1

PM-1 3½"x4" PHOTOGRAPHIC MICROCOPY TARGET
NBS 1010a ANSI/ISO #2 EQUIVALENT

National Library
of Canada

Bibliothèque nationale
du Canada

Canadian Theses Service

Service des thèses canadiennes

Ottawa, Canada
K1A 0N4

NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Reproduction in full or in part of this microform is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30, and subsequent amendments.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

La reproduction, même partielle, de cette microforme est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30, et ses amendements subséquents.

**KNOWLEDGE ACQUISITION AND SYSTEM DESIGN FOR AN
EXPERT ADVISORY SYSTEM FOR STATISTICAL ANALYSIS
(EASSA)**

**FOR ELECTRONIC MAPPING AND
GEOGRAPHICAL INFORMATION SYSTEMS**

By Aining Zhang

**A thesis submitted to
the Faculty of Graduate Studies and Research
in partial fulfillment of
the requirements for the degree of**

Master of Arts

Department of Geography

Carleton University

Ottawa, Ontario

August 2, 1991

© copyright

1991, Aining Zhang

National Library
of Canada

Bibliothèque nationale
du Canada

Canadian Theses Service Service des thèses canadiennes

Ottawa, Canada
K1A 0N4

The author has granted an irrevocable non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of his/her thesis by any means and in any form or format, making this thesis available to interested persons.

The author retains ownership of the copyright in his/her thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without his/her permission.

L'auteur a accordé une licence irrévocable et non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de sa thèse de quelque manière et sous quelque forme que ce soit pour mettre des exemplaires de cette thèse à la disposition des personnes intéressées.

L'auteur conserve la propriété du droit d'auteur qui protège sa thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

ISBN 0-315-70934-0

Canada

**The undersigned recommend to the Faculty of Graduate Studies
and Research acceptance of the thesis**

**KNOWLEDGE ACQUISITION AND SYSTEM DESIGN FOR AN
EXPERT ADVISORY SYSTEM FOR STATISTICAL ANALYSIS
(EASSA)**

**FOR ELECTRONIC MAPPING AND
GEOGRAPHICAL INFORMATION SYSTEMS**

**submitted by Aining Zhang
in partial fulfillment of the requirements for
the degree of Master of Arts**

Thesis Supervisors

Chair, Department of Geography

Carleton University
Sept. 16, 1991

LIST OF FIGURES

Figure 1	EASSA's Working Environment.....	45
Figure 2	EASSA's Knowledge Structure.....	66
Figure 3	Frame of Three-Phase Deduction.....	77
Figure 4	Frame Structure of Statistical Methods - Examples.....	134
Figure 5	Frame Representation for Statistical Knowledge - An Example.....	135
Figure 6	A Rule Representation Example for Advice-Oriented Knowledge -- Rules for Selecting a Statistical Method	138
Figure 7	Top Level Structure of EASSA.....	142
Figure 8	Interface Design of EASSA.....	144
Figure 9	Components of EASSA's Inference Engine.....	146
Figure 10	Components of EASSA's Knowledge Module.....	147
Figure 11	Components of Frame Management Module.....	147
Figure 12	EASSA's Top Level Reasoning Chart..	149
Figure 13	Reasoning for Analysis Issues.....	152
Figure 14	Reasoning for Statistical Methods.....	155
Figure 15	Reasoning for Sampling Modes.....	157
Figure 16	Reasoning for Hypotheses, Assumptions, Analysis Procedures, Result Interpretations, Graphic Methods, and Limitations.....	158

ACKNOWLEDGEMENTS

I would especially like to thank Dr. D. R. F. Taylor, for his guidance throughout the M. A. program.

The research was accomplished under an Memorandum of Understanding between Carleton University and Energy, Mines, and Resources Canada (EMR). I wish to thank the Geographical Services Division (GSD) of EMR for providing facilities and partial funds. I especially wish to thank Dr. E. M. Siekierska, the Assistant Director of GSD and the adjunct advisor of the M. A. thesis, for her guiding me to this research domain, and for her continued support throughout the course of the research. I also wish to express my gratitude to Mr. R. Groot, the Director of GSD, for his encouragement, support, and constructive discussions.

The research involved interviews of a large number of domain experts. I would especially like to thank Dr. M. Goodchild, for reviewing the Questionnaires and giving valuable comments. I would also like to express my grateful thanks to Dr. J. Clarke, to Dr. L. A. Lapalme-Roy, to Dr. A. Kobayashi, and to Ms. J. Frappier, for completing the lengthy Questionnaires and providing expertise. Thanks are also given to Dr. M. Fox, for reviewing the formulated statistical knowledge, and to Dr. G. Decker, for his concerns and expert advice. The cooperation of Dr. C. Armenakis, Dr. I. Wallace, Ms. D. Chapman, Mr. S. Palko, Mr. P. Harker, Ms. D. Lacasse, and Mr. P. Lloyd are also acknowledged.

In addition, I would like to thank Dr. I. Wallace along with Ms. H. Anderson, for their help during my studies in the M.A. program.

Special thanks go to Dr. L. A. Lapalme-Roy for proof-reading the manuscript.

TABLE OF CONTENTS

ACCEPTANCE SHEET	ii
ABSTRACT	iii
ACKNOWLEDGEMENTS	iv
LIST OF FIGURES	vi
LIST OF APPENDICES	vii
1. INTRODUCTION	1
1.1 ORIGIN OF THE TOPIC.....	1
1.2 AREAS OF CONCERN.....	3
1.3 CURRENT DEVELOPMENT	6
1.4 OBJECTIVES OF THIS RESEARCH.....	8
1.5 STRATEGY FOR THE RESEARCH	9
1.6 METHODOLOGY OVERVIEW.....	12
1.7 PRESENTATIONS AND THESIS STRUCTURE.....	14
2. BACKGROUND OVERVIEW	16
2.1 INTRODUCTION.....	16
2.2 STATISTICS IN GEOGRAPHY	16
2.2.1 DEVELOPMENT OF STATISTICS	16
2.2.2 TYPES OF STATISTICS	17
2.2.3 STATISTICAL METHODS AND GEOGRAPHICAL APPLICATIONS	18
2.3 SPATIAL STATISTICS.....	20
2.3.1 FUNDAMENTAL TOPICS.....	21
2.3.1.1 POINT PATTERN ANALYSIS.....	21

2.3.1.2 SPATIAL AUTOCORRELATION.....	23
2.3.1.3 SPATIAL SURFACES.....	26
2.3.1.4 MOSAIC.....	27
2.3.2 CONCERNS WITH CLASSICAL STATISTICS.....	28
2.3.2.1 ON AUTOCORRELATION.....	28
2.3.2.2 ON RANDOM SAMPLES.....	29
2.4 EXPERT SYSTEM TECHNOLOGY.....	30
2.4.1 CHARACTERISTICS OF EXPERT SYSTEMS.....	30
2.4.2 BASIC COMPONENTS OF EXPERT SYSTEMS.....	32
2.5 ELECTRONIC ATLASES AND THE ELECTRONIC ATLAS OF CANADA.....	33
2.5.1 ELECTRONIC ATLASES.....	33
2.5.1.1 CONCEPT AND DEVELOPMENT.....	33
2.5.1.2 CHARACTERISTICS.....	35
2.5.2 THE ELECTRONIC ATLAS OF CANADA.....	36
2.5.2.1 CARTOGRAPHIC FUNCTIONS.....	37
2.5.2.2 QUERY AND ANALYTICAL FUNCTIONS.....	38
2.5.2.3 DATA INPUT.....	38
2.5.2.4 DATA STORAGE.....	39
2.5.2.5 DATA SETS.....	39
2.6 SUMMARY.....	40
3. EASSA'S FUNCTIONALITY.....	42
3.1 INTRODUCTION.....	42
3.2 TASK ANALYSIS AND FUNCTIONAL SCOPE.....	42
3.3 USER REQUIREMENT ANALYSIS AND ADVISORY ISSUES.....	45
3.4 ADVISORY MODES.....	48
3.5 STATISTICAL TOPICS.....	50

3.6	SELF-EXPLANATION CAPABILITIES	53
3.7	AUTOMATIC KNOWLEDGE ACQUISITION CAPABILITIES.....	53
3.8	SUMMARY.....	55
4.	KNOWLEDGE ACQUISITION AND DEVELOPMENT	57
4.1	INTRODUCTION.....	57
4.2	PROBLEM DECOMPOSITION.....	57
4.2.1	THE CONCEPT OF PROBLEM-SOLVING SPACE.....	58
4.2.2	PROBLEM DECOMPOSITION OF EASSA.....	59
4.2.2.1	TOP LEVEL PROBLEM-SOLVING SPACE.....	59
4.2.2.2	SECOND LEVEL PROBLEM-SOLVING SPACE.....	60
4.2.2.3	THIRD LEVEL PROBLEM-SOLVING SPACE.....	61
4.2.2.4	THE COMPOSITE PROBLEM-SOLVING SPACE OF EASSA.....	64
4.3	STRUCTURED KNOWLEDGE SCHEMAS.....	64
4.3.1	DOMAIN KNOWLEDGE LAYER.....	65
4.3.1.1	ANALYSIS ISSUE KNOWLEDGE SCHEMA.....	65
4.3.1.2	STATISTICAL KNOWLEDGE SCHEMA	68
4.3.1.3	SPATIAL SAMPLING KNOWLEDGE SCHEMA.....	71
4.3.1.4	ATLAS CONTEXT KNOWLEDGE SCHEMA	74
4.3.2	ADVICE-ORIENTED KNOWLEDGE LAYER	76
4.3.2.1	DEDUCTIVE REASONING FRAME.....	76
4.3.2.2	ADVICE-ORIENTED INFERENCE SCHEMAS	77
4.3.3	TASK-ORIENTED KNOWLEDGE LAYER.....	83
4.3.4	META KNOWLEDGE.....	83
4.3.4.1	KNOWLEDGE FOR DEALING WITH TEMPORAL DIFFERENCES.....	84
4.3.4.2	KNOWLEDGE FOR MULTIPLE STATISTICAL METHOD SELECTIONS	85
4.3.4.3	KNOWLEDGE FOR MAKING EXPLANATIONS	85

4.4	KNOWLEDGE EXTRACTION AND DEVELOPMENT	86
4.4.1	STATISTICAL KNOWLEDGE EXTRACTION	87
4.4.1.1	SELECTIONS OF STATISTICAL METHODS	87
4.4.1.2	KNOWLEDGE FORMULATION	93
4.4.2	DEVELOPMENT OF ANALYSIS ISSUES	93
4.4.3	SAMPLING KNOWLEDGE FORMULATION	95
4.4.3.1	METHODOLOGY FOR FORMULATION	96
4.4.3.2	UNIVARIATE ANALYSIS	96
4.4.3.3	BIVARIATE ANALYSIS	97
4.4.3.4	MULTIVARIATE ANALYSIS	99
4.5	KNOWLEDGE VALIDATION	100
4.5.1	OBJECTIVES OF KNOWLEDGE VALIDATION	101
4.5.2	METHODOLOGY FOR KNOWLEDGE VALIDATION	102
4.5.2.1	INTERVIEWS OF EXPERTS	102
4.5.2.2	CASE STUDIES	104
4.5.2.3	EXTENDED REVIEW OF LITERATURE	105
4.5.2.4	KNOWLEDGE REFINEMENT	105
4.5.3	REFINEMENT OF STATISTICAL KNOWLEDGE	106
4.5.3.1	CONSIDERATIONS OF SPATIAL CONTEXT	106
4.5.3.2	CHANGES TO STATISTICAL METHODS	107
4.5.3.3	CHANGES IN THE KNOWLEDGE STRUCTURE	110
4.5.3.4	OTHER CHANGES	112
4.5.4	ON ANALYSIS ISSUE KNOWLEDGE	112
4.6	SUMMARY	113
5.	KNOWLEDGE FORMALISM	115
5.1	INTRODUCTION	115

5.2	EASSA'S KNOWLEDGE MODELS	115
5.2.1	DOMAIN KNOWLEDGE MODELS.....	116
5.2.2	ADVICE-ORIENTED KNOWLEDGE MODELS.....	119
5.2.3	TASK-ORIENTED KNOWLEDGE MODELS	123
5.3	EASSA'S KNOWLEDGE REPRESENTATION METHODS.....	125
5.3.1	REPRESENTATION METHOD DESIGN.....	126
5.3.1.1	PRESENTLY USED REPRESENTATION METHODS	126
5.3.1.2	METHOD SELECTIONS AND CONSIDERATIONS	131
5.3.2	EASSA'S KNOWLEDGE REPRESENTATION METHODOLOGY	133
5.3.2.1	FRAME REPRESENTATION OF DOMAIN-ORIENTED KNOWLEDGE.....	133
5.3.2.2	RULE REPRESENTATION FOR ADVICE-ORIENTED KNOWLEDGE	138
5.4	SUMMARY	140
6.	SYSTEM DESIGN AND BEHAVIOR DEMONSTRATION	141
6.1	INTRODUCTION.....	141
6.2	SYSTEM STRUCTURE OF EASSA	141
6.2.1	TOP LEVEL STRUCTURE	141
6.2.2	SECOND LEVEL STRUCTURE	143
6.2.2.1	INTERFACE MODULE.....	143
6.2.2.2	INFERENCE ENGINE MODULE.....	145
6.2.2.3	KNOWLEDGE BASE MODULE.....	146
6.2.2.4	FRAME MANAGEMENT MODULE.....	146
6.3	REASONING PROCEDURES OF EASSA	148
6.3.1	TOP LEVEL REASONING	148
6.3.2	SECOND LEVEL REASONING	151
6.3.2.1	REASONING FOR ANALYSIS ISSUES.....	151
6.3.2.2	REASONING FOR STATISTICAL METHODS	154

6.3.2.3 REASONING FOR SAMPLING MODES.....	156
6.3.2.4 REASONING FOR OTHER ADVISORY ISSUES.....	156
6.4 DEMONSTRATION OF THE SYSTEM'S BEHAVIORS.....	159
6.5 SUMMARY.....	160
 7. SUMMARY, CONCLUSIONS AND FURTHER DEVELOPMENT ...	161
7.1 SUMMARY.....	161
7.2 CONCLUSIONS.....	164
7.3 FURTHER DEVELOPMENT.....	166
7.4 FINAL COMMENT.....	168
 REFERENCES USED IN STATISTICAL KNOWLEDGE FORMULATION.....	169
REFERENCES CITED.....	170
REFERENCES CONSULTED.....	177

LIST OF FIGURES

Figure 1	EASSA's Working Environment.....	45
Figure 2	EASSA's Knowledge Structure.....	66
Figure 3	Frame of Three-Phase Deduction.....	77
Figure 4	Frame Structure of Statistical Methods - Examples.....	134
Figure 5	Frame Representation for Statistical Knowledge - An Example.....	135
Figure 6	A Rule Representation Example for Advice-Oriented Knowledge -- Rules for Selecting a Statistical Method	138
Figure 7	Top Level Structure of EASSA.....	142
Figure 8	Interface Design of EASSA.....	144
Figure 9	Components of EASSA's Inference Engine.....	146
Figure 10	Components of EASSA's Knowledge Module.....	147
Figure 11	Components of Frame Management Module.....	147
Figure 12	EASSA's Top Level Reasoning Chart..	149
Figure 13	Reasoning for Analysis Issues.....	152
Figure 14	Reasoning for Statistical Methods.....	155
Figure 15	Reasoning for Sampling Modes.....	157
Figure 16	Reasoning for Hypotheses, Assumptions, Analysis Procedures, Result Interpretations, Graphic Methods, and Limitations.....	158

LIST OF APPENDICES

APPENDIX 1: STATISTICAL KNOWLEDGE FOR EASSA.....	180
APPENDIX 2: PRELIMINARY SAMPLING KNOWLEDGE.....	238
APPENDIX 3: CASE DEMONSTRATIONS.....	246

1. INTRODUCTION

1.1 ORIGIN OF THE TOPIC

Most modern geographical information systems (GIS) include complete sets of functions for entering, storing and retrieving spatially indexed information. In contrast, the methods of analysis that are usually included in standard GISs are often limited. They are mostly based on computer related cartographic technology, such as line intersection, point-in-polygon, polygon overlay, and spatial nearness (proximity) between objects (Openshaw et al., 1990; Burrough, 1990). Statistics, though well known as powerful analytical tools, are often limited to simple descriptions on the attributes of the entities. As Goodchild (1987) commented: "In part, this situation is a reflection of demand and is self-fulfilling.....In part, it is behavioral, since much of the current widespread enthusiasm for GIS is undoubtedly due to the attractiveness of their graphical output and their use of high technology."

Such analysis functions can satisfy the needs of GIS users in analyzing simple issues, for example, Ottawa-Hull urban area land use (Marble and Peuquet, 1986). Nevertheless, spatial analysis can be far more sophisticated than this, hence a need to incorporate sophisticated statistical tools into GISs. Molenaar (1989) regarded spatial analysis as one of four main application domains of GISs, the others being spatial planning, implementation of planning, and environmental monitoring and management. Examples of spatial analysis include the study of major contributory factors to cancer occurrences, comparison of average income as well as income variation between regions, or classification of cities into several groups for marketing decision making. It is possible to use GISs for these analyses, however, this possibility has never become true in current standard GISs.

The integration of a large range of statistics into GISs would also make it possible to satisfy the need "to close the gap between the technical sophistication of spatial analysis and the relative crudeness of its applications" (Openshaw et al., 1987). While there have been a number of systems, such as SAS and SPSS, that have dealt with statistical analysis, the spatial context (see Section 1.2) makes general statistical systems inadequate to support the analysis (Goodchild, 1987). A GIS, however, is a distinct environment with both geographical information and spatial operation capabilities or potential capabilities, hence it is especially supportive for conducting spatial analysis.

The Electronic Atlas of Canada (EA), being developed in the Geographical Services Division (GSD) of Energy, Mines, and Resources Canada (EMR), is a system which shares the main principles of general GISs, but is characterized by its built-in maps of the National Atlas of Canada, and by a user friendly interface for creative and cognitive interaction of users and maps (Groot, 1990). The potential benefits of incorporating statistics into GISs motivated a plan in GSD, that is, to integrate statistical functions into the EA system. Thus, not only would spatially distributed phenomena be displayed, but significant hypotheses could be generated and confirmed; not only could individual themes be depicted, but the interrelationships between them could be explored and revealed. Also, information newly derived by analysis as well as captured data would be queried and visualized.

However, statistics, especially spatial statistics and multivariate techniques, are not widely used by GIS users, not only because they are not provided in current systems, but also because of their complexity. Therefore, there is a need to develop a guiding system that could act as an advisor and make the complex techniques available to all users. Robinson and Frank (1987) addressed this need:

"Spatial analysis is an extremely complicated field where expertise in spatial statistics, data quality, data structure, and interpretation interact to help determine the correct combination of techniques to apply to a particular problem. The development of such advisors would bring about practical improvements in geographic decision making and formalize geographic knowledge in a manner that makes it accessible to nonspecialists."

It is this need that initiated the development of the Expert Advisory System for Statistical Analysis (EASSA) for electronic mapping and GISs, in particular for the Electronic Atlas of Canada.

1.2 AREAS OF CONCERN

The development of EASSA is an interdisciplinary research topic, which involves statistics, spatial statistics, spatial analysis, expert system methodology, GIS as well as electronic mapping methodology, and cartographic modeling of geographical phenomena. Geography is also involved in the system development.

Statistics. In order to derive advice on statistical analysis, the system has to possess statistical knowledge sufficient to derive advice, as conducted by a human statistician. Therefore, it is of vital importance to formulate statistical knowledge in the system development. It is this type of knowledge that constitutes the core of the system.

Spatial statistics and spatial analysis. The data dealt with differ from general data in that they are not only measures of attributes, but also measures of spatial locations. This characteristic makes it difficult to directly apply classical statistical methods, since these approaches often have assumptions that are often violated in the spatial context. Spatial

statistics and spatial analysis are the domains that specifically address these problems, and hence must be involved in the formulation of statistical knowledge.

Expert system. The advisory system aims to emulate human statisticians' thinking behaviors in order to generate a variety of analytical advice. Emulation of human beings' intelligent behaviors is the domain of artificial intelligence, of which expert system technology is a branch. This technology deals with symbolic reasoning, uncertainty problems involved in the reasoning, self-explanations for behaviors, and auto-knowledge acquisition. All of these issues need to be involved in the development of the statistical advisory system.

GIS and electronic mapping. EASSA will work within the environment of the Electronic Atlas of Canada, a GIS-like system with emphasis on interactive cartographic functions. Based on EASSA's advice, the user will use sampling tools provided by the EA to obtain analytical data for analysis. Data sampling for map-based analysis differs from general statistical sampling in two major aspects. First, it involves the data models of the system. For example, bivariate or multivariate statistical analysis require bivariate or multivariate data sets sampled from two or several maps with the same sampling points. For raster data, there is not much difficulty in conducting this kind of sampling. However, problems may arise when the data are modeled in vector format, since some data may not be available in some sampling points. Secondly, the definitions of some statistics have to take into account the capabilities or potential capabilities of data queries in a GIS or electronic mapping environment. For example, the distances between general cities and municipalities may be used as a statistic in city evaluation and related statistical analysis. However, whether it is feasible will depend on whether the GIS system can or has the potential to provide the required sampling means. Therefore, there is a need to fully understand the capabilities and potential capabilities of GIS and electronic mapping

technology (refer to Section 2.5), particularly the Electronic Atlas of Canada, in order to develop a practical advisory system.

Cartographic modeling. The advisory system will deal with cartographic models of geographical phenomena. General GISs possess only a few types of cartographical models such as choropleth maps, nominal maps. The Electronic Atlas of Canada, however, is an electronic version of the National Atlas of Canada, and thus contains almost all kinds of cartographic models, including various cells of geographical features, from point, line, areal, to volume, and involves various cartographic representation methods, from quantitative to qualitative, and from analytical to synthetic. The methods chosen for specified analyses are to large degree dependent on the cartographical models used for representing the phenomena.

Geography. The National Atlas of Canada, as the source of maps for the EA, embraces a variety of geographical themes, including some in physical, environmental, social, cultural, economic, and resource geography. A major reason for adding statistical functions to the EA is to explore the interrelationships between various geographical phenomena. However, not every pair of phenomena can be used for correlation analysis, since they may essentially not correlate. For example, there is no logical relationship between library expenditure and cancer rate, though they may correlate with each other when correlation analysis is applied. Therefore, there is a need for the expert advisory system to provide advice on which geographical phenomena have potential correlation relationships. The same applies to other statistical issues such as regression analysis and factor analysis. Thus, geographical knowledge has also to be taken into account in system development.

1.3 CURRENT DEVELOPMENT

Enhancement of the analytical functionality in GISs has become an increasingly important GIS research topic in recent years. Goodchild (1987) addressed the spatial perspective on geographical information systems, emphasizing the importance of spatial analysis capabilities of GISs. NCGIA (National Center for Geographical Information Analysis, USA) put spatial analysis and spatial statistics in five top research priorities in their 1989 Research Plan (Goodchild, 1990). SPANS, the Spatial Analysis System developed by TYDAC Inc. of Canada, puts emphasis on modeling (TYDAC, 1990), which is particularly welcome by GIS users, though there are essentially little spatial statistics in the current product. A few small systems, e.g. the IDRISI Version 3.1 (Eastman, 1990), developed at Clark University, have added some preliminary spatial statistical functions, such as spatial autocorrelation analysis and trend surface analysis. The attempt to strengthen data manipulation capabilities has also lead to the development of data transfer routines between GISs (e.g. SPANS, CGIS) and statistical packages (e.g. SAS, SPSS). There seems to have been more activities in Europe. Openshaw et al. (1987) developed a Geographical Analysis Machine, which emphasizes the interaction of spatial coverages. A prototype "Geographic Correlates Exploration Machine" (GCEM) was further developed by Openshaw et al. (1990), in order to search for relationships and to generate hypotheses for subsequent testing. Kehris (1990) incorporated autocorrelation statistics into the ARC/INFO environment; applications to epidemiological research have been carried out by Gatrell and Dunn (1990), using ARC/INFO as well as GCEM. Oliver and Webster (1990) integrated geostatistical techniques (Kriging) into a GIS analysis context with special applications to soil mapping. Similar research was carried out by Okx et al. (1990), integrating GIS and geostatistics in soil pollution studies. Haslett et al. (1990) developed SPIDER, an interactive statistical tool for the analysis of spatially distributed data in a GIS context, and the authors argue that the system can provide a very rich environment for

pursuing exploratory statistical analysis of spatial data. Bailey (1990) devised INFO-MAP, an unsophisticated system with the simplest data management capabilities, but possessing a powerful range of standard and spatially oriented statistical functions, and the ability to interactively present results through choropleth, contour, proportionate symbol, or proximally shaded maps. A few efforts have also been made in other countries. Wang Zheng et al. (1990, China) incorporated a range of multivariate statistical models into PURSIS (Peking University Real-world Spatial Information System), and commented that the models have proven valuable in regional planning. A Geographical Information, Analysis and Modeling System (GIAMS), developed in Karlsruhe (Kilchenmann, 1990), is comprised of several tools, including the GIS ARC/INFO tool, the relational database management system ORACLE, the image processing system ERDAS, the computer aided cartographic system UNIRAS, and the statistical analysis system SAS. All these systems shared the same aim, namely, to enhance analytical capabilities, but some were more focused on dealing with the spatial characteristics of data, while other put the emphasis on statistical functionality.

