spikes MOEE data. Figure 4.10 illustrates a cross section between golden spikes: OGS-PJB-OGS-82 and GSC-BH-S-KING. Boreholes from MOEE data which intercept this

golden spike profile (that is, cross-section) are also inserted. This illustrates sediment differences in golden spikes and MOEE data.

Clay is rare sediment in the golden spikes cross section where the MOEE boreholes occur. Hence, it is less likely to have significant clay occurrence in the MOEE data in this environment. The MOEE data should exhibit trace presence of gravel at depths between 317m and 288m. The lens length of sand in MOEE data is overstated compared to its occurrence in golden spikes. There are few occurrences of silt which are similar in both data but this is limited to only five MOEE boreholes and their lens widths vary considerably. Also the pattern of sediment states shown in GSC-BH-S-KING at the immediate surface (between 289m and 280m) is characterized with unconformity. This pattern is not represented in MOEE boreholes.

Figure 4.10: Sediment profile between two golden spikes and intercepted boreholes from MOEE data (horizons between sediment contacts are in metres)



This illustrates unique sediment differences between golden spikes and MOEE data. Sediment states in the MOEE data can be modified at specific depths in order to replicate sediment distribution pattern from golden spikes. Hence, this section provides a means of simulating sediment states from golden spikes into MOEE data.

In sum, rough sets use major aquifer properties (porosity, permeability, grain size and specific yield) to appropriately group borehole geologic units. The sediment grouping enables better classification of geologic units into different aquifer suitability levels. This sediment clusters were applied to MOEE data in order to assess sediment variation in the subsurface environment. Sediment variation assessment was, however, preceded with the assessment of GSC classification system which was used for standardizing the MOEE data. The GSC scheme was assessed using rough sets. The assessment results unveil significant extent of the GSC classification and provide a metadata equivalent for the MOEE borehole database. The metadata information such as: percent accuracy of borehole characterization, inconsistent sediments in each borehole, accuracy of sediment identification, etc provide relevant ingredients for uncertainty assessment.

Further, transition probability was used to simulate sediment state and transition for both golden spikes and MOEE data. This technique incorporates sediment variation from golden spikes into MOEE data. Both comparable and conflicting patterns were described in order to enhance sediment distribution in the MOEE data. This method showed great potential for sediment simulation from one borehole data into another. Apparent limitations, however, such as presence of consistent sediment requirements were also described.

CHAPTER 5:
CONCLUSIONS AND FURTHER WORK

The preceding two chapters have discussed data uncertainties, methods and outputs illustrating specific uncertainties in two data: census data and borehole data. This chapter brings together observed uncertainties in these two case studies and provides concluding remarks and recommendations for further work. The chapter also provides an overview for integrating these unique case studies into one research. The section below discusses research implications for integrating the two case studies.

5.1 INTEGRATING BOTH CASE STUDIES

The research techniques and the data used in this study raised many questions. One question is: why use data that are different in many ways. For example, census data are used to measure socio-economic and demographic information, and borehole data are used to characterize subsurface geology. There are many factors which influenced the choice of data and their subsequent uncertainties - which are examined in this study.

First, uncertainties in both case studies are different and they reflect their respective data distribution. Census data are different from borehole data, so also are their subsequent uncertainties. The scale problem in census data is different from erroneous sediment description in borehole data. These differences provide a computational opportunity to assess different uncertainties with one analysis tool (i.e. rough sets). Second, data availability also affected the research technique. The data used in this study were made accessible by research communities who are ready to provide data for research without cost.

Third, most methods for assessing uncertainty are plagued with the tendency to examine problems in a specific data or area of application. Fuzzy sets, for example, have experienced huge applications in environmental and physical systems (Lagacherie

et al. 1996; Fortin and Edwards 2001; Allen et al. 2002; Carranza and Hale 2001; Dragicevic et al. 2001) but little applications in human systems. One may find many articles on uncertainty but they are often specific to particular data property or analytical process. For example, Warren et al. (2003) developed a technique for representing and propagating uncertainty through predictive model for species richness in order to estimate different levels of biodiversity. Their method was unique and may be useful for ecosystem models. But a broad perspective of geographic uncertainty is needed to enhance integrity of geographic analysis. For example, Goodchild (1989) discussed 'modeling error in objects and fields'. He described a generic technique that is sensitive to the nature of data and its uncertainty when minimizing error. This method provides a broad perspective for data uncertainty and accounts for basic properties of data which may be subject to uncertainty. Hence, the use of rough sets in this study was to provide a broad technique for reducing uncertainty that is sensitive to data characteristics and their uncertainties.

However, rough sets have not gained popularity in GIScience partly because many researchers tend to focus narrowly on specific data or area of application. In this study, the utility of rough sets was applied to reduce uncertainty in census data and borehole data. The technique was not only to minimize uncertainty in these data but to demonstrate rough sets as a knowledge base tool for characterizing uncertainty irrespective of the data or area of application under study. Hence, this research have show-cased the use of rough sets as a broad uncertainty characterization tool. A brief outline of uncertainties in each case study is described below.

The study has described uncertainties which plagued census data and borehole data. It recognized that these uncertainties result in marginal data quality and erroneous geographic information. These uncertainties threaten the integrity of GIScience applications, but this study provided a technique for reducing the effects of uncertainty in both data independently. For the first case study, rough sets enhanced spatial characterization of the relationship between recent immigrants and deprivation indices by mitigating the scale problem. The rough sets technique used can be applied

to any spatially grouped data for which scale distortion is encountered. Hence, for census data, the effects or extents of uncertainty resultant from scale issue are minimized using rough sets.

In the second case study, rough sets and transition probability were used to assess GSC classification scheme. The techniques replicate sediment variation by enhancing geologic understanding in the subsurface environment. This case study enriched borehole data of marginal quality by providing metadata information and integrating accurate sediment distribution. Hence, for borehole data, the implications of uncertainty ensuing from erroneous sediment identification and description were reduced using rough sets and transition probability.

The following sections discuss research findings and conclusions separately for both case studies in: census data and borehole data.

5.2 CONCLUSIONS AND FURTHER RESEARCH: FIRST CASE STUDY

The study has identified problems associated with common spatial analysis tools and illustrated the utility of rough sets in spatial data analysis. Common geospatial data characteristics such as spatial dependency, spatial autocorrelation and heterogeneity of geographic phenomena confound most analytical tools and discourage representation of data distributions into derived models. The study used rough sets to accommodate different data characteristics. So, outputs from the rough sets method showed the preservation of data distribution during spatial analysis. This ensured relative accuracy and reduced excess discrepancies between data distribution and model characteristics. Appropriate treatment of conflict within input data variables and convenient representation of multiple data distribution were recognized as key merits in rough sets analysis. The rough sets tool did not improve nor create new data, but its worth is in reasoning with data to develop models which replicate single or multiple data characteristics. These properties enhance spatial integrity in attribute space.

Next, the scale dimension of MAUP has been explored to estimate scale transition parameters in order to translate data characteristics across different scales. Different patterns of data distribution have been examined at multiple census scales: CSD, CT and DA. Scale transitions from large resolution data to small areas are necessary because policy decisions are often made for small areas using large resolution data. Accuracy indicators were also computed to describe the thresholds within which these census data estimates can be applied satisfactorily across scales.

Finally, the spatial relationships of recent immigrant and deprivation index were derived and were characterized with model strength and magnitude. The strength of recent immigrant and deprivation index relationships computed at CSD, CT and DA levels were 0.600, 0.549 and 0.468 respectively. The variation in the strength of this relationship at the various census scales showed the erroneous conclusions which can be reached based on any specific census scale used. The steady decline in the strength of this relationship showed that census household or individual level data would have a smaller correlation index. While the DA model cannot be used to uniquely describe reality (i.e. individual level pattern), the individual level model is not far from it. This quantitative method of neighbourhood characterization may not be recognized for its validity of using indicators to describe human phenomena. However, for quantitative analysis of spatial aggregate data, rough sets minimized the scale problem and better characterized spatial relationships.

A number of questions, however, remain unanswered and require further investigation. Census household level data estimates should be analyzed with known data to validate the actual residual and evaluate the rough sets approximation process. Census data used in this study is subject to a single aggregation pattern homogenized for specified area units, and it is apparent that different census grouping will unearth different data patterns. Subsequent studies should extend this analysis to disaggregate data at different census resolutions to explore the aggregation pattern over multiple resolutions and also to compute transformation parameters to translate between different scales. Finally, the rough sets technique estimated deprivation levels for small

census areas using large resolution census data, but the spatial specifications (or location) for these small areas remained unknown.

5.2.1 Research Contributions: First Case Study

Research findings observed in this study have both practical and theoretical relevance. Research contributions for the first case study are itemized below:

- Data distributions for multiple variables were retained during spatial analysis. Rough sets provided a means of translating socio-economic characteristics across different census scales: CSD, CT and DA. This sequence of analysis may be applied in other research areas, for example, population health studies. First, to harmonize nonlinear patterns among multiple variables, which may characterize the same health neighbourhood differently. Second, to describe health scenarios at different area units. In other words, population health analyst may assess the effects of data aggregation on their analysis.
- Openshaw (1984a) suggested that accuracy of parameter estimates be computed in order to examine the effects of data aggregation. The first case study has computed accuracy measures for different levels of census aggregation (CSD, CT and DA). Estimated scale sensitivity measure enhanced spatial analysis operation with minimum scale distortions. Rough sets method provided a unique technique of simulating topological relationships for attribute data so we may know the amount of distortion introduced by data aggregation.
- Considering the city of Burnaby, the relationship between recent immigrants and deprivation index for different census scales: CT (0.5405) and DA (0.5590) have very close outputs. Hence, for compact study areas characterized with homogeneity and randomness, spatial relationships at different census units (e.g. CT, DA) can be identical. This could minimize the uncertainty of choosing which census units (or scales) to characterize deprivation levels. So, regarding the choice of census scale for particular policy implementation,

there is a limiting threshold within which derived models could yield very close results in order to inform census scale specification.

The section below discusses conclusions and future work for the second case study.

5.3 CONCLUSIONS AND FURTHER WORK: SECOND CASE STUDY

The effects of uncertainty examined for the second case study were aimed at enhancing the use of borehole data of marginal quality for accurate geological inquiry. The sequence of methods applied on borehole data addressed sediment variation and erroneous description problems.

First, the GSC standardization scheme was assessed using MOEE data. The assessment examined sediment variation as perceived by private well drillers (pwd) in MOEE data. This sediment distribution pattern was examined against outputs from the GSC classification system. The assessment identified sediments with high percentage error (e.g. silt) and computed accuracy measures for sediment descriptions within each borehole.

Second, transition probability simulation characterized sediment distribution in the subsurface using both golden spikes and MOEE data separately. The method compared sediment state and transition patterns in golden spikes and MOEE data in the vertical direction. The T-PROG simulation was a further step for enhancing and validating MOEE data beyond the GSC standardization scheme assessment. For example, silt occurred as the most abundant sediment in error when standardized by GSC scheme. Clay, on the other hand, occurred as the overall predominant sediment in the MOEE data. But, the GSC scheme has reduced its occurrence of error, and it was the most standardized sediment. From T-PROG simulation, clay and silt were identified for exhibiting the most conflicting sediment state and transition pattern compared with other sediment distributions in golden spikes. Hence, though clay had less error during GSC scheme assessment, its significant occurrence in error was exposed through transition probability simulation.

Third, T-PROG simulated sediment distributions in the subsurface, but specific depth and spatial information where these sediment state and transitions occurred remained unknown. So, this final stage showed a simple profile between two golden spikes. Boreholes in the MOEE data which intersected the golden spike profile were examined to estimate depth and spatial information where sediment differences occurred.

Further questions, however, remain for future investigation. Fuzzy sets should be used to describe gradual sediment state and transition. But, specific data on sediment transition which describe boundary properties are needed. Further, sediment state and transition simulation were focused in the vertical direction. Sediment distribution pattern should also be examined for the horizontal direction. Also sediment differences observed were sampled from both golden spikes and MOEE data. The same approach should be extended for other golden spike sub-clusters and MOEE data. Sediment distributions cannot be replicated from one golden spike cluster to another, but specific sediment states and transitions which are consistently in error could be identified.

5.3.1 Research Contributions: Second Case Study

In the second case study, which focused on sediment identification and description problems in borehole data, below are its research contributions:

- Outputs from GSC standardization assessment and sediment variability should support training programs for private well drillers and enhance borehole data quality. The output for the GSC scheme assessment provided a metadata equivalent for the MOEE database.
- Accuracy measures, which indicate the reliability of sediment distribution for specific boreholes, should form part of the GSC or MOEE borehole database, so researchers may quantify the level of uncertainty when using this data.
- Sediment state and transitions simulated using transition probability are valuable for estimating depth information for water wells in order to reduce

the cost of drilling and provide productive water wells. The methods applied on borehole data represent a unique approach to enhance data of marginal quality using high quality data.

To conclude, the analysis process for both census and borehole data emphasize the need to develop knowledge base techniques uniquely for different uncertainties, however these methods should be designed to resist distortions in scale and data distributions.

5.4 FINAL CONCLUSIONS

The sources and effects of uncertainty have been examined in two disparate data: census data and borehole data. Scale issues in census data and sediment identification and description problems in MOEE borehole data have linked these data under a single umbrella of uncertainty. Hence, this study focused on providing tools to reduce the effect of these uncertainties in order to enhance geographic inquiry.

In census data, applying rough sets to spatial analysis has provided a scale sensitivity measure to translate geographic relationships over multiple census scales. So while the rough set tool cannot eradicate uncertainty during spatial transition, it does provide accuracy thresholds within which rough sets estimates apply. It also enhances data distribution retention across census scales.

In borehole data, rough sets enabled sediment grouping using aquifer-supporting properties. These sediment clusters highly facilitated the GSC standardization scheme assessment and T-PROG simulation. The sequence of techniques employed enhanced the quality of MOEE data in accurate geological inquiry. The utility of rough sets and transition probability is not limited to ORM southern Ontario alone, but also any aquifer with borehole data of marginal quality. However, the ORM provided a unique opportunity in order to enrich data of marginal quality from high quality data.