Expert system technology has been increasingly applied to various domains of GISs. These systems can be classified into four major groups (Robinson and Frank, 1987). The first group includes the systems that assist in map design; to this group belong the MAPEX for automatic map generalization (Nickerson and Freeman, 1984), AUTONAP for automatic name placement (Freeman and Ahn, 1986), MAP-AID (Robinson and Jackson, 1985), ACES (Pfefferkorn et al. 1986) and CES (Muller et al., 1986) for map design, MaPKBS (Jankowski and Nyerges, 1989; Nyerges and Jankowski, 1989) for map projection selection, etc. A recent development is found in a frame-based advisory system for cartographical symbol design for thematic maps (Muller and Wang, 1990). The second category involves terrain/feature extraction; examples are FES (Goldberg et al., 1984; Goodenough et al., 1987), CERBERUS (Engle, 1985), SPAM (MaKeown et al., 1989),

DNESYS (Qian et al., 1990), etc. The third group includes the systems used for geographical database management, to which belong ORBI (Pereira et al., 1982), KBGIS-II (Smith et al., 1987), SRAS (Robinson and Wong, 1987), LOBSTER (Egenhofer and Frank, 1990), etc. The last category refers to geographical decision support systems, such as URBYS (Tanic, 1986) for assisting in evaluating site suitability for land-use activities, and ASPENEX (Morse, 1987) for assisting in the management of aspen. (after Huang, 1991).

Although there is no expert system to guide users in utilizing statistics or spatial statistics in a GIS environment, several prototype statistical expert systems have been developed in non-GIS fields. MUSE is such a system developed by Dambroise and Massotte (1986). STATEX (Muzard et al., 1991), developed at Communication Canada, is another example, which involves basic statistical functions with a user-friendly iconic interface.

In summary, the enhancement of system capabilities in spatial analysis and the incorporation of expert system technology into GISs are two trends in GIS development. Efforts have also been made in non-GIS fields to apply expert system technology in helping users conduct conventional statistical analysis. However, no work has been found in the development of expert systems for statistical analysis in a GIS environment.

1.4 OBJECTIVES OF THIS RESEARCH

(1) *Framework development for EASSA.* It is an original research topic to integrate expert system technology and statistics into an electronic mapping or GIS environment, in particular the Electronic Atlas of Canada. Therefore, the first objective of this research is to develop a framework both for physical configuration of the integration and for conceptual structure of the system.

(2) *Knowledge Acquisition.* The key task in knowledge engineering or expert system development is to develop high quality expertise. In order to act as a human statistician to give advice on spatial statistical analysis, EASSA needs a great deal of statistical knowledge, which will constitute the core of the system. The acquisition of statistical knowledge is therefore set as the major objective of the research. The research also requires the development of expertise on spatial queries, which are needed to serve as sampling tools in the analysis processes, on geographical analysis issues, which are intended to facilitate users using the EA information for analysis, and on the map contexts of the EA, which provides a basis for statistical method selections.

(3) *Knowledge Formalism.* The third objective of this research is to concisely model the knowledge to be used in EASSA, and design an appropriate method to represent the knowledge. This is particularly important in EASSA's development, since the system is knowledge-intensive, and a wide range of knowledge has to be involved. An efficient system can only be based on a clearly defined knowledge set that can facilitate knowledge auto-acquisition, and on an appropriate knowledge representation approach that provides a clear structure of knowledge organization and that facilitates the reasoning of the system.

(4) *System Design.* The design of EASSA's logical structure and reasoning framework is set as the last objective of the research. By following the design, and using the related knowledge as well as the knowledge representation method, a knowledge engineer should have no difficulty in implementing the prototype system of EASSA.

1.5 STRATEGY FOR THE RESEARCH

The development of a complete statistical advisory system for GIS- or EA-based analysis is

an extremely large project, since statistics form a large domain, spatial statistics are even more complex, and the formulation of a set of geographical analysis issues necessitates a wide range of knowledge on geographical analysis. Technically, the system's functionality could be varied from a simple knowledge-based system to a system with advanced learning capabilities. Therefore, there is a need to set a research strategy.

Starting with classical statistics but taking spatial context into account. Classical statistics provide powerful tools for various information analyses. Considerations should also be given to spatial statistics, since they are pertinent approaches to analysis in spatial contexts. However, it is not possible to include all statistics in a system, and hence both necessity and feasibility need to be considered in the determination of the ranges of statistics for EASSA. The necessity is determined from the users of the system, which can be divided into two categories: the map interpreters and the map makers of the National Atlas. The former could be geographers, environmental specialists, resource managers, or general map readers. The classical statistics would be useful for all of them, but only some of them, e.g., geographers, would be interested in pure spatial analysis. The latter category of users, i.e., map makers, includes both geographers who work on the compilations of map contents, and cartographers who work on map design and representations. Most classical methods are helpful for them, while specific spatial approaches are not necessary, except for a few geostatistical techniques. From the viewpoint of feasibility, classical statistics have been utilized over a half century, and are hence more mature than pure spatial techniques, which were spread in the 1970's. Based on both necessity and feasibility, therefore, classical statistics should be ranked first in priority, and spatial analysis and spatial statistics second. However, classical statistics have to be adapted to the spatial information environment, since the assumptions made in classical methods are frequently violated in spatial contexts. Also, the system should move toward more spatial analysis in

its future development, through the development of the domain of spatial statistics and spatial analysis itself.

Using multiple methods for expertise development. EASSA's knowledge base will be composed of several types of knowledge, including statistics, analysis issues, sampling knowledge, and Atlas contexts. The statistical knowledge forms a basic expertise and hence is emphasized in this research. The approach to knowledge acquisition is to extract the knowledge from texts. The techniques for data sampling involve advanced query functions of GISs, some of which have not yet been available in methodologies and in existing systems. Therefore, expertise on this knowledge should involve the summary of existing methods and the design of new query functions. The formulation of analysis issues involves a wide range of theoretical and empirical geographical expertise, hence there is high cost in knowledge acquisition. However, analysis issues are not the indispensable expertise in EASSA. Only an attempt has therefore been made to formulate this knowledge, leaving completion of the knowledge for further development. The knowledge auto-acquisition function of the system will be used to acquire this knowledge during consultations. The Atlas contexts deal with the ways in which the Atlas maps are stored in the EA, and the manners in which geographical themes are modeled (refer to 4.3.1.4). This knowledge also remains for auto-acquisition by the system, since it is not fundamental for the prototype development, it contains a set of facts with simple syntax, which are easily collected automatically, and also, it is not currently available.

Incremental development of the system's functionality. The development of the system will start with the advisory system which aims to give advice on statistical analysis, move to a statistical expert system which integrates both reasoning and data processes into a single system, and then expand towards a distributed decision support system which consists of a variety of subsystems used for various decision making tasks. In the phase of

the advisory system development, the system's functionality will also be expanded step-by-step. The first system will involve fewer statistical methods, and will have simple knowledge auto-acquisition mechanism capable of collecting facts. More statistics will be included, and a full capability for knowledge acquisition may be considered in its further development.

1.6 METHODOLOGY OVERVIEW

Development of expert systems is also termed knowledge engineering. A number of researchers address development methodologies (Payne and McArthur, 1990; Kidd, 1987; Waterman, 1986). The methodologies adopted in the development of EASSA abide by the basic principles presented in the literature dealing with development methodologies, but with consideration of the domain requirements.

Problem Identification. First of all, the problem domain was identified, namely, the problems, their domains, the subproblems, the expertise needed, the extent to which the system will solve the problems, etc. The identification was conducted by examining some statistical software, by analyzing the themes and data characteristics of the National Atlas of Canada, and by reviewing literature on statistics and geographical analysis. This resulted in an outline of the advisory system.

Knowledge Schema Construction. After the problems were identified, a set of knowledge schemas were constructed, by decomposing the problems into multiple levels, and finding the knowledge elements which are needed in problem-solving at various levels. These schemes actually constitute the knowledge structure of the system.

Knowledge Extraction. Related knowledge sources were selected from the literature, and the knowledge was extracted or further developed in terms of the knowledge schemas. The emphasis of the knowledge extraction process was put on the statistical domain, especially classical statistics, for which there have been a large number of texts illustrating principles and methodologies. Spatial sampling knowledge was summarized according to analytical requirements. An attempt was also made to identify analysis issues that are of common interest and make sense. This was done based on the themes of the National Atlas, general geographical principles, and common understanding of geographical analysis.

Knowledge Validation. Since the knowledge was mostly extracted from literature, it was necessary to interview domain experts to ensure its applicability. Two questionnaires were designed, one for statistics and one for analysis issues. They were sent to about twenty-five experts for evaluations. The knowledge was refined according to the experts' responses. Then, reinterviews were carried out, and refinement was conducted again. The emphasis of validation was also put on statistical knowledge, because of the focus of the research.

Knowledge Modeling. The acquired knowledge represents a part of the knowledge instances to be used in the system. The system's knowledge as a whole, however, needs to be further conceptualized, so as to be presented concisely, precisely, and formally. This is especially useful in knowledge auto-acquisition. The conceptualization was carried out by modeling the knowledge using mathematical notation.

Knowledge Representation. The mathematical models are conceptual descriptions of the knowledge. Yet physically, the models as well as the corresponding knowledge have to be represented in appropriate ways in the expert system. This is the issue of knowledge representation. A review of existing knowledge representation methods was carried out, in

order to optimally represent the intensive knowledge involved in EASSA. Approaches were then designed for each type of knowledge of EASSA.

System Design. Knowledge formalism, from knowledge modeling to representations, has provided a conceptual design for EASSA. System design deals specifically with the logical structure and the reasoning procedures of the system. It aims to develop diagrams and specifications, following which a knowledge engineer should be able to implement the system.

1.7 PRESENTATIONS AND THESIS STRUCTURE

There are two major submissions of the research: the research report (thesis), and a set of acquired and validated statistical knowledge (Appendix 1).

The statistical knowledge was formulated by reviewing about fifteen statistical texts and other related literature. About forty statistical methods are included in the knowledge set, covering a comprehensive statistical domain, ranging from elementary to advanced, from univariate to multivariate, from parametric to non-parametric, and from classical to spatial. The emphasis was placed on the classical techniques. The knowledge for each method includes fourteen items, involving its applicability, principles, methodology for usage, etc.

The thesis presents the research project. The first chapter gives an introduction to the project, describing why and how the project was carried out, its main topics, and the research completed. The next chapter provides an overview of the background knowledge of the project, discussing both major problem domains (statistics) and technical issues (expert systems, EA). Chapter 3 presents the expert advisory system, including advisory issues, advisory modes, statistical methods, and other system capabilities, such as self-

explanation and learning. Knowledge acquisition, a key issue of this research, is addressed in detail in Chapter 4, from problem decomposition, construction of EASSA's knowledge schemas, through knowledge extraction, to knowledge validation. The validated knowledge is then formalized in the next chapter, by means of mathematical modeling and knowledge representation schemes. Based on the formalized knowledge as well as the system functionality, Chapter 6 presents a thorough logical design for EASSA and case demonstrations of system's external behavior. The limitations of the system and further development are discussed in Chapter 7. In the end, a summary and conclusions are given, and the directions for further development are recommended.

In addition to Appendix 1, two other appendices are presented. Appendix 2 deals with sampling knowledge; fourteen spatial query modes are used in acquiring analytical data from the EA. Appendix 3 presents case demonstrations by the advisory system. The system's external advisory behaviors are demonstrated through two cases, namely, a randomness test for spatial distribution of life expectancy, and an equality test of average income between two regions.

2. BACKGROUND OVERVIEW

2.1 INTRODUCTION

In the development of EASSA, several areas must be dealt with. First, the system aims to act as a human statistician, hence, statistics are the key domain of the project. Secondly, spatial statistics has to be involved, since the statistical analysis deals with spatial information. Thirdly, the system is going to emulate human experts behavior in solving problems, and expert system technology is hence desirable. Finally, the statistical analysis is to be carried out in the environment of GISs, particularly of the Electronic Atlas of Canada, therefore, it is necessary to have a general understanding of electronic atlases as well as the Electronic Atlas of Canada. This chapter presents an overview of the concerned domains. Cartographical models and geography are also needed in this project, but are not included here because of space limitations and because the literature and research in these fields is more extensive than in the issues described.

2.2 STATISTICS IN GEOGRAPHY

Collections of facts and figures are constantly encountered both in the pursuit of professional goals and in every day life. Statistics, in the sense of a subject of study, is a scientific discipline that provides methods to assess such data (Devore and Peck, 1986). It is a branch of mathematics with its own concepts, methods and values, which find many applications in geography.

2.2.1 DEVELOPMENT OF STATISTICS

Statistics was initiated with verbal descriptions in the mid-eighteenth century, and was

considered to be a part of politics. During the course of the nineteenth century, statisticians began to devote more and more time to the organized collection and presentation of numerical data, and statistics was quickly developed and increasingly applied to many disciplines such as agriculture, industry, and censuses. However, it was not until about 1950 that geographers began to take advantage of statistical techniques for geographical analysis (Williams, 1984). The ensuing "Quantitative Revolution" virtually initiated a new way of thinking. There was a period of time when every geographical problem seemed to be statistical in nature. This enthusiasm is waning somewhat as geographers become more fully aware of the assumptions and limitations of these techniques. Since the 1970's, more attention has been paid to overcoming the restrictions of statistics and to seeing problems in a more holistic manner. Williams (1984) commented:

"The Quantitative Revolution, while it may have over-reached itself, has been, on balance, a great blessing. It has revitalized geography, giving it new directions, and providing it with powerful tools."

2.2.2 TYPES OF STATISTICS

Statistics can be categorized into two basic types: descriptive statistics and inferential statistics.

Descriptive Statistics. A large array of figures makes tedious reading and is difficult to interpret, but if the same figures are presented in the form of graphs or histograms, their essentials may be grasped virtually at a glance. Alternatively, the figures can be summarized by using percentages, averages, and other numerical measures. These form the subject matter of descriptive statistics, which attempt to provide methods for organizing and simplifying data so that their significance is easily comprehensible. Some of the more

recently developed descriptive techniques fall under the heading of what statisticians now refer to as exploratory data analysis (EDA). These include ingenious plots and visual displays designed to uncover some of the subtler relationships in the data and reveal more clearly any hidden messages and patterns (Devore and Peck, 1986). These also involve a method of examining a set of data from various angles, and of synthesizing information about the system of study that is revealed by each analysis (Mather, 1976).

Inferential Statistics. The entire collection of objects, or measurements about which information is desired is called the population of interest. Usually, a population is so large that only a part or subset of the population, referred to as a sample, can be investigated. It is also possible that populations are infinite or not easily defined, hence only samples are available. However, no sample can be completely trusted. There must inevitably be some uncertainty about any inference based on a limited number of observations. Thus, using sample information to draw conclusions about a population or several populations forms the second branch of statistics, inferential statistics, which has been the major preoccupation of modern statistics. Contrary to exploratory data analysis, an advance of inferential statistics refers to confirmatory methods, which provide estimates of the parameters of *a priori* models, and allow significance tests of hypotheses relating to the values taken by these parameters (Mather, 1976).

2.2.3 STATISTICAL METHODS AND GEOGRAPHICAL APPLICATIONS

Statistics provides a number of analysis methods, almost all of which can be used in geographical analysis.

The most elementary statistical methods are basic descriptive statistics. The first group of descriptive statistics are tabular and pictorial methods for describing data, such as

contingency tables, frequency distributions, histograms, etc. The second group refers to numerical summary measures, such as measuring the central tendency of a data set, measuring variability in a data set, and describing the distribution of a data set. Summarizing bivariate data forms the third group of the elementary techniques, to which belong scatter plots, correlation analysis, and fitting a line to bivariate data. All of these methods have been extensively used in summarizing geographical data, either for information analysis, or for map compilation.

More advanced techniques are the inferential techniques dealing with one or two samples or two variables. Among these, single sample tests are used to test hypotheses on means and proportions by using the simplest student t-test or z-test. Two sample tests compare means, proportions, distributions, and variations between two samples, for which a large number of tests, either parametric or nonparametric, are devised to deal with different data characteristics. Examples are t-test, Mann-Whitney U-test, Chi-square analysis, etc. Several methods are devoted to the analysis of the relationship of bivariate data; to this category belong linear or non-linear regression analysis, and a number of correlation analyses, such as Pearson's product moment coefficient, and Spearman's correlation coefficient, all of which deal with various data characteristics. The common basis for all of these methods is the formulation of a hypothesis and the test of the hypothesis using sampling distribution knowledge. These methods are widely used in geography in comparison analysis and simple relationship analysis of geographical phenomena (e.g. R.K. Wilkinson and E.M. Sigsworth, 1972).

The most sophisticated approaches of statistics are the multivariate analysis methods, which handle interrelationships among more than two variables. The most often used technique is the multiple regression analysis, which has been widely employed to explore dependent relationships among multiple geographical phenomena (Cook and Pocock, 1983). Cluster

analysis and discriminant analysis are used to categorize multivariate objects or multiple variables into several groups; their major applications in geography are regionalizations and resource classifications. Factor analysis and principal component analysis deal with the simplification of multivariate interdependent relationships, and have been adopted to analyze individual geographical systems. An example is given in Table 19 of Appendix 1. Multivariate analysis of variance can be used to compare multiple samples, assess the effects of multiple factors on a response variable, and examine interactions among multiple variables. Its applications in geography can be found in geographical comparison analysis among multiple objects with respect to multiple factors; for example, it can be used to analyze whether mean energy consumptions differ by income per capita in different regions. In addition, there are also some less used developed techniques such as canonical correlation analysis and character analysis (Levine, 1977; Kilchenmann, 1990).

A technique especially useful in geography and cartography is the univariate classification approach, which deals with the classification of univariate objects into several groups. A number of methods have been developed for this purpose, which can be generalized into three categories. The first category refers to as regular series, including equal step, equal frequency, equal membership, mean-and-standard-deviation, and nested-means. The second category includes systematically unequal series, such as arithmetic, geometric and reciprocal methods. Irregular series, including graphic techniques and iterative techniques, form the third category. Experiments have shown that, among these approaches, the iterative techniques are the most effective ones in terms of classification accuracy evaluations (Currie, 1989).

2.3 SPATIAL STATISTICS

The spatial nature of geographical phenomena makes it insufficient to summarize the data

using classical statistics such as means and standard deviations. Instead, it is necessary to analyze spatial patterns and to reduce the patterns to a few clear and useful summaries. This is referred to as 'spatial data analysis'. Using statistics to conduct spatial data analysis thus forms the domain of 'spatial statistics', which draws statistical inferences by comparing the summaries of spatial analysis with what might be expected from theories of how the pattern might have originated and developed (Ripley, 1981). Spatial statistics is a relatively young domain, which originated with point pattern analysis in the 1960's, and has begun to serve modeling of spatial processes in recent decades. Only the fundamental topics, as well as the links between spatial statistics and classical statistical methods, are addressed in this section.

2.3.1 FUNDAMENTAL TOPICS

Spatial analysis and spatial statistics are usually driven by data types and specific problems. Yet, fundamental studies in spatial science are mainly concentrated on such topics as point pattern analysis, spatial autocorrelation and proximity analysis, spatial surface smoothing and interpolation, and mosaic problems (Ripley, 1981).

2.3.1.1 POINT PATTERN ANALYSIS

In all cases, an object will occupy a finite area in the plane. Nevertheless, in some processes, objects will occupy only a very small part of the plane, and it is then convenient to think of the objects as being located at different points in the plane, or of forming a spatial point pattern. The analysis of point pattern was initiated in the 1960's (for review see Ripley, 1981; Rogers, 1974; Getis and Boots, 1978). It was the origin of spatial data analysis, and has been applied to a wide range of geographical phenomena over past three

decades (Goodchild, 1986). The main interest in analyzing point patterns is found in detecting spatial processes. Density analysis is also of concern.

There are basically two approaches to the analysis of point patterns. The first approach is the quadrat count method, which samples small zones in the plane and counts the number of objects occurring in each such zone. The distance method is the other approach, which involves two kinds of measurements: nearest neighbor (NN), which measures the distance from a selected individual to the first, second, ..., n th nearest individual; and point-to-individual (PI), measuring the distance from a selected point in the plane to the nearest, second nearest, ..., n th individual. The nearest neighbor distances convey information on the spatial pattern, whether taken from randomly selected locations (or objects) or considered for all objects in a mapped area. Also, the observations are effectively independent providing that the number of observations taken is not too large relative to the total study area. The method is therefore considered more appropriate for mapped point pattern analysis. (Cliff and Ord, 1981).

With these measurements, spatial processes can be inspected, by first proposing hypothetical processes with theoretical frequency distributions, and then testing the hypotheses using test statistics. A standard benchmark of process for point patterns is the Poisson process, which is usually used as a "no dependence" null hypothesis.

Against the null hypothesis, two patterns are examined with the quadrat method: the generalized Poisson process and the compound Poisson process. A variety of compound and generalized Poisson distribution models have been developed to test 'good fit' of processes for the quadrat method (Cliff and Ord, 1981; for procedures see Ord, 1972; and Rogers, 1974). Some examples are the negative binomial distribution, the Polya-Aeppli

distribution, the Thomas distribution, etc. Chi-square is usually used to compare the observed distributions and the theoretical distributions.

With the distance method, three major departures from Poisson processes are analyzed in point patterns: clustering, heterogeneity, and inhibition. Estimations for these models' parameters are largely unsolved problems (Ripley, 1981). A variety of test statistics for randomness test are available, such as those developed by Clark and Evans, Hopkins, Holgate, Besag and Gleaves, etc. (Cliff and Ord, 1981; Ripley, 1981). Each statistic has its drawbacks, hence the statistical test for the distance method still remains a research topic.

If sampled at random, the estimation of the overall intensity of a process may be of concern. An approach to this is to take sample quadrats and use a mean estimator. Stratified random sampling should be adopted if different parts of the study area are thought to have very different intensities; edge effects should also be considered in sampling. However, quadrat sampling may be time-consuming, and distance methods are therefore preferred both for convenience and for gaining some information about the spatial pattern. An unbiased intensity estimator for the distance methods is a pooled estimator of NN and PI measurements. Some estimators have been developed to allow for clusters (Warren and Batcheler, 1979).

2.3.1.2 SPATIAL AUTOCORRELATION

Another important data type in spatial contexts is the areal or lattice data type, which is usually analyzed by using the techniques for detecting spatial autocorrelations. The early text devoted to this subject was by Cliff and Ord (1973). Since then, a number of texts have emerged to further illustrate the topic as well as its applications. The most synthetical

one is "Spatial processes - Models and Applications" written by Cliff and Ord (1981). Ripley (1981), Getis and Boots (1978), and Upton and Fingleton (1985) have also included discussions of this subject as part of more general review of spatial statistics. Two recent publications specifically addressing this topic are found in Goodchild (1986) and Griffith (1987).

In its most general sense, spatial autocorrelation deals with the degree to which objects or activities at some place on the earth's surface are similar to other objects or activities located nearby (Goodchild, 1986). Its existence is reflected in the proposition which Tobler (1970) has referred to as the "first law of geography: everything is related to everything else, but near things are more related than distant things". Two spatial objects, for example two forests, may be similar or dissimilar in attributes, and their proximity will determine how similar they are. Spatial autocorrelation compares two sets of similarities. If features which are similar in location also tend to be similar in attributes, then the pattern as a whole is said to show positive spatial autocorrelation. Conversely, negative spatial autocorrelation exists when features which are close together in space tend to be more dissimilar in attributes than features which are further apart. Zero autocorrelation represents a case where attributes are independent of location. Thus, spatial autocorrelation can be interpreted as a descriptive index, which measures aspects of the way things are distributed in space; it can also be seen as a casual process, which measures the degree of influence exerted by something over its neighbors.

The most commonly used spatial autocorrelation measures are Moran's index and Geary's index for interval or ratio attributes, and joint-count statistic for nominal attributes. Other indices based on ranks have also been devised for ordinal attributes, but these are less used. All these indices measure the degree of autocorrelation in terms of both attribute similarities and locational similarities. A concept implied in the measures is scale, which

can be defined as spatial order adjacency. To make the concept explicit in the inspection of spatial patterns, two graphic methods are typically used: the correlogram which shows the performance of the correlation index across scale, and the variogram which represents the variance or mean squared difference between attributes at a given proximity.

An important factor in spatial autocorrelation analysis is to make statements in a larger context, by generalizing from the sample to the larger population. However, in most cases, the set of objects used as a sample in measuring spatial autocorrelation is the entire set found within the study area boundary at some given point in time, since the way the indices are defined makes it difficult to take a random sample. Thus, "Great imagination has gone into turning what appears to be a population into a sample" (Upton and Fingleton, 1985). Consequently, two popular perspectives have been employed. One is randomization^o, which treats the order of arrangement of the data over a geographical surface as important, and then assumes that the observed map is a single sample drawn from the set of all possible data permutations for the given map. The second perspective is resampling (or normality), which essentially has to do with stratified random sampling, where the areal units of a map represent a stratification, and data for each unit are treated as a random sample from a population in that areal unit (Goodchild, 1986; Griffith, 1987). Whether to apply a randomization or a resampling approach in a given situation, Goodchild (1986) commented, "..... is best answered by one's conceptualization of the sampling process, or the mechanism by which the sample was drawn from the population.... If the clock were set back and the system allowed to redevelop.... resampling would be appropriate if the hypothetical value were drawn independently and were different from the real values, while randomization would be appropriate if the effect of rerunning the clock were simply to redirect the same values to other parts of the map. It seems that resampling is more reasonable in most contexts, despite the inherent disadvantage that assumptions must be

made about the statistical distribution (normality) from which the hypothetical values are drawn."

2.3.1.3 SPATIAL SURFACES

Spatial surface analysis deals with three main aspects: inferring values at particular places between sampling points; showing the spatial variation of variables; and converting spatially continuous data into maps. A number of techniques have been developed for these purposes since the 1950's. They fall into two main groups: global methods which emphasize the global trends of spatial variation, and local approaches which are devoted to the modeling of spatial variations in details.

A typical global approach is Trend Surface Analysis (TSA), a family of polynomial methods, which was widely used in geography in the 1960's and the 1970's. The technical basis of TSA is the multiple regression approach, which ensures that whatever the nature of the data, the trend surface is always the best-fitting polynomial surface (in the least square sense) of the given order. The major shortcomings of TSA are: outliers would greatly affect the fitted surface; local details could be omitted; and, usually, it can only serve as a technique to display global trends, since otherwise, several assumptions are difficult to meet in applications, and the method is thus not reliable.

Kriging, the most commonly accepted local interpolation technique, embraces a set of methods which includes simple and ordinary kriging, co-kriging, universal kriging, and disjunctive kriging. These are based on the regionalized variable theory developed by Matheron (Matheron, 1971), and named after D. G. Krige who originated the technique (Ripley, 1981). There are several reasons why this technique is believed to be the best one for local spatial surface analysis. First, it is based on a model of stochastic spatial variation

tha. fits well with reality. Secondly, it is practical, since its assumptions seem to be readily satisfied in

practice (Oliver and Webster, 1990). The last reason is that it leads to statistically sound estimates with minimum and known variance. These properties make it appropriate for inferring values between sampling points and for mapping spatially continuous phenomena. The technique is still under development; recent research mainly emphasizes the estimation of the variogram, which is used as a tool in the analysis, and the use of this approach for mapping purposes (Chung et al., 1988).

2.3.1.4 MOSAIC

Point data, lattice or areal data, and spatially continuous data, as discussed above, are analyzed by means of point pattern analysis, spatial autocorrelation analysis, and spatial surface analysis respectively. However, there exists another type of spatial patterns that are neither point patterns, nor lattice or areal data, nor continuous surface, but rather, that are composed of a set of discrete objects with two or three phases forming a mosaic (Ripley, 1981). Patterns of vegetation provide two-dimensional examples; three dimensional examples are bone and tissue, dislocations in metals, etc. Analysis of this class of spatial patterns started in the 1960's, and was of more interest to geologists. Nevertheless, the rapid development of computer technology makes it possible for wide availabilities in both geoscience and geography.

Practical measurements for this type of data can be made either on planar slices or linear probes through a three-dimensional specimen. Stereology is the theory of recovering three-dimensional information from one- or two-dimensional samples. Image analyzers were devised to scan a section and to record information at each point of a very fine grid

(typically, 1024×1024). The information is usually grey level. It may also be an indication of whether the density of the image exceeds a threshold (Ripley, 1981). Random set theory was developed by Matheron to formalize the measurements which can be made by image analyzers and to provide some models for random sets (Ripley, 1981).

The techniques for analyzing all of the data types provide a basis for constructing models of spatial processes, and for studying the interactions or reactions of geographical variables over space. Discussions on the model construction issues are found in Cliff and Ord (1981), Ripley (1981), Upton and Fingleton (1985) etc. Space and mathematical complexity have been limiting factors, as well as the technical difficulties of model calibration. Goodchild (1986) commented that this was an area of great potential for spatial analysis.

2.3.2 CONCERNS WITH CLASSICAL STATISTICS

Hypothesis testing is a prerequisite to inductive theory construction about spatial processes. Most spatial analysis have been carried out using classical or conventional methods of statistical inference. However, two difficulties are frequently encountered in the application of the conventional methods to the spatial context, namely, autocorrelation and non-random samples.

2.3.2.1 ON AUTOCORRELATION

Most of the conventional methods were developed on the basis of assumptions including that of independent observations. However, autocorrelation is a feature of spatially located data. In the usual situation of positive spatial autocorrelation, an observation carries less information than an independent observation, since it is partly predictable from its

neighbors. Thus, increasing the number of sampling points in a study area may, in contrast to the position with independent observations, provide little or no extra information. This makes it possible to frequently violate the assumption, and serious inferential error is thus probable (Cliff and Ord, 1981).

Studies have been conducted to overcome these difficulties. Cliff and Ord (1981) pointed out that an appropriate strategy might be: (1) before carrying out any analysis, one of the measures of autocorrelation should be used to determine whether there is any spatial autocorrelation in the data. If the data are spatially independent, then proceed with the analysis using the conventional models. (2) if it is concluded that the independent assumption can not be sustained, two main alternatives are possible: to modify the appropriate technique to allow for the correlate observations, or, to remove the correlation from the data - make them spatially independent - so that the conventional models can be applied.

Efforts have been made in the modification of the classical techniques. In terms of the characteristics of spatial autocorrelation, several frequently used models have been rewritten, including t-test, product moment correlation, analysis of variance, linear regression, and spatial variate differencing (Cliff and Ord, 1981; Griffith, 1987). However, the adaptations of other conventional models to the spatial context still remain to be done.

2.3.2.2 ON RANDOM SAMPLES

Another assumption, made in inferential statistics, is that samples under study are assumed to be random. In a spatial context, this assumption is frequently violated, either because of

the limited study areas, or because of the intention to reflect spatial structures. The approaches have been discussed in Section 2.3.1.2.

2.4 EXPERT SYSTEM TECHNOLOGY

As mathematics has created a new world for study through its powerful capabilities of precise simulations for both real and abstract spaces, artificial intelligence (AI), a conceptual breakthrough technology in the field of computer science (Waterman, 1986), is attempting to originate another universe in which the complex problems of human thinking are emulated. Expert systems, a subarea of AI technology, focus on solving problems usually solved by human experts. This technique has been drawing growing attention because of the rapid development of computer technology and the increasing possibilities of solving complex problems with computers.

2.4.1 CHARACTERISTICS OF EXPERT SYSTEMS

Expert systems are computer programs that help solve real-world problems that are normally solved by domain experts. They are differentiated from conventional systems in that they have some characteristics, which conventional systems usually do not have (Waterman, 1986).

Symbolic Reasoning. Expert systems emphasize solving problems which are usually solved by human experts by choosing symbols to conceptualize problems, and applying various strategies and heuristics to accomplish inferences with these concepts.

Depth. An expert system operates effectively in a narrow but real-world domain containing sophisticated problems.

Dealing with Uncertainty. A typical problem of symbolic reasoning is uncertainty. Expert systems are capable of handling this problem by taking several factors into account, and producing multiple solutions with indications of certainty factors.

Self-explanation Capability. Like a human expert, an expert system can not only provide solutions to domain problems, but can also justify its behavior by answering users' questions, such as 'Why is additional information requested?' and 'How is the solution found?'.

Ease to Code and Update Problem-solving Knowledge. Conventional systems organize knowledge on two levels: data and program; the problem-solving approaches are embedded in the programs. Expert systems, however, organize knowledge in two different parts: domain knowledge, and common inference mechanism. The problem-solving knowledge is treated as a type of data, coded in declarative form, and stored in a knowledge base that is separated from the controlling mechanism (inference). Thus, coding and updating problem-solving knowledge requires involvement of controlling sentences or modifications of programs for conventional systems; but only the descriptions of knowledge and the changes of knowledge bases are necessary for expert systems. This, obviously, makes coding and updating much easier to carry out.

Automatic Knowledge Acquisition. Another characteristic of expert systems is the capability of knowledge auto-acquisition. Although it is a challenging task in expert system development, and the capability is not present in most of existing expert systems, it is appealing to users, and the expert system technique has great potentials to make it possible.

2.4.2 BASIC COMPONENTS OF EXPERT SYSTEMS

An expert system is normally composed of four elements: user interface, inference engine, knowledge base, and knowledge acquisition mechanism.

Human-machine Interface. The human machine interface prompts users for required information, invokes the inference engine to reason users' consultations, shows users conclusions, and upon request, gives justifications for its behavior.

Inference Engine. The inference engine is a reasoning mechanism, which uses goals or data given by users to match rule patterns. By doing so, it decides which rules should be fired, and chains all rules that are needed in reaching a particular solution. Two fundamental approaches for the inference are forward chaining, where inferences are driven by data, and backward chaining, where inferences are driven by goals.

Knowledge Base. This is regarded as the core of an expert system. Domain expertise can be classified into two types: rules and facts. Rules may be represented as production rules that are directly readable, or as first order logic that is usually used in logic programming languages. Several methods can be used to represent facts. The mostly common ones are semantic nets and frames, which both provide hierarchical structures together with a number of additional capabilities.

Knowledge Acquisition Module. This module is the mechanism to extract knowledge from reasoning procedures or from experts, store the knowledge into a knowledge base or update the knowledge base, and check consistency of the knowledge base. This element provides an advanced function, also called machine learning, for expert systems.

2.5 ELECTRONIC ATLASES AND THE ELECTRONIC ATLAS OF CANADA

Cartography is undergoing a process of rapid change as a result of the impact of the new technologies of the information era. Siekierska and Taylor (1990) stated that a new field of cartography, electronic mapping, is emerging in the process, and it deals with maps and other cartographic products for use on electronic media; one of such products is electronic atlases. The Electronic Atlas of Canada (EA), which is the basis of the thesis topic, is the outcome of the combination of this field and the National Atlas program of Canada. This section provides an overview of the concept, development, and characteristics of electronic atlases as well as various functions of the EA.

2.5.1 ELECTRONIC ATLASES

2.5.1.1 CONCEPT AND DEVELOPMENT

The concept of electronic atlases was formulated in the mid 1980's. Siekierska and Taylor (1990) described this concept as follows:

"The electronic atlas is a new form of cartographic presentation and can be defined as an atlas developed for use primarily on electronic media. The electronic atlas is one product of an Electronic Mapping System (EMS) which is a system for the development and/or the use of electronic maps."

A number of electronic atlases as well as electronic atlas systems have been developed. Their capabilities vary with the amount of user interaction and the extent of analytical functions. One of the first products was the Digital Atlas of the World produced by

Delorme Mapping Systems of Freeport, Maine in 1986 (Taylor, 1987). It contains world-wide coverages of fundamental geographical features for 16 basic scales, and can manipulate and edit the large data base in near real time, but the analytical capabilities are limited. Similar products are the Electronic Atlas of Arkansas (Smith, 1987), developed at the University of Arkansas, and the Electronic World Atlas, developed by Electromap, Inc. (Siekierska and Taylor, 1990). The best known electronic mapping product might be the Domesday Project (Openshaw and Mounsey, 1987), which is a multimedia product based on a large variety of information in forms of maps and images, but is a large static data base with limited analytical functions. Some inexpensive products are being developed using Hypercard TM, taking advantage of the technology in making a number of non-sequential associations among different elements of data and representing these in a variety of visual forms. Examples are HyperAtlas and MacAtlas developed by Apple Computers, and "La Francophonie nord-américaine à la carte" and "Mines et minéraux à la carte" developed at Laval University (Raveneau et al, 1991). Recent development of electronic atlases tends to put emphasis on analytical capabilities. Two of the present efforts are CORINE (Mounsey and Briggs, 1988), an electronic mapping system on the environment being developed for the European Electronic Community, and the PC Atlas of Sweden initiated in 1988 (Arnberg and Justusson, 1989). The Digital Chart of the World, being financed by the Defence Mapping Agency of the USA, is believed to be one of the most significant electronic mapping and GIS projects of recent times. Siekierska and Taylor (1990) commented: "The system will not only allow the user to perform a number of analytical GIS functions, but it will also form a building block for a variety of electronic atlases, and will thus be a significant advance in the development of electronic mapping and electronic mapping systems". The Electronic Atlas of Canada (Siekierska, 1983; Siekierska and Palko, 1986; Siekierska and Taylor, 1990) is another electronic atlas project which has drawn much attention from the electronic mapping community.

2.5.1.2 CHARACTERISTICS

Electronic atlases as well as electronic mapping systems have several characteristics which conventional atlases do not have. First, computer technology has greatly enhanced the capability of cartographic presentations. The electronic media permits the display of time dependent information by using animation techniques. The "scroll" function can be used to portray the continuity of areas which usually have to be subdivided into individual map sheets in conventional cartography. The "blink" function can significantly facilitate map legibility and simplify the task of visually finding small map elements that are difficult to recognize.

The second advantage of electronic atlases is that the new technology used in the product can significantly facilitate quantitative analysis of geographical information. The digital maps stored in the electronic media can be used directly for numerical analysis; the electronic mapping systems, on the other hand, may provide powerful tools to collect information from one or more maps (overlay analysis), and furthermore, to display the results of the analysis. Various analysis become possible in such an environment, such as spatial surface analysis, autocorrelation analysis, point pattern analysis, simulation modeling, etc. It is especially attractive to planning and decision making, in which spatial information is involved.

The third characteristic of electronic atlases is that most electronic atlases are supported by spatial data base management modules, which allow information updating to be done much more efficiently on electronic maps than on conventional maps. The integration of remote sensing into electronic atlases will allow this process to be even more efficient and effective.

Another advantage of electronic atlases and electronic mapping systems over conventional cartography is the great flexibility of map design, which allows the user to select themes, symbols, colors, etc., in order to make a specialized map. Furthermore, with the development of computer science and artificial intelligence, it will be possible not only for cartographers, but also for general users to design maps and visualize spatial information in an electronic atlas environment, through the use of expert system technology.

Electronic atlases as well as electronic mapping systems have similarities with GISs, in the sense that they both deal with spatial information, from data input and data storage, through to data retrieval and data output. However, they have different emphases. EMSs give primary importance to the presentation and display of spatial information, while in most GISs, primary importance is on information retrieval and the analysis of data. Siekierska and Taylor (1990) pointed out that an efficient electronic mapping system should have all of the functions of a GIS, but, in addition, it should also have the capability of creating, storing and presenting material in a variety of different visual formats on electronic media. Another major difference between a GIS and an electronic atlas is the incorporation in the electronic atlas of a set of official electronic maps, which constitute the authoritative geographical summary of the region mapped. However, visualization, which is an emerging field of computer graphics dealing with methods of displaying information, is drawing more attention in the GIS field, hence narrowing the gap between GISs and EMSs. (Siekierska and Taylor, 1990).

2.5.2 THE ELECTRONIC ATLAS OF CANADA

The concept of the electronic Atlas of Canada (EA) was initiated in the Geographical Services Division of the Canada Center for Mapping in the early 1980's. The prototype system has been developed to demonstrate the potential of applying computer technology to

National Atlas information. After extensive evaluation of the system by potential users, a decision was taken to further develop the system's capabilities and to upgrade the prototype into a full-capability production system (Siekierska and Taylor, 1990). The system's major functionality as well as basic aspects for data handling is introduced in this section.

2.5.2.1 CARTOGRAPHIC FUNCTIONS

Several special cartographic functions are provided in the system (Siekierska and Taylor, 1990). One is the map structuring function which permits the user to select information stored in the system's database, and to combine it in order to create customized maps. The second one is the map design function, which allows the user to interactively design a map by specifying the display criteria for a given subject, and selecting symbols and colours. For choropleth maps, the system calculates histograms to facilitate the interactive selection of class intervals. Another important feature of the cartographic functions is the colour manipulation capability. Colours can be assigned to particular map elements in a static way, which means that individual colours may be selected, or in a dynamic way, meaning that the user selects start and end colours of a given scale, and all intermediate colours are calculated automatically by the system. The dynamic function allows for the use of animation or blinking of selected features in map displays.

In addition, the system provides generalization and enlargement capabilities, corresponding to the varied scales of the National Atlas of Canada (from 1:2M to 1:7.5M). The generalization function provides generalized map elements, resymbolizing those which reach visibility threshold; the enlargement capability allows the user to display additional details that were visually suppressed at the smaller scales, when enlarging selected portions of a map.

2.5.2.2 QUERY AND ANALYTICAL FUNCTIONS

The EA provides two types of queries: attribute query and spatial query. Attribute query allows the user to select elements based on their ability to satisfy a specified set of criteria regarding the element's attributes. All types of attributes can be examined, including numeric, textual, and special (logical, position, time, and set). Users can graphically select, on the electronic display, a map element about which they wish more information (Siekierska and Taylor, 1990). The spatial query function permits the user to select map elements based upon their relationships with other objects. This function deals with all kinds of map elements (point, line, polygon), and responds to queries on both nearness and inclusion of either an entire object, or a part of an object (Cardinal Consulting, 1990).

The analytical functions provided in the EA include attribute synthesis functions and a range of spatial metric calculation functions (Cardinal Consulting, 1990). The attribute synthesis functions deal with statistical calculations, including means, totals, standard deviations, and frequency histograms, based on the attributes of spatial objects. The metric calculation functions allow the user to perform various calculations involving length, area, and distance of spatial objects.

On the basis of the cartographic and analytical capabilities: the system permits two types of simulations (Siekierska and Taylor, 1990). One involves historical reconstruction, for example, the territorial evolution of Canada. The other is the "prediction" of the future evolution of certain processes, for example, the depletion of oil reserves.

2.5.2.3 DATA INPUT

The EA was intended to be an analytical and map design component of the National Atlas

Information System (NAIS). Thompson (1989) stated: "NAIS is a program of the Geographical Services Division, Energy, Mines and Resources Canada, initiated in response to the need to produce multi-theme maps and geographic information of Canada, to a consistent standard, in a digital environment, to serve the needs of the federal government". It is intended to include three principal data bases: the National Toponymic Data Base, the National Cartographic Data Base, and the National Atlas Data Base (Thompson, 1989). The supporting systems for NAIS are ARC/INFO and ORACLE. In such an environment, the Electronic Atlas coverages are generated from the output of ARC/INFO using an ARC-CCSM translator, developed at the Nova Scotia College of Geographical Sciences.

2.5.2.4 DATA STORAGE

The Electronic Atlas database has been specially designed to store and handle cartographic data (Siekierska and Taylor, 1990). The data are organized using three types of files: positional files which organize the spatial positions of objects with fundamental topological properties; attribute files which store the thematic values associated with each geographical object; and cartographic specification files containing sets of symbols and relevant rules, such as generalization rules and colour manipulation rules, for data display. All the three types of files are utilized to derive output files, namely the display and analysis files. The output files are organized in raster mode, which allows for flexible overlay of multiple maps.

2.5.2.5 DATA SETS

Information found in the National Atlas of Canada forms the potential data sets for EA. The present prototype system has incorporated a representative selection of topics, ranging

from the economic geography maps, such as Oil and Gas Fields, Oil and Gas Pipelines, Oil Refineries, Coal Mines and Coal Bearing Formations, through physical geography maps such as Soil Capability for Agriculture and Wind Risk Erosion, to human geography maps, such as Indian and Inuit Settlements and Communities. The scales range from 1:7.5M to 1:25M.

2.6 SUMMARY

This chapter gives a brief introduction to the domains important to the EASSA project. Statistics is a branch of mathematics which deals with general survey data. There are two basic categories of statistical methods: descriptive and inferential. The former organizes and simplifies data so that their significance is comprehensible. The latter draws conclusions from samples to the whole populations under studies. The complexity of statistics ranges from basic descriptions, to univariate analysis, to bivariate analysis, and to multivariate analysis. Spatial statistics are the outcome of the development of spatial analysis and the application of statistics in spatial sciences. Its fundamental topics concern point pattern analysis, autocorrelation analysis, spatial surface analysis, mosaic problems, etc. Two special considerations have to be taken into account when applying conventional statistics to a spatial context: spatial autocorrelation and the formulation of spatial random samples. Expert system technology, a branch of artificial intelligence technology, is devoted to the development of systems capable of solving problems which are usually solved by domain experts. The major characteristics of the technology are its effectiveness in symbolic reasoning, declarative knowledge representations, uncertainty handling, self-explanations, and automatic knowledge acquisition. The Electronic Atlas of Canada, developed at the Geographical Services Division, Canada Center for Mapping, is an outcome of applying modern computer technologies into the National Atlas program of

Canada. The system has general GIS functions, but emphasizes the capability of interactive map design and the presentation of a set of verified national scale thematic maps.

3. EASSA'S FUNCTIONALITY

3.1 INTRODUCTION

An expert system for statistical analysis in a GIS environment may behave in different ways: it may terminate with advice or carry out whole processes; the focus may be given to spatial statistical methods, to multivariate techniques, or to basic analysis. An advisory system also varies with advisory topics and behaving modes. This chapter discusses all these issues and presents a framework of EASSA.

3.2 TASK ANALYSIS AND FUNCTIONAL SCOPE

Statistical analysis in a GIS environment shares the same principles as conventional statistics, but is differentiated from the latter in that two additional issues have to be taken into account: the spatial data characteristics and the data operation capacity of the GIS. The tasks of conducting statistical analysis in GISs can be summarized as follows.

Statistical Issues. A statistical analysis usually begins with a requirement of the analysis proposed by users, for example, spatial interpolation to transform a point map into a continuous map. It is a statistical expert's responsibility to select the appropriate method to meet the user's need. Data organization and statistical computing should then be carried out, and this should be followed by an interpretation and analysis of the results. For some analyses, the populations/samples and the analyzed results need to be visualized to facilitate user understanding.

Map-based Issues. The Electronic Atlas will have a multitude of geographic information themes with different measurement scales and varied spatial dimensions. Majority of the

maps are national coverages; for the densely populated areas, regional coverages at large scales may be provided. Some of these themes are closely related, and some are not. An expert should be able to understand such a data environment, in order to flexibly use the information, and to select and conduct analyses that make sense.

Technical Issues. A major difference between conventional statistics and spatial statistics is that the former work with a data matrix divided into rows for objects and columns for variables, but the latter also has a spatial context. This makes a significant difference in data sampling techniques. For example, a quality analysis of water bodies may use water bodies as sampling points. However, some variables involved in the analysis, e.g air quality index and temperature, may not be available in these points in the EA, since they may be modeled in points (weather stations) that do not fit with the water bodies. On the other hand, an entire population may have to be analyzed in spatial based analysis, in order to represent the spatial structure. In the other words, not only attributes, but locational information has to be considered in conducting statistical analysis within a spatial context. This particularity requires a number of technical supporting tools to capture information for use in analysis. Moreover, the GIS environment makes it possible to revisualize the analyzed results in maps.

With these tasks, EASSA may function in several ways:

(1) EASSA could complete the whole process from analysis issue determination, data source capturing, through methodology issues, practical computing, to results display and interpretations. By consulting the advisor, users could get the conclusions of the analysis directly.

maps are national coverages; for the densely populated areas, regional coverages at large scales may be provided. Some of these themes are closely related, and some are not. An expert should be able to understand such a data environment, in order to flexibly use the information, and to select and conduct analyses that make sense.

Technical Issues. A major difference between conventional statistics and spatial statistics is that the former work with a data matrix divided into rows for objects and columns for variables, but the latter also has a spatial context. This makes a significant difference in data sampling techniques. For example, a quality analysis of water bodies may use water bodies as sampling points. However, some variables involved in the analysis, e.g air quality index and temperature, may not be available in these points in the EA, since they may be modeled in points (weather stations) that do not fit with the water bodies. On the other hand, an entire population may have to be analyzed in spatial based analysis, in order to represent the spatial structure. In the other words, not only attributes, but locational information has to be considered in conducting statistical analysis within a spatial context. This particularity requires a number of technical supporting tools to capture information for use in analysis. Moreover, the GIS environment makes it possible to revisualize the analyzed results in maps.

With these tasks, EASSA may function in several ways:

(1) EASSA could complete the whole process from analysis issue determination data source capturing, through methodology issues, practical computing, to results display and interpretations. By consulting the advisor, users could get the conclusions of the analysis directly.

(2) EASSA could sample data from the Electronic Atlas, dealing with geographical analysis and statistical methodology issues in terms of the data properties, and have additional statistical and graphic software to conduct computing and display tasks.

(3) Or, EASSA could take over human experts' symbolic reasoning behavior and provide users with advice on analysis issues and methodologies, while leaving data sampling for the Electronic Atlas, and statistical computing and results display for special statistical and graphic software.

The first way is ideal, in that human intervention would be minimized. However, it would be a great risk as an initial system, because the system would have to be of large scale, different programming languages would have to be involved for different tasks, and a long period of time would have to be invested. This ideal would also lack flexibility, since the system would be restricted to the Electronic Atlas. The second way shares similar drawbacks, even though the heavy statistical computing and data display load are released. The last way, though the least synthetic, is the most promising one, since it could lead to a relatively easy implementation and a flexible expert advisory system. The cost would also be minimized because the system would take advantage of existing software.

EASSA is defined as a system that responds to users' requests for advice, on (spatial) statistics related issues and methodologies, by symbolic reasoning, while not manipulating data directly. Following the reception of advice, users would be expected to handle the whole analysis procedure, by working on the Electronic Atlas and statistical & graphic software (combined or separated). Figure 1 shows the basic environment for the analysis.

Fig. 1 EASSA's working environment

3.3 USER REQUIREMENT ANALYSIS AND ADVISORY ISSUES

Based on the defined functional scope of EASSA, the kinds of advice the system should give to users should be identified. An analysis of user requirements is thus necessary, and this should be combined with consideration of statistical analysis procedure.