APPENDIX A

Apendix A1: Descriptive values of porosity for a range of geological materials

Material	Porosity (per cent)	Material	Porosity (per cent)
Coarse gravel	28	Loess	49
Medium gravel	32	Peat	92
Fine gravel	34	Schist	38
Coarse sand	39	Siltstone	35
Medium sand	39	Claystone	43
Fine sand	43	Shale	6
Silt	46	Till – mainly sand	31
Fine-grained sandstone	33	Till – mainly silt	34
Clay	42	Tuff	41
Medium grained sandstone	37	Basalt	17
Limestone	30	Gabbro (weathered)	43
Dolomite	26	Granite (weathered)	45
Dune sand	45		
Adapted from Water Supply Paper 1839-D by permission of the United States Geological Survey (Brassington 1988, p53)			

Apendix A2: Descriptive values of specific yield for a range of geological materials

Material	Specific Yield (per cent)	Material	Specific Yield (per cent)
Coarse gravel	23	Limestone	14
Medium gravel	24	Dune sand	38
Fine gravel	25	Loess	18
Coarse sand	27	Peat	44
Medium sand	28	Schist	26
Fine sand	23	Siltstone	12
Silt	8	Till – mainly silt	6
Clay	3	Till – mainly sand	16
Fine-grained sandstone	21	Till – mainly gravel	16
Medium grained sandstone	27	Tuff	21
Adapted from Water Supply Paper 1662-D by permission of the United States Geological Survey (Brassington 1988, p53)			

Apendix A3: List of descriptive porosities and hydraulic conductivities for unconsolidated sediments and rocks

Geological Material	Grain size (mm)	Porosity	Hydraulic conductivity, K (metres per day)
Unconsolidated Sediments			
Clay	0.0005 – 0.002	45 – 60	$< 10^{-2}$
Silt	0.002 – 0.06	40 – 50	$10^{-2} – 1.0$
Alluvial sands	0.06 – 2.0	30 – 40	1.0 – 500
Alluvial gravels	2.0 – 64	25 – 35	500 – 10 000
Consolidated Sedimentary Rocks			
Shale	Small	5 – 15	$5 \times 10^{-8} – 5 \times 10^{-6}$
Sandstone	Medium	5 – 30	$10^{-4} – 10$
Limestone	Variable	0.1 – 30	$10^{-5} – 10$
Igneous and Metamorphic Rocks			
Basalt	Small	0.001 – 1	0.0003 – 3
Granite	Large	0.0001 – 1	0.0003 – 3
Slate	Small	0.001 – 1	$10^{-8} – 10^{-5}$
Schist	Medium	0.001 – 1	$10^{-7} – 10^{-4}$
Reproduced from S248 by permission of the Open University (Brassington 1988, p56)			

Apendix A4: Hydraulic conductivities in metres/day for various rocks

Hydraulic Conductivity in m/d								
10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-1}	1	10	10^{2}	10^{3} 10^{4}
Relative Hydraulic Conductivity								
Very low			Low			Moderate		High Very high
Represented Materials Unconsolidated deposits								
Massive clay		Silt, clay and mixtures of sand, silt and clay		Fine sand		Clean sand & sand & gravel		Clean gravel
Consolidated Rocks								
Massive igneous & metamorphic rocks		Laminated sandstone, shale & mudstone		Clean sandstone & fractured igneous & metamorphic rocks		Vesicular & scoriaceous basalt & cavernous limestone & dolomite		
Adapted from the Groundwater Manual by permission of the United States Department of the Interior (Brassington 1988, p56)								

Apendix A5: Geologic units present in MOEEE data identified by the GSC

GSC_mat_code	Description
1	Bedrock
10	Fill
11	Covered
1-1	Limestone
1-2	Shale
1-3	Granite
1-4	Dolomite
1-5	Pot_Bedrock
1-6	Sandstone
1-7	Limestone_Shale_Inter
2	Sand_Diamicton
3	Silt_Diamicton
4	Clay_Diamicton
5	Gravel
6	Sand
7	Silt
8	Clay
9	Organic
99	Unknown

Apendix A6: ORM geologic unit approximation into aquifer supporting groups

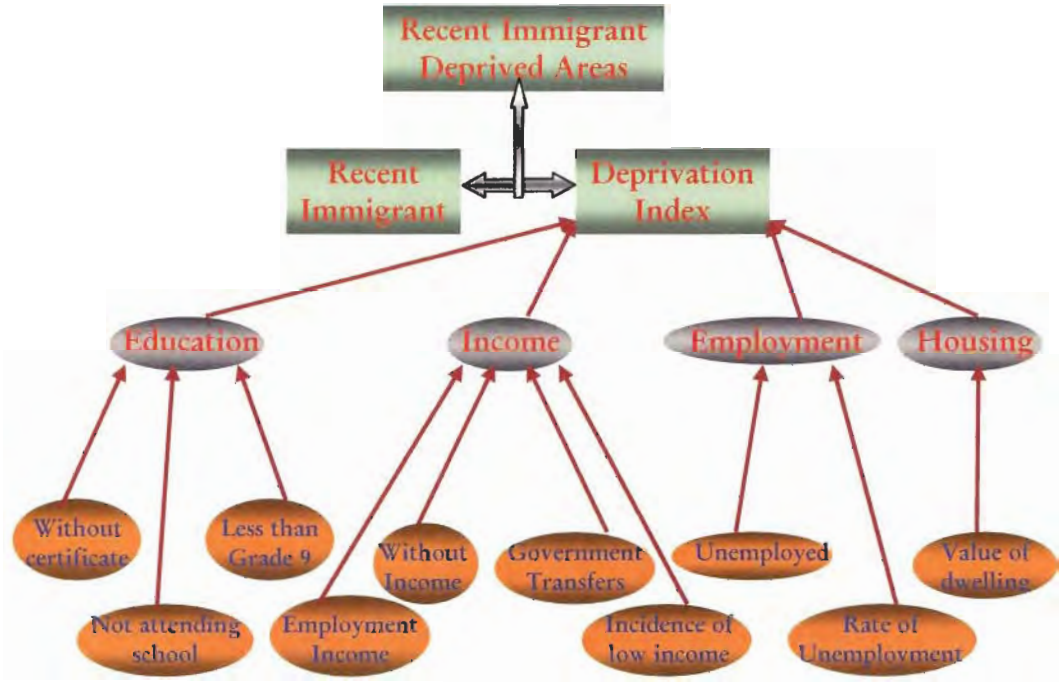
Set Category	Modified Upper Approximation set	Lower Approximation set
Very Good Aquifer Indicators (VGAI)	Gravel	
Good Aquifer Indicators (GAI)	Sand_Diamicton	Sand
Moderate Aquifer Indicators (MAI)	Silt_Diamicton	Silt
Poor Aquifer Indicators (PAI)	Fill	Clay
	Clay_Diamicton	
Non–Aquifer Indicators (NAI)	Organic	Bedrock, Pot_Bedrock
		Limestone, Shale
		Sandstone
		Limestone_Shale_Inter
		Granite, Dolomite

APPENDIX B

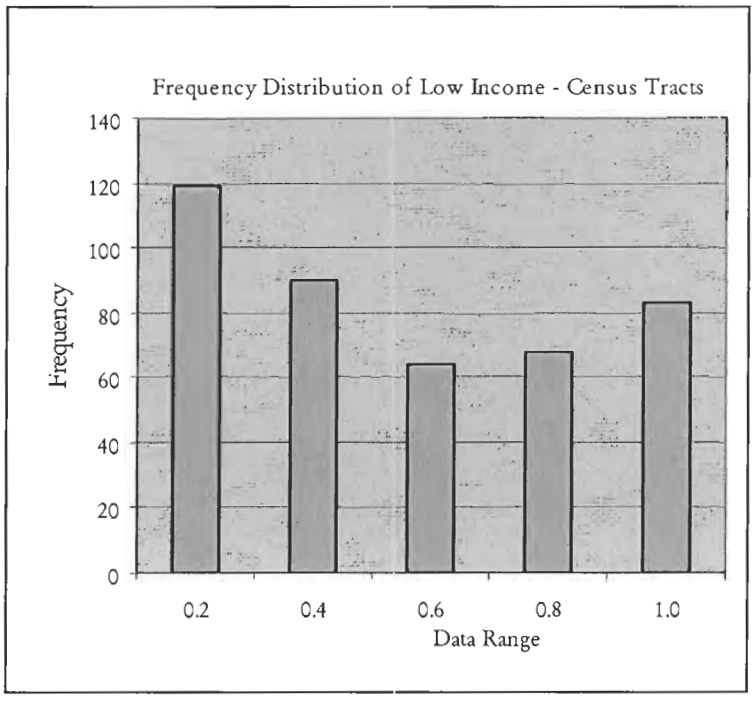
Apendix B1: **Individual variable constituents of major deprivation categories**

Deprivation Indicator Categories			
Education	Employment	Housing	Income
Percent of 15 to 24 years population not attending school	Unemployed population 15 years and over by labour force activity %	Average value of dwelling $	Employment income %
Population 20 years and over by highest level of schooling–less than grade 9(%)	Unemployment rate		Government transfer payments %
Without high school graduation certificate	Unemployed population 25 years and over by labour force activity %		Population 15 years and over without income %
	Unemployment rate		Incidence of low income in 2000 %

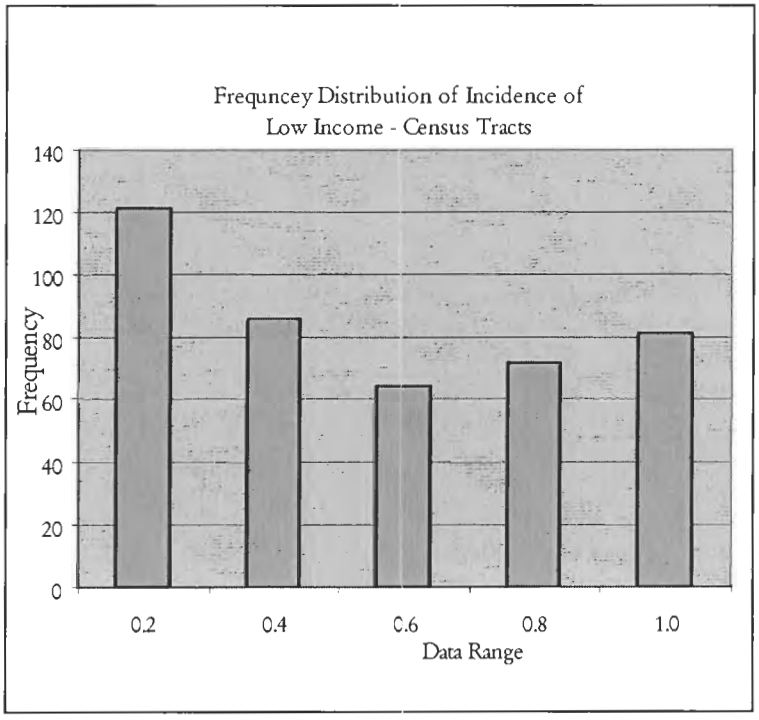
Apendix B2: **Deprivation index derivation process with no assumption of spatial dependency on Recent immigrants**



Apendix B3: **Frequency distribution chart for low income**



Apendix B4: **Frequency distribution chart for Incidence of low income**



Apendix B5: **Frequency distribution table for Low Income**

Data Range	Frequency
0.00 – 0.20	119
0.21 – 0.40	90
0.41 – 0.60	64
0.61 – 0.80	68
0.81 – 1.00	83
Total	424

Apendix B6: **Frequency distribution table for Incidence of Low Income**

Data Range	Frequency
0.00 – 0.20	121
0.21 – 0.40	86
0.41 – 0.60	64
0.61 – 0.80	72
0.81 – 1.00	81
Total	424

Apendix B7: **Sample deprivation indicator values at DA resolution grouped into CT for set approximations**

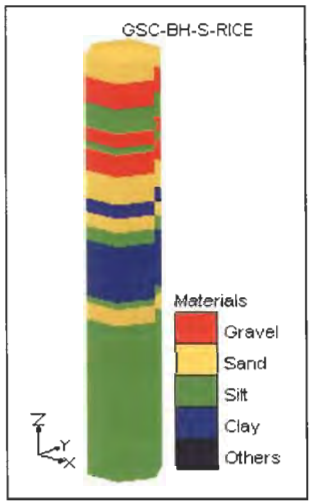
DAUID[5]	CTUID[6]	Education	Average/ Median Education	Employment	Average/ Median Employment	Housing	Average/ Median Housing	Income	Average/ Median Income
0015	0132.00	0.261	0.271/ 0.261	0.626	0.493/ 0.602	0.857	0.787/ 0.857	0.438	0.465/ 0.438
0016		0.139		0.251		0.967		0.269	
0017		0.412		0.602		0.536		0.687	
0004	0133.01	0.175	0.117/ 0.131	0.563	0.288/ 0.208	0.652	0.805/ 0.792	0.278	0.324/ 0.310
0006		0.021		0.172		0.715		0.340	
0007		0.087		0.195		0.869		0.400	
0008		0.183		0.220		0.985		0.279	
0009	0133.02	0.645	0.477/ 0.477	0.167	0.287/ 0.287	0.236	0.411/ 0.411	0.742	0.492/ 0.492
0011		0.309		0.407		0.586		0.242	
0001	0250.02	0.190	0.274/ 0.190	0.365	0.434/ 0.365	0.906	0.895/ 0.906	0.246	0.267/ 0.266
0002		0.482		0.665		0.935		0.266	
0003		0.150		0.272		0.845		0.290	

[5] DAUID is a unique identifier for a DA
[6] CTUID is a unique identifier for a CT

APPENDIX C

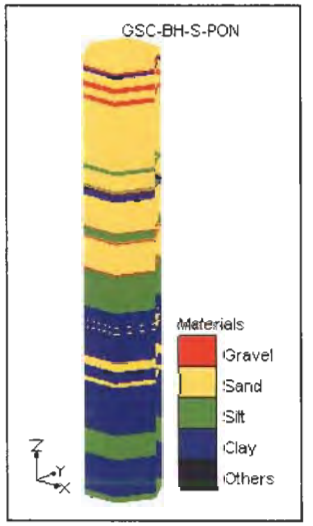
Appendix C1: Vertical T-PROG simulation output and borehole plot for Group 1 golden spike

			Transition Rates				Embedded Transition Probabilities			
Material	Proportion	Lens Length	Gravel	Sand	Silt	Clay	Gravel	Sand	Silt	Clay
Gravel	0.148	8.833	-0.113	0.038	0.075	0.000	8.833	0.333	0.667	0.000
Sand	0.189	8.500	0.030	-0.118	0.058	0.030	0.250	8.500	0.500	0.250
Silt	0.497	17.873	0.022	0.022	-0.055	0.011	0.400	0.401	17.873	0.198
Clay	0.166	15.000	0.000	0.034	0.033	-0.067	0.000	0.500	0.500	15.000
			Embedded Transition Frequencies				Maximum Entropy Factors			
Material			Gravel	Sand	Silt	Clay	Gravel	Sand	Silt	Clay
Gravel			8.833	0.084	0.139	0.000	8.833	1.017	1.240	0.000
Sand			0.084	8.500	0.129	0.084	1.022	8.500	0.829	1.672
Silt			0.139	0.129	20.021	0.064	1.245	0.433	20.021	1.063
Clay			0.000	0.084	0.064	15.000	0.000	1.670	1.032	15.000
Background material: silt			lag: 0.3m		Golden spikes: GSC-BH-S-RICE					



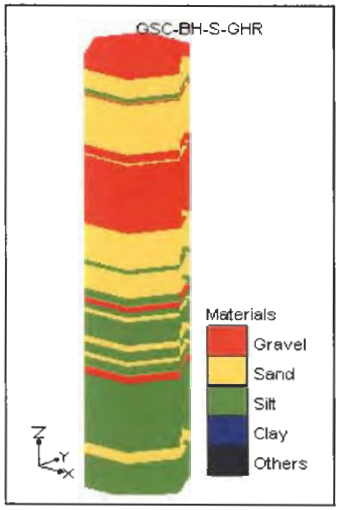
Appendix C2: Vertical T-PROG simulation output and borehole plot for Group 2a golden spike

			Transition Rates				Embedded Transition Probabilities			
Material	Proportion	Lens Length	Gravel	Sand	Silt	Clay	Gravel	Sand	Silt	Clay
Gravel	0.044	1.029	-0.972	0.447	0.251	0.275	1.029	0.429	0.286	0.286
Sand	0.411	6.751	0.075	-0.161	0.030	0.056	0.476	6.751	0.182	0.341
Silt	0.187	5.300	0.066	0.059	-0.189	0.064	0.400	0.200	5.300	0.400
Clay	0.358	7.588	0.000	0.099	0.033	-0.132	0.000	0.750	0.250	7.588
			Embedded Transition Frequencies				Maximum Entropy Factors			
Material			Gravel	Sand	Silt	Clay	Gravel	Sand	Silt	Clay
Gravel			1.029	0.102	0.061	0.061	1.029	0.793	1.321	0.990
Sand			0.163	6.147	0.061	0.123	1.314	6.147	0.738	0.806
Silt			0.061	0.061	5.300	0.061	1.456	0.348	5.300	1.250
Clay			0.000	0.184	0.061	7.588	0.000	1.431	1.226	7.588
Background material: sand			lag:0.3m		Golden spikes: GSC-BN-S-PON					