The Electronic Atlas will present extensive geographic information involving various realms of physical and human geography. The users could include professionals from various fields such as geography, cartography, economics, politics, or the physical sciences; potential users also include anyone interested in national geographical information. Different users would have varied application perspectives and different statistical and geographical backgrounds, therefore their requirements for analysis as well as for advice would also differ. With regards to these factors as well as the EA environment, the potential users of EASSA may be summarized into six groups.

User Group 1: users who are only interested in the analyses based on national atlas information and general analysis issues. The requirements of this group of users for

EASSA may include the whole analysis accomplishment from analysis issue recommendation to results interpretation and graphic display. With the help of EASSA, these users would hope to quantitatively explore the geographical information presented in the Atlas, although they may not possess specific knowledge of statistics, geography and cartography.

User Group 2: users who are interested in the National Atlas information, but who would like to explore issues that are not included in EASSA's knowledge base. A helpful advisory system for this group of users should be able to accept their analysis requirements, check whether the analysis can be conducted based on Atlas information, and, if it can, provide methodological advice.

User Group 3: users who want to analyze themes beyond the information of the EA, but with analysis issues suggested by the system. This is especially the case in the expansion of EA information. For this group of users, the system is required to provide advice on analytical themes, accept map-related information from them, and then guide them through the analysis.

User Group 4: users who attempt to use analysis issues recommended by the system, but partially use the EA information and partially provide analytical themes for analysis. In this case, the system is expected to recommend analysis issues, to tell the user what information the EA has, and what themes, as well as map characteristics, should be input, and then to provide advice on the analytical methodology.

User Group 5: the difference between this and the fourth group of users is that the analysis issues are provided by users. By incorporating the information both from the EA and from the user into the analysis, a wider range of issues can be analyzed. Thus, EASSA's

responsibility would be the displays of the EA information, the acceptance of user's information including both analysis issues and map contexts, and the guidance for the analysis.

User Group 6: users who want to propose analysis issues and to conduct the analyses based on the map themes provided by themselves. What this group of users need is that the system provide a spatial statistical analysis tool so that they could analyze any statistical issues with any maps.

The requirements of all the users can be generalized into two aspects: the issues that can be analyzed and the manners in which analysis is conducted. The latter includes several issues. The first one is the selection of appropriate statistical methods for specific analytical requirements, in terms of spatial data characteristics. The second issue deals with the determination of proper spatial sampling techniques according to the cartographic models and spatial distributions of features. Following these suggestions, the user should be able to organize analytical data, and use an appropriate statistical procedure. However, most analysis procedures need human interactions. For example, a regression analysis involves a series of residual analyses, which, to large extent, depend on the user's needs and his/her feedback. Therefore, the method of handling the analysis should be further suggested. This should be followed by an interpretation of the major analytical results, since many results are described in technical terms, which may not be understandable to general users. To do this, showing users the relevant hypothesis would be helpful, for most inferential statistics conclude with acceptance or rejection of the hypothesis. Graphics as well as maps are useful tools in helping the result interpretation, hence should also be addressed in consultation. In addition, no inferential statistical method is without an assumption, and most statistical techniques have limitations. Advice on the assumptions and limitations

could help the user to better understand the analysis. Based on these analyses, EASSA should be able to give advice on the following issues:

- . Significant analysis issues based on the National Atlas of Canada (WI - What Issue)
- . Statistical methods appropriate for a specific analysis (WS - What Statistical method)
- . Query modes (sampling techniques) that should be used for the analysis (WQ - What Query)
- . Hypothesis that may be generated for the analysis (WH - What Hypothesis)
- . Assumptions which should be met (WA - What Assumptions)
- . Handling of the analysis (HH - How to Handle)
- . Interpretation of the analyzed results (HI - How to Interpret)
- . Graphic methods that can be used to visualize the data and results (WG - What Graphics)
- . Limitations of the statistical method (WL - What Limitations)

3.4 ADVISORY MODES

With the advisory issues listed above, EASSA should behave in multiple modes, in order to satisfy the various groups of users. The advisory modes are outlined as follows, with respect to the groups of users.

Issue-Atlas-dependent mode. This advisory mode will serve the first group of users in the user category (Section 3.3). It is an issue and Atlas dependent mode, since the users are not required to provide any information, except for making choices upon prompts by the system. Both the analysis issues and knowledge on analytical data are supplied by the system's built-in knowledge base.

Issue-independent-Atlas-dependent mode. This mode serves the second group of users. It supposes that users propose their own analytical issues based on the maps in the EA. The system checks the feasibility of the analysis and gives relevant advice. The feasibility judgement will be made based on the data characteristics and the statistical method requirements, and not on geographical knowledge or logic.

Issue-dependent-Atlas-independent mode. This is the mode for use by the third group of users. It is issue dependent, meaning that users will consult with the system on analytical issues. It is Atlas independent, since the users are not going to use the maps in the EA, rather, they want to carry out the analysis based on their own maps. In this mode, the users will be expected to answer questions such as data measurement scales and spatial distributions of the relevant themes.

Issue-dependent-Atlas-semi-dependent mode. This is the advisory mode for the fourth group of users. It will provide users with analysis issues, so is issue dependent. There will be two sources of thematic information: the EA and the users, therefore, it is semi-dependent on the Atlas. The difference of this mode from the preceding mode is that the analytical themes presented in the EA will also be used in analysis.

Issue-independent-Atlas-semi-dependent mode. This mode is for the fifth group of users, who propose analysis issues and use both the EA information and their own map themes for analysis. Only the way in which analysis issue is generated makes difference between this mode and the preceding mode.

Issue-Atlas-independent mode. This will be an appropriate mode for the last group of users. The analysis will not be restricted to the analytical issues known by the system, nor to the maps possessed by the EA, instead, the users will give their own analysis issues,

based on their own maps. The system will respond to them with advice on analysis methodologies as well as related issues.

With such advisory modes, EA3SA will be able to function both as a specific advisor for statistical analysis for the Electronic Atlas of Canada, and as a general statistical advisory system in a GIS environment. The first, second, fourth, fifth modes described above are subject to the former, while the third and the last modes belong to the latter.

3.5 STATISTICAL TOPICS

A wide range of conventional statistics is available, ranging from basic to advanced, and from univariate and bivariate to multivariate. Some of them can be directly adopted for analysis of spatial data, such as factor analysis and cluster analysis. Some of them, however, need to be adapted to the spatial context, such as regression analysis and correlation analysis. In addition, a number of spatial statistics and geostatistical techniques are available which are more pertinent to the EA analysis environment as well as to application requirements.

The specific statistics that will be involved in the advisor should thus be considered. This is directly related to the question of what analyses will be carried out in the EA environment. It would be ideal if a full range of methods could be included and various analytical requirements be satisfied. Nevertheless, potential analytical requirements are too extensive to do so. The strategy adopted here is to consider the most frequently used conventional methods together with some basic spatial statistical techniques. There may still be too many methods for the initial system, hence the methods are further separated into two groups: one for map-makers and the other for map interpreters. The reason for this categorization is that there are essentially two types of EA users: map-makers

including both geographers and cartographers who are working in the National Atlas program, and general atlas readers. For completeness and systematic approach, all the related knowledge is acquired, as one of the major efforts of this research. In practical implementations, however, incremental development will be a wise strategy. The whole range of statistical topics to be involved in EASSA is described as follows:

(1) *Statistical summary*. This includes the basic descriptions such as central tendency, data variability, frequency distribution and histogram. They are basic statistical techniques which are used in, or before, almost every analysis.

(2) *Equality test*. It involves the comparisons between two samples or multiple samples in means, variability, frequency distributions and proportions. Both parametric and non-parametric methods will be involved, and spatial autocorrelation will be considered. For map interpreters, these are simple but interesting analyses based on maps.

(3) *Bivariate relationship analysis*. This set of techniques deals with the most frequently concerned analysis topics: whether and how two variables are correlated. Both correlation analysis and regression analysis are included in this category, taking spatial autocorrelation into account. They are interesting to both map interpreters and map makers, but especially to the former.

(4) *Multivariate relationship analysis*. Bivariate analysis can explore the relationship between two variables. Geographical phenomena, however, usually appear in such a way that multiple variables are mutually related. Therefore, multiple regression analysis, which explores the relationships between a dependent variable and a number of independent variables, may be of especial interest to map interpreters. Also, they are useful tools for the selection of variables in map compilations. Another frequently used multiple relationship

analysis is factor analysis, which is used to find common factors in a multiple variable system, and hence can help map interpreters to better understand the nature of the analyzed phenomena. Map compilers may also benefit from this analysis in the selection of representative themes to model geographical systems.

(5) *Univariate classification*. This deals with the techniques that are especially useful in map design. The techniques involved will accommodate both operational efficiency and the optimization of class intervals.

(6) *Multivariate classification*. This set of techniques deals with the classification of objects into groups based on multiple variables. They have been widely used in image classification in the field of remote sensing. In a GIS context, particularly in the Electronic Atlas, they can be used to classify objects on the basis of map overlay. For example, cities can be classified into several levels according to the characteristics involved in an analysis. They are also powerful techniques in regionalization. Both discriminant and cluster techniques are useful in GIS-based analyses, with and without prior classification knowledge. They can help map makers create synthesized maps based on existing individual maps. They can also facilitate map interpreters by synthesizing the information on multiple maps.

(7) *Randomness test*. This includes a set of spatial statistical methods, which are basic but important in spatial analysis. They provide measures for spatial autocorrelation and point distributions, and can be used to carry out hypothesis testing about the randomness of spatial distributions. Spatial autocorrelation analysis is especially important, since on the one hand, it is a basic descriptive statistic of spatial statistics, and on the other hand, it serves as a testing tool in adapting conventional statistics into spatial contexts. Therefore, both map readers and map makers would use the techniques.

(8) *Spatial surface analysis*. This includes the techniques which might be the most interesting analysis tools in GIS contexts. They can either provide an overview of the spatial distributions of features, or infer data for any spatial location based on sampling points. Contouring maps can be created on the basis of this technique.

3.6 SELF-EXPLANATION CAPABILITIES

A domain expert should not only be able to provide advice when consulted, but also be capable of explaining WHY and HOW. The explanations for WHY usually deal with the reasons a certain input from users is needed; for example, WHY a user is asked to provide the spatial dimension of features. The explanations for HOW tell users how a decision is made; for example, HOW a certain statistical method is selected for the analysis. These are the explanation capabilities usually required of general expert systems, and will also be included in EASSA.

In addition to the explanations for WHY and HOW, EASSA will also answer questions about WHAT. A characteristic of an expert advisory system for statistical analysis is that a large number of statistical terms have to be involved in the advice. For example, Type I and Type II errors may be mentioned in the advice on analysis limitations. For some users, they are basic concepts, but for others, explanations may be required.

3.7 AUTOMATIC KNOWLEDGE ACQUISITION CAPABILITIES

Automatic knowledge acquisition is usually regarded as an optional function of expert systems. The reason for this function is obvious: an expert should not only be able to use knowledge to solve real world problems, but also be capable of learning new knowledge,

in order to expand the capabilities or to update the knowledge. However, it is an option since the necessity of a system to learn varies with the domain problems; for example, an expert system for forecasting has a few requirements to have learning capability, because climate knowledge is relatively stable, while the acquisition function may be highly desirable in an urban planning expert system, since a new plan is made based on the change in the urban pattern. Most existing expert systems for use in GISs lack the acquisition capability. In addition to domain properties, another important reason for the lack is that the application of expert system technology into GIS is still in the preliminary stage, and past emphases were primarily on domain problem solving.

Thus, whether or not EASSA should have the acquisition capability largely depends on the necessity of having the learning capability and the feasibility of implementation. A variety of knowledge will be involved in EASSA's advisory activities, including statistical knowledge, spatial sampling knowledge, Atlas context knowledge, and geographical analysis knowledge (discussed in the next chapter). Updating and expansion of knowledge makes sense for all of these issues. However, a more practical need is to automatically acquire the knowledge on Atlas context and geographical analysis issues, since they are more needed and easier to implement.

The main priority to automatically acquire the Atlas context knowledge has resulted from the fact that cartographic models of the EA are still under development in GSD, and the Atlas contents will be kept updated and expanded. The necessity of automatically acquiring analysis issues lies in the nature of the knowledge. There may be a great number of potential analysis issues, based on the information of the National Atlas of Canada. The issues vary with application requirements, and hence there is no possibility to exhaust them. An approach is therefore to use some available information for prototype

development, while setting up a mechanism to automatically acquire the relevant knowledge in the use of the system.

The implementation of the mechanism for acquiring these two types of knowledge is relatively simple, since only facts are concerned, and the syntaxes are hence easier to formulate. In fact, all Atlas-independent or Atlas-semi-dependent modes (see Section 3.3) can be directly used as the channels for the system to accept the Atlas context knowledge, while the Issue-independent modes are appropriate channels for acquiring analysis issue knowledge. Further discussion of this is provided in Chapter 6.

3.8 SUMMARY

An outline of EASSA has been given in this chapter, from the perspective of system functionality. In summary, EASSA will be an expert advisory system for statistical analysis in the EA and GIS environment. It will provide users advice on statistics related issues and methodologies, by symbolic reasoning, while not manipulating data directly. Several advisory issues will be concerned, including significant analysis issues, statistical methods, sampling techniques, statistical hypotheses, analysis assumptions, analysis procedures, result interpretations, graphic representations, and limitations. Multiple advisory modes will be provided, dealing with varied user requirements. A user can consult the system either based on the EA maps or on their own maps, and using either analysis issues recommended by the system, or proposing their own analysis topics. A set of frequently used statistical methods will be involved, ranging from basic to advanced, and from univariate to multivariate. Several spatial statistical techniques will also be included, providing powerful tools for both map readers and map makers. The system will also be able to explain WHY, HOW and WHAT, upon a user's requests. In addition, the

system will be able to automatically acquire the knowledge on Atlas context and geographical analysis issues.

4. KNOWLEDGE ACQUISITION AND DEVELOPMENT

4.1 INTRODUCTION

Knowledge acquisition is a key in knowledge engineering, hence a focus of this project. The objective is clear: to acquire and develop the knowledge that is required to derive the intended advice and to give advice in various modes (cf. Section 3.3 and Section 3.4). To do this, it was first necessary to analyze the reasoning strategies human experts use in giving advice for statistical analysis in a spatial context, and to decompose the complex problems into simple problems. The knowledge schemas, as well as their structure, were then formulated. Based on the schemas, knowledge acquisition was carried out and expertise developed. This was followed by the process of knowledge validation, in order to ensure the quality of the acquired or developed expertise. This chapter discusses the methodologies and considerations for each of these steps. The structured knowledge schemas as well as the knowledge are also presented. In order to provide a complete picture for the knowledge schemas and the relevant knowledge, the final results are presented in the respective sections, and changes during the validation process are described in the section dealing with knowledge validation.

4.2 PROBLEM DECOMPOSITION

The definition and study of domain problems as well as problem-solving strategies are always the first step in the accomplishment of knowledge engineering. Statistical analysis in the EA or GIS contexts is a rather complex domain, which involves several fields (as discussed in Section 1.2). An approach to the analysis of problems and problem-solving methods was therefore devised for this study. It decomposes problems and problem-solving strategies using the concept of problem-solving space, so as to facilitate the

clarification of human experts' reasoning strategies, and provide a basis for the formulation of knowledge schemas as well as their structure.

4.2.1 THE CONCEPT OF PROBLEM-SOLVING SPACE

Definition of problem-solving space: let A_{ijk} be a set of problem-solving elements; $i = 1, 2, 3, \dots, \alpha$, $j = 0, 1, 2, 3, \dots, \beta$, $k = 1, 2, 3, \dots, \gamma$; where i is the level of problem-solving with $i = 1$ being the top level, j is its parent's element index with $j = 0$ being used for the top problem-solving level, k indicates the k th problem-solving element at the current problem-solving level, then the Cartesian product (or cross product; Grimaldi, 1985) of all universes of A_{ijk} :

$$PSS = \prod_{k=1}^{\gamma} A_{ijk} \quad (4-1)$$

is a problem-solving space of the i th problem-solving level with parent index j . The location of a node in space is a process of problem-solving reasoning. It represents a space in which a particular problem is solved. The elements constituting such a space are the knowledge sources involved in the problem-solving.

Definition of composite problem-solving space: let PSL_i be the i th level of problem-solving, PSS_{ij} be the j th problem-solving space at i th problem-solving level, and $PSL_i(\{PSS_{ij} | j = 1, 2, \dots, \beta\})$ be a compound problem-solving space including all problem-solving spaces at the i th problem-solving level, then a composite problem-solving space is a set of compound problem-solving spaces which are involved in fulfilling a complete problem-solving task. A composite problem-solving space is expressed as follows:

$$CPSS = \{ PSL_i(\{PSS_{ij} | j = 1, 2, \dots, \beta\}) | i = 1, 2, \dots, \alpha \} \quad (4-2)$$

Expression (4-2) indicates that a complete problem-solving task can be vertically decomposed into any number of problem-solving levels, depending on the complexity of the problems. Moreover, each level can be horizontally decomposed into several problem-solving spaces, each of which deals with a particular problem.

4.2.2 PROBLEM DECOMPOSITION OF EASSA

EASSA's problem-solving task can be decomposed using the concepts of problem-solving space, and then summarized using the concept of composite problem-solving space. A complete problem-solving task for EASSA is to give advice on statistical analysis in an EA or GIS environment, according to users' requirements. Following human experts' reasoning behaviors in fulfilling this task, we can vertically decompose the task into three levels. A single level can be horizontally decomposed into several individual problem-solving spaces, based on the complexity of the problems.

4.2.2.1 TOP LEVEL PROBLEM-SOLVING SPACE

The top level problem-solving deals with the determination of advisory strategies (modes) regarding to various user requirements. The knowledge sources in relation to this level are the information input by users, and the advice given by the system. The top level problem-solving space is thus a 2-dimensional space of the Cartesian product:

$$TS = UI * SA \quad (4-3)$$

where UI denotes a set of user inputs, and SA represents a set of advice from the system. Any node in this space corresponds to an advisory mode, and the positioning of such a node implies a reasoning process on the advisory strategy.

4.2.2.2 SECOND LEVEL PROBLEM-SOLVING SPACE

A further decomposition of the problem-solving space should be considered for each element in the first level problem-solving space. Thus, there are two candidates for decomposition: the user inputs UI and the system advice SA, as expressed in expression (4-3). It is not necessary to decompose the set of user information, since an expert system should not expect the user to conduct reasoning during consultations. Users should only be responsible for proposing analysis requirements upon the system's prompts, and to carry out the analysis by following the system's advice. However, the set SA has to be further decomposed, for several reasoning processes are contained in the notion.

An intermediate decomposition of SA leads to the second level problem-solving space, as expressed below:

$$SA = WI * WS * WQ * WA * HH * HI * WG * WL \quad (4-4)$$

where WI denotes the set of analysis issues recommended by the system; WS and WQ represent the set of appropriate statistical methods and spatial sampling schemes (or advanced spatial query modes) respectively; WA refers to the set of assumptions applied to the statistical methods; HH, HI and HV denote the sets of analysis procedures, result interpretations, and graphical representation methods respectively; the last element, WL, represents the set of limitations of the analysis. Thus, a point in this Cartesian space

corresponds to a set of advice appropriate for analysis, while determining such a point involves a process of reasoning.

4.2.2.3 THIRD LEVEL PROBLEM-SOLVING SPACE

In order to successfully position a point in the space, each element in the second level problem-solving space should be further decomposed. It is thus necessary to analyze the manner in which each piece of advice is derived, and to identify the knowledge sources that are used in the derivation. This results in the generation of the third level problem-solving space. At this level, horizontal decomposition is desirable, since the elements in the second level space involve different problem-solving sub-spaces. The analysis of the relevant reasoning procedure shows that, despite diversity, the knowledge sources in relation to the third level problem-solving spaces can be summarized into the following four categories:

- . AI -- Analysis Issues that are of common interest and make sense, or that are proposed by users;
- . AC -- Atlas Contexts about spatial data characteristics either of the National Atlas or of maps prepared by users;
- . SK -- Statistical Knowledge that is suitable for spatial contexts;
- . SM -- knowledge on Sampling Modes suitable for the EA or GIS environment.

Different combinations of these knowledge sources comprise the third level problem-solving spaces, as described below.

Third level problem-solving space 1 -- {WI, WS}: To determine the issues that may be analyzed as regards users' interests in both the Atlas and non-Atlas contents, the system needs to have expertise on analysis issues (AI), Atlas contexts (AC), and statistics (SK).

With the first two types of knowledge, the system can derive analysis topics that are of common interest and make sense, and that involve maps which are found in the EA or prepared by the user. The statistical knowledge allows the system to make judgements on whether or not the analysis can be conducted within the analysis environment. Thus the Cartesian product of AI, AC and SK constitutes this problem-solving space:

$$WI = AI * AC * SK \quad (4-5)$$

This is also the problem-solving space used for selections of appropriate statistical methods. Statistical method selections depend primarily upon the analysis issues; for example, to test the equality in average income between two regions requires the use of two sample testings. Further consideration may be given to analysis purposes, for example, to test equality of average income per capita between two regions, or to test the equality of variability of income per capita between the regions. Data characteristics should then be examined, since the selection between parametric and non-parametric methods depends on the frequency distribution of the data sets and the measurement scales of the data. Other data characteristics may also need to be considered; for example, the selection of methods for two sample testing depends upon whether or not spatial autocorrelation exists (refer to Appendix 1, Table 4 and Table 5).

Third level problem-solving space 2 -- WQ: Selections of query or sampling modes for specified analyses are determined from several factors. The most important factor is the type of cartographic models used to represent the map, since sampling on a continuous map, for example, would be significantly different from that on a symbol map. Before examining data models, however, the analytical themes included in analysis issues have to be considered. In addition, it is certainly necessary to take properties of various sampling

modes into account. The third level space for selections of sampling methods can thus be formulated as follows:

$$WQ = AI * AC * SK * SM \quad (4-6)$$

Third level problem-solving space 3 -- { WH, WA, HI, WG, WL }: This is a multiple problem-solving space, in which hypotheses are generated (WH), assumptions are made (WA), interpretations of results are identified (HI), appropriate graphic representation methods are generated (WG), and limitations are indicated (WL). These advisory issues share the same problem-solving space, since they are all derived from two knowledge sources: analysis issues and statistical knowledge. The former provides analytical themes, which need to be involved in the statements of hypotheses, assumptions, interpretations, graphic methods, and limitations. The latter is the major knowledge source in dealing with these issues. Expression 4-7 formalizes the multiple problem-solving space.

$$\{ WH, WA, HI, WG, WL \} = AI * SK \quad (4-7)$$

Third level problem-solving space 4 -- HH: Advice on the handling of a specific analysis is derived in this space. The identification of analytical themes is the first step involved, so the space should include the knowledge source of analysis issues. Statistical knowledge is desirable, since statistical analysis procedures vary with statistical methods. Another knowledge source involved in the procedure is sampling schemes, for the command to be used in EA to obtain samples depends on which sampling modes are to be used. Thus, this space includes three elements:

$$HH = AI * SK * SM \quad (4-8)$$

4.2.2.4 THE COMPOSITE PROBLEM-SOLVING SPACE OF EASSA

In summary, a complete problem-solving task for EASSA can be vertically decomposed into three levels: the top level dealing with the interaction between users and the system, the second level devoted to the derivation of various types of advice, and the third level intended to generate the advice required by the second level. Horizontally, only the third level is further decomposed into four problem-solving spaces. Thus, the composite problem-solving space of EASSA can be summarized as follows:

$$\text{CPSS} = \{\text{PSL}_1(\text{TS}), \text{PSL}_2(\text{SA}), \text{PSL}_3(\{\text{WI}, \text{WS}\}, \text{WQ}, \{\text{WH}, \text{WA}, \text{WI}, \text{WG}, \text{WL}\}, \text{HH})\}$$

(4-9)

where PSL_1 refers to the top problem-solving level, with one space TS incorporating users and the system behaviors; PSL_2 refers to the second problem-solving level, with one space SA for generating the set of advice required by the top level space; PSL_3 denotes the third problem-solving level, with 4 spaces dealing with four groups of advisory issues, including the analysis issues and statistical methods, the sampling modes, the hypotheses, assumptions, result interpretations, graphic methods and limitations, and the analysis procedures.

4.3 STRUCTURED KNOWLEDGE SCHEMAS

The analysis of the problem-solving space enables the identification of the knowledge needed for EASSA, and of the manner in which it could be structured to ensure flexibility. Firstly, three types of knowledge should be used for EASSA's problem-solving: task-oriented knowledge for coordinating the users' and the system's behavior for consultation

task; advice-oriented knowledge for appropriately deducing required advice, and domain knowledge for supporting the advice deductions. Secondly, the types of knowledge can be, and actually are structured in layers, which allows for a flexible usage of knowledge and a clear reasoning procedure. There is also a type of knowledge, needed by the advisor, dealing with advisory behavior. That is meta knowledge, such as how to make explanations to users upon request, and how to make an optimal decision in case several answers are available. Figure 2 shows the primary structure of EASSA's knowledge. The domain knowledge lies at the bottom of the architecture, constituting a basis for the advisor. The advice-oriented knowledge, above the domain layer, is used to retrieve domain knowledge and formulate advice. Based on this layer is the task-oriented knowledge that incorporates users' information and advice-oriented knowledge to fulfill specific tasks. The meta knowledge, however, goes through all the three layers, and provides solutions whenever additional problems occur in any layer of reasoning. This section provides a detailed discussion of the knowledge schemas.

4.3.1 DOMAIN KNOWLEDGE LAYER

In the domain layer is found a human expert's domain-specific knowledge that constitutes the basis for deducing relevant advisory issues. This knowledge basis has four major components: geographical analysis issues, statistics, spatial sampling modes, and Atlas contexts (cf. Section 4.2.2.3)

4.3.1.1 ANALYSIS ISSUE KNOWLEDGE SCHEMA

Analysis issues will be presented by the advisor, in order to facilitate the users' accessing of geographical themes found in the EA, and to guide general users to a better understanding of geography. An analysis issue can also be input by users, so as to satisfy

statistical hypothesis, on the interpretation of the results, on the selection of suitable graphic methods, and on the limitations of the analysis.

An additional item of the knowledge schema is the factor of method priorities. This needs to be applied to the situations where several methods are intended to deal with the same statistical topic. For example, multiple linear regression, curvilinear regression, Poisson regression, etc. can all be used to analyze the relationships between multiple variables. However, they should be ranked with different priorities according to the simplicity of the methods, since simple methods are usually preferred if they are sufficient to solve the particular problems. The issue is further discussed in Section 4.3.4.2.

4.3.1.3 SPATIAL SAMPLING KNOWLEDGE SCHEMA

Williams (1984) gives a systematical description of spatial sampling techniques, which involve simple random samples, stratified random samples, cluster samples, nested samples, systematic samples, and systematic unaligned samples. The random sampling approach is thought to be the most suitable one for statistical analysis because of its probability characteristic. In the electronic map context, however, this is only suitable for statistical analysis based on themes with areal or continuous spatial distributions (2-cell or 3-cell) presented in a raster format. Unfortunately, this is not the case in the Electronic Atlas environment, since a large portion of the themes involved are of 0-cell or 1-cell distributions, and bivariate or multivariate analyses more likely involve the 0-cell or 1-cell themes. Moreover, the raster access of map themes has not yet been available in the EA system (though it will be implemented in the future). Therefore, there is a need to select spatial sampling methods that are suitable for statistical analysis in the Electronic Atlas or in a GIS environment. Since the sampling method itself is not the emphasis of this thesis, only knowledge related to the proper use of sampling methods is addressed. The

varied user requirements. The knowledge schema of analysis issues comprises of five knowledge items: the statistical issue, the functional explanation, the analytical themes, the major geographical category of the themes, and additional information.

The item known as statistical issue refers to the statistical topics discussed in Section 3.5. They present a high level outline of statistics, rather than statistical approaches themselves. Thus, users who do not understand statistics or spatial statistics very well will be able to conveniently select the analysis issues related to their interests.

The next item is functional explanations, which explains what a statistical topic deals with, and what sense it makes. Such explanations can provide elementary users with further knowledge. For advanced users, they are information sources about the statistical approaches that can be utilized within the current analysis environment.

Analytical themes constitute the third item of the knowledge schema. It refers to the map themes involved in the analysis issues. The significance of this item of knowledge is twofold. First, it facilitates users' utilization of the extensive geographical information included in the EA or the National Atlas, since few users would be familiar with the content of the Atlas. Secondly, it provides geographical knowledge involved in geographical analysis. For example, multiple variables that exist in a single geographical system may be used for factor analysis, and multiple variables which are found in differed geographical systems may be used for multiple regression analysis (Mather, 1976). The understanding of the interrelationship among variables requires the comprehension of geography.

The item geographical category is used as an index in the organization of analysis issues. Geography of Canada in the 5th edition of National Atlas is subdivided into 44 geographical realms. For simplicity, the realms are grouped into three categories: physical

and environmental geography, social and cultural geography, and economic and resource geography. More than one geographical realm may be involved in an analysis issue. For example, a bivariate correlation analysis may be carried out between infant mortality, which belongs to human geography, and mean daily temperature, which is a theme of physical geography. Therefore, there is a need to distinguish the themes involved by using the concepts of principal analytical themes and minor analytical themes. *A principal analytical theme is the one that is of major concern in the analysis; the minor analytical themes play minor roles in the analysis.* For example, infant mortality is regarded as the principal theme, and temperature takes the minor position. It is worth noticing, however, that whether a theme is principal or minor is not only determined from the the themes involved, but also from the perspective of the analysis.

The last item of the knowledge schema deals with an option, namely, additional information. It is an option since not every analysis issue needs it.

4.3.1.2 STATISTICAL KNOWLEDGE SCHEMA

Statistics are the most fundamental knowledge of the expert advisor. Of the nine advisory issues to be provided by EASSA, eight require this knowledge source. The knowledge schema of statistics consists of fifteen items: the statistical issue, the statistical method, the functional explanation, examples, the analysis principle, the data measurement scale, the hypothesis, the assumptions, the uncertainty description on the method suitability, the analysis procedure, the major results interpretation, the graphic representation methods, the limitations, and the method priority. An "instance" of such a scheme corresponds to a particular statistical method, as shown in Appendix 1.

and environmental geography, social and cultural geography, and economic and resource geography. More than one geographical realm may be involved in an analysis issue. For example, a bivariate correlation analysis may be carried out between infant mortality, which belongs to human geography, and mean daily temperature, which is a theme of physical geography. Therefore, there is a need to distinguish the themes involved by using the concepts of principal analytical themes and minor analytical themes. *A principal analytical theme is the one that is of major concern in the analysis; the minor analytical themes play minor roles in the analysis.* For example, infant mortality is regarded as the principal theme, and temperature takes the minor position. It is worth noticing, however, that whether a theme is principal or minor is not only determined from the the themes involved, but also from the perspective of the analysis.

The last item of the knowledge schema deals with an option, namely, additional information. It is an option since not every analysis issue needs it.

4.3.1.2 STATISTICAL KNOWLEDGE SCHEMA

Statistics are the most fundamental knowledge of the expert advisor. Of the nine advisory issues to be provided by EASSA, eight require this knowledge source. The knowledge schema of statistics consists of fifteen items: the statistical issue, the statistical method, the functional explanation, examples, the analysis principle, the data measurement scale, the hypothesis, the assumptions, the uncertainty description on the method suitability, the analysis procedure, the major results interpretation, the graphic representation methods, the limitations, and the method priority. An "instance" of such a scheme corresponds to a particular statistical method, as shown in Appendix 1.

The statistical issue, corresponding to the item involved in the analysis issue scheme, refers to the statistical topic for which the statistical method is used. The item of statistical method is an identification of the statistical approach being described. Following these items are the functional explanations, examples, and analysis principles. They are included in the scheme because of their usefulness in making explanations. The item of functional explanation indicates what the statistical approach is used for. For users not familiar with statistics, it may be difficult to understand the concepts by reading the explanation, therefore, examples of applications of the approach to particular analyses are provided. The basic principle of the approach is indicated so that interested users be better able to understand the method, as well as the analysis.

The most important knowledge for statistics, however, is the suitability of statistical methods to measurement scales, to spatial dimensions, to normality of frequency distributions, and to spatial dependencies, since they constitute the basic criteria in the selections of appropriate statistical methods. Data may be measured in four different scales: nominal, ordinal, interval, or ratio. The measurement scale is the most deterministic factor in the method selection, because few statistical issues can be dealt with by single statistical methods. Statistical approaches are also distinguished as parametric or non-parametric, according to their capabilities to deal with data with normal or non-normal frequency distributions. The consideration of spatial dependency is a special requirement of conducting statistical analysis in a spatial context. Conventional statistics can be adopted directly if there appears to be no significant spatial dependency (autocorrelation) in an analytical data set; otherwise, the use of modified approaches with respect to the conventional methods is suggested. In terms of commonly used descriptions of statistical methods, the suitability of measurement scales is set as an independent item in the statistical knowledge schema, while the suitabilities of frequency distributions and spatial dependencies are included in the item - assumptions.

In addition to normality and spatial dependency, the assumptions may also involve sample size, linearity, etc., depending on the particular statistical approach. The spatial distribution of features is another factor relevant to the use of statistical methods. There are normally four types of spatial distributions: 0-cell which refers to point features, 1-cell to line features, 2-cell to areal features, and 3-cell to volume features (Dent, 1985). This factor is also included in the item of assumptions, considering that it is only related to spatial statistical techniques such as point pattern analysis and spatial surface analysis.

However, assumptions are rarely completely satisfied; rather, they are usually met or violated to some degree. Therefore, certainty factors need to be introduced. The statistical knowledge schema should include the item of the suitability descriptions of statistical methods, together with certainty factors indicating the degree to which the methods are reliable (see Section 4.3.4.2).

The handling of an analysis is another key issue in conducting statistical analysis. On the one hand, the complexity of analysis procedures varies with statistical methods. For the methods which involve more interactions, advice on this issue is especially helpful (regression analysis is an example). On the other hand, the spatial context makes this item of knowledge indispensable, since analysis in such a context has to involve the testing of spatial dependency, the sampling of spatial objects, and, moreover, the mapping techniques needed to display the analyzed results. The procedure hence becomes more complex than non-spatial statistics.

The statistical knowledge schema should also include the items of hypotheses, result interpretations, appropriate graphic representation methods, and limitations of the statistical methods. All of these are needed in order to generate the advice on the establishment of a

statistical hypothesis, on the interpretation of the results, on the selection of suitable graphic methods, and on the limitations of the analysis.

An additional item of the knowledge schema is the factor of method priorities. This needs to be applied to the situations where several methods are intended to deal with the same statistical topic. For example, multiple linear regression, curvilinear regression, Poisson regression, etc. can all be used to analyze the relationships between multiple variables. However, they should be ranked with different priorities according to the simplicity of the methods, since simple methods are usually preferred if they are sufficient to solve the particular problems. The issue is further discussed in Section 4.3.4.2.

4.3.1.3 SPATIAL SAMPLING KNOWLEDGE SCHEMA

Williams (1984) gives a systematical description of spatial sampling techniques, which involve simple random samples, stratified random samples, cluster samples, nested samples, systematic samples, and systematic unaligned samples. The random sampling approach is thought to be the most suitable one for statistical analysis because of its probability characteristic. In the electronic map context, however, this is only suitable for statistical analysis based on themes with areal or continuous spatial distributions (2-cell or 3-cell) presented in a raster format. Unfortunately, this is not the case in the Electronic Atlas environment, since a large portion of the themes involved are of 0-cell or 1-cell distributions, and bivariate or multivariate analyses more likely involve the 0-cell or 1-cell themes. Moreover, the raster access of map themes has not yet been available in the EA system (though it will be implemented in the future). Therefore, there is a need to select spatial sampling methods that are suitable for statistical analysis in the Electronic Atlas or in a GIS environment. Since the sampling method itself is not the emphasis of this thesis, only knowledge related to the proper use of sampling methods is addressed. The

knowledge schema is described in this section, and more detailed discussion is given in Section 4.4.3 for knowledge formulation.

It is necessary to explain the concept of sampling base. For any univariate analysis, sampling only involves the arrangement of sampling units (or points). For bivariate and multivariate analysis, however, sampling needs to be carried out in the bivariate or multivariate spaces, but with the same sampling units. This requires the appropriate determination of sampling units. For example, for the correlation analysis between average life expectancy and mean daily temperature, weather stations may be used as sampling units, and both life expectancy and temperature can be sampled from these points (the functions of overlay and inclusion of the EA can be used to sample the life expectancy data from the points of weather stations; see Cardinal Consulting, 1990). However, if the analytical variables for a correlation analysis are life expectancy and income per capita, census areas may be the best sampling units. Therefore, it is necessary to choose a variable, and use the units of the variable as the sampling basis. The concept of sampling base is thus defined. *A sampling base refers to the sampling units on which samples are taken.* Sampling bases are normally selected from sets of units in which analytical themes are represented; they can also be user defined regions in particular situations.

The selection of the sampling base becomes the first task in determining sampling modes. The major factors considered in the selection are the spatial dimensions of analytical features. For the current example, weather stations are chosen as the sampling base, because they are point features more suitable for sampling. Thus, the sampling knowledge schema should include the spatial dimension of analytical themes modeled in maps or computer-based systems.

In addition to the spatial dimension of cartographic or data models, the spatial dimension of the analytical features in the real world is also important in determining the sampling base. For example, if two themes, both modeled with 0-cell, are analyzed for correlation, and in the real world, only one of them is 3-cell, then the theme with non-3-cell in nature should be taken as the sampling base, and the theme with 3-cell in nature should be transformed to 3-cell for sampling. However, if none of them is 3-cell in the real world, the correlation analysis may not be possible, since samples cannot be carried out, unless the points of different themes fit each other, or more sophisticated sampling techniques are devised. Therefore, the spatial dimension of themes in the real world should also be included in the sampling knowledge schema. For differentiation, the spatial dimensions of themes in the real world are called spatial distributions in this thesis (as already used in Section 4.3.1.2), while the spatial dimensions of themes in databases or on maps retain the original names.

Another factor in close relation to the selection of sampling modes is statistical methods. The influence of statistical methods on sampling method selections is twofold. First, the types of statistics, i.e. univariate, bivariate or multivariate, is related to the selection of sampling modes. For example, the determination of sampling bases for univariate analysis is usually simpler than that for multivariate analysis (further discussion is given in 4.5.3). Second, varied statistical methods may require different sampling measures with the same geographical theme. For example, cities can be analyzed either with a point pattern, or with spatial autocorrelation of certain variables of the cities. The point pattern analysis treats cities as a nominal variable and calculates, for example, the distances of the nearest neighbors as statistics. The joint-count statistic for autocorrelation analysis, however, needs joint counts between pairs of polygons derived from the cities. The sampling methods hence completely differ from each other. Therefore, statistical methods to be used in analysis should be taken into account in the sampling knowledge schema.

In some cases, principal analytical themes should be taken into account in the selection of sampling modes. For example, if two point maps are involved in an analysis, and the point sets do not fit each other, then distance should be used as a measure for one of the point maps, and the principal analytical theme needs to be identified, so as to determine which set of points is measured in distance. Measures of analytical themes may also be involved in the consideration of sampling methods. A typical example is that both density and length, which require different sampling methods, can be used in the measure of roads; which is used in an analysis, should be specified by users.

In addition, a brief description of a sampling method should be included in the knowledge schema. This is necessary, since the sampling approaches involved in the spatial context are not widely provided in current systems, and the descriptions could act as a preliminary guide for the implementation of the sampling tools and the further development of the project.

In summary, the sampling knowledge schema includes six items: the spatial dimensions of analytical themes, the spatial distributions of analytical themes, the statistical method(s) for which the sampling technique is used, the principal analytical theme of analysis, the measures of analytical themes, and the brief description of the sampling method. The fourth and fifth items only apply to some sampling cases.

4.3.1.4 ATLAS CONTEXT KNOWLEDGE SCHEMA

Whether or not an issue can be analyzed depends on the map themes and their data characteristics, either in the Electronic Atlas or in user prepared maps. The schema of Atlas context knowledge is composed of six items: the geographical theme, the spatial dimension of the theme (refers to the map or the database), the spatial distribution of the theme (refers

to the real world), the data measurement scale, the temporal indicator of the collected data, and the temporal suitability for analysis.

The geographical theme is the basic component of a map. It refers to the theme involved in the analysis issue knowledge. The spatial dimension and distribution of a theme are mainly involved in the selection of statistical methods and the determination of sampling modes, as discussed in Sections 4.3.1.2 and 4.3.1.3. They are included in the statistical knowledge schema and the sampling knowledge schema for the suitability judgement of statistical or sampling methods. For a particular analysis, however, the knowledge sources on the dimension and distribution are provided by the set of Atlas context knowledge. Likewise, data measurement scales are regarded as the most important factor in the selection of statistical methods; the knowledge source on this aspect is also provided by the Atlas context knowledge set.

The temporal indicator provides the date when the mapped data were surveyed. This factor represents a facet of data accuracy and quality, which should be considered in conducting a bivariate or multivariate statistical analysis (see Section 4.3.4.1 for details).

In addition, the temporal stabilities of map themes for specified analyses need to be contained in the knowledge schema. The reason for this is illustrated in Section 4.3.4.1. This item may contain six sub-items. The first two sub-items are the number of years during which a variable is believed to be stable for general use, and a certainty factor for this stability. The next two are the same two sub-items, but for high accuracy analysis. The last two are the same two sub-items, but for low accuracy analysis. These knowledge sources provide bases for dealing with different analysis requirements for time accuracy.

4.3.2 ADVICE-ORIENTED KNOWLEDGE LAYER

The advice-oriented knowledge layer is composed of inference knowledge, related to various types of advice to be given by EASSA. It refers to the second level problem-solving space as well as the third level problem-solving (see Section 4.2.2). The domain layer knowledge, as discussed in the previous section, provides the basis for the third level problem-solving. Thus, the layer of advice-oriented knowledge deals with the appropriate utilization of the layer of domain knowledge. The inference knowledge is advice-dependent, since the types of domain knowledge used in inferences vary with advice. However, in order to make the inferences clear, we can abstract the inferences into a simple frame, and then apply it to various types of advice-oriented reasoning.

4.3.2.1 DEDUCTIVE REASONING FRAME

The proper use of knowledge in several domains suggests a kind of inference, namely, deductive reasoning, in which conclusions are inferred using existing knowledge. Therefore, the three-phase theory of deduction, a typical deductive logic, can be employed to abstract EASSA's reasoning knowledge.

The three-phase deductive logic consists of three steps of reasoning: setting the major premise, finding the minor premise, and drawing the conclusion. Figure 3 presents the frame of deductive logic. A major premise is usually a rule, which consists of a set of premises and a set of conclusions. A minor premise is usually composed of a set of facts that correspond to the premises included in the major premise. The conclusion of the three-phase deduction is drawn by testing or matching whether the facts presented in the minor premise meet the premises appearing in the rule of the major premise. If the matching is successful, the conclusion contained in the major premise applies to the subject referred to

by the minor premise, and the deduction comes to a positive conclusion; otherwise, a negative conclusion is reached with the deduction.

Fig. 3 Frame of three-phase deduction

4.3.2.2 ADVICE-ORIENTED INFERENCE SCHEMES

Using deductive logic, we can summarize the inferences involved in the derivation of various types of advice in the form of three components: the major premise, the minor premise, and the conclusion. Filling this frame will lead to the formulation of inference knowledge schemas.

The Knowledge schema for Deriving WI (what issue can be analyzed)

The first step in the derivation of analysis issues is to generate a hypothesis of such an issue. This refers either to a retrieval of an analysis issue from the analysis issue set of EASSA, or to an analysis issue input by users. Deductive reasoning is then carried out to infer whether or not the issue can be analyzed. The major premise of the deductive scheme

involves analysis issues and statistical knowledge. The set of premises contained in the major premise are as follows: themes involved in the hypothetical analysis issue and measurement scales suitable for the statistical topic involved in the issue. The minor premise of the scheme refers to the map contexts corresponding to the items of the major premise. All the knowledge involved is either retrieved from the domain knowledge layer (implying the use of Atlas maps), or input by users (implying that maps are provided by users). A conclusion on whether the issue can be analyzed is then made by matching the major premise and the minor premise. An example is presented below.

Example: what issue can be analyzed?

Hypothetical analysis issue: a bivariate relationship analysis may be conducted between two themes, namely the elected party and the employment growth rate (Analysis issue #1).

Major premise: if the themes are the elected party and the employment growth rate, and measurement scales are nominal, ordinal, interval or ratio, then the hypothetical issue can be analyzed.

Minor premise: the themes of elected party and of employment growth rate are in the EA, and the measurement scales are nominal for the elected party and ratio for the employment growth rate.

Conclusion: the hypothetical issue can be analyzed.

The Knowledge schema for Deriving WS (the appropriate statistical method)

The selection of statistical methods is based on a specific analysis issue, on the map

contexts involved, and on the statistical knowledge. Before the deduction is carried out, a hypothetical statistical method needs to be generated. The major premise consists of three items: the analysis purpose, the data measurement scales, and the assumption satisfactions. The analysis purpose is a further identification of the statistical issue involved; for example, whether to look at how two variables are correlated, or if they are correlated, when looking at the bivariate relationship analysis. The item of assumption satisfactions indicates the properties that the data set should possess. The conclusion of the major premise is the hypothetical statistical method. The minor premise of the scheme includes the analysis purpose input by the users, the measurement scales of the analytical themes involved, and the testing results on whether and how the data sets meet the assumptions of the major premise. A positive conclusion is drawn if the matching between the major premise and the minor premise succeeds; otherwise, the hypothetical statistical method cannot be applied. It should be noted that certainty factors and priority factors may need to be considered with some statistical techniques (refer to Sections 4.3.1.3 and 4.4.1).

Example: which statistical method is appropriate for Analysis issue #1?

Hypothetical statistical method: product moment correlation analysis.

Major premise: if the analysis purpose is to determine whether two variables are correlated, and the measurement scales of the two variables are both interval or ratio, and the assumptions of product correlation analysis are met, then product moment correlation analysis is an appropriate statistical method.

Minor premise: the analysis purpose is to determine whether two variables are correlated, and the measurement scales of the two variables involved are nominal and ordinal (ratio/interval classified) respectively, and assumptions can be checked if necessary.

Conclusion: the hypothetical method is not suitable for the analysis (because the measurement scales of the analytical themes are nominal and ordinal, which do not meet the requirement of the hypothetical statistical method, i.e. interval or ratio scale, included in the major premise).

Knowledge schema for Deriving WQ (appropriate sampling modes)

The major premise of the deductive frame includes five premises, namely the spatial dimensions of analytical themes, the spatial distributions of analytical themes, the statistical method used is the analysis, the principal theme of the analysis, and the measures of the analytical themes (some of the premises may be empty in some cases). The conclusion contained in the major premise indicates the sampling base and the sampling method that are suitable when these premises are satisfied. The minor premise provides the facts on the analytical themes as well as the statistical method involved, corresponding to the items of the major premise. If the minor premise meets the premises contained in the major premise, then the conclusion of the major premise applies to the conclusion of the deduction; otherwise the conclusive sampling base and the sampling method in the major premise are not appropriate for the analysis, and other solutions are explored.

Example: which sampling modes are appropriate for Analysis issue #1?

Major premise: if the statistical analysis involve two variables, and spatial dimension of the two variables are 0-cell and 2-cell respectively, and the measures of the themes do not include density, then the sampling base is the 0-cell theme, the sampling scheme for the 0-cell theme is Mode (9), and the sampling scheme for the 2-cell theme is Mode (13) (cf. Appendix 2, Rule-14).

Minor premise: the statistical analysis is bivariate, the analytical themes are elected party and employment growth rate, the spatial dimensions of the two variables are 2-cell and 0-cell respectively, and no density is included in the analytical measures.

Conclusion: the sampling base is the 0-cell theme; the sampling scheme for the elected party is Mode (13), and the sampling scheme for the employment growth rate is Mode (9).

Knowledge schemas for Deriving WH, WA, HL, WG, and WL (what hypothesis, assumptions, result interpretation, graphic methods, limitations)

Deductions for these types of advice are simple. For each of the advisory issues, the premises of the major premise are found both a set of analytical themes and a statistical method; the conclusion of the major premise is a statement on the corresponding advisory issue. The minor premise presents the themes being analyzed and the statistical method being used. In cases where the matching succeeds, the conclusion of the major premise becomes the conclusion of the deduction.

Example: what null hypothesis is generated for Analysis issue #1?

Major premise: if analytical themes are the elected party and the employment growth rate, and the statistical method is the Chi-square test, then the hypothetical null hypothesis is: the variation of the elected party is independent of the employment growth rate.

Minor premise: the analytical themes are the elected party and the employment growth rate, and the statistical analysis is Chi-square test.

$$\frac{d}{dt} \left(\frac{1}{2} m v^2 \right) = \frac{d}{dt} \left(\frac{1}{2} m \dot{x}^2 + \frac{1}{2} m \dot{y}^2 + \frac{1}{2} m \dot{z}^2 \right)$$
[illegible]

Conclusion: the null hypothesis for the analysis is: the variation of the elected party is independent of the employment growth rate.

Knowledge schema for Deriving HH (the handling of an analysis)

The major premise of the deductive frame involves three aspects: the analytical theme(s), the statistical method, and the sampling mode(s). The conclusion of the major premise is an analysis procedure which deals with these issues. The minor premise presents the theme(s) being analyzed, and the statistical method and sampling mode(s) being utilized. The conclusion of the major premise applies to the conclusion of the deduction if the inference comes to a positive conclusion.

Example: How should Analysis issue #1 be handled?

Major premise: if the statistical method is Chi-square test, and analytical themes are THM A and THM B, and the sampling modes are Mode A and Mode B, then the analysis procedure is as follows: 1) use Mode A to draw a sample from THM A data set, and use Mode B to draw a sample from the THM B data set; 2) use 'Contingency table' to display the cell counts as well as the marginal totals; 3) set the null hypothesis and a significant level α (0.01 - 0.1); 4) use 'Chi-square test' to execute the analysis.

Minor premise: the statistical method is Chi-square test, and the analytical themes are elected party and employment growth rate, and the sampling modes are Mode (13) and Mode (9).

Conclusion: the analysis procedure is as follows: 1) use Mode (13) to draw a sample from the elected party data set, and use Mode (9) to draw a sample from the employment growth

rate data set; 2) use 'Contingency table' to display the cell counts as well as the marginal totals; 3) set the null hypothesis and a significant level α (0.01 - 0.1); 4) use 'Chi-square test' to execute the analysis.