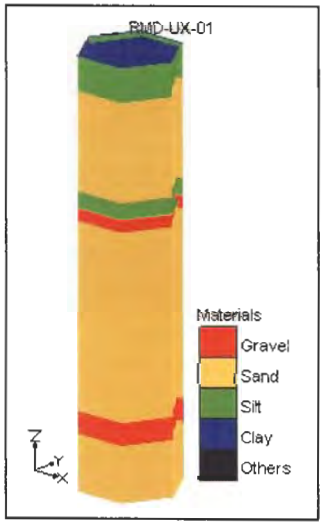
Appendix C3: Vertical T-PROG simulation output and borehole plot for Group 2b golden spike

			Transition Rates				Embedded Transition Probabilities			
Material	Proportion	Lens Length	Gravel	Sand	Silt	Clay	Gravel	Sand	Silt	Clay
Gravel	0.246	5.637	-0.177	0.122	0.056		5.637	0.667	0.333	
Sand	0.357	5.434	0.042	-0.184	0.142		0.222	5.434	0.778	
Silt	0.397	6.054	0.072	0.090	-0.163		0.442	0.558	6.054	
Clay										
			Embedded Transition Frequencies				Maximum Entropy Factors			
Material			Gravel	Sand	Silt	Clay	Gravel	Sand	Silt	Clay
Gravel			5.637	0.176	0.076		5.637	1.372	0.664	
Sand			0.088	5.434	0.292		0.682	5.434	1.179	
Silt			0.164	0.204	6.241		1.050	0.840	6.241	
Clay										
Background material: silt			lag: 0.3m		Golden spikes: GSC-BH-S-GHR					



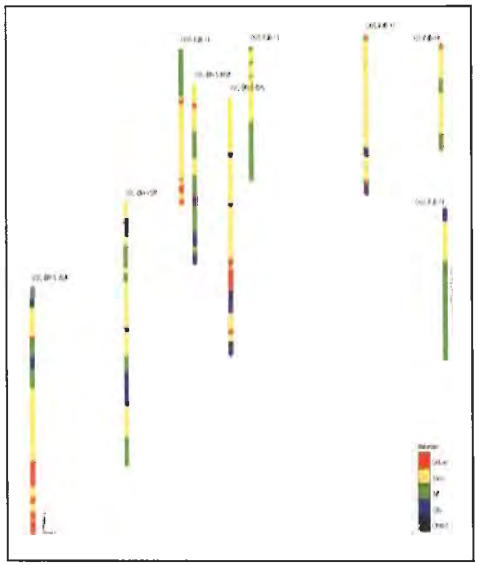
Appendix C4: Vertical T-PROG simulation output and borehole plot for Group 3 golden spike

			Transition Rates				Embedded Transition Probabilities			
Material	Proportion	Lens Length	Gravel	Sand	Silt	Clay	Gravel	Sand	Silt	Clay
Gravel	0.082	3.400	-0.294	0.294	0.000	0.000	3.400	1.000	0.000	0.000
Sand	0.795	22.068	0.016	-0.052	0.017	0.020	0.346	22.068	0.218	0.436
Silt	0.118	5.000	0.100	0.100	-0.200	0.000	0.500	0.500	5.000	0.000
Clay	0.005	0.300	0.000	1.197	2.136	-3.333	0.000	0.000	1.000	0.300
			Embedded Transition Frequencies				Maximum Entropy Factors			
Material			Gravel	Sand	Silt	Clay	Gravel	Sand	Silt	Clay
Gravel			3.400	0.246	0.000	0.000	3.400	1.698	0.000	0.000
Sand			0.102	22.670	0.095	0.159	0.830	22.670	0.863	0.000
Silt			0.144	0.095	5.000	0.000	1.841	0.883	5.000	0.000
Clay			0.000	0.015	0.144	0.300	0.000	0.000	2.698	0.300
Background material: sand			lag: 0.3m		Golden spikes: RMD-UX-01					



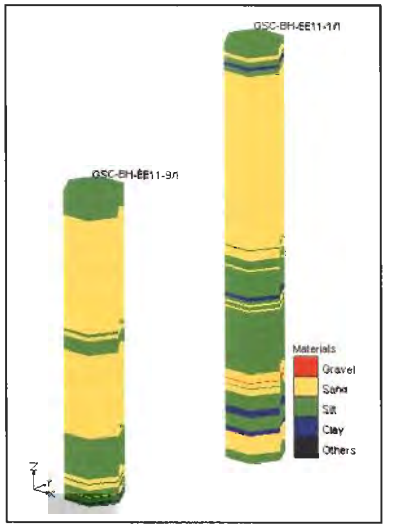
Appendix C5: Vertical T-PROG simulation output and borehole plot for Group 4 golden spikes

			Transition Rates				Embedded Transition Probabilities			
Material	Proportion	Lens Length	Gravel	Sand	Silt	Clay	Gravel	Sand	Silt	Clay
Gravel	0.068	3.210	-0.312	0.169	0.092	0.051	3.210	0.571	0.238	0.190
Sand	0.538	12.762	0.029	-0.084	0.028	0.028	0.346	12.762	0.362	0.292
Silt	0.296	11.609	0.016	0.055	-0.086	0.015	0.217	0.522	11.609	0.261
Clay	0.098	4.265	0.008	0.181	0.045	-0.234	0.045	0.727	0.227	4.265
			Embedded Transition Frequencies				Maximum Entropy Factors			
Material			Gravel	Sand	Silt	Clay	Gravel	Sand	Silt	Clay
Gravel			3.210	0.106	0.046	0.036	3.210	0.889	1.256	0.797
Sand			0.133	12.503	0.135	0.113	1.092	12.503	0.914	0.874
Silt			0.046	0.126	11.609	0.055	0.962	0.578	11.609	0.833
Clay			0.009	0.150	0.046	4.265	0.190	1.090	0.795	4.265
Background material: sand					lag: 0.3m					
Golden spikes: GSC-BH-VSR, OGS-PJB-14, OGS-PJB-15, OGS-PJB-16, OGS-PJB-17, OGS-PJB-18, OGS-PJB-19, GSC-BH-S-AUR, GSC-BH-S-BAL, GSC-BH-S-MSR										



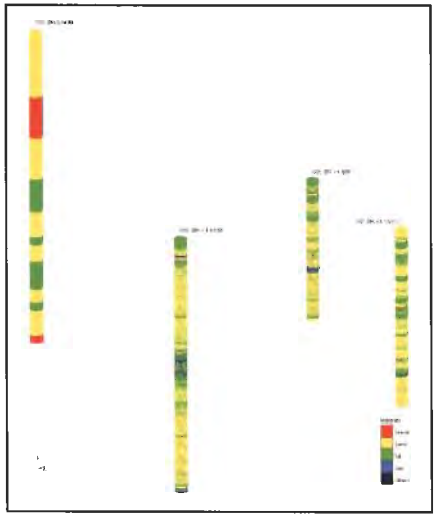
Appendix C6: Vertical T-PROG simulation output and borehole plot for Group 5 golden spikes

			Transition Rates				Embedded Transition Probabilities			
Material	Proportion	Lens Length	Gravel	Sand	Silt	Clay	Gravel	Sand	Silt	Clay
Gravel	0.002	0.130	-7.692	7.692	0.000	0.000	0.130	1.000	0.000	0.000
Sand	0.593	3.858	0.021	-0.276	0.237	0.017	0.081	3.858	0.876	0.044
Silt	0.373	2.081	0.000	0.364	-0.481	0.117	0.000	0.737	2.081	0.263
Clay	0.032	0.598	0.000	0.472	1.200	-1.671	0.000	0.167	0.833	0.598
			Embedded Transition Frequencies				Maximum Entropy Factors			
Material			Gravel	Sand	Silt	Clay	Gravel	Sand	Silt	Clay
Gravel			0.130	0.031	0.000	0.000	0.130	5.735	0.000	0.000
Sand			0.031	3.934	0.333	0.016	7.082	3.934	0.975	0.640
Silt			0.000	0.333	2.081	0.120	0.000	0.887	2.081	0.937
Clay			0.000	0.016	0.120	0.598	0.000	1.187	0.832	0.598
Background material: sand					lag: 0.3m					
Golden spikes: GSC-BH-EE11-1/1, GSC-BH-EE11-9/1										



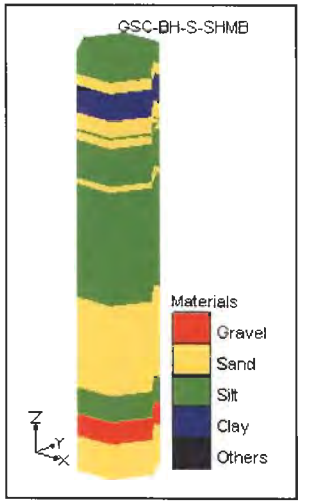
Appendix C7: Vertical T-PROG simulation output and borehole plot for Group 6a golden spikes

			Transition Rates				Embedded Transition Probabilities			
Material	Proportion	Lens Length	Gravel	Sand	Silt	Clay	Gravel	Sand	Silt	Clay
Gravel	0.040	9.840	-0.102	0.068	0.033	0.000	9.840	0.500	0.500	0.000
Sand	0.650	1.478	0.006	-0.759	0.515	0.237	0.009	1.478	0.449	0.541
Silt	0.284	0.663	0.000	0.835	-1.507	0.672	0.000	0.750	0.663	0.250
Clay	0.027	0.078	0.000	9.358	3.391	-12.749	0.000	0.338	0.662	0.078
			Embedded Transition Frequencies				Maximum Entropy Factors			
Material			Gravel	Sand	Silt	Clay	Gravel	Sand	Silt	Clay
Gravel			9.840	0.002	0.001	0.000	9.840	1.051	0.766	0.000
Sand			0.003	1.319	0.214	0.170	2.028	1.319	2.798	0.708
Silt			0.000	0.236	0.663	0.101	0.054	1.501	0.663	0.948
Clay			0.000	0.151	0.121	0.078	0.000	0.303	0.456	0.078
Background material: sand					lag: 0.3m					
Golden spikes: GSC-BH-S-NOBL, GSC-BH-S-VGHN, GSC-BH-V4-18/110, GSC-BH-V4-4/158, GSC-BH-V4-9/86										



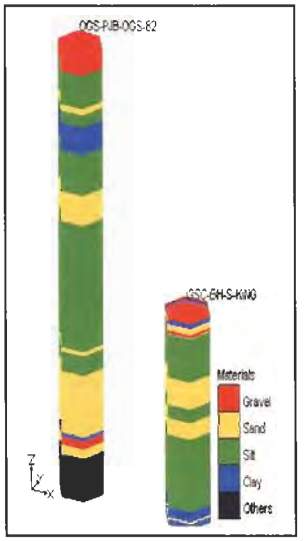
Appendix C8: Vertical T-PROG simulation output and borehole plot for Group 5 golden spikes

			Transition Rates				Embedded Transition Probabilities			
Material	Proportion	Lens Length	Gravel	Sand	Silt	Clay	Gravel	Sand	Silt	Clay
Gravel	0.052	8.000	-0.125	0.125	0.000	0.000	8.000	1.000	0.000	0.000
Sand	0.398	10.167	0.000	-0.098	0.082	0.016	0.000	10.167	0.800	0.200
Silt	0.484	15.436	0.014	0.054	-0.068	0.000	0.209	0.832	15.436	-0.041
Clay	0.065	10.000	0.000	0.100	0.000	-0.100	0.000	1.000	0.000	10.000
			Embedded Transition Frequencies				Maximum Entropy Factors			
Material			Gravel	Sand	Silt	Clay	Gravel	Sand	Silt	Clay
Gravel			8.000	0.083	-0.005	0.000	8.000	1.371	0.000	0.000
Sand			0.000	10.167	0.385	0.083	0.000	10.167	0.902	1.363
Silt			0.078	0.302	15.418	-0.005	4.360	0.899	15.418	0.000
Clay			0.000	0.083	-0.005	10.000	0.000	1.378	0.000	10.000
Background material: silt					lag: 0.3m					
Golden spikes: GSC-BH-S-SHMB										



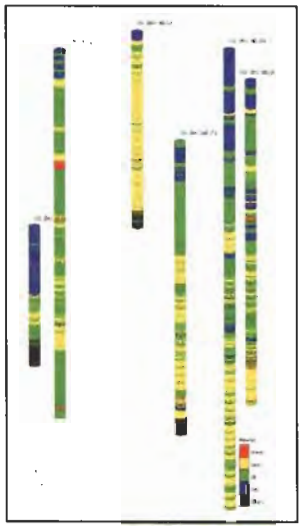
Appendix C9: Vertical T-PROG simulation output and borehole plot for Group6c golden spikes

			Transition Rates					Embedded Transition Probabilities				
Material	Proportion	Lens Length	Gravel	Sand	Silt	Clay	others	Gravel	Sand	Silt	Clay	others
Gravel	0.052	4.200	-0.238	0.076	0.060	0.102	0.000	4.200	0.250	0.250	0.500	0.000
Sand	0.248	8.190	0.012	-0.122	0.073	0.025	0.012	0.100	8.190	0.600	0.200	0.100
Silt	0.553	19.807	0.011	0.032	-0.056	0.012	0.000	0.201	0.588	19.807	0.212	-0.001
Clay	0.081	4.483	0.037	0.105	0.081	-0.223	0.000	0.200	0.600	0.200	4.483	0.000
Others	0.066	22.000	0.000	0.000	0.045	0.000	-0.045	0.000	0.000	1.000	0.000	22.000
			Embedded Transition Frequencies					Maximum Entropy Factors				
Material			Gravel	Sand	Silt	Clay	others	Gravel	Sand	Silt	Clay	others
Gravel			4.200	0.038	0.023	0.075	0.000	4.200	0.670	0.821	2.1508	0.000
Sand			0.038	8.190	0.186	0.075	0.038	0.513	8.190	1.320	0.66	2.295
Silt			0.061	0.186	20.93	0.050	-0.004	0.000	1.298	20.93	1.0036	0.000
Clay			0.038	0.113	0.050	4.483	0.000	1.228	0.916	0.527	4.4833	0.000
Others			0.000	0.000	0.033	0.000	22.000	0.000	0.000	0.000	0	22.00
Background material: silt			lag: 0.3m		Golden spikes: GSC-BH-S-KING, OGS-PJB-OGS-82							



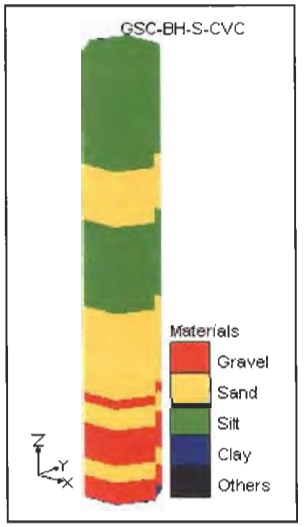
Appendix C10: Vertical T-PROG simulation output and borehole plot for Group 7a golden spikes

			Transition Rates					Embedded Transition Probabilities				
Material	Proportion	Lens Length	Gravel	Sand	Silt	Clay	others	Gravel	Sand	Silt	Clay	others
Gravel	0.008	0.620	-1.613	0.247	1.366	0.000	0.000	0.620	0.200	0.800	0.000	0.000
Sand	0.314	0.513	0.002	-1.951	1.366	0.565	0.018	0.000	0.513	0.643	0.344	0.013
Silt	0.472	0.578	0.025	0.940	-1.907	0.902	0.039	0.015	0.486	0.578	0.481	0.017
Clay	0.178	0.294	0.000	0.933	2.457	-3.398	0.008	0.000	0.351	0.644	0.294	0.005
Others	0.029	1.122	0.000	0.000	0.892	0.000	-0.892	0.000	0.000	1.000	0.000	1.122
			Embedded Transition Frequencies					Maximum Entropy Factors				
Material			Gravel	Sand	Silt	Clay	others	Gravel	Sand	Silt	Clay	others
Gravel			0.620	0.001	0.005	0.000	0.000	0.620	0.558	2.013	0.000	0.000
Sand			0.000	0.513	0.185	0.107	0.004	0.154	0.513	0.953	0.7868	0.784
Silt			0.006	0.196	0.581	0.185	0.007	1.612	1.057	0.581	0.7748	1.241
Clay			0.000	0.099	0.192	0.294	0.001	0.010	0.735	0.794	0.2943	0.199
Others			0.000	0.000	0.012	0.000	1.122	0.000	0.000	1.035	0	1.122
Background material: silt			lag: 0.3m		Golden spikes: GSC-BH-C34B-14, GSC-BH-C34B-17A,							
Golden spikes:GSC-BH-C34B-21, GSC-BH-C34B-28A, GSC-BH-C34B-29, GSC-BH-C48-4A												