4.3.3 TASK-ORIENTED KNOWLEDGE LAYER

This layer of knowledge includes the strategies for incorporating users' information and the advisor's behavior in fulfilling specific tasks. Its working space refers to the first level problem-solving space as shown by Expression 4-3 in Section 4.2.2. Obviously, the knowledge schema should consist of two aspects of knowledge: the user's information, and the system's advice. The user information includes selection of analysis issues, analysis issues proposed by users, analytical themes as well as map contexts input by users, and statistical functions (or analysis purpose) chosen by users. The system's advice includes the nine types of advisory issues (Section 3.3), ranging from analysis issues, statistical methods, sampling schemes, etc., to result interpretations and method limitations. According to the preceding analysis on user requirements as well as advisory modes (Section 3.3 and 3.4), there should be six types of combinations of the user information and the system's advice, corresponding to the Issue-Atlas-dependent mode, the Issue-independent-Atlas-dependent mode, the Issue-dependent-Atlas-independent mode, the Issue-dependent-Atlas-semi-dependent mode, the Issue-independent-Atlas-semi-dependent mode, and the Issue-Atlas-independent mode.

4.3.4 META KNOWLEDGE

Meta knowledge deals primarily with the use of the expertise of an expert system. There are basically four types of meta knowledge needed for performing EASSA's functionality: knowledge for determining whether an analysis can be conducted with data sources

collected at different dates, knowledge to select statistical methods with considerations to both priority and uncertainty factors, and knowledge to provide appropriate explanations upon users' requests.

4.3.4.1 KNOWLEDGE FOR DEALING WITH TEMPORAL DIFFERENCES

Maps used for analysis are mostly thematic maps based on surveyed data. Data of different themes were unlikely to be captured at the same time; for example, information in the National Atlas is actually composed of a large range of data source date. However, all bivariate and multivariate analyses involve multiple variables, and it is believed that an analysis can be performed if the multiple themes involved were captured within tolerable time differences. Thus a question arises: what are the tolerable time differences with which the related themes can be analyzed in a single issue? This further raises such an issue: what are the relative stabilities of different spatial phenomena (as mentioned in Section 4.3.2.4). For example, climate factors are statistically not very sensitive to yearly changes, while employment rates may drop dramatically in a few months. Based on the temporal suitability knowledge of various themes, it is possible to formulate the tolerable time differences for an analysis.

Thus, EASSA should possess the capability to judge whether an analysis can lead to a meaningful solution from the viewpoint of temporal characteristics. To do this, a special mechanism, namely, a type of meta knowledge, needs to be devised, in order to combine stability values as well as uncertainty factors of multiple themes involved in an analysis, thereby to infer the conclusion on whether or not the time differences permit a meaningful analysis. This conclusion should have multiple solutions, each associated with a certainty factor.

4.3.4.2 KNOWLEDGE FOR MULTIPLE STATISTICAL METHOD SELECTIONS

Several statistical methods may be suitable for a particular analysis. Some methods can provide higher reliability, while others may be more accurate from the point of view of statistical methods. For example, the product moment correlation analysis is believed to be more accurate in the sense of statistical principles than the Spearman's correlation analysis, which, in turn, is more accurate than the Chi-square test. A higher accuracy is usually preferred in analysis, and a priority factor should hence be assigned to each of the methods (as mentioned in Section 4.3.2.2). However, the methods with higher accuracy may have more restrictions for data characteristics; for example, the product moment correlation requires a large sample or an at least roughly normal frequency distribution. As discussed in Section 4.3.2.2, this imposes uncertainty on the utilization of the method, since the satisfaction of assumptions is usually fuzzy (see the item Uncertainty Descriptions of Table 4 in Appendix 1). Thus, there is a need to develop a mechanism for the system to incorporate both priority factors and uncertainty factors into the selection of statistical methods. This mechanism is what this type of meta knowledge refers to.

4.3.4.3 KNOWLEDGE FOR MAKING EXPLANATIONS

The capability of self-explanation is an important function of an expert system. It is also a typical issue handled by meta knowledge. Two types of explanations are involved with EASSA: those for dynamic solutions and those for static concepts. The former deals with the explanations on WHY the user is prompted to input certain information and on HOW advice is derived, while the latter provides explanations for WHAT a particular statistical concept means (refer to Section 3.3).

For making explanations on WHY and HOW, the system needs to have mechanisms to trace the reasoning chains. Upon a user's request, the system thus provides a set of reasoning rules, in relation to the solutions, by retrieving the chains. To explain WHAT, on the other hand, the system should possess a set of explanations corresponding to the statistical concepts. The system should be able to quickly retrieve the related knowledge, and explain the concept to users.

Knowledge on this issue can be developed at two different levels. One is to involve only the general analysis procedures of various statistical methods, and the other refers to the statistical and graphic software used for the analysis.

4.4 KNOWLEDGE EXTRACTION AND DEVELOPMENT

The structured knowledge schemas described in the preceding section provide a framework for knowledge acquisition and expertise development. Considering the enormous amount of knowledge needed in the development of EASSA, the thesis work was focused on the extraction of statistical expertise, which is believed to be the most fundamental and complex knowledge involved. A preliminary effort was also made to extract analysis issues that are of common interest, and have potential significance from the perspective of geography. Also, a framework of sampling method selections was developed, based on the characteristics of spatial features and the requirements of statistical analysis. Atlas context knowledge (as discussed in 3.7) remains to be automatically acquired by the system after data models have been developed by the staff in the Geographical Services Division. The knowledge at the advice-oriented level and the task-oriented level deals only with inferential and strategic issues, which were discussed in the preceding section. Therefore, only the extraction and development of three types of knowledge are addressed in this section,

namely, statistical knowledge, analysis issue knowledge, and sampling knowledge; emphasis is placed on statistical knowledge.

4.4.1 STATISTICAL KNOWLEDGE EXTRACTION

The statistical issues to be involved in EASSA have been discussed in Section 3.5. They include statistical summary, equality test, bivariate relationship analysis, multivariate relationship analysis, classification, randomness test, and spatial surface analysis. However, the statistical approaches that should be selected in order to accomplish these analyses must be identified. For each of the approaches, knowledge should be formulated in terms of the statistical knowledge schemas. Thus, work of statistical knowledge extraction involves two aspects: the selection of statistical methods, and the knowledge formulation of the statistical methods involved.

4.4.1.1 SELECTIONS OF STATISTICAL METHODS

Statistical Summary. A set of data can be statistically summarized in several ways. Central tendency, one of the most generally used descriptions, has three different measures: mean, median, and mode. They are all collected into the knowledge base, since they are especially suitable for different data measurement scales (ratio or interval, ordinal, and nominal). Data variability and skewness are also included in the knowledge set for statistical summary. For the description of data variability, variation ratios, percentiles, standard deviations, and coefficients of variation are included, with respect to the four types of data measurement scales. The last set of data summary techniques involved in the system includes the histogram, the frequency table, and the contingency table. They provide good tools to describe data distributions. The related knowledge is shown in Table

1, 2, and 3 in Appendix 1 (all tables referred to in this section are included in this appendix).

Equality Tests. Equality tests are devoted to the statistical comparison of two or several samples/populations. Two sample tests form the basic type of equality tests. Several measures can be compared between two samples: means, variabilities, frequency distributions, and proportions. The T-test and the Whitney U-test are used for mean comparisons; the first one is suitable for normal data sets, and the second one for non-normal data sets (Tables 4 and 5). The modified t-test can be used to deal with spatial autocorrelation (Table 4). The F-test is suitable for variability comparisons, and its modified version can cope with spatial dependency (Table 12). The Chi-square test and the Lilliefors' test are referred to as the goodness-of-fit tests; the former is suitable for general frequency distribution comparisons, while the latter is particularly used for the normality test (Tables 8 and 11). The Chi-square test is also suitable for comparisons of categorical and proportional populations. In addition to the two sample tests, there are a series of methods dealing with multiple sample tests. Mean comparisons can be carried out among multiple groups or samples, by using the single factor analysis of variance (ANOVA) for parametric (Table 6) or the Kruskal-Wallis test for non-parametric (Table 7). The Friedman's test is another non-parametric equivalent ANOVA (Table 10), but dealing with randomized block design. It is particularly useful in map analysis (see examples in Table 10). Moreover, multi-factor ANOVA can be used to compare means of multiple samples/populations under multiple factor categories (Table 9), or to assess the effects of multiple factors on a response variable (see the example in Table 9). These equality testing methods provide a rich tool box for sample comparisons, and hence have all been included in EASSA's statistical knowledge set. However, it is worth indicating that only the modified methods for t-test, F-test, and ANOVA have been found to deal with spatial

dependency. As well, no non-parametric equivalent methods were found for variability comparisons and the multi-ANOVA.

Bivariate Relationship Analysis. There are basically three kinds of bivariate relationship analyses in statistics: independence test, correlation analysis, and regression analysis. The independence test analyzes the independence between two or more categories for categorical or nominal data by using the Chi-square test (Table 13). Three methods are most frequently used in correlation analysis, namely, Pearson's product-moment correlation, Spearman's rank correlation, and Kendal's Tau (Tables 14, 15 and 16). The use of the first method is restricted to the interval or ratio data with linear relationships; the last two are both used for ordinal data, but are suitable for large samples and small samples respectively. Of these methods, the modified approach of the Pearson's product moment correlation and bivariate regression deals with the spatial dependency. The bivariate regression analysis actually includes several methods, such as linear regression, curvilinear regression, and Poisson regression (Table 17). The first two are used for linear relationships and curvilinear relationships respectively; the third is suitable for binary variables. An important problem encountered in applying regression methods to spatial contexts is that the assumptions made in regression analysis are likely to be violated in spatial contexts. Therefore, the emphasis in the knowledge extraction of the regression analysis was on the approaches to various assumption violations, such as spatial autocorrelation and heterogeneous variance.

Multivariate Relationship Analysis. There are two types of multivariate techniques for analyzing relationships. One is multivariate regression analysis, referred to as dependency analysis. This is believed to be the most useful technique in statistical analysis (Myers, 1990). It is used to formulate a model that statistically represents the dependency relationships between a dependent variable and a set of independent variables. Considering

methodological complexity, only the multiple linear regression model is described in the knowledge set (Table 18). The emphasis is also put on various diagnoses and corresponding approaches to the violations of model assumptions. Several topics are involved, such as autocorrelation, heterogeneous variances, influential points, multicollinearity, and non-normality. The other type of multivariate technique is factor analysis, which is referred to as the interdependency analysis. It can be used to identify a set of descriptive concepts which summarize the interrelationships of interactive multiple variables existing in a single system. Since it is a relatively advanced technique, the analysis procedures and results interpretations are emphasized in knowledge formulation (Table 19). Principal Component Analysis is also regarded as a technique for interdependency analysis (Mather, 1976). It attempts to determine a set of orthogonal variables that represent the original variable dimensions. However, it is included as a step and an instance in factor analysis in the EASSA knowledge set (as treated in some software packages, e.g. SPSS). This is based on the nature of the method and the objective of the analysis. The spatial dependency problem is neglected in the factor analysis, since no successful efforts were found to deal with this issue.

Univariate Classification. This is a statistical method primarily used in map design. It aims to classify data into several groups, so that values within a group are more similar than those in different groups. There are a large number of classification schemes available. Currie (1989) summarizes them in four categories. The first category, which refers to the regular series, includes the equal step, the equal frequency, the equal membership, the mean-and-standard-deviation method, and the nested-means. The second category includes systematically equal series, such as the arithmetic method, the geometric method, and the reciprocal method. The third category refers to irregular series, including such graphic techniques as the frequency histogram and the cumulative frequency curve, and such iterative techniques as Jenks' procedure, Monmonier's procedure, and disjoint cluster

analysis. Other techniques in the last category are principal analysis and linkage analysis. Of these methods, only five have been selected for the usage in EASSA, namely, the Jenks' and the Monmonier's iterative procedures (Table 21), the equal step and the equal frequency methods, and the histogram method (Table 20). The two iterative procedures are taken as the major approaches for choropleth mapping, since they have been evaluated as the best methods that provide consistently good classes regardless of data distributions (Currie, 1989). However, iterative techniques usually take more time to accomplish a classification task than other methods. Essentially, for data sets with uniform or normal distribution, the equal step and equal frequency approaches can achieve good results, but take a much shorter time. Therefore, they are taken as alternatives, when applicable, to obtain equal step classes and equal frequency classes respectively. The histogram is a visual way that simply displays the frequency distribution of a data set; the classes are selected based on the natural breaks in the histogram. The graphic displaying method is actually also an assistant method for other classification approaches.

Multivariate Classification. Similarly to the univariate classification, the multivariate classification also deals with classifying objects into several groups, but these involve multiple variables with each object. There are two different types of techniques for this purpose: cluster analysis and discriminant analysis. Cluster analysis groups objects without prior knowledge of the groups. The objects can be either cases or variables. The most frequently used method is the hierarchical clustering method. With respect to different clustering formulation criteria, there are mainly two types of clustering approaches, namely the divisive approach and the agglomerative approach. Unlike cluster analysis, discriminant analysis classifies multivariate objects using prior knowledge. There are also several different techniques used for this purpose. Only the agglomerative clustering method (Table 22) and the linear discriminant function (Table 23) are involved in EASSA in the present stage.

Randomness Tests. Randomness tests are one of the fundamental topics of spatial statistics. As discussed in Section 2.3, progressive research has been undertaken on spatial processes. However, only autocorrelation analysis and the nearest neighbor distance method are included in the knowledge set. This is due to the strategy stated in Section 1.5, namely, starting with conventional methods, then moving to more spatial techniques. For autocorrelation analysis, joint-count statistic (Table 24) and Moran's I statistic (Table 25) are used as indices. The former deals with nominal variables, and the latter with interval or ratio data. Moran's statistic is selected because it is believed to be more advantageous than Geary's statistic, which is another commonly used autocorrelation measure, though they are both satisfactory (Goodchild, 1986). The nearest neighbor method (Table 26) is chosen as the method to test whether or not a point pattern conforms to a uniform or Poisson process. This may have little meaning, but it provides a tool to deal with random point samples required for other analyses. The distance measure rather than the quadrat measure is chosen for this analysis, since it is believed to perform better. It conveys information on the spatial pattern, whether taken from randomly selected locations (or objects) or for all objects in a mapped area; this does not apply to the quadrat method (Cliff and Ord, 1981).

Spatial Surface Analysis. This is another important topic of spatial statistics. It deals with the interpolation of spatial surface using sampling points. Several techniques have been devised for this purpose, such as trend surface analysis, moving average method, tessellation and triangulation techniques, and Kriging technique (stochastic process prediction). Only Kriging and trend surface analysis are chosen for use in EASSA, since the former is regarded as the best method for local interpolation, and the latter is appropriate for global interpolation (refer to Section 2.3.1.3).

With the statistical methods described in this section, various thematic maps can be analyzed from many different perspectives. They also provide a possibility for both map makers and map interpreters to make adequate use of statistics, especially conventional statistics, to acquire needed information.

4.4.1.2 KNOWLEDGE FORMULATION

For all of the methods discussed above, knowledge has been formulated by following the knowledge schemas discussed in the preceding section. To do so, a fairly large number of texts were reviewed. These books range from the basic to the advanced, from univariate or bivariate to multivariate, from conceptual to practical, and from conventional to spatial. The statistical texts in the bibliography of the thesis presents an overview, in which Williams (1984) and Jay and Roxy (1986) were used as the major reference for univariate and bivariate conventional statistics, Myers (1990) and Mather (1976) as major references for multivariate techniques, and Norusis (1985) and Velleman (1989) for practical references. Major references for spatial statistics were Cliff and Ord (1981), Ripley (1981), Goodchild (1986), and Griffith (1987). Other texts were used as minor references. For spatial statistics, some recent papers were also referred to.

4.4.2 DEVELOPMENT OF ANALYSIS ISSUES

The analysis issues were developed, rather than extracted, since there is no specific literature that systematically addresses this topic. Considering the extensiveness of the geographical themes presented in the National Atlas, and the endless analysis's potential of the themes, a general principle was determined before the task was carried out. As already mentioned in the preceding sections, the principle was to focus on the analyses that are of

common interest, and make sense from the geographical perspective. Subject to this principle, work was carried out by using the following methodology.

Firstly, basics on the geography of Canada were reviewed, in order to understand major geographical features in physical, environmental, social, cultural, economic, and resource geography. The fundamental distribution characteristics of geographical phenomena were also examined, and the relationships between the phenomena were inspected. For example, the basic topographical distribution in Canada, and the phenomena closely related to this distribution were studied. This step provided a basis for further analysis of geographical distributions represented in the National Atlas of Canada.

Secondly, an overview of statistical methods, their functions, and the required data characteristics was conducted, in order to formulate an outline on the issues that can be analyzed, and on the types of data required. This was actually carried out together with the preliminary review of statistical knowledge.

Thirdly, several maps out of different geographical categories were selected for analysis, in order to determine the major constraints imposed on the map-based analysis, and the statistics that could be conducted for maps in the National Atlas. The major constraint factors in determining the applicabilities of statistical methods were found to be the data measurement scales presented in maps, the map representation methods used to present the geographical themes, and the spatial distributions of the geographical themes in the real world. Essentially, this partly provided a basis for the formulation of the statistical knowledge schema and the sampling knowledge schema. Taking into account both the constraint factors and the requirements of data characteristics of various statistical methods, the potential statistical topics were then formulated. It was found that almost every conventional statistical method can be adopted to analyze the geographical information

presented in the National Atlas. This was also a knowledge source for the determination of statistical issues discussed in Section 3.5.

Fourthly, about fifty existing maps of the National Atlas were analyzed, in order to generate analysis issues that fit both the requirements of statistical methods, and the characteristics of the maps. This was carried out on a map by map basis, and by following both the statistical topic scheme, formulated in the preceding step, and the knowledge schema discussed in Section 4.3.1.1. The analysis showed that almost every map of the National Atlas can be analyzed using some statistical methods.

Finally, analytical information was extended to the maps under production, and to the potential themes to be included in the National Atlas. Thus, it is possible to formulate more issues that are more of common concern; for instance, analyses of the state of the environment, and analyses more closely related to human issues, such as life expectancy and housing area per resident.

The formulated analysis issue knowledge (Zhang, 1990), because of its preliminary stage, is not included in the thesis.

4.4.3 SAMPLING KNOWLEDGE FORMULATION

The basic sampling knowledge schema has been formulated in Section 4.3.1.3. With respect to the factors included in the scheme, the preliminary knowledge for the selection of sampling approaches has been summarized in rules, as presented in Appendix 2. Although the sampling technique itself is not of major concern, it has to be addressed in knowledge formulation, since, essentially, the formulation of sampling knowledge deals with the categorization of sampling approaches in terms of the factors included in the knowledge

schema. The methodology for the categorization, as well as the detailed analysis for each category, are given in this section.

4.4.3.1 METHODOLOGY FOR FORMULATION

For convenience of analysis, the sampling cases were first categorized into three groups: univariate analysis, bivariate analysis, and multivariate analysis. Within each group, the cases were first analyzed in terms of the spatial dimensions of the analytical features. Then, consideration was given to the spatial distribution of the features, and to the statistical methods for which the sampling modes are intended to serve. An important issue in relation to the determination of sampling methods was the selection of sampling bases appropriate for particular cases.

In order to simplify the situations, two assumptions are made for the sampling knowledge. First, 1-cell themes (linear features) are all treated as nominal or ordinal. Secondly, volume variables (3-cell distributed variables) are represented as points (0-cell model) in the system database. These assumptions are reasonable since, on the one hand, we are not going to involve the complex network analysis in the initial advisory system, and on the other hand, the point cartographic model seems more suitable for representing 3-cell themes in a primarily vector based system.

4.4.3.2 UNIVARIATE ANALYSIS

For univariate analysis, there are three types of situations: the spatial dimension of the analytical theme is 0-cell, 1-cell, and 2-cell. In the 0-cell case, measures of samples vary with statistical methods. For example, statistical summary requires that samples be captured by regions; the nearest neighbor distance method needs distance measures

between points; and autocorrelation may require spatial proximity. In the 1-cell case, only length or density in specified regions are required for statistical summary. Two different sampling approaches are involved in the 2-cell case, namely, values in specified regions, and spatial proximity. The former is used for statistical summary and equality test; the latter for autocorrelation analysis. Thus, in addition to the spatial dimension of analytical themes, statistical methods are the major factor in the selection of sampling approaches. Sampling bases refer to the analytical themes themselves, except that values, densities, or lengths within specified regions are required, where the regions are regarded as sampling bases.

4.4.3.3 BIVARIATE ANALYSIS

For bivariate analysis, the first issue to be dealt with is the selection of sampling bases (as indicated in Section 4.3.1.3). According to the spatial dimensions of analytical themes, there could be six types of cases: 0-0, 0-1, 0-2, 1-1, 1-2, and 2-2 where the number refers to the spatial dimension of an analytical theme.

For the 0-0 case, there could be three different situations: one and only one of the themes is 3-cell in the real world, (i.e. 3-cell spatial distribution); two themes both have the 3-cell spatial distribution; and none of them is of the 3-cell distribution. For the first situation, the theme with 3-cell distribution should be transferred into 3-cell, and the other theme used as a sampling base. For the second situation, if the two sets of points fit each other, no preprocessing is needed, and the points are used as a sampling base directly; otherwise, one of them should be transformed into 3-cell in the system, and the other regarded as the sampling base. The third situation only allows the sampling of the distances between the two sets of points, where the specification of the principal theme is of vital importance

(refer to 4.3.1.1). In all cases, once the sampling base is determined, the two themes are sampled from the same base.

In the case of 0-1, only two types of sampling are involved: distance from points to lines, and the number of lines connected to points. The choice depends on the analytical themes stated in the analysis issues. The 0-cell theme is assumed to be a non-3-cell theme in nature. The sampling base in both cases is referred to the 0-cell theme.

For the 0-2 case, there are two different situations: the 0-cell theme is chosen as a sampling base; or the 2-cell theme is the sampling base. The former applies to the situation where the 0-cell theme is 3-cell distribution in nature, or where theme values rather than densities are used for analysis. The latter is applied when densities are required in analysis issues.

The 1-1 case refers only to a simple sampling approach: to sample densities or lengths for each of the 1-cell themes within specified regions. The sampling base to be used is essentially a 2-cell space, i.e. the specified region. The 1-1 case is similar to the 1-2 case, where the 2-cell theme is used as a sampling base, and densities or lengths are sampled within the given areas for the 1-cell theme. The difference is that the other sample consists of the values of the 2-cell theme.

The last case, 2-2, has two different situations: the two sets of polygons fit each other, or they do not. If they fit, values for each theme are taken, and sampling becomes simple. If not, two approaches are applicable for respective purposes: if the statistical issue being used is not classification, generate random points over the area and use the points as a sampling base; or, if the samples are used for classifications, overlay the two themes.

4.4.3.4 MULTIVARIATE ANALYSIS

For multivariate sampling, the key issue is the determination of sampling bases. Once the sampling base is selected, the sampling approaches discussed in the case of bivariate analysis apply to the sampling of each sampling target involved in a multivariate analysis. To identify sampling bases, three cases are distinguished: one and only one of the analytical themes involved is 0-cell, more than one theme is 0-cell, and, there is no 0-cell.

For the first case, the 0-cell theme is chosen as the sampling base, no matter which cells the other themes are. A case is neglected here, namely, the possibility that the measure on the 0-cell theme required for the analysis is density. It is not accounted for, since point features are limited in the National Atlas, and the possibility to use point density for multivariate analysis seems very low.

In the case where more than one theme is 0-cell, complex judgement is needed to determine the sampling base. Firstly, whether the multiple sets of points fit each other should be checked; if yes, the points constitute the sampling base. If the first check is negative, a further check should be conducted on how many themes that are modeled in 0-cell are also 0-cell in the real world. If the number is one, or the number is more than one and the number of point sets fit each other, then use the points as the sampling base, transfer all other 0-cell themes (that are 3-cell in the real world) into 3-cell, so that they can be sampled by the sampling base, and the bivariate sampling approaches apply to the sampling of various analytical themes. If more than one 0-cell themes are 0-cell in the real world, and the points do not completely fit each other, check if the analytical measures for the non-fitted themes are distance. If yes, check if the principal analytical theme is 0-cell. If yes, use the analytical theme as the sampling base, transfer all 0-cell themes that are 3-cell in the real world into 3-cell for sampling, and the bivariate sampling approaches are applied to the

sampling of various themes. If the principal analytical theme is not 0-cell, or there is a pair of point themes which are 0-cell in the real world, and neither of which is measured in distance, then the sampling can not be done. If none of the themes is 0-cell in the real world (meaning all 0-cell themes are 3-cell in the real world), then transfer all the 0-cell themes into 3-cell, except for one that is remained to be the sampling base, and all bivariate sampling approaches apply to the sampling of each analytical theme involved.

If no 0-cell theme appears in the analytical themes involved in the multivariate analysis, check whether there is one and only one 2-cell theme; if yes, use it as the sampling base. If more than one theme is 2-cell, check if the polygons fit each other. If yes, use them as the sampling base. Otherwise, if a classification analysis is carried out, use overlay to get the sampling base; if the statistical method is not classification, use random points as the sampling base. All sampling approaches to bivariate analyses apply to the sampling of each analytical theme.