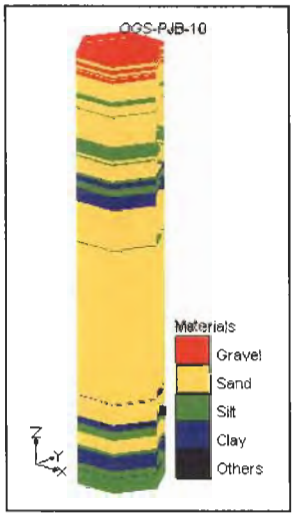
Appendix C11: Vertical T-PROG simulation output and borehole plot for Group 7b golden spikes

			Transition Rates				Embedded Transition Probabilities			
Material	Proportion	Lens Length	Gravel	Sand	Silt	Clay	Gravel	Sand	Silt	Clay
Gravel	0.149	8.843	-0.113	0.076	0.001	0.036	8.843	0.667	0.000	0.333
Sand	0.402	17.871	0.042	-0.056	0.014	0.000	0.750	17.871	0.250	0.000
Silt	0.447	37.915	0.000	0.025	-0.027	0.002	-0.001	0.954	37.915	0.047
Clay	0.003	0.454	0.000	0.000	2.203	-2.203	0.000	0.000	1.000	0.454
			Embedded Transition Frequencies				Maximum Entropy Factors			
Material			Gravel	Sand	Silt	Clay	Gravel	Sand	Silt	Clay
Gravel			8.843	0.241	-0.043	0.121	8.843	0.950	0.000	3.266
Sand			0.362	17.871	0.064	0.000	1.419	17.871	0.805	0.000
Silt			-0.043	0.185	61.192	-0.004	0.000	1.633	61.192	0.000
Clay			0.000	0.000	0.117	0.454	0.000	0.000	0.000	0.454
Background material:					lag: 0.3m					
Golden spikes: GSC-BH-S-CVC										

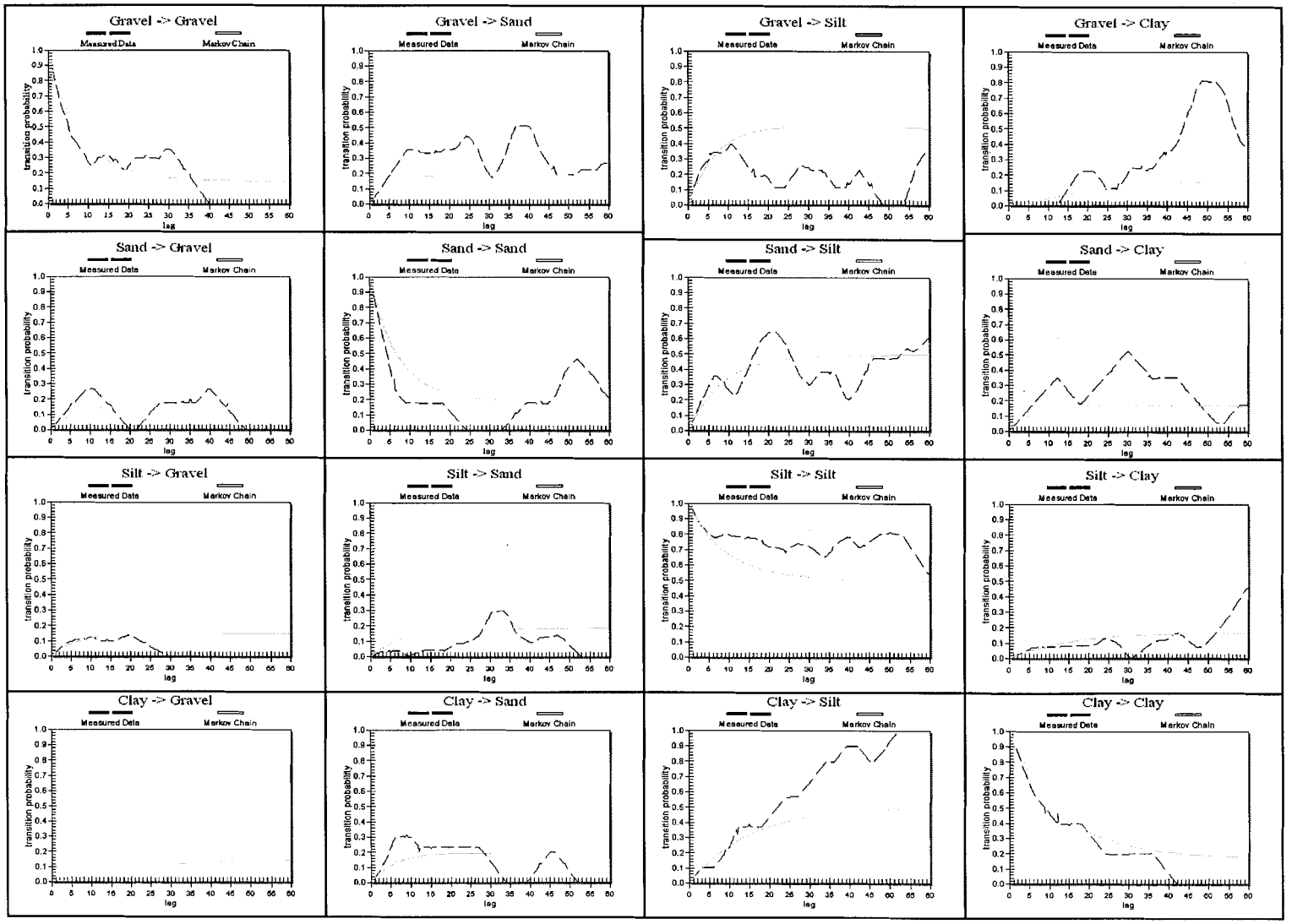


Appendix C12: Vertical T-PROG simulation output and borehole plot for Group 8 golden spikes

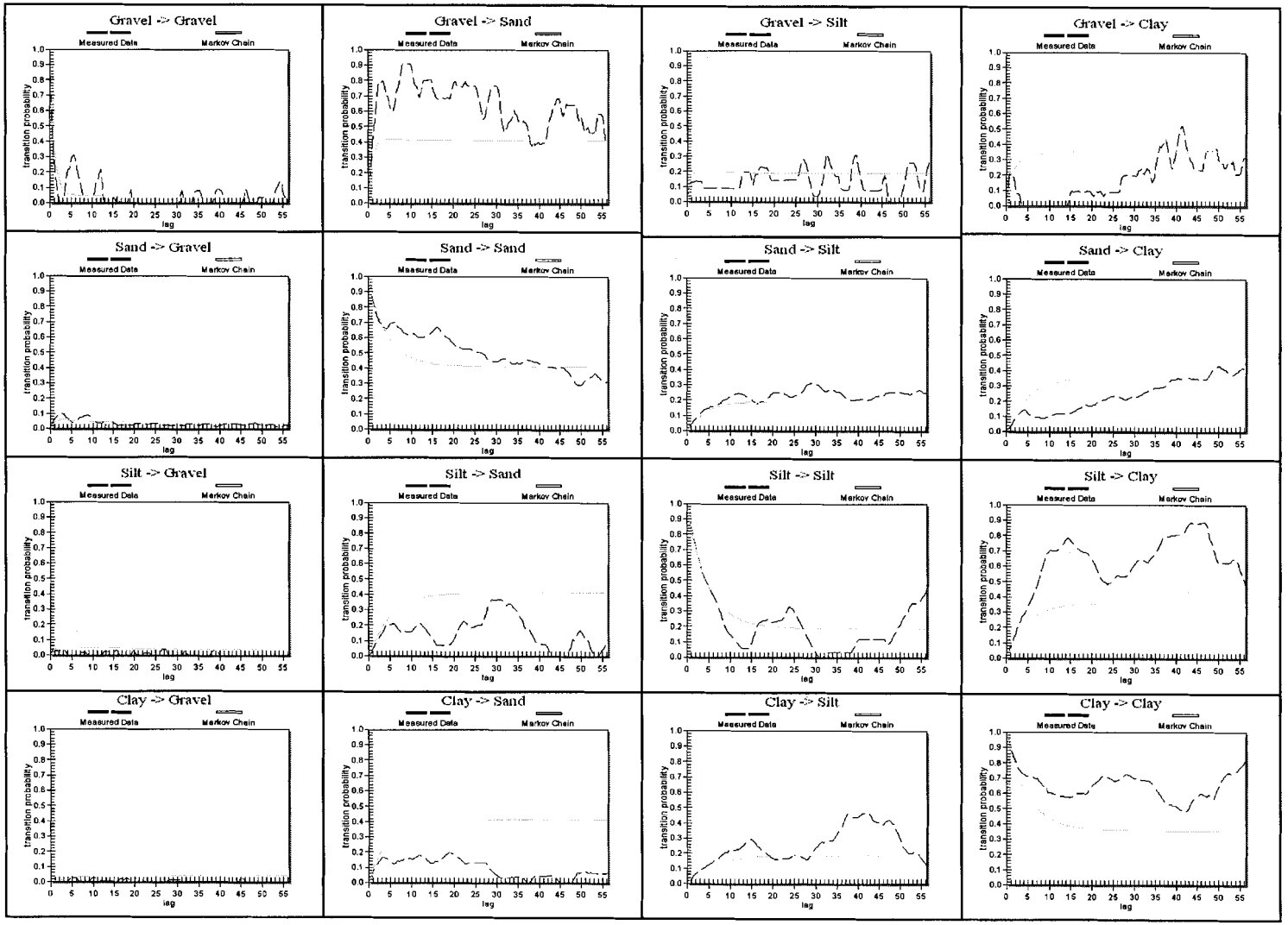
			Transition Rates				Embedded Transition Probabilities			
Material	Proportion	Lens Length	Gravel	Sand	Silt	Clay	Gravel	Sand	Silt	Clay
Gravel	0.038	0.914	-1.094	1.094	0.000	0.000	0.914	1.000	0.000	0.000
Sand	0.704	5.458	0.059	-0.196	0.106	0.031	0.322	5.458	0.555	0.124
Silt	0.161	1.440	0.000	0.453	-0.694	0.241	0.000	0.600	1.440	0.400
Clay	0.097	1.600	0.000	0.243	0.383	-0.625	0.000	0.333	0.667	1.600
			Embedded Transition Frequencies				Maximum Entropy Factors			
Material			Gravel	Sand	Silt	Clay	Gravel	Sand	Silt	Clay
Gravel			0.914	0.121	0.000	0.000	0.914	2.006	0.000	0.000
Sand			0.121	5.479	0.202	0.052	1.558	5.479	0.999	0.597
Silt			0.000	0.202	1.440	0.125	0.000	0.831	1.440	1.570
Clay			0.000	0.052	0.125	1.600	0.000	0.615	1.502	1.600
Background material:					lag: 0.3m					
Golden spikes: OGS-PJB-10										



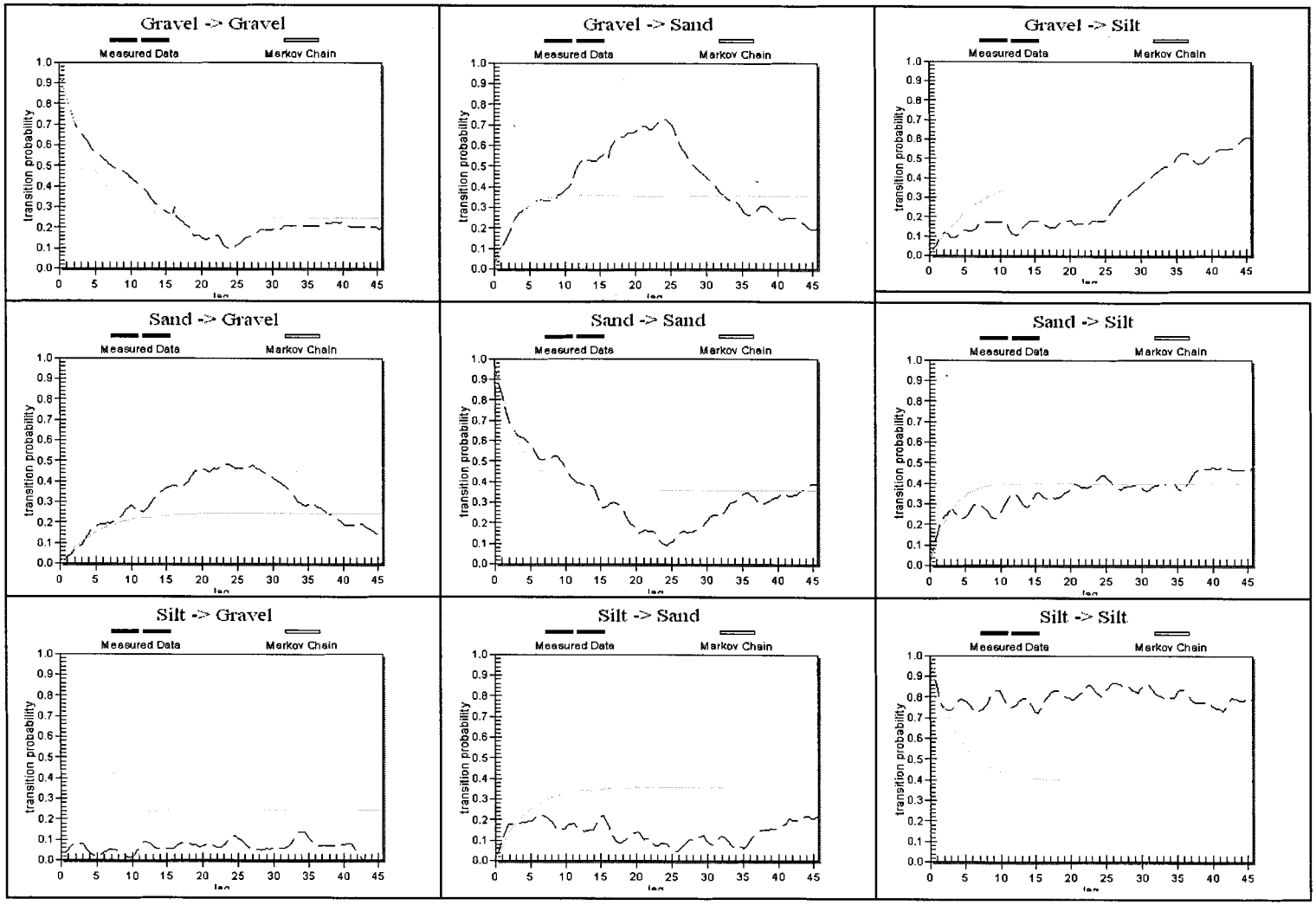
Appendix C13: Markov chain graphs for group 1 golden spikes



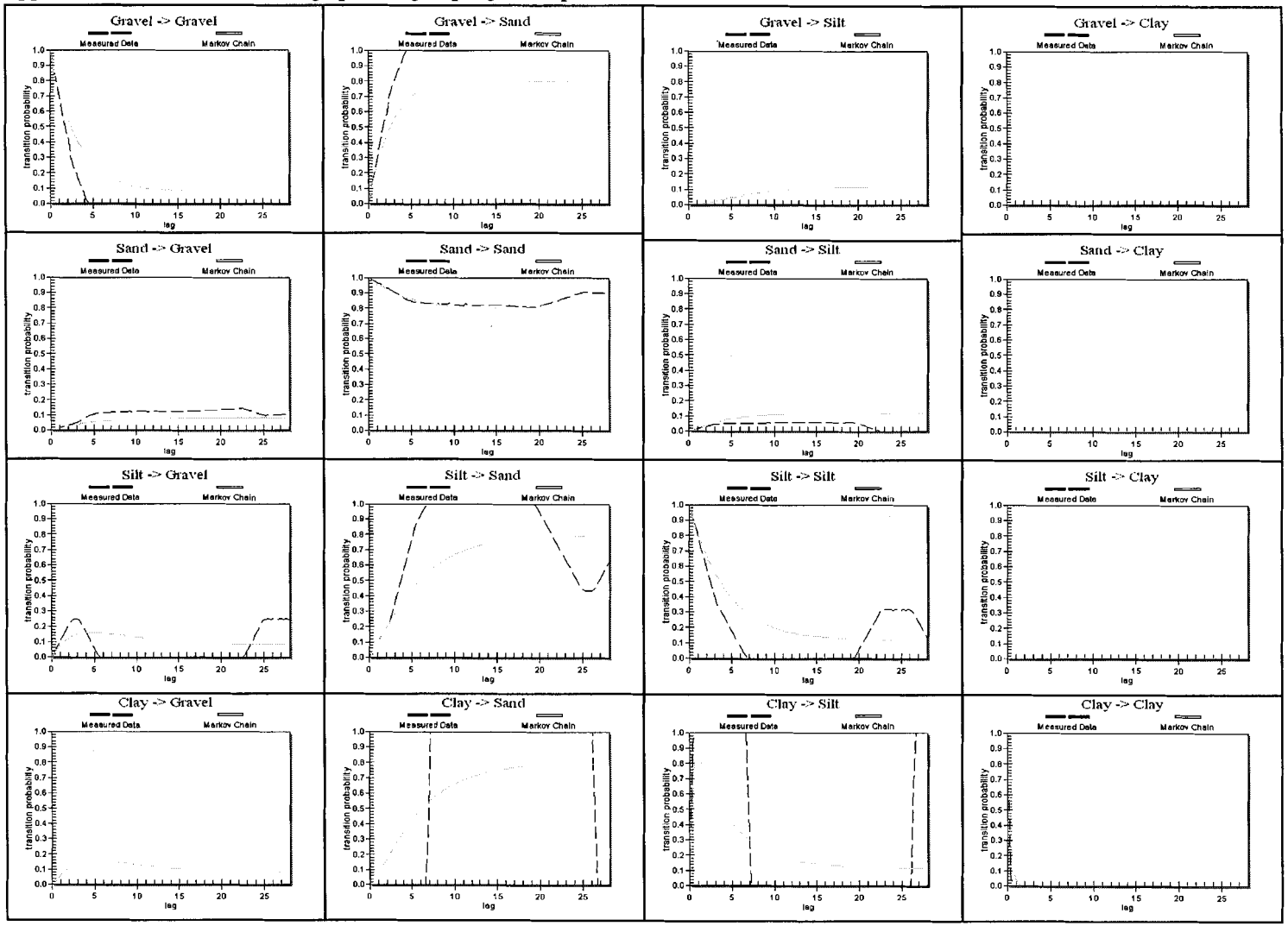
Appendix C14: Markov chain graphs for group 2a golden spikes



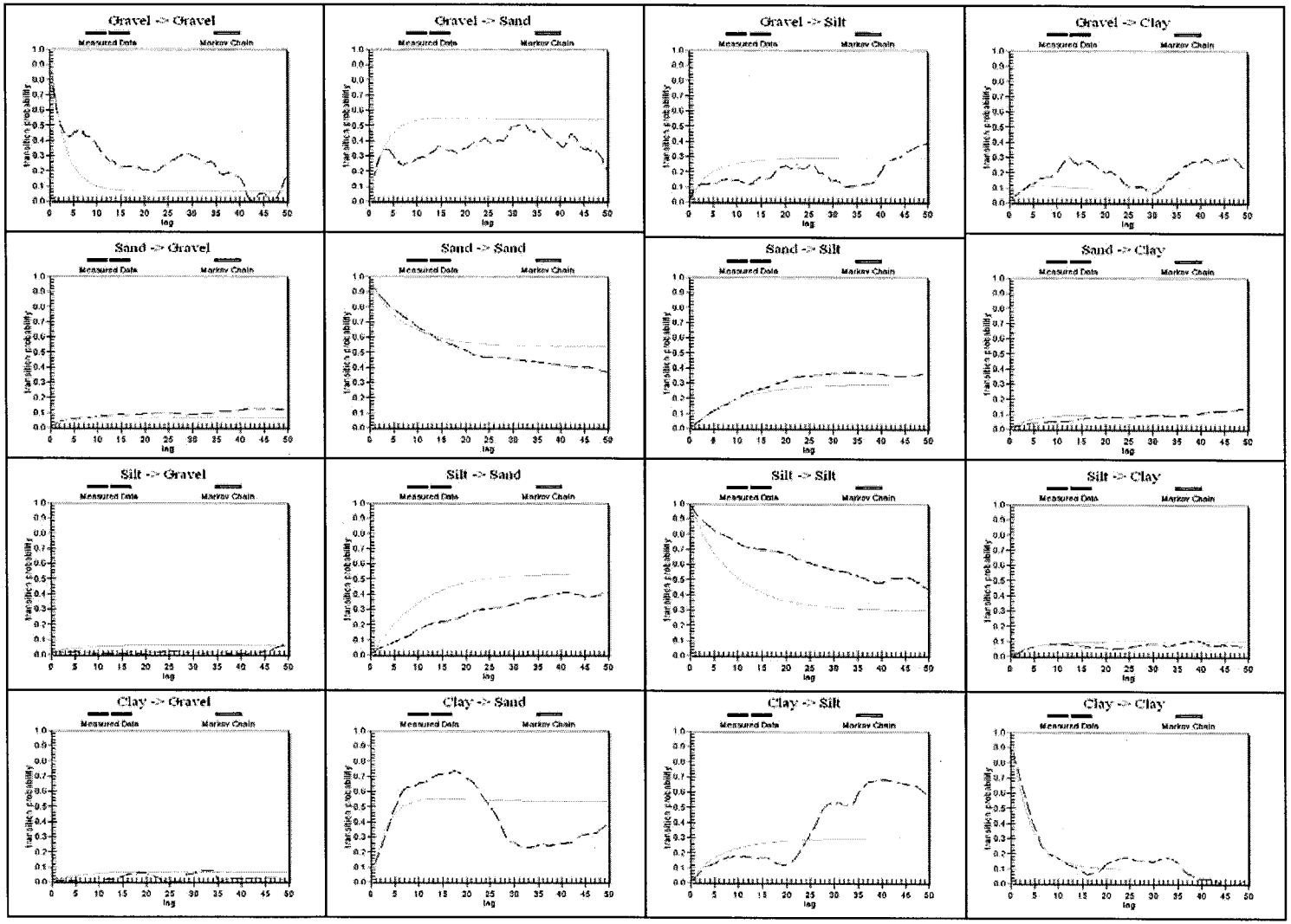
Appendix C15: Markov chain graphs for group 2b golden spikes



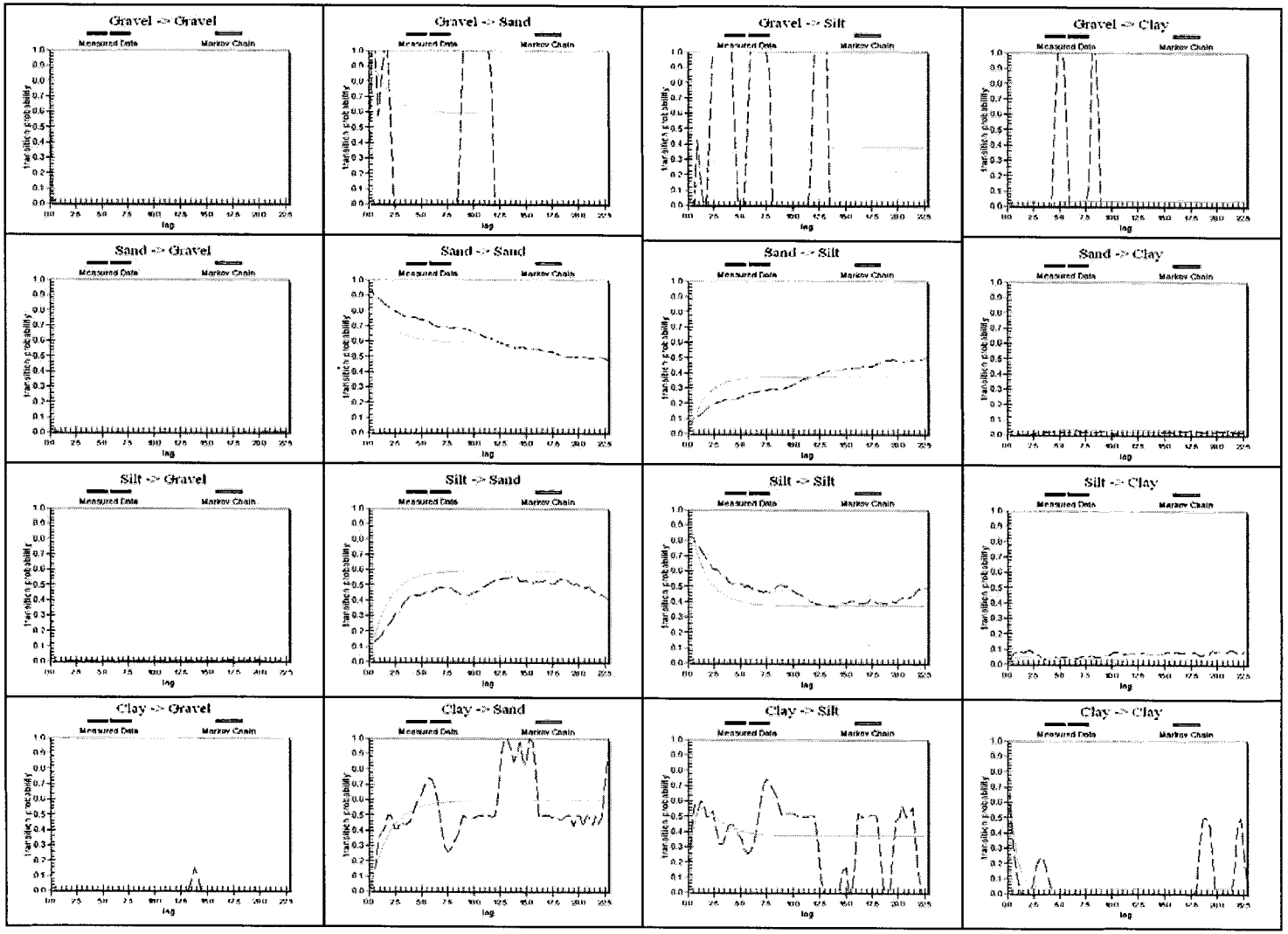
Appendix C16: Markov chain graphs for group 3 golden spikes



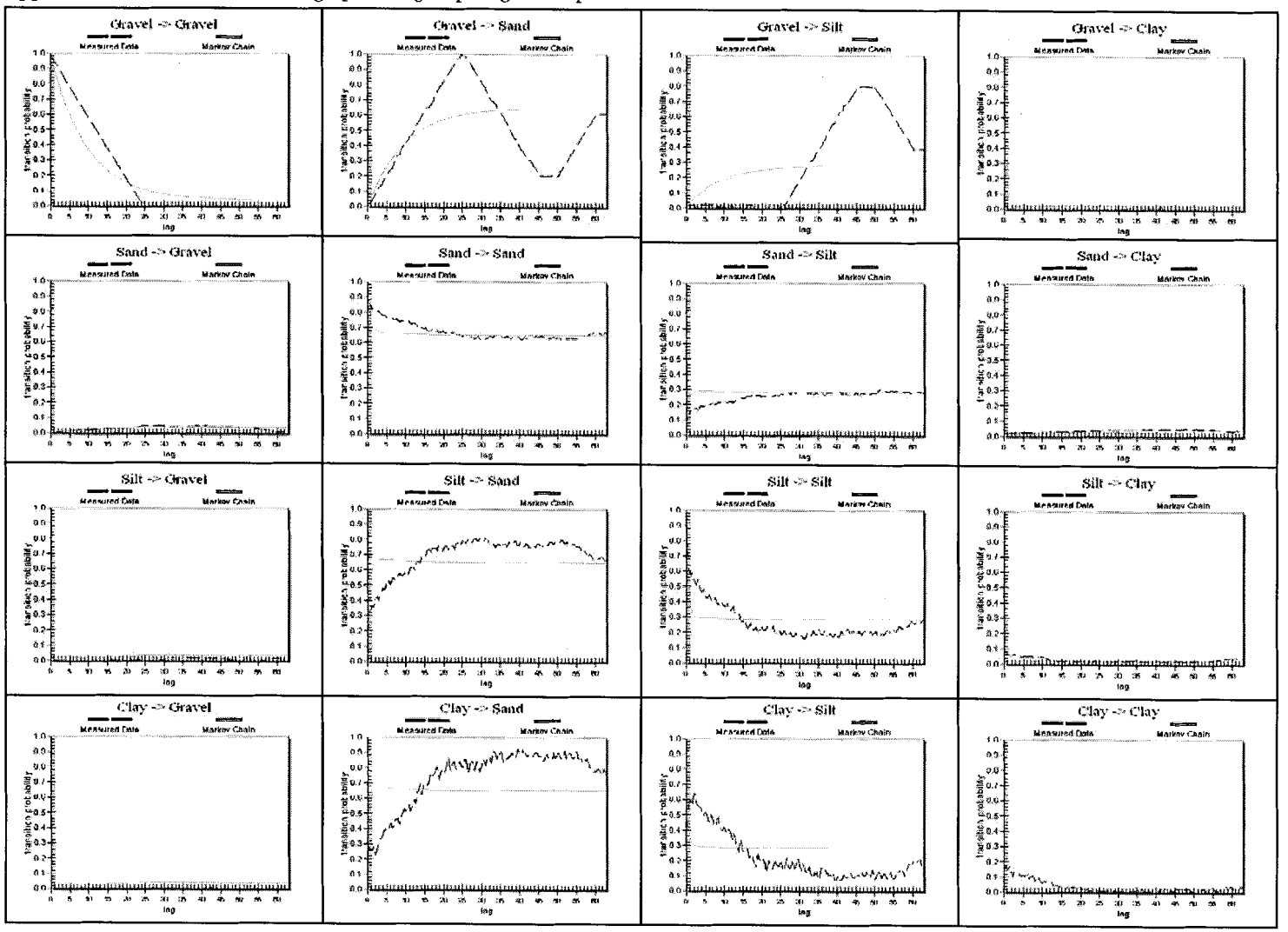
Appendix C17: Markov chain graphs for group 4 golden spikes