4.5 KNOWLEDGE VALIDATION

The knowledge extracted or developed in the knowledge acquisition phase needs to be validated, in order to ensure the quality of the expertise, and make it more suitable for the system's behaviors. The knowledge involved with this issue is of three types: statistical methods, analysis issues, and sampling knowledge, as discussed in the preceding section. Because of the large amount of work involved, only the type of statistical knowledge was validated in the thesis work. This section presents the objectives of knowledge validation, the methodologies adopted in the process, and the refinement of the statistical knowledge. Additionally, basic ideas on the refinement of analysis issues are given.

4.5.1 OBJECTIVES OF KNOWLEDGE VALIDATION

Knowledge validation is a process involved in knowledge engineering. The purpose of validating EASSA's knowledge is multifold.

(1) *To reduce the subjectivity of knowledge.* Formulation of statistical knowledge is based on a review of numerous existing books, and on the knowledge possessed by the knowledge engineer. Although there are a large number of statistical books that can be used in the formulation, none perfectly fits our interests. Thus, it is inevitably subjective to decide which books should be selected as major references, which as minor references, and how a combination should be made. The formulation of the analysis issue knowledge is even more subjective, since there is no book that can be directly referred to for this purpose. A validation process is therefore necessary. The subjectivity may be reduced, to a certain extent, through interactions with domain experts.

(2) *To make the expertise more complete.* Statistics is a large domain. Although most of the the expertise can be found in books, interactions with experienced experts may enable the researcher to formulate a more complete and practical knowledge set.

(3) *To find holes in the existing knowledge structure.* The knowledge extraction for EASSA was based on a scheme which was designed in terms of a preliminary conceptual design of the system. In such a scheme, holes were unavoidable. These holes were expected to be eliminated by the validation process.

(4) *To avoid mistakes.* The knowledge extraction for statistical methods involved a fairly large volume of knowledge, which made it very difficult to avoid mistakes. Through validation, the mistakes may be corrected.

(5) *To get more professionals involved with the project.* EASSA is going to be a system that helps users to use statistics, either in the EA environment, or independently of the EA. However, statistics are presently not widely used by GIS users, hence the necessity of informing concerned professionals about the project, before the system is implemented. At the same time, helpful feedback for system development can be obtained from the users. The knowledge validation is an opportunity to accomplish this purpose.

4.5.2 METHODOLOGY FOR KNOWLEDGE VALIDATION

The methodology for knowledge validation is adopted from general knowledge engineering approaches, but taking into account the particular situation of the EASSA project. Several basic procedures are involved: interviewing experts, case studies, further literature review, knowledge refinement, reinterviews, and further refinement.

4.5.2.1 INTERVIEWS OF EXPERTS

The knowledge deals with domains such as statistics, spatial analysis, and various realms of geography, including physical and environmental geography, social and cultural geography, and economic and resources geography. In order to ensure the representativeness of the expertise, about twenty-five experts were interviewed, covering all these domains, and involving both the academic world and application agencies.

A questionnaire package for the interview was designed and formulated. It consisted of four components: Questionnaire 1 dealing with statistical methods, Questionnaire 2 dealing with statistical analysis issues, an introduction to the EASSA project, and lists of existing and potential themes of the National Atlas of Canada. The experts were required to fill

either one of the questionnaires or both, according to their specialties. A five class rating scale was designed for the experts' evaluations. Several face-to-face interviews have been conducted to further discussion.

Of the interviewed experts, seven specialize in statistics. The reason why the number is small is that the subjectivity of the statistical knowledge is much lower than that of the analysis issue knowledge, and less input from experts can basically meet the needs of the validation. Four of these experts gave responses, one filling both of the Questionnaires, one filling the statistical Questionnaire, and the others giving comments and suggestions on the knowledge as well as the project. The other eighteen experts interviewed were geographers. About half of them gave responses to the analysis issues: five of them filled the Questionnaire of analysis issues, and the other five gave comments and suggestions.

The interactions with the domain experts showed that the project is interesting and meaningful. Several geographers with special interests in statistics believe that such a system could be very helpful for general users and students using statistics, and could also be helpful for the domain professionals in recalling infrequently used statistical methods. Another important input from the interactions was that the project might involve too wide domains and too many functions, which might reduce the possibility of success. For the analysis issue knowledge, responses were of two different types. One type was optimistic, believing that such a type of knowledge can offer users much convenience in using the information of the Electronic Atlas, and the summarization of the knowledge is possible. The other type of response, however, doubted if the analysis issues can be summarized with such extensive geographical information, varied data quality, and infinite application potentials.

4.5.2.2 CASE STUDIES

Case studies have been conducted, partly in order to find holes in the knowledge structure. Two cases were tested against the modes to be used by users. One was the spatial autocorrelation analysis of life expectancy; the other was the equality test of average income per capita between two regions. Various types of advice were generated by emulation, involving analysis issues, statistical methods, sampling modes, hypotheses, assumptions, analysis procedures, result interpretations, graphic methods, and limitations. The self-explanation capability of the system was also demonstrated.

An emphasis in the case studies was to check whether the designed knowledge schemas, as well as their structure, are necessary and sufficient to fulfil the tasks to be undertaken by the expert advisor. For this purpose, a dialog between the system and the user was carried out by following both the intended functionalities and the formulated knowledge as well as the three layer structure. Since the meta knowledge on temporal issues has not yet been developed, the temporal differences in the analyzed data were neglected in the studies.

Some weaknesses of the conceptual system were exposed in the case studies. First, the functional explanations on some statistical methods seemed to be insufficient for general users; this may be improved by adding examples for each of statistical methods. Secondly, the development of a multi-level explanation mechanism was suggested, since users with varied statistical background and different mathematical interests would require the explanations in different detail. It is particularly necessary to enable the system to explain some statistical concepts that are not easily understood. Thirdly, it was found that for optimal selections of statistical methods, it is necessary to assign priority factors to some methods that are suitable for the same type of analysis. The case studies, on the other

hand, proved that the basic structure (the three layer structure) is suitable for the system's behavior.

The case studies were also intended to demonstrate the capabilities and behaviors of the expert advisor. Therefore, the topic is further addressed in Chapter 7 dealing with the design and behavior demonstration of the system.

4.5.2.3 EXTENDED REVIEW OF LITERATURE

The decision to focus on the validation of the statistical knowledge was actually made after receiving feedback from the questionnaires and carrying out the case studies. This was because of the amount of work that needed to be done for the validation of both the statistical methods and the analysis issues; this undoubtedly exceeded the capacity of the thesis work.

Both discussions with experts and case studies suggested that more detailed knowledge is needed for advice and explanations, and, more importantly, spatial contexts should be taken into consideration. Therefore, an extended review of literature was carried out. The reviewed book list was extended from seven to thirteen, and some recent papers were referenced as well. The emphasis of the extended review was placed on the spatial statistical methods and the approaches to adapting conventional methods to the spatial context.

4.5.2.4 KNOWLEDGE REFINEMENT

Based on the results of the interviews, case studies, and extended literature review, the statistical knowledge was refined. The changes focused mainly on three aspects:

considering the spatial issues in most of the classical methods, adding a number of statistical methods to the knowledge set, and adding or cancelling some knowledge items in the knowledge structure. At the same time, many questions, which were proposed in the original knowledge set as possible problems, were solved. The next section gives a detailed report on these changes.

4.5.3 REFINEMENT OF STATISTICAL KNOWLEDGE

After carrying out the series of methods for knowledge validation, the statistical knowledge was improved. Although the spatial context is a focus of the refinement, a strategy was still followed, namely, starting with the conventional methods, while taking the spatial context into account. Thus, those methods specifically for spatial analysis will not be considered for the initial system, except for some approaches that are very useful for the electronic mapping purpose. In addition, although there were suggestions to restrict the number of methods to one or two and develop the whole analysis functions in the EA environment, the idea guiding the refinement was still to develop a relatively thorough knowledge set, since the attempt is to develop an advisory system first. This section presents the considerations of spatial contexts, changes to statistical methods, and changes in the knowledge structure.

4.5.3.1 CONSIDERATIONS OF SPATIAL CONTEXT

As introduced in Section 2.2.2, statistics can be differentiated into two branches: descriptive statistics and inferential statistics. Descriptive statistics are those techniques used to organize and summarize data in general data handling problems. They can be used for spatial data summary without specific problems. Inferential statistics involve generalization from samples to populations. The generalizations are always based on two

assumptions: random samples, and independent observations. These assumptions, however, are often violated in a spatial context. Nevertheless, inferential statistics are much more powerful and more interesting to researchers. In fact, over seventy percent of the statistics to be used in EASSA belong to this branch.

Solutions, therefore, have to be found to deal with assumption violations. A number of books have been published to this purpose. The most popular one might be 'Spatial process: models and applications' by Cliff and Ord (1981). Several other books address these problems as well, such as Griffith (1987) and Ripley (1981).

The approaches proposed in these books have been introduced to the statistical knowledge set of EASSA from two channels: sampling assumptions, and autocorrelation analysis. Sampling assumptions are used to constitute random samples in inferential statistics. Two alternatives can be chosen: randomization assumption and normality assumption (see Tables 24 - 26 for details; refer to Cliff and Ord, 1981, for further explanations). The autocorrelation analysis as well as modified statistical models are used to deal with the violation of the independent observation assumption required in equality test, correlation analysis, regression analysis, and trend surface analysis.

4.5.3.2 CHANGES TO STATISTICAL METHODS

About ten methods are added to the statistical knowledge set, based on the suggestions of experts and further reviews of literature. Meanwhile, several methods were removed, because the responses from specialists showed that they were less used in geographical analysis.

Methods Added

Three types of spatial analysis methods were added to the knowledge set: the nearest neighbor method, the autocorrelation analysis approaches, and the Kriging technique.

Nearest Neighbor Distance Method. This is one of the basic spatial statistical methods used to test the uniform process of point patterns. It is included in the knowledge set since it is needed to verify random point samples for other statistical analysis involved in the system. The nearest neighbor distance method is used to replace the quadrat method in the original knowledge set. The reason for this replacement is that, as discussed in 4.5.1.1, the nearest neighbor distances convey information on the spatial pattern, whether taken from randomly selected locations (or objects) or considered for all objects in a mapped area. When sampling at random, the observations are effectively independent, provided that the number of observations taken is not too large relative to the total study area. For mapped data, even though the n distances will not form n independent observations, their test can be used to search for randomness under the Poisson null hypothesis, since the distribution is close to that for independent observations (Cliff and Ord, 1981). Compared to this method, the quadrat approach is thought to be poor for the requirement both to reflect spatial structures and to provide random samples.

Autocorrelation Analysis. This is a test tool to be used in several inferential statistics. Also, it is an independent analysis method used to explore spatial processes for various types of variables with 0-cell, 2-cell or 3-cell distributions. The National Atlas involves a large number of such variables, hence, the method is very useful. In order to deal with different data measurements, two autocorrelation indices are included in the refined statistical knowledge set, namely, Moran's index, suitable for interval or ratio data, and joint-count index, for nominal data.

Kriging. Spatial surface analysis is one of the most useful functions in cartography and map analysis. A traditional analysis method for this purpose is trend surface analysis, which was implied in the regression analysis in the unrefined knowledge set. The trend surface analysis has two major shortcomings: it is based on assumptions that are rarely met in practice; and it neglects local variation because it uses pure mathematical models. Kriging is a set of spatial surface analysis methods devised to overcome these difficulties. Its assumption is readily met in spatial analysis. Also, local variation can be easily detected. Only the simple kriging approach is added to the knowledge set, because of the relative simplicity of this method.

In addition, trend surface analysis is separated from multiple regression analysis because of its particular use. The reasons why the method is not cancelled are that it is more efficient than Kriging, and, also, it is useful when the purpose of the analysis is restricted to an inspection of the general spatial trend, namely, to finding a 'best fit'.

Other classical statistical methods which have been modified are regression analysis, factor analysis, cluster analysis, discriminant analysis, and equality test. For regression analysis, the curvilinear method, data transformation techniques, as well as a modified regression model for autocorrelation problems have been added, in order to deal with situations where assumptions are violated. Both statistical analysis and non-statistical analysis are involved in factor analysis, so that both variables and cases can be analyzed as objects. The same applies to cluster analysis; both case cluster and variable cluster methods are included. In addition to the linear function in discriminant analysis, the quadratic function is also involved in the refined knowledge set in order to cope with different data characteristics. Two modified models are added to equality tests, one for dealing with the problem proposed in the original knowledge set, and the other for autocorrelation cases. In

addition, Lilliefors' test is added to deal with the normality test, the F-test is added for the variability comparison of two samples, and the Chi-square test is used in population distribution comparison.

Methods cancelled

Several methods were removed: distribution description, randomness tests for sequential distributions, and equality tests between two sequentially distributed populations and between paired populations. They were all evaluated at low rates in the completed questionnaires.

Besides, the Wilcoxon rank sum test has been replaced by the Mann-Whitney U-test, which is regarded as the best non-parametric method for two sample testing.

4.5.3.3 CHANGES IN THE KNOWLEDGE STRUCTURE

In order to derive various types of advice, fourteen items of knowledge for each statistical method were formulated in the original knowledge set. The interviews, case studies, and literature review suggested the addition of four items and the removal of three others.

Knowledge Items Added

(1) *Examples*. For each method, examples can help users understand the goals of the method. They will be obtained, upon request, during users consultations.

(2) *Methodology*. Most of interviewed experts thought that a brief explanation of the principles for a recommended method would be welcomed by users.

(3) *Certainty Factor Descriptions.* All inferential statistical methods have assumptions, but these assumptions are rarely fully met in applications. It is therefore necessary to assign certain factors to various degrees to which the statistical methods are applicable. This item is added to the knowledge structure, but the knowledge is still not complete for some methods because of the lack of sufficient reference or of the shortage of the researches.

(4) *Priority Factor.* As indicated in 4.3.4.2, several statistical methods may be suitable for some analysis, but different methods have varied accuracy in terms of statistical principles. When multiple methods can be used for an analysis, the method providing the highest accuracy should be recommended. This can be realized by assigning priority factors to the methods. This item is therefore added to the statistical knowledge schema.

Knowledge Items Cancelled

(1) *Spatial Distribution.* The item of spatial distribution, included in the original statistical knowledge schema, dealt with differentiating two types of methods: the methods for sequential distribution, and the methods for non-sequential distributions. The interviews suggested that sequential methods do not have high values in most geographical analyses. Therefore, the item, as well as a number of sequential methods, has been removed.

(2) *Test statistics.* The item of test statistics is taken out because it is not a heuristic knowledge in method selections, nor information important for the users. Essentially, it is included in the item of Analysis Processes.

(3) *Frequency distribution.* The frequency distributions of analytical data sets constitute an important factor in statistical method selections, but in general statistical texts, it is included

in the assumptions of methods. This trend is followed in the refined statistical knowledge set.

4.5.3.4 OTHER CHANGES

Knowledge for Analysis Processes, Result Interpretations, and Possible Problems (changed into Limitations) are explained in more detail in the refined knowledge set, especially for multivariate techniques. The solutions available for some problems proposed in the original version are presented in the item Analysis Processes. In addition, a few mistakes found in the original knowledge set were corrected.

4.5.4 ON ANALYSIS ISSUE KNOWLEDGE

As previously mentioned, the analysis issue knowledge is more subjective than the statistical knowledge, since there is no book that can be referred to directly. This requires a larger number of statistical investigations. Unfortunately, the number of responses with evaluations is too small to conduct this analysis. Several face-to-face discussions were held in order to collect information relating to this issue. The responses varied with the domains that the experts are working on. In general, experts in the academic world tend to believe that such knowledge is useful for users as *a priori* knowledge, and that it is possible to generate it, while several geographers suggested that such knowledge was too difficult to generate.

The author's opinion is that it is useful knowledge, but more investigations are needed. A better way to do it is to start with part of the issues, e.g. the set for regression analysis. After they are proven to be valuable, the issues can then be expanded to various statistical methods.

4.6 SUMMARY

This chapter has described the work done in the acquisition and development of expertise involved in EASSA's advisory activities. Firstly, the problems to be solved by EASSA were identified and decomposed, using the concept of problem-solving space. EASSA's problem-solving processes are thus vertically decomposed into three levels, and the third level is further decomposed into five subspaces. On the basis of the decompositions, the structured knowledge schema of the system was formulated. It is composed of three layers of knowledge, with respect to the three levels of problem-solving spaces. The bottom layer, referred to as the domain knowledge layer, includes four types of knowledge. The first type is analysis issues, which indicate statistical topics, as well as the analytical geographical themes suitable for the topics. The second type is statistical knowledge, which includes statistical methods dealing with various statistical topics involved in EASSA, as well as a number of items of statistical knowledge related to each of the methods. The third type of knowledge deals with data sampling. Several factors related to the selection of map-based sampling techniques, as well as the sample measures, form the sampling knowledge schema. The last type refers to the Atlas context knowledge, which presents a number of characteristics of electronic maps. Above the bottom layer is the advice-oriented knowledge layer. It deals with various inferences, as regards the types of advice to be given by the system, through the appropriate use of the domain layer of knowledge. At the top layer is the task-oriented knowledge layer, which deals with the cooperation between users and the system, in order to fulfil various tasks with respect to the multiple advisory modes to be provided by the system. In addition, meta knowledge, including that for solving such problems as uncertainty of statistical method selections, temporal differences of analytical themes, and explanations, is added to resolve additional problems encountered in the derivation of advice. Based on the structured schemas, the

knowledge including analysis issues, statistical knowledge, and sampling knowledge was extracted and developed. This was followed by a knowledge validation process, which involved interviews of domain experts, case studies, extended literature review, knowledge refinement, and a reiterative process. The focus of the validation was on the statistical knowledge. The refinement of the analysis issues and sampling knowledge remains for later research.

5. KNOWLEDGE FORMALISM

5.1 INTRODUCTION

EASSA's knowledge schemas were discussed in the previous chapter, and the primary knowledge was formulated to a certain extent. In order to make knowledge precise, concise, and complete, this chapter firstly describes the hierarchically structured knowledge using mathematical languages; EASSA's knowledge models are thus formalized. Secondly, the knowledge formalism is accomplished from the perspective of knowledge representation in the system.

5.2 EASSA'S KNOWLEDGE MODELS

The layer structure of EASSA's knowledge implies that EASSA's knowledge models must be hierarchically constructed. Vertically, there are three levels of models, one for each of the three layers of knowledge. The bottom level is comprised of four models dealing with the analysis issues, statistical methods, sampling approaches, and Atlas contexts respectively. The intermediate level includes five knowledge models, which correspond to the five types of inferences for deriving advice. The top level model refers to the task-oriented knowledge for fulfilling particular tasks. For simplicity, this section only presents the models of the various types of knowledge; Section 4.3 should be referred to for detailed explanations.

The mathematical notations used in the model descriptions are as follows (Grimaldi, 1985; Frost, 1986):

{ }

set symbol

$()$	tuple symbol or brackets
$x \in A$	element x is a member of set A
$x \notin A$	element x is not a member of set A
$A \supseteq B$	B is a subset of A
ϕ	empty or null element
$\exists$	existential quantifier; for some
$\forall$	universal quantifier; for all
$p \wedge q$	the conjunction of p and q ; p and q
$\rightarrow$	implies
$\vdash$	infers
$= (x, y)$	$x = y$
$> (x, y)$	$x > y$

5.2.1 DOMAIN KNOWLEDGE MODELS

Analysis Issues Knowledge Model

EASSA's analysis issue knowledge AI is an infinite set of 5-tuples, where each of the elements represents an item of the analysis issues. AI can be expressed as follows:

$$AI = \{ (si, fe, gr, (thm_1, thm_2, \dots, thm_j, \dots, thm_n), nt); i = 1, 2, \dots \} \quad \text{Model-1}$$

where $si \in SI$, SI represents a set of statistical issues, and $SI = \{\text{statistical summary, equality test, bivariate relationship analysis, multivariate relationship analysis, univariate classification, multivariate classification, randomness test, spatial surface analysis...}\}$; $fe \in FE$, FE denotes a set of function explanations for SI , and fe is the functional explanation for the statistical topic si ; $gr \in GR$, GR denotes a set of geographical realms, and $GR =$

{physical environmental geography, social and cultural geography, economic and resource geography}; $\text{thm}_i \in \text{THM}$, THM denotes a set of tuples of analytical themes for various analyses, thm_i is the principal theme ascribed to gr, $\{\text{thm}_j | j = 2, 3, \dots, n\}$ are minor themes for the i th analysis, n is the number of analytical themes involved in the analysis, where themes are both those included in the Electronic Atlas of Canada and those listed in its perspective; $\text{nt} \in \text{NT}$, NT denotes a set of additional information or notes for all analyses, and nt is the notes for the i th analysis.

Statistical Knowledge Model

EASSA's statistical knowledge SK is an infinite set of 14-tuples, with each tuple being an item of statistical knowledge. SK can be expressed as:

$$\text{SK} = \{ (si, st, sf, em, hd, ms, hp, as, ud, ap, ri, gm, lm, mp) ; i = 1, 2, \dots \} \quad \text{Model-2}$$

where $si \in \text{SI}$, and SI denotes the set of statistical issues as defined for Model-1; $st \in \text{ST}$, ST denotes a set of statistical methods categorized by SI, and $\phi \notin \text{ST}$; $sf \in \text{SF}$, SF denotes a set of functional explanations for all of the statistics, and $\phi \notin \text{SF}$; $em \in \text{EM}$, EM denotes a set of examples for all of the statistical methods, and $\phi \notin \text{EM}$; $hd \in \text{HD}$, HD denotes a set of analysis principles of the statistical methods, and $\phi \notin \text{HD}$; $ms \in \text{MS}$, MS denotes a set of data measurement scales, and $\text{MS} = \{\text{nominal, ordinal, interval, ratio}\}$; $hp \in \text{HP}$, HP denotes a set of hypotheses for statistical tests, and $\phi \notin \text{HP}$; $as \in \text{AS}$, AS denotes a set of assumptions for the statistics, and $\phi \notin \text{AS}$; $ud \in \text{UD}$, UD denotes a set of uncertainty descriptions on the applicabilities of the statistical methods resulting from the testing of assumption satisfactions, and $\phi \notin \text{UD}$; $ap \in \text{AP}$, AP denotes a set of analysis processes for the statistics, and $\phi \notin \text{AP}$; $ri \in \text{RI}$, RI denotes a set of analytical result interpretations, and $\phi \notin \text{RI}$; $gm \in \text{GM}$, GM denotes a set of graphic methods to visualize data and statistical

results, and $\phi \in GM$; $lm \in LM$, LM denotes a set of limitations of the statistical methods, and $\phi \in LM$; and $mp \in MP$, MP denotes a set of statistical method priorities dealing with preference of statistical methods, and $\phi \in MP$.

Atlas Context Knowledge Model

EASSA's Atlas context knowledge AC is an infinite set of 7-tuples, with each element being an item of Atlas context knowledge. AC can be expressed as:

$$AC = \{ (thm, gr, sd, rd, ms, ds, ts)_i \mid i = 1, 2, 3, \dots \} \quad \text{Model-3}$$

where $thm \in THM$, and THM represents a set of geographical themes in the Electronic Atlas; $gr \in GR$, and GR denotes a set of geographical realms as defined in Model-1; $sd \in SD$, SD denotes a set of spatial dimensions of themes of THM , and $SD = \{0\text{-cell}, 1\text{-cell}, 2\text{-cell}, 3\text{-cell}\}$; $rd \in RD$, RD denotes a set of spatial distributions of themes of THM , with the same elements as SD ; $ms \in MS$, and MS denotes a set of measurement scales as defined in Model-2; $ds \in DS$, and DS denotes a set of dates when the themes of THM were surveyed; and $ts \in TS$, and TS denotes a set of temporal stabilities of the themes of THM .

Sampling Mode Knowledge Model

EASSA's sampling knowledge SM consists of an infinite set of rules which can be expressed in a predicate logic (Frost, 1986):

$$\begin{aligned} & \exists(thm_1, thm_2, \dots, thm_n) \exists(sd_1, sd_2, \dots, sd_n) \exists(rd_1, rd_2, \dots, rd_n) \exists st \exists tm \\ & \supseteq (AC, \{ (thm_i, gr_i, sd_i, rd_i, ms_i, ds_i, ts_i) \mid i = 1, 2, \dots, n \}) \\ & \rightarrow \exists sb \exists(sm_1, sm_2, \dots, sm_n) \exists(se_1, se_2, \dots, se_n) \end{aligned} \quad \text{Model-4}$$

where $thm \in THM$, THM is as defined in Model-1, and $(thm_1, thm_2, \dots, thm_n)$ are all themes involved in an analysis; AC is the set of map contexts as defined in Model-3; $(sd_1, sd_2, \dots, sd_n)$ are n spatial dimensions of the analytical themes; $(rd_1, rd_2, \dots, rd_n)$ are n spatial distributions of the analytical themes; st is the statistical method applied to the analysis; $tm \in TM$, and TM is a set of measures of THM ; sb is the spatial dimension of the sampling base selected for the analysis; $(sm_1, sm_2, \dots, sm_n)$ are the sampling modes selected to sample the n analytical themes respectively; and $(se_1, se_2, \dots, se_n)$ are the sampling explanations which correspond to the n sampling modes.