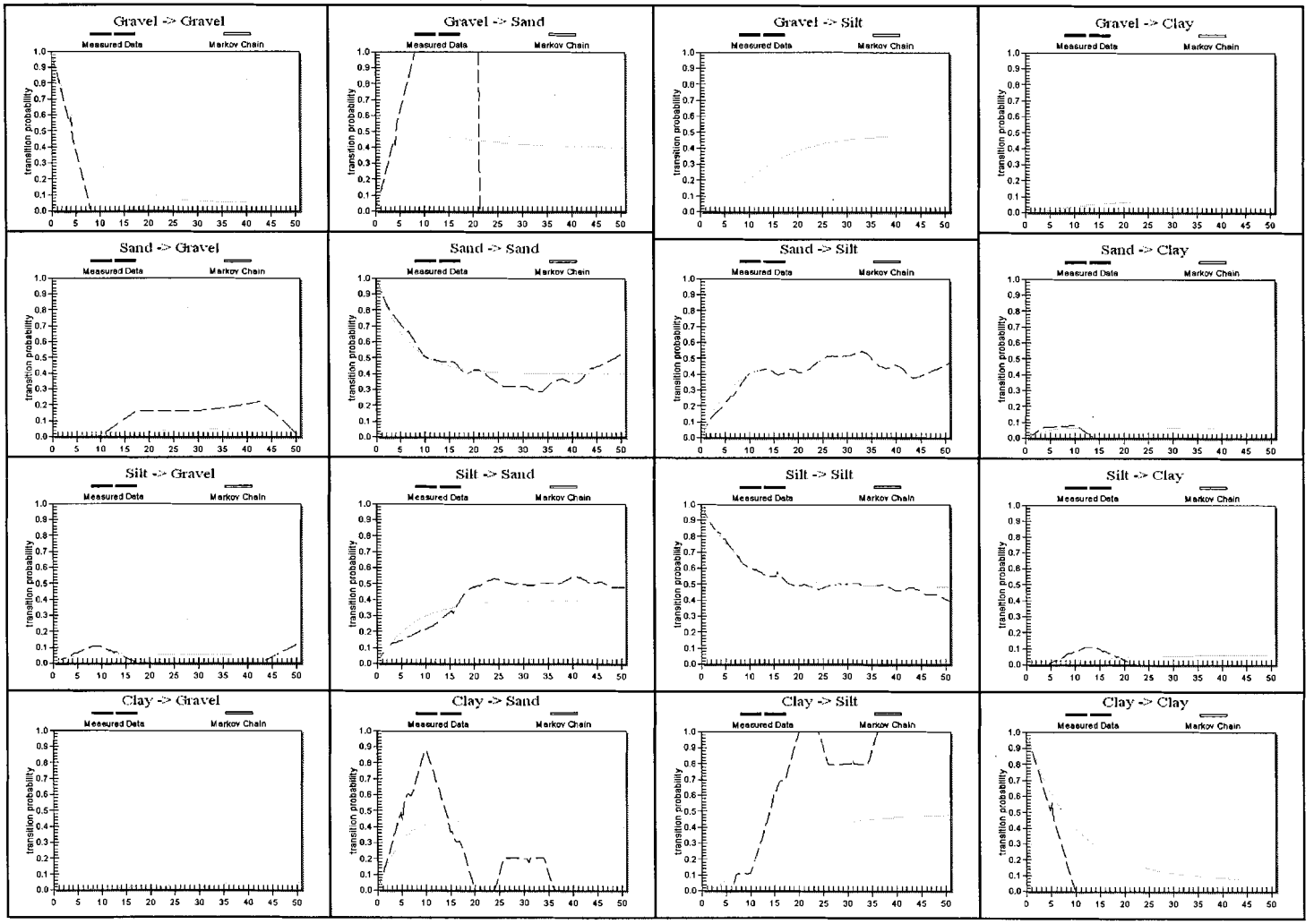
Appendix C18: Markov chain graphs for group 5 golden spikes



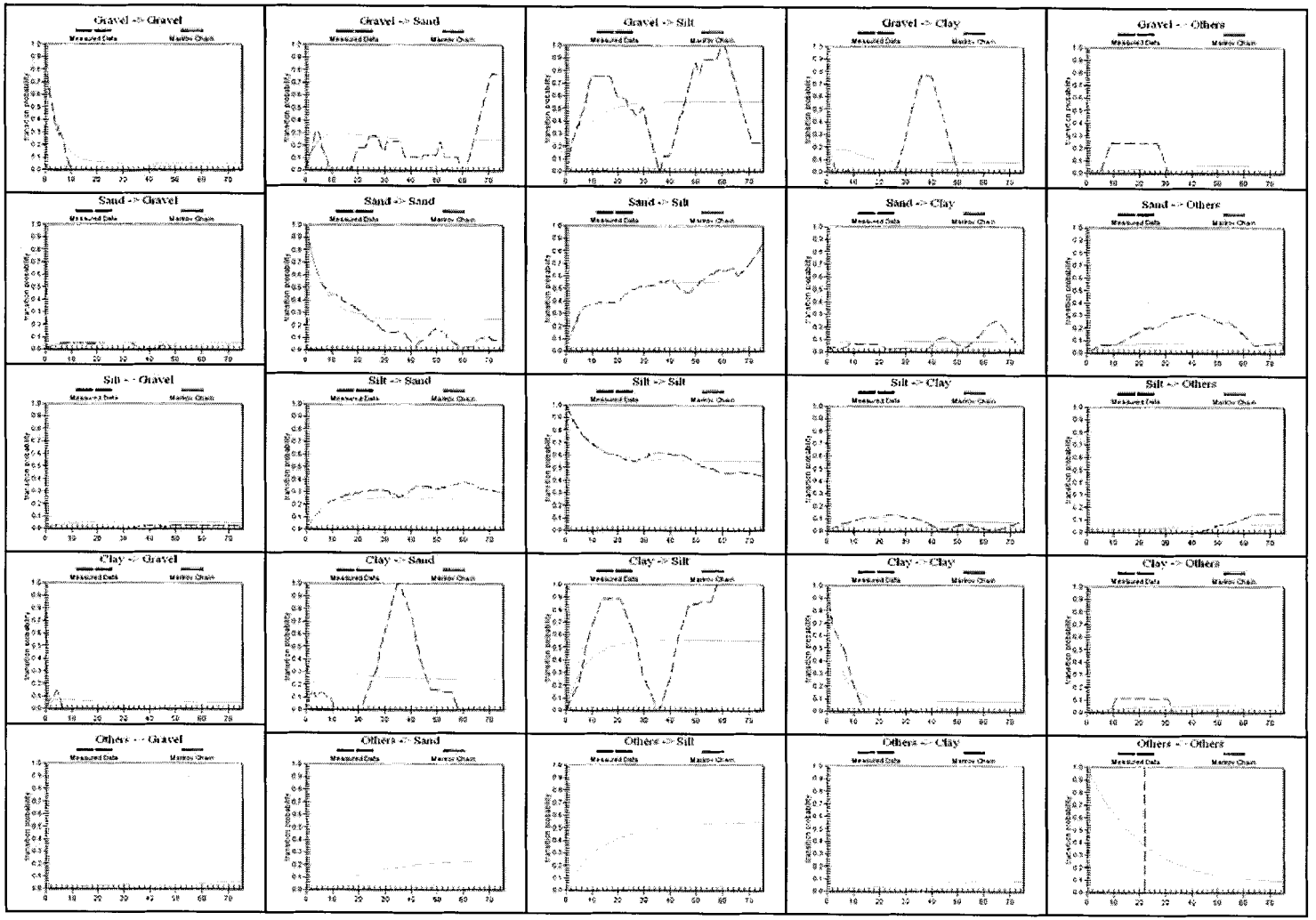
Appendix C19: Markov chain graphs for group 6a golden spikes



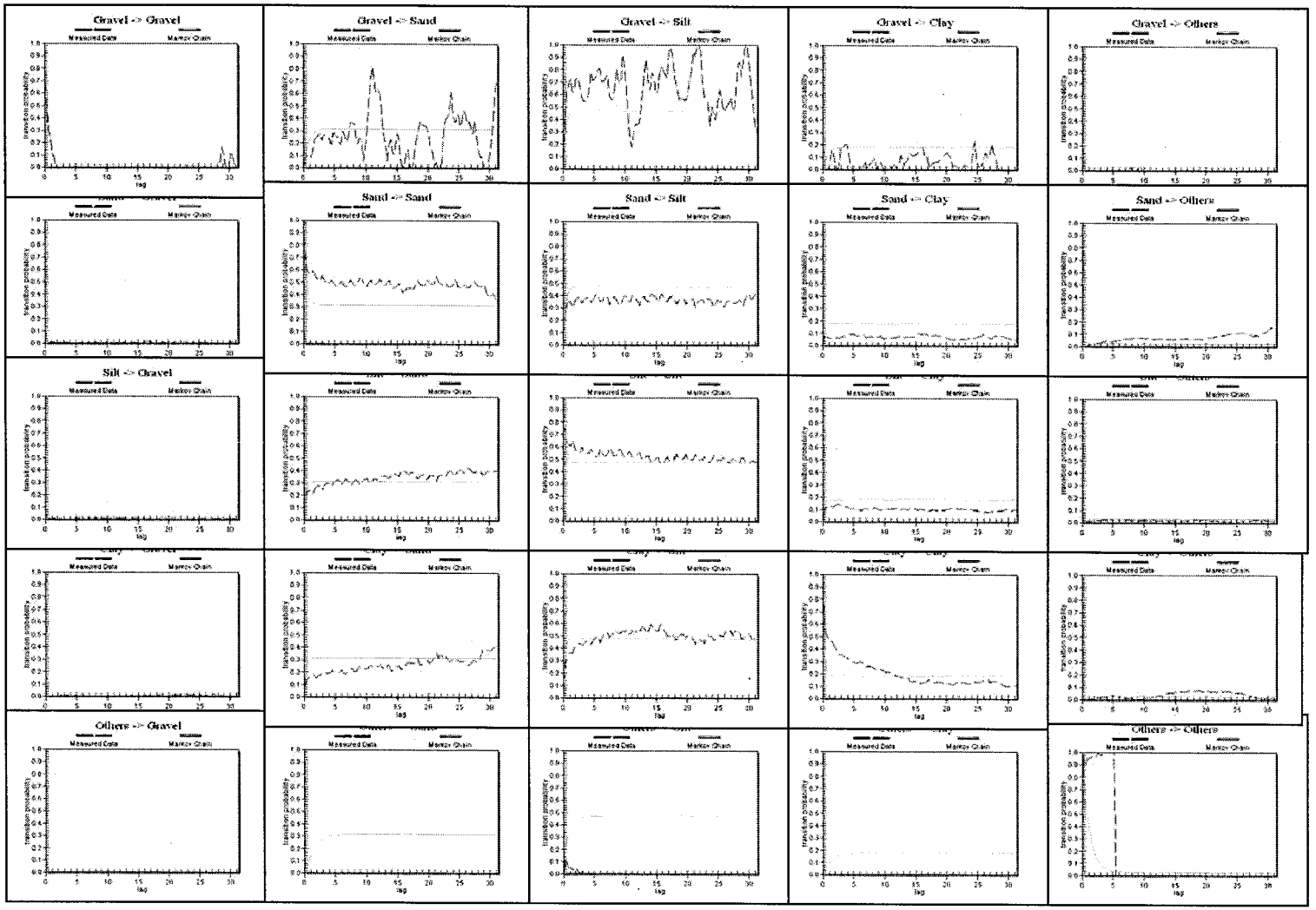
Appendix C20: Markov chain graphs for group 6b golden spikes



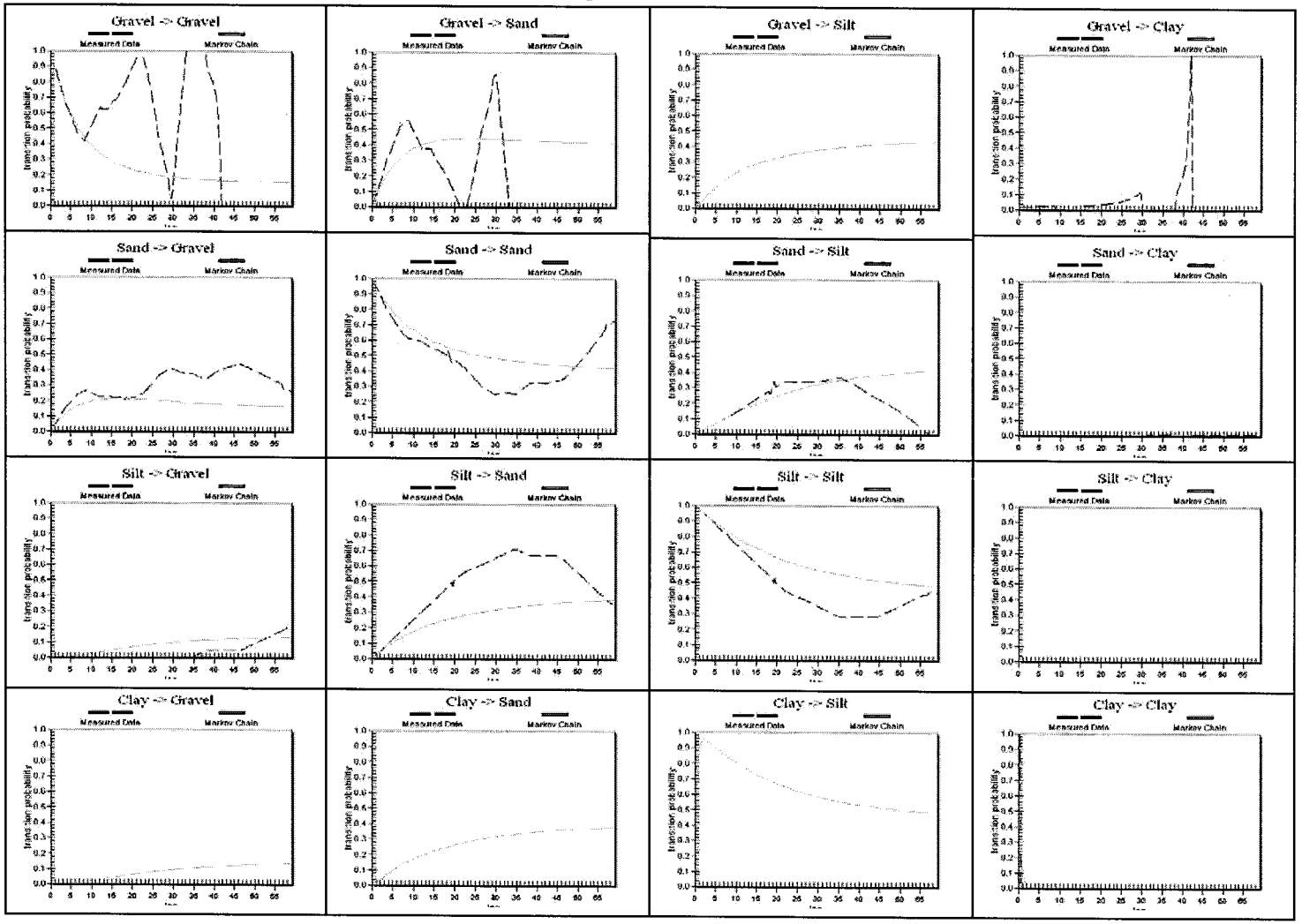
Appendix C21: Markov chain graphs for group 6c golden spikes



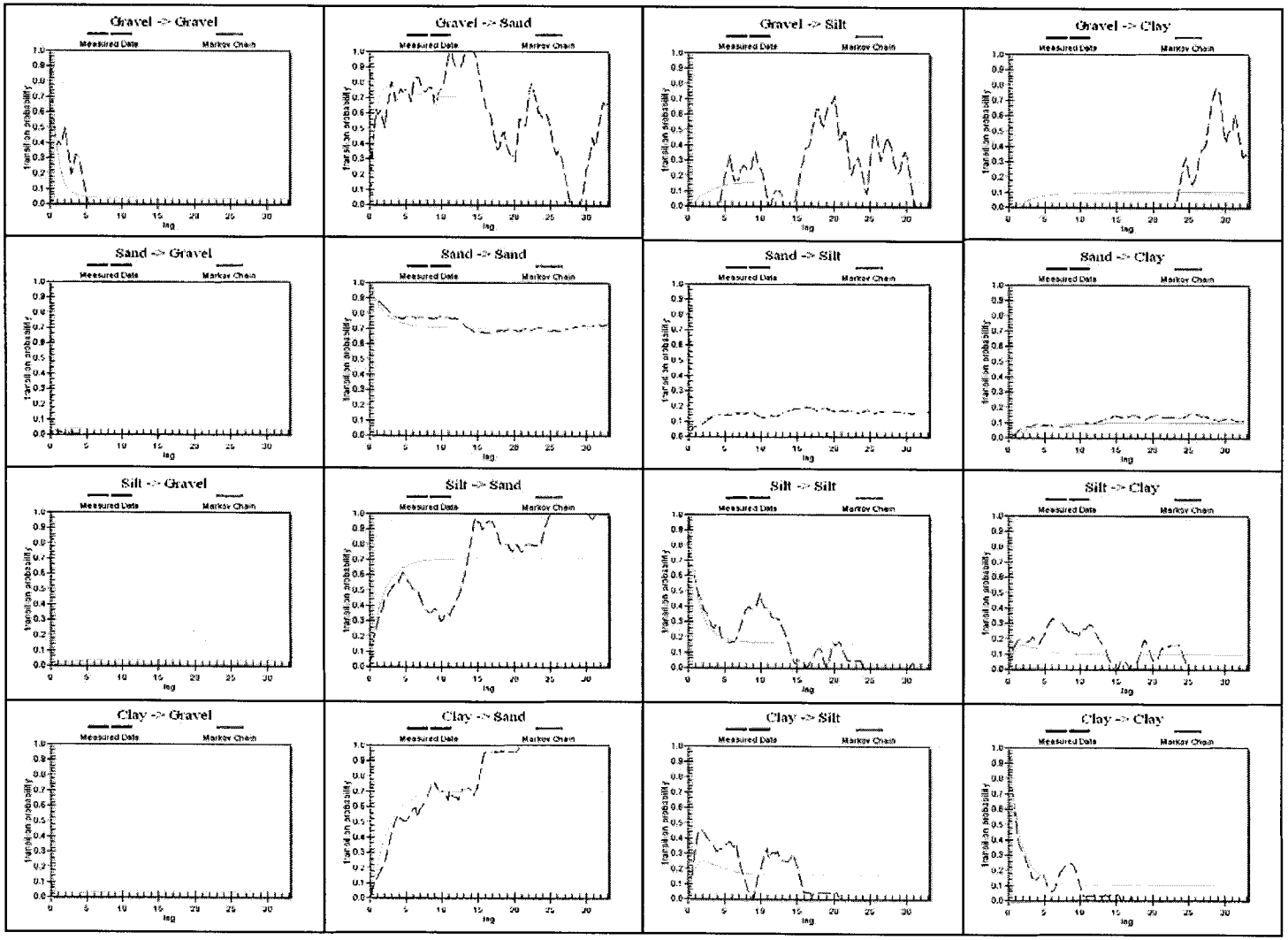
Appendix C22: Markov chain graphs for group 7a golden spikes



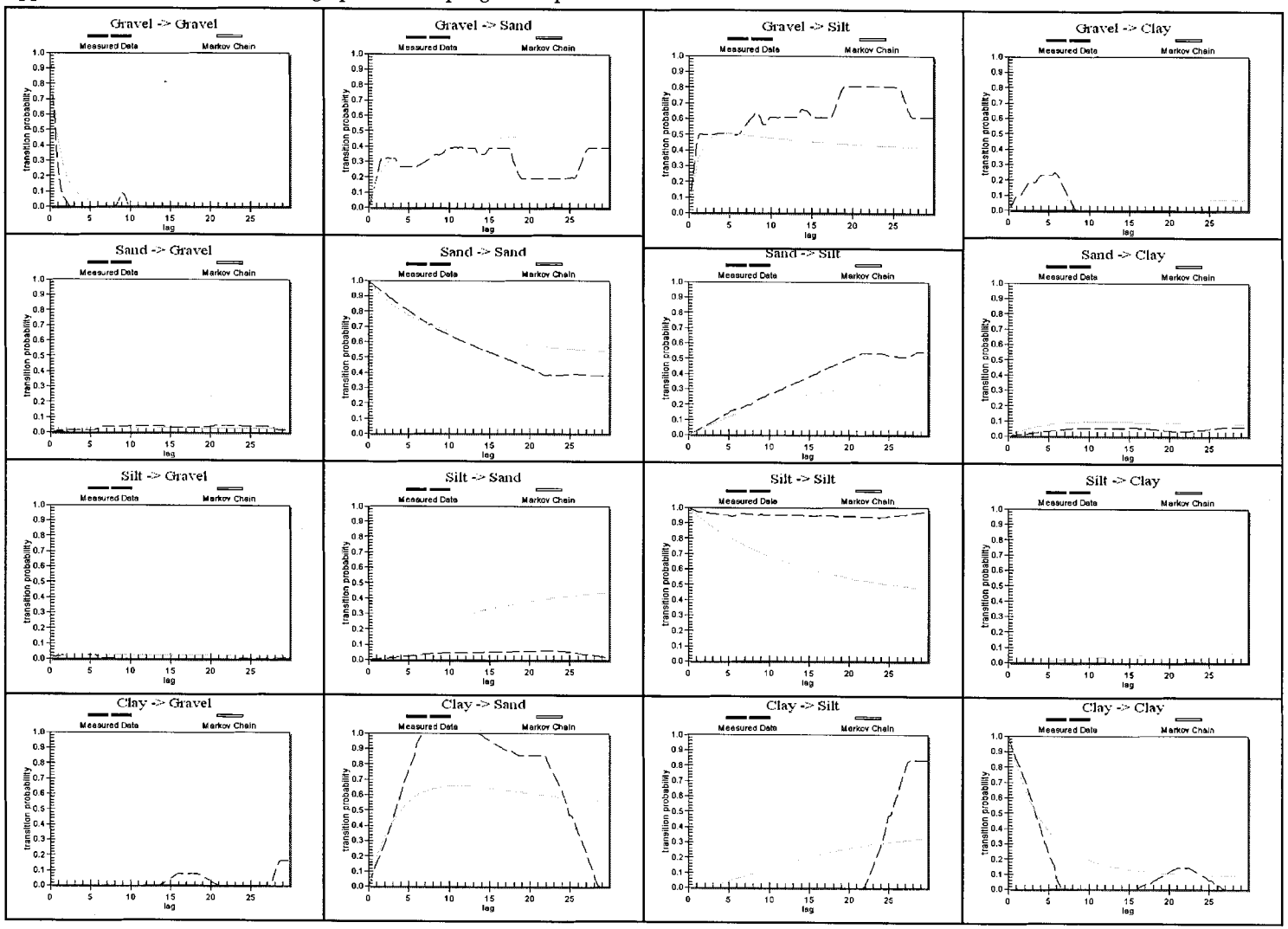
Appendix C23: Markov chain graphs for group 7b golden spikes



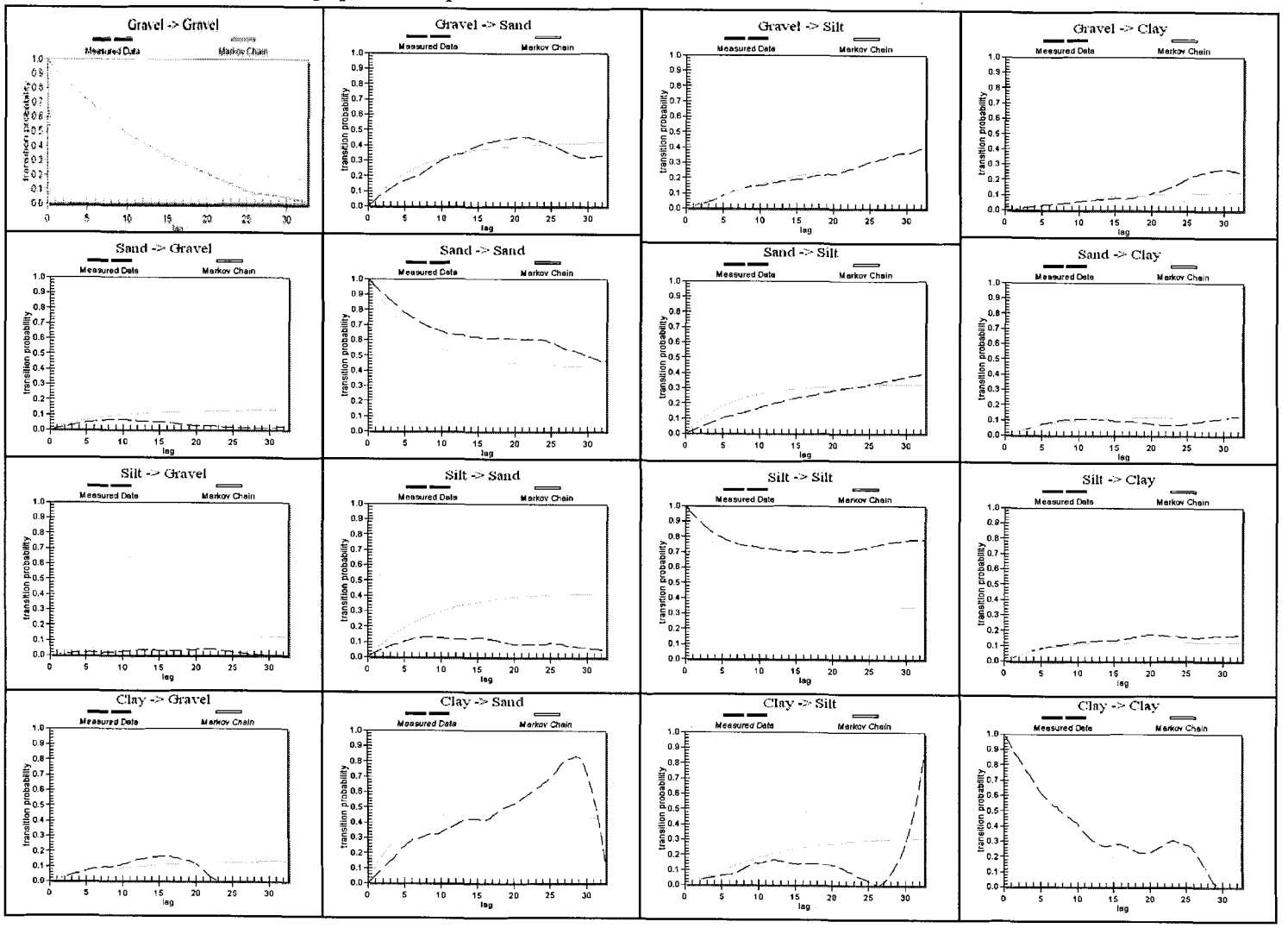
Appendix C24: Markov chain graphs for group 8 golden spikes



Appendix C25: Markov chain graphs for sample golden spikes



Appendix C26: Markov chain graphs for sample MOEE data



REFERENCE LIST

Abbaspour, R. A., Delavar, M. R. and Batouli, R. 2003. The Issue of Uncertainty Propagation in Spatial Decision Making. *The 9th Scandinavian Research Conference on Geographical Information Science*, 57 - 65.

Allan, Brimicombe. 2003. *GIS, Environmental Modelling and Engineering*. London: Taylor & Francis Group.

Allen, D. M., Schuurman, N. and Zhang, Q. 2002. Application of Fuzzy Logic for Aquifer Architecture Modelling.

Anderson, P. Mary, and William W. Woessner. 1992. *Applied Groundwater Modeling - Simulation of Flow and Advective Transport*. San Diego, California: Academic Press Inc.

Barber, G. M. 1998. *Elementary Statistics For Geographers*: Guilford, New York.

Bardossy, Andras. 2004. *Introduction to Geostatistics*. University of Stuttgart 2003 [cited 11th August 2004]. Available from http://www.warem.uni-stuttgart.de/study/program03/downloads/download702e/geostatistics_screen.pdf.

Barnett, P. J, D. R Sharp, H. A. J Russell, T. A Brennand, G. Gorrell, F. Kenny, and A. Pugin. 1998. On the origin of the Oak Ridges Moraine. *Canadian Journal of Earth Sciencies* 35:1152-1167.

Bevington, R. P., and D. K. Robinson. 1992. *Data Reduction and Error Analysis for the Physical Sciences*. Second ed. New York: McGraw-Hill, Inc.

Bittner, T. and Stell, J. G. 2002. Vagueness and Rough Location. *GeoInformatica* 6(2):99 - 121.

Bonissone, P. P. 1997. *Approximate Reasoning Systems: Handling Uncertainty and Imprecision in Information Systems.* In Uncertainty Management in Information Systems: From Needs to Solution, Motro, A. and Smets, P. (Eds.), Kluwer Academic: The Netherlands.

Bosc, P. and Prade, H. 1997. *An Introduction to the Fuzzy set and Possibility Theory-Based Treatment of Flexible Queries and Uncertain or Imprecise Databases.* In Uncertainty Management in Information Systems: From Needs to Solution, Motro, A. and Smets, P. (Eds.), Kluwer Academic: The Netherlands, p 285 - 324.

Brassel, K., F. Bucher, E. M. Stephan, and A. Vckovski. 1995. Completeness. In *Elements of spatial data quality*, edited by S. C. Guptill and J. L. Morrison. New York: Elsevier Science Ltd.

Brassington, R. 1988. *Field Hydrogeology.* Milton Keynes, Open University Press & Halsted Press; New York.

Brunsdon, C. 2001. A Bayesian Approach to Schools' Catchment-based Performance Modelling *Geographical and Environmental Modelling* 5(1): 9 - 22.

Burrough, Peter A. 1996. Natural Objects with Indeterminate Boundaries. In *Geographic Objects with Indeterminate Boundaries*, edited by P. A. Burrough and A. U. Frank. London: Taylor & Francis.

Burrough, A. Peter, and Rachael A. McDonnell. 1998. *Principles of Geographical Information Systems.* New York: Oxford University Press.

Buttenfield, B., and M. K Beard. 1994. Graphical and Geographical Components of Data Quality. In *Visualization In Geographical Information Systems*, edited by H. M. Hearshaw and D. J. Unwin. Chichester: John Wiley & Sons Ltd.

Carle, Steven F. 1999 T-PROGS: Transition Probability Geostatistical Software Version 2.1, Hydrologic Sciences Graduate Group University of California, Davis.

Carle, S. F. and Fogg, G. E. 1997. Modelling Spatial Variability with One and Multidimensional Continuous-Lag Markov Chains. *Mathematical Geology* 29(7): 891 - 918.

Carranza, E. J. M. and Hale, M. 2001. Geologically Constrained Fuzzy Mapping of Gold Mineralization Potential, Baguio District, Philippines. *Natural Resources Research*, 10(2): 125 -136. Kluwer Academic Publishers.

Clarke, D. G, and D. M Clark. 1995. Lineage. In *Elements of spatial data quality*, edited by S. C. Guptill and J. L. Morrison. New York: Elsevier Science Ltd.

Davis, Benjamin. 2003. *Choosing a method for poverty mapping*: Agriculture and Economic Development Analysis Division, FAO, UN.

Diggle, P. J. and Ribeiro Jr. P. J. 2002. Bayesian Inference in Gaussian Model-based Geostatistics *Geographical and Environmental Modelling* 6(2): 129 - 146.

Dolgoff, Anatole. 1996. *Physical Geology*. Lexington, MA: D. C. Heath and Company.

Dowd, P. A. and Pardo-Iguzquiza, E. 2002. The Incorporation of Model Uncertainty in Geostatistical Simulation. *Geographical and Environmental Modelling* 6(2): 147 - 169.

Dragicevic, Suzana, J. Danielle Marceau, and Claude Marois. 2001. Space, time, and dynamics modeling in historical GIS databases: a fuzzy logic approach. *Environment and Planning B: Planning and Design* 28:545-562.

Dragicevic, Suzana, N. Schuurman, and J. M FitzGerald. 2004. The Utility of exploratory spatial data analysis in the study of Tubercolosis incidences in urban Canadian population. *Cartographica* 39 (2):29-39.

Drummond, J. 1995. Positional Accuracy. In *Elements of spatial data quality*, edited by S. C. Guptill and J. L. Morrison. New York: Elsevier Science Ltd.

Dubois, D. and Prade, H. 1988. *Possibility Theory: An Approach to Computerized Processing of Uncertainty*, New York: Plenum.

Dubois, D. and Prade, H. 1980. *Fuzzy Sets and Systems: Theory and Applications,* Academic Press Int.: New York.

Duckham, M. and Sharp, J. (forthcoming, 2004) *Uncertainty and geographic information: computational and critical convergence.* In Re-presenting objects, Unwin, D. and Fisher, P.F. (Eds.), Wiley: New York.

Duckham, M., Keith, M., Stell, J. and Worboys, M. 2003. Formal Approach to Imprecision In Geographic Information. *Computer, Environment and Urban Systems* 25:89-103.

Dungan, J. L. 2002. Towards a Comprehensive View of Uncertainty in Remote Sensing Analysis. In *Uncertainty in Remote Sensing and GIS*, edited by G. M. Foody and P. M. Atkinson. Chichester: John Wiley & Sons Ltd.

Dutton, Geoffrey. 1989. Modeling locational uncertainty via hierarchical tesselation. In *Accuracy of Spatial Databases*, edited by F. M. Goodchild and S. Gopal. London: Taylor & Francis.

Elfeki, A. and Dekking, M. 2001. A Markov Chain Model for Subsurface Characterization: Theory and Applications. *Mathematical Geology* 33(5): 569 - 589.

Foody, Giles M. and Atkinson, Peter M. 2002. *Uncertainty in Remote Sensing and GIS*: John Wiley & Sons, England.

Fortin, M. and Edwards, G. 2001. Delineation and Analysis of Vegetation Boundaries. In *Spatial Uncertainty in Ecology: Implications for Remote Sensing and GIS Applications*, Hunsaker, C. T., Goodchild, M. F., Friedl, M. A. and Case, T. J. (Eds.), Springer-Verlag, New York.

Fotheringham, A. S. and Wong, D. W. S. The modifiable areal unit problem in multiavriate statistical analysis. *Environment and Planning A* 23: 1025-1044.

Goodchild, F. Michael. 1989. Modeling error in objects and fields. In *Accuracy of Spatial Databases*, edited by F. M. Goodchild and S. C. Guptill. London: Taylor & Francis Ltd.

Goodchild, F. Michael. 1995. Attribute Accuracy. In *Elements of spatial data quality*, edited by S. C. Guptill and J. L. Morrison. New York: Elsevier Science Ltd.

Goodchild, M. F., Buttenfield, B. and Wood, J. 1994. *Introduction to Visualizing Data Quality.* In Visualization in Geographic Information Systems, Hearnshaw, H. M. and Unwin, D. J. (Eds.), John Wiley & Sons: England.

Griffith, D. A., Wong, D.W.S and Whitefield, T. 2003. Exploring Relationships between the Global and Regional Measures of Spatial Autocorrelation. *Journal of Regional Science* 43(4): 686 - 710.

Groundwater Modeling System 5.0. Environmental Modeling Systems, Inc., Utah, USA.

Guptill, S. C. 1995. Temporal Information. In *Elements of spatial data quality*, edited by S. C. Guptill and J. L. Morrison. New York: Elsevier Science Ltd.

Harvey F., Kuhn W., Pundt H., Bishr Y., Riedemann C. 1999. Semantic Interoperability: A central issue for sharing geographic information. *The Annals of Regional Science*: 213-232.

Henrion, M., Suermondt, H. J. and Heckerman, D. E. 1997. *Probabilistic and Bayesian Representations of Uncertainty in Information Systems: A Pragmatic Introduction.* In Uncertainty Management in Information Systems: From Needs to Solution, Motro, A. and Smets, P. (Eds.), Kluwer Academic: The Netherlands.

Horner, W. Mark and Murray, T. Alan. 2002. Excess Commuting and the Modifiable Areal Unit Problem. *Urban Studies* 39 (1):131-139.

Howard, R. A. 1971. *Dynamic Probabilistic Systems Volume 1: Markov Models.* John Wiley & Sons Inc., New York.

Hunsaker, C. T., Goodchild, M. F., Friedl, M. A. and Case, T. J. 2001. *Spatial Uncertainty in Ecology: Implications for Remote Sensing and GIS Applications*: Springer-Verlag, New York.

Hunter, Gary J. 1998. Managing Uncertainty in GIS, *NCGIA Core Curriculum in GIScience*, [cited May 11, 2004]. Available from: http://www.ncgia.ucsb.edu/giscc/units/u187/u1871.html, posted February 03, 1998.

Jarvis, P.G. 1995. Scaling processes and problems. *Plant, Cell, and Environment*, 18: 1079 -1089.

Jelinski, Dennis E. and Jianguo Wu. 1996. The modifiable areal unit problem and implications for landscape ecology. *Landscape Ecology* 11 (3):129 -140.

Jiang, B. 1998. Visualisation of Fuzzy Boundaries of Geographic Objects Graphics. *Cartography - The Journal* 27(2): 41 - 46.

Jones, Norman L. 2003. *Seepage & Groundwater Modeling*. Dept. of Civil and Environmental Engineering, Brigham Young University 2003 [cited August 10th 2004]. Available from http://class.et.byu.edu/ce547/.

Kainz, W. 1995. Logical Consistency. In *Elements of spatial data quality*, edited by S. C. Guptill and J. L. Morrison. New York: Elsevier Science Ltd.

Kassim, S. and Rothman L. 2003. *Immigrant Poverty in Canada: Focus on Toronto.* Campaign2000 [cited October, 20 2003]. Available from http://www.fsatoronto.com.

Katzberg, J. D. and Ziarko, W. 1994. *Variable Precision Rough Sets with Asymmetric Bounds.* In Rough Sets, Fuzzy Sets and Knowledge Discovery, Ziarko, W. P (Eds.),Springer-Verlag and British Computer Society: London, p 167 - 177.

Kazemipur, A. and Halli, S.S. 1998. Plight of Immigrants: The Spatial Concentration of Poverty in Canada. *Canadian Journal of Regional Science* XX (1-2):11-28.

Keukelaar, J. 1999. *A Visual Programming Language for the Analysis of Uncertain Spatial Data*. [Cited January 11, 2004]. Available from: http://www.nada.kth.se/utbildning/forsk.utb/avhandlingar/lic/990608.pdf.

Klinkenberg, Brian. 2003. *Geography 470 Advanced Issues in GIS: Developing an understanding* [cited January, 20 2004]. Available from http://www.geog.ubc.ca/courses/geog470.

Klinkenberg, Brian. 2004. Uncertainty in a GIS. *In Geog 516: Graduate GIS Seminar* [cited May 11, 2004]. Available from: http://www.geog.ubc.ca/courses/geog516/notes/Uncertainty.ppt.

Kuhn, W. 2001. Ontologies in support of activities in geographical space. *International Journal of Geographical Information Science* 15 (7):613-631.

Kuhn, W. 2003. Semantic reference systems. *International Journal of Geographical Information Science* 17(5): 405-409.

Lagacherie, P., Andrieux, P. and Bouzigues, R. 1996. Fuzziness and Uncertainty in Soil Boundaries: From Reality to Coding in GIS. In *Geographic Objects with Indeterminate Boundaries*, Burrough, P. A. and Andrew, U. F. (Eds.), Taylor and Francis. p 275 -286.

Lark, R. M. 2000. Regression analysis with spatially autocorrelated error: simulation studies and application to mapping of soil organic matter. *International Journal of Geographic Information Science* 14 (3):247-264.

Lemay, Philippe. 1999. *The Statistical Analysis of Dynamics and Complexity In Psychology: A Configural Approach.* PhD Thesis, Faculté Des Sciences Sociales Et Politiques, Université De Lausanne.

Levin, Harold L. 1981. *Contemporary Physical Geology*. Philadelphia: Saunders College Publishing.

Ley, David, and Smith Heather. 1997. Immigration and Poverty in Canadian Cities. *Canadian Journal of Regional Science* 20 (12):29-51.

Lo, C. P. and Yeung, A. K. W. 2002. *Concepts and Techniques of Geographic Information Systems*, Keith C. Clarke. New Jersey: Lo, C.P (Chor Pang).

Logan, C. 2005. Borehole Drilling costs, May 6, personal conversation.

Logan, C., Russell, H. A. J. and Sharpe, D. R. 2001-D1. Regional three-dimensional stratigraphic modelling of the Oak Ridges Moraine area, southern Ontario. *Geological Survey of Canada:* 19.

Manslow, J. F. and Nixon, M. S. 2002. On the Ambiguity Induced by a Remote Sensor's PSF. In *Uncertainty in Remote Sensing and GIS*, Foody, G. M. and Atkinson, P. M. (Eds.), John Wiley & Sons, England.

Marceau, J. Danielle. 1999. The scale issue in social and natural sciences. *Canadian Journal of Remote Sensing* 25 (4):347-356.

Morrison, J. L. 1995. Spatial Data Quality. In *Elements of spatial data quality*, edited by S. C. Guptill and J. L. Morrison. New York: Elsevier Science Ltd.

Motro, A. 1997. *Sources of Uncertainty, Imprecision and Inconsistency in Information Systems.* In Uncertainty Management in Information Systems: From Needs to Solution, Motro, A. and Smets, P. (Eds.), Kluwer Academic: The Netherlands.

Nakaya, Tomoki. 2000. An information statistical approach to the modifiable areal unit problem in incidence rate maps. *Environment and Planning A* 32 (1):91-109.

Natural Resources, Canada. 2005. Oak Ridges Moraine 2003 [cited March 28 2005]. Available from http://sts.gsc.nrcan.gc.ca/orm_dcp/index_e.asp?CaId=2&PgId=3.

Openshaw, S. 1984a. The Modifiable Areal Unit Problem. *Concepts and Techniques in Modern Geography* (CATMOG), no. 38.

Openshaw, S. 1984b. Ecological fallacies and the analysis of areal census data. *Environment and Planning A* 16: 17-31.

Openshaw, S. 1989. *Learning to live with errors in spatial databases.* In The Accuracy of Spatial Databases, Goodchild, M. F. and Gopal, S. (Eds.), Taylor & Francis, London, pp. 263 - 76.

Pang, Alex. 2001. Visualizing Uncertainty in Geo-spatial Data. Santa Cruz: University of California.

Pawlak, Zdzislaw. 1982. Rough sets. *International Journal of Computer and Information Sciences* 11:341-356.

Pawlak, Zdzislaw, Jerzy Grzymala-Bausse, Roman Slowinski, and Wojciech Ziarko. 1995. Rough Sets. *Emerging Technologies; Communication of the ACM* 38 (11):89-95.

Piatetsky-Shapiro, G. 1997. *Knowledge Discovery and Acquisition from Imperfect Information.* In Uncertainty Management in Information Systems: From Needs to Solution, Motro, A. and Smets, P. (Eds.), Kluwer Academic: The Netherlands.

Plewe, B. S. 2002. The Nature of Uncertainty in Historical Geographic Information. *Transactions in GIS* 6(4): 432 - 456.

Plewe, B. S. 2003. Representing Datum-level Uncertainty in Historical GIS. *Cartography and Geographic Information Science* 30(4): 319 - 334

Price, Michael. 1985. *Introducing Groundwater.* London: George Allen & Unwin Ltd.

Raubal, Martin. 2001. Ontology and epistemology for agent-based wayfinding simulation. *International Journal of Geographical Information Science* 15 (7):653-665.

Reynolds, Harold David. 1998. *The Modifiable Area Unit Problem: Empirical Analysis by Statistcal Simulation.* Doctor of Philosophy, Graduate Department of Geography, University of Toronto, Toronto.

Rokos, D., Petrou, M. and Desachy J. 2004. *Multi-Sources Information Fusion for Satellite Images Classification* [Internet]. National Technical University of Athens: Laboratory of Remote Sensing 2004 [cited February, 10 2004]. Available from http://www.survey.ntua.gr/main/labs/rsens/DeCETI/IRIT/MSI-FUSION/index.html.

Russell, H. A. J, C Logan, T. A Brennand, M. J. Hinton, and D. R Sharp. 1996. Regional geoscience database for the Oak Ridges Moraine project (southern Ontario). *Current Research 1996-E; Geological Survey of Canada*:191-200.

Russell, H. A. J., Brennand, T. A., Logan, C. and Sharpe, D.R. 1998-E. Standardization and assessment of geological descriptions from water well records, Greater Toronto and Oak Ridges Moraine areas, southern Ontario. *Geological Survey of Canada*: 89 -102.

Salge, F. 1995. Semantic Accuracy. In *Elements of spatial data quality*, edited by S. C. Guptill and J. L. Morrison. New York: Elsevier Science Ltd.

Schuurman, Nadine. 2002. Flexible Standardisation: Making Interoperability Accessible to Agencies with Limited Resources. *Cartography and Geographic Information Science* 29 (4):343-353.

———. 2004. *GIS: A short Introduction.* Oxford: Blackwell.

Sharp, D. R, L. D. Dyke, S. E. Hinton, H. A. J. Russell, T. A. Brennand, P. J. Barnett, and A. Pugin. 1996. Groundwater prospects in the Oak Ridges Moraine area, southern Ontario: application of regional geological models. *Current Research; Geological Survey of Canada*:181-190.

Sharp, D. R., P. J. Barnett, H. A. J. Russell, T. A. Brennand, and G. Gorrell. 1999. Regional geological mapping of the Oak Ridges Moraine, Greater Toronto Area, southern Ontario. *Current Research; Geological Survey of Canada*:123-136.

Shaw, G. and Wheeler, D. 1994. *Statistical Techniques in Geographic Analysis.* Second ed. London: David Fulton.

Shibli, A. R. Syed. 2004. *Conditional Simulation* 2003 [cited 11th August 2004]. Available from http://www.ai-geostats.org.

Skinner, Brian, J, and Stephen Porter, C. 1989. *The Dynamic Earth: an introduction to physical geology.* New York: John Wiley & Sons Inc.

Smets, P. 1997. *Imperfect Information: Imprecision and Uncertainty.* In Uncertainty Management in Information Systems: From Needs to Solution, Motro, A. and Smets, P. (Eds.), Kluwer Academic: The Netherlands.

Theobald, M. David. 2001. Topology revisited: representing spatial relations. *International Joounal of Geographic Information Science* 15 (8):689-705.

Tobler, W. R. 1970. A computer movie simulating urban growth in the Detroit region. *Economic Geography* 46:234 - 240.

Tood, D. K. 1964. *Ground Water Hydrology.* John Wiley & Sons; New York.

Tolman, C. F. 1937. *Ground Water.* Mcgraw-Hill Books; New York.

Tranmer, M. and Steel, D. G. 1998. Using census data to investigate the causes of the ecological fallacy. *Environment and Planning A* 30:817-831.

Veregin, H. 1999 Data quality parameters, in P.A. Longley, M.F. Goodchild, D.J. Maguire and D.W. Rhind (Eds), *Geographical information systems*, 177-189, New York, Wiley.

Warren, Anthony J., Michael J. Collins, Edward A. Johnson, and Peter F. Ehlers. 2002. Managing Uncertainty in a Geospatial Model of Biodiversity. In *Uncertainty in Remote Sensing and GIS*, edited by G. M. Foody and P. M. Atkinson. Chichester: John Wiley & Sons Ltd.

Wicander, Reed, and James Monroe, S. 1995. *Essentials of Geology*. New York: West Publishing Company.

Worboys, Mike. 1998. Imprecision in Finite Resolution Spatial Data. *GeoInformatica* 2 (3):257-279.

Worboys, M. F. and Clementini, E. 2001. Integration of Imperfect Spatial Information. *Journal of Visual Languages & Computing* 12(1): 61 - 80.

Zhang, J and Goodchild, M. F. 2002. *Uncertainty in geographical information*. New York: Taylor & Francis, 2002.

Zlatanova, Siyka, Alias Abdul Rahman, and Wenzhong Shi. 2004. Topological models and frameworks for 3D spatial objects. *Computers & Geosciences* 30:419-428.