5.2.2 ADVICE-ORIENTED KNOWLEDGE MODELS

Model for Inference of Analysis Issues

Let AI , SK , and AC be defined as Model-1, Model-2, and Model-3 respectively, then the deductive inference of WI for recommending an analysis issue can be expressed as:

$$\forall ai (F1(ai) \rightarrow G1(ai)), \quad F1(aii) \vdash G1(aii)$$

$F1(ai)$ is a set of constraints defined as:

$$=(ai, (si, fe, gr, (thm_1, thm_2, \dots, thm_n), nt)) \wedge \in (ai, AI)$$

$$\wedge \exists ACi (=(ACi, \{(thm_i, gr_i, sd_i, rd_i, ms_i, ds_i, ts_i) \mid i = 1, 2, \dots, n\}))$$

$$\wedge \supseteq (AC, ACi)$$

$$\wedge \exists sk (=(sk, (si, st, sf, em, hd, ms, hp, as, ud, ap, ri, gm, lm, mp))) \wedge \in (sk, SK)$$

$$\wedge =(ms, ms_i)$$

$G1(ai)$ means that ai is recommended

Model-5

where ai is any analysis issue considered by EASSA; aii is an analysis issue instance; and the italics show the elements used in the deduction (this also applies to the following models).

Model for Statistics Selections

Let AI, SK, and AC be defined as Model-1, Model-2, and Model-4 respectively, and sfu denotes the choice of existing statistical functions by users, then the deductive reasoning of WS for selecting an appropriate statistical method can be expressed as:

$$\forall st (F2(st) \rightarrow G2(st)), \quad F2(sti) \vdash G2(sti)$$

$F2(st)$ is a set of constraints defined as:

$$=(ai, (si, fe, gr, (thm_1, thm_2, \dots, thm_n), nt) \wedge \in (ai, AI)$$

$$\wedge \exists sk (=(sk, (si, st, sf, em, hd, ms, hp, us, ud, ap, ri, gm, lm, mp)) \wedge \in (sk, SK)$$

$$\wedge \exists ACi (=(ACi, \{(thm_i, gr_i, sd_i, rd_i, ms_i, ds_i, ts_i) \mid i = 1, 2, \dots, n\})$$

$$\wedge \supseteq (AC, ACi))$$

$$\wedge = (sfu, sf)$$

$$\wedge = (ms, ms_i)$$

$$\wedge (\eta(us, mp) \rightarrow st)$$

Model-6

where η is a function dealing with the synthesis of the uncertainty factors of assumption satisfactions and the priority factor of the statistical method; st is any statistical method considered for method selection; and sti is a statistical method instance considered for an analysis.

Model for Sampling Mode Selections

The deductive reasoning of WQ for selecting a set of appropriate sampling modes for the set of analytical themes involved in an analysis can be expressed as:

$$\begin{aligned} & \exists(thm_1, thm_2, \dots, thm_n) \exists(sd_1, sd_2, \dots, sd_n) \exists(rd_1, rd_2, \dots, rd_n) \exists st \exists tm \\ & \supseteq(AC, \{ (thm_i, gr_i, sd_i, rd_i, ms_i, ds_i, ts_i) \mid i = 1, 2, \dots, n \}) \\ & \rightarrow \exists sb \exists(sm_1, sm_2, \dots, sm_n) \exists(se_1, se_2, \dots, se_n), \end{aligned}$$

$$\begin{aligned} & \supseteq(AC_i, \{ (thm_i, gr_i, sd_i, rd_i, ms_i, ds_i, ts_i) \mid i = 1, 2, \dots, n \}) \quad \text{I--} \\ & \exists sb_i \exists(smi_1, smi_2, \dots, smi_n) \exists(sei_1, sei_2, \dots, sei_n) \end{aligned}$$

Model-7

The first formula represents the inference of sampling modes for general case, where $thm \in THM$, THM is as defined in Model-1, and $(thm_1, thm_2, \dots, thm_n)$ are all themes involved in an analysis; AC is the set of map contexts as defined in Model-3; $(sd_1, sd_2, \dots, sd_n)$ are n spatial dimensions of the analytical themes; $(rd_1, rd_2, \dots, rd_n)$ are n spatial distributions of the analytical themes; st is the statistical method applied to the analysis; $tm \in TM$, and TM is a set of measures of THM ; sb is the spatial dimension of the sampling base selected for the analysis; $(sm_1, sm_2, \dots, sm_n)$ are the sampling modes selected to sample the n analytical themes respectively; and $(se_1, se_2, \dots, se_n)$ are the sampling explanations which correspond to the n sampling modes. The second formula indicates the inference for an instance; all the symbols correspond to those of the first formula, but have additional 'i's which mark the instance.

Models for Generating Hypotheses, Assumptions, Interpretations, Graphic Methods, and Limitations

Let AI, SK, and WS be defined as Model-1, Model-2, and Model-6 respectively, then the deductive inference of WH, WA, HI, WG, and WL, for generating statistical hypotheses, assumptions, result interpretations, graphic representation methods, and limitations respectively, can be expressed as:

$$\forall ai \exists st (F3(ai, st) \rightarrow (ahp, aas, ari, agm, alm)),$$

$$F3(a_{ii}, st_i) \vdash (ah_{pi}, aa_{si}, ar_{ii}, ag_{mi}, al_{mi})$$

$F3(ai, st)$ is a set of constraints defined as:

$$=(ai, (si, fe, gr, (thm_1, thm_2, \dots, thm_n), nt)) \wedge \in (ai, AI)$$

$$\wedge (WS \rightarrow st)$$

$$\wedge (sk, (si, st, sf, em, hd, ms, hp, as, ud, ap, ri, gm, lm, mp)) \wedge \in (sk, SK)$$

$$\wedge (Aggregate((thm_1, thm_2, \dots, thm_n), hp, as, ri, gm, lm), (ahp, aas, ari, agm, alm))$$

Model-8

where ai and st denote general analysis issues and statistical methods; a_{ii} and st_i denote a particular analysis issue and the statistical method used for the analysis; (ahp, aas, ari, agm, alm) refer to the generated hypothesis, assumptions, result interpretations, graphic methods, and limitations respectively for general cases; and (ah_{pi}, aa_{si}, ar_{ii}, ag_{mi}, al_{mi}) refer to the same items for the analysis issue instance a_{ii} as well as st_i.

Model for Generating Analysis Procedures

Let AI, SK, WS, and WQ be defined as Model-1, Model-2, Model-6, and Model-7

respectively, then the deductive reasoning of HH for generating analysis procedures can be expressed as:

$$\forall ai \exists st \exists (sm_1, sm_2, \dots, sm_n) (F4(ai, st, (sm_1, sm_2, \dots, sm_n)) \rightarrow aap),$$

$$F4(ai, st, (smi_1, smi_2, \dots, smi_n)) \vdash aapi$$

$F4(ai, st, (sm_1, sm_2, \dots, sm_n))$ is a set of constraints defined as:

$$=(ai, (si, fe, gr, (thm_1, thm_2, \dots, thm_n), nt)) \wedge \in (ai, AI)$$

$$\wedge (WS \rightarrow st)$$

$$\wedge =(sk, (si, st, sf, em, hd, ms, fd, hp, as, ud, ap, ri, gm, lm, mp)) \wedge \in (sk, SK)$$

$$\wedge (WQ \rightarrow (sm_1, sm_2, \dots, sm_n))$$

$$\wedge =(Aggregate ((thm_1, thm_2, \dots, thm_n), ap, (sm_1, sm_2, \dots, sm_n)), aap)$$

Model-9

where app is inferred analysis procedure for general cases, and aapi is its instance.

5.2.3 TASK-ORIENTED KNOWLEDGE MODELS

Definitions of UI and SA

Definition- . Let AIS be a set of users' selections of analysis issues, chosen from the analysis issue list suggested by the system ; let SFU be a set of statistical functions chosen by users; let AIP be a set of analysis issues provided by users; and let MCP be a set of map contexts provided by users. Then the users' interactive information with EASSA is defined as:

$$UI = \{ AIS, SFU, AIP, MCP \}$$

Definition-2: Let WI, WS, WQ, WH, WA, HI, WG, WL and HH be defined as Model-5, Model-6, Model-7, Model-8, and Model-9 respectively, then EASSA's advice is defined as:

$$\begin{aligned} SA &= (WI, WS, WQ, WH, WA, HH, HI, WG, WL) && \text{if } WS \neq \emptyset \wedge WQ \neq \emptyset; \text{ or} \\ SA &= \text{Cannot be analyzed} && \text{if } WS = \emptyset \vee WQ = \emptyset \end{aligned}$$

These definitions are applied to the following models.

Model for Issue-Atlas-dependent Task

Let $ais \in AIS \in UI \wedge sfu \in SFU \in UI \wedge sa \in SA$. Then, the knowledge oriented to a Issue-Atlas-Dependent task is a set of ordered tuples:

$$IAD = \{ (ais, sfu, sa)_{\gamma} \mid \gamma = 1, 2, \dots, k \} \quad \text{Model-10}$$

Model for Issue-independent-Atlas-dependent Task

Let $aip \in AIP \in UI \wedge sfu \in SFU \in UI \wedge sa \in SA$. Then, the knowledge oriented to a Issue-Independent-Atlas-Dependent task is a set of ordered tuples:

$$IIAD = \{ (aip, sfu, sa)_{\gamma} \mid \gamma = 1, 2, \dots, k \} \quad \text{Model-11}$$

Model for Issue-dependent-Atlas-independent and Issue-dependent-Atlas-semi-dependent Tasks

Let $ais \in AIS \in UI \wedge sfu \in SFU \in UI \wedge mcp \in MCP \in UI \wedge sa \in SA$. Then, the

knowledge oriented to Issue-Dependent-Atlas-Independent or Issue-Dependent-Atlas-Semi-dependent tasks is a set of ordered tuples:

$$IDAI/IDAS = \{ (ais, sfu, mcp, sa)_{\gamma} \mid \gamma = 1, 2, \dots, k \} \quad \text{Model-12}$$

Model for Issue-Atlas-independent and Issue-independent-Atlas-semi-dependent Tasks

Let $aip \in AIP \in UI \wedge sfu \in SFU \in UI \wedge mcp \in MCP \in UI \wedge sa \in SA$. Then, the knowledge oriented to Issue-Atlas-Independent or Issue-Independent-Atlas-Semi-dependent task is a set of ordered tuples:

$$IAI/IIAS = \{ (aip, sfu, mcp, sa)_{\gamma} \mid \gamma = 1, 2, \dots, k \} \quad \text{Model-13}$$

5.3 EASSA'S KNOWLEDGE REPRESENTATION METHODS

The previous section provides a conceptual formalism for EASSA's knowledge. The knowledge should be further logically formalized, before being implemented in the expert advisor. This deals with the design of knowledge representation methods. A large amount of knowledge is necessary for the system. As discussed in the previous chapter, about forty statistical methods are to be found in the system; each method has fifteen items of knowledge, and some of the items may be lengthy; the information on the Atlas involves hundreds of themes, each dealing with several facets; and the analysis issues will constitute another large set of information. Therefore, the design of appropriate knowledge representation methods is of crucial importance in EASSA's development.

5.3.1 REPRESENTATION METHOD DESIGN

For optimal design of knowledge representation methods for EASSA, a brief review of commonly used methods was carried out. Based on this, as well as on the characteristics of EASSA's knowledge, solutions are proposed.

5.3.1.1 PRESENTLY USED REPRESENTATION METHODS

The most widely used knowledge representation methods for expert systems are rules, semantic nets, and frames. The model-based method and schema-based method are relatively new approaches. First order logic is widely used in logic programming, which is one of the ways to develop expert systems. Neural network is another artificial intelligence technique that can be used for problem-solving, and is thus also accounted for as a knowledge representation method.

Production Rules

Rules, also called production rules (Frost, 1986; Darwish and Jain 1988), are the most popular knowledge representation technique. They are expressed as IF-THEN statements, and provide a formal way of representing recommendations, directives, or strategies. Rules can be executed in two ways: forward chaining and backward chaining. Forward chaining refers to the inference chains formed from successive execution of rules, by matching their IF portions and executing their THEN portions. Backward chaining conducts inference by matching the THEN portion, and then checking the IF portion.

The major advantage of this method is that it provides a clear and understandable representation of inference logic. Also, uncertainty manipulations can be easily carried out. Its major shortfall is that it can not clearly represent structured knowledge.

Semantic Nets

Semantic nets (Frost, 1986; Niemann, et al. 1990) refer to a knowledge representation method based on a network structure. A semantic net consists of nodes connected by arcs that describe the relations between the nodes. The nodes stand for objects, concepts or events; arcs can be defined in a variety of ways such as *is-a* and *has-part*, which establish a property of inheritance and a hierarchy in the net.

Inferences in semantic nets are realized through inheritance, which allows the properties of a node N_a to be inherited by the set of nodes $\{N_i\}$ which are connected to the N_a with *is-a* or *has-part* arcs. Therefore, the method is a useful way to represent knowledge in domains that use well-established taxonomies to simplify problem-solving. They have also been successfully used in natural language understanding.

Frames

A frame is a data structure for representing a stereotyped situation (Frost, 1986; Payne and McArthur, 1990). It is a network of nodes and relations organized in a hierarchy, where the topmost nodes represent general concept and the lower nodes represent more specific instances of those concepts. The hierarchy also provides inheritance to the structure. The difference between frames and semantic nets is that in a frame system, the concept at each node is defined by a collection of attributes and the names of those attributes; the attributes are called slots. Each slot can have procedures attached to it, and these are executed when

the information is needed or changed. Slots can also have conditions which restrict the slot values within some ranges, and can also have links between the same class of frames, these being used to speed up matching. These properties do not apply to semantic nets. Three types of deductive inference can be made using frames: to infer the existence of an object, to infer the generic values of the object, and, to deduce the values of the object through analogy (see Frost, 1986).

The frame method has many advantages. First, the structure is intuitively appealing. The clustering of properties relating to an object is 'natural' in the sense that people are quite familiar with the approach, and use it in everyday discourse. Second, the method accommodates a taxonomy of knowledge, and provides a clear structure for the knowledge network. Third, it can activate procedural as well as declarative knowledge, which can greatly enhance the flexibility of the structure. In addition, default values, generic properties, and analogical reasoning are accommodated, and all of these increase the capacity of the structure.

One of the major shortfalls of the method is that as yet, there is no formal theory of frames. Consequently, the mechanisms for inference and consistency checking are not based on well-defined semantics. Also, the mechanism for reasoning with uncertainty knowledge is based on *ad hoc* expressions with no formal foundation.

Schema-based Method

Using schemas to represent knowledge models (Havens, 1985; Mulder, et al., 1988) is another approach to knowledge representations. A schema represents a class by specifying its membership as a small number of subclasses that satisfy a set of constraints. The constraints are based on two orthogonal relationships in knowledge organization:

composition and specialization. Complex objects have internal structure, which can be represented as composition constraints among their parts. Likewise, specialization constraints segment classes by type. A schema class has a variety of properties, such as having a set of other schema classes from the knowledge base which constitute components for the class, having a set of constraints over component classes, and having a set of procedures called methods.

The method is used mostly in model-based image vision. An instantiated schema represents a specific image model. A matching between a feature and the schema is carried out by confirming a hypothesis through a hypothesis tree. The inference is implied in this process.

Logic Programming Languages

Logic programming languages (Frost, 1986; Payne and McArthur, 1990) are variants of predicate logic. They can also be attributed to knowledge representation methods, since they are often used to develop expert systems directly. A typical logic programming language is Prolog, which consists of Horn clauses, predicates, operators, and grammar. Only backward chaining is implied in the execution; forward chaining can be developed based on the backward chaining mechanism. Another logic programming language, LISP, provides a mechanism for both forward and backward chaining, but their executions need organization.

The major advantage of logic programming languages in the representation of knowledge lies in their great flexibility. In fact, they are usually used in the programming of expert system shells which use the various knowledge representation methods previously introduced. However, more knowledge would be necessary in using these languages than

using expert system shells or development environments, since some mechanisms such as the ones for explanations and for object management, which are usually provided in expert system shells and development environments, need to be developed in the use of programming languages.

Neural Network

Neural Network (Stanley, 1989; Bailey and Thompson, 1990) is another artificial intelligence approach dealing with human problem-solving. An Artificial Neural Network (ANN) is composed of a large number of simple, interconnected 'processors' (neurons) working in parallel within a network. The ANN has the capability to develop an internal representation of a signal pattern that is presented as input to the network. This automatic programming or 'learning' is accomplished through the dynamic adjustment of network interconnection strengths (adaptive weights) associated with each neuron. This process, termed back propagation, uses a desired outcome pattern and a defined input (training set) to initiate feedback to the neural network. The network cycles through the training set until the synapse weights are such that the network correctly relates the defined input to the desired output. When presented with new data, the internal synapse weights excite or inhibit the firing of specific processing units (neurons). The pattern of these neuron firings segregates the input signals into one of the output classes.

The most apparent advantage of the neural network is that the learning algorithm is not programmed into the network. This eliminates the algorithm and computer code development necessary in expert system development. This approach appears to be very promising for problems that involve the association of elements in a set with elements in another set. Pattern recognition and spectral and textual classification belong to this class.

However, they are not regarded as a good method for problem-solving dealing with sophisticated reasoning.

Others

There are also other representation methods such as conceptual dependency and script (Frost, 1986). The conceptual dependency method regards the universe as consisting of a large set of ontological concepts. The basic units of meaning are structured sub-sets called conceptualizations, which can be linked to each other to build an infinitely complex structure. The method is used in several text understanding systems. The script structure represents a commonly occurring sequence of events. It is used in newspaper understanding, natural language processing, etc.

5.3.1.2 METHOD SELECTIONS AND CONSIDERATIONS

EASSA will have three layers of knowledge: domain-oriented, advice-oriented, and task-oriented knowledge, plus meta knowledge which is involved with all these layers.

Domain Knowledge Representations. The domain-oriented knowledge includes knowledge on statistics, analysis issues, Atlas contexts, and sampling schemes (Section 4.3.1). The first three types of knowledge possess hierarchical structures. Statistical knowledge is first categorized by statistical issues, then by purpose, and, last, by methods (Appendix 1). Analysis issue knowledge is a tree structure dealing with statistical issues and geographical realms. Atlas contexts also require a hierarchical structure that consists of geographical realms and themes. Obviously, frames and schema-based methods are the most appropriate for the representation of these types of knowledge. Of these two approaches, the frame method is preferred, since it is more widely used, and more

techniques and tools are available. Also, the frame method provides a convenient way to include procedures and rules into individual objects. Thus, assumption testing and assumption satisfactions involved in statistical methods can be represented in procedures and rules respectively, as daemons* of frames. Since sampling modes are present in rule form, though having inherent hierarchy as well, either heuristic rules or Horn clauses will be employed to represent the knowledge; the choice between heuristic rules and Horn clauses will depend on whether an expert system shell or a logic programming language is selected for the system implementation.

Advice-oriented Knowledge Representation. The advice-oriented layer knowledge deals with the derivation of various types of advice (Section 4.3.2). It is characterized by deductive inference, inferring from the specifications of an analysis and domain knowledge, to the advice on performing the analysis. Production rules, therefore, can best represent this knowledge. Horn clauses are also applicable, if the implementation tool of the system is a logic programming language rather than an expert system shell.

Task-oriented Knowledge Representation. The task-oriented knowledge is used to control the interface between users and the system (Section 4.3.3). It incorporates the users' information into the system's operations, and ensures proper accomplishment of a specific task. The suitable knowledge representation approach is the rule method or Horn clauses.

Meta Knowledge Representation. The meta knowledge includes the knowledge for giving explanations and for dealing with fuzzy problems involved with advice derivations. This knowledge can be divided into two types: one type is directly related to individual statistical methods, and is hence local (e.g. functional explanation); and the other type is not

* Including procedures and rules

subject to any particular item of domain knowledge, and is therefore global (e.g. explanation of a general statistical concept). The best way to represent the first type of meta knowledge is to place the knowledge in corresponding frames. For the second type of knowledge, it is necessary to further distinguish it as descriptive and deductive. For example, the explanations of statistical concepts are descriptive, while the handling of uncertainty factors and priority factors in statistical method selections may deal with rules, and hence is deductive. Descriptive meta knowledge is best represented in independent frames, while deductive meta knowledge should be represented in rules or Horn clauses.

5.3.2 EASSA'S KNOWLEDGE REPRESENTATION METHODOLOGY

In this section, two examples of knowledge representation methods applied to EASSA are given, one being frame representation for domain knowledge, and the other being rule application for advice-oriented knowledge.

5.3.2.1 FRAME REPRESENTATION OF DOMAIN-ORIENTED KNOWLEDGE

Three top level frames will be used to represent the three types of domain knowledge, namely, statistics, analysis issues, and Atlas contexts. Each top frame is followed by various levels of subframes. The way to represent these types of knowledge can be demonstrated from the two aspects: frame structures, and frame instances; the representation of statistical knowledge is used as an example.

Figure 4 shows a partial structure of statistical methods organized in frames. Only two statistical issues are included, namely, statistical summary and equality test. The tree structure illustrates the hierarchical organization through frames.

Fig. 4 Frame structure of statistical methods - examples

Figure 5 portrays an example of the representation of a statistical method - student t-test. All upline frames are also represented, since they carry information which is shared by the method of t-test.

Fig. 5 Frame representation for statistical knowledge - an example

Frame name: statistical methods

Slot 1 Name: AKO (A Kind Of)
Value: remove statistical methods.

Frame name: equality test

Slot 1 Name: AKO
Value: statistical methods
Slot 2 Name: functional explanation
Value: assert equality between samples
Slot 3 Name: assumptions
Value: random sample

Frame name: two sample test

Slot 1 Name: AKO
Value: equality test
Slot 2 Name: functional explanation
Value: assert equality between two samples

Frame name: two sample mean comparison

Slot 1 Name: AKO
Value: two sample test
Slot 2 Name: functional explanation
Value: assert mean equality between two samples
Slot 3 Name: example
Value: do people in Ontario and Quebec differ in mean income?
Slot 4 Name: null hypothesis H_0
Value: there is no significant difference in means between the two populations
Slot 5 Name: result interpretation
Value: The result, p-value, is the smallest value to reject the null hypothesis H_0 . If p is smaller than the significant level α , then H_0 is rejected, meaning that the two populations significantly differ in means
Slot 6 Name: graphic methods
Value: boxplot; distribution curves for each of the samples.

Frame name: parametric methods for asserting mean equality between two samples

Slot 1 Name: AKO
Value: two sample mean comparison
Slot 2 Name: how it's done

Value: Assume two samples are drawn from two independent populations. Calculate the samples' means and the mean differences between the samples. The sampling distribution of the differences will be normal either when the populations have normal distribution, or when the sample sizes are infinite. With the knowledge of sampling distribution, a significance test can then be conducted.

Slot 3 Name: measurement scale

Value: interval; ratio

Slot 4 Name: assumptions

Value: normal or roughly normal distributions

Slot 5 Name: if needed

Value: assumption satisfaction check (sample size, autocorrelation, normality, sample standard deviation)

Slot 6 Name: analysis procedures

Value: 1) take a random sample of each population using suggested sampling mode

2) conduct normality test

3) calculate standard deviation for each of the samples

4) conduct autocorrelation test; choose 'Normality assumption' if the data sets have normal or roughly normal distributions; otherwise, use 'Randomization assumption'.

5) set null hypothesis H_0 , and significant level α (0.01 - 0.1) at which you can accept the hypothesis

6) input sample sizes, standard deviations of the samples, conclusions of the autocorrelation test, and the normality test

7) use the suggested type of t-test to get p-value

8) get interpretation of the result

9) use graphic methods to display the data sets if necessary

Slot 7 Name: limitations

Value: 1) type I or type II error exists

2) no appropriate statistical method has been found when data are autocorrelated but have a large departure from normal distribution, or the standard deviation of the two groups significantly differ

Frame name: student t-test

Slot 1 Name: AKO

Value: parametric methods for asserting mean equality between two samples

Slot 2 Name: assumptions

Value: 1) independent observations

2) the same sample size

Slot 3 Name: method priority factor

Value: PF = 1

Frame daemons for parametric methods for asserting mean equality between two samples:

IF no autocorrelation, and normal or roughly normal, and $n_1=n_2$, and $s_1=s_2$,
THEN t-test with CF=95.

IF no autocorrelation, and very approximately normal, and $n_1=n_2$, and sample size is very large (infinite),
THEN t-test with CF=90.

IF no autocorrelation, and very approximately normal, and $n1=n2$, and $s1=s2$,
and $n1>30$,
THEN t-test with $CF=80$.

IF no autocorrelation and normal and $n1=n2$ and $s1\neq s2$,
THEN t-test with $CF=70$.

IF no autocorrelation, and non-normal, and $n1\neq n2$
THEN t-test with $CF=-70$.

IF no autocorrelation, and $n1\neq n2$, and $s1\neq s2$,
THEN t-test with $CF=-90$.

IF no autocorrelation, and $n1\neq n2$, and $s1\neq s2$, and $n1\rightarrow\infty$,
and $n2\rightarrow\infty$,
THEN t-test with $CF=70$.

IF normal or roughly normal, and $n1\neq n2$, and $s1\neq s2$, and no autocorrelation,
THEN Modified t-test (1) with $CF=95$.

IF normal or roughly normal, and autocorrelation,
THEN Modified t-test (2) with $CF = 95$.

where $n1$ and $n2$ denote sample sizes of the two samples; $s1$ and $s2$ represent the standard deviations of the two samples; and CF denotes certainty factor.

Several rules will be applied in the frame operations:

(1) Inheritance rule. A property that is shared by all the same level of frames should be stored in their parent frame. Their inheritance relationship is represented by AKO (A Kind Of) pointers. The property will be automatically inherited by all of the children frames of a parent frame. For instance, the assumption of random sample included in the top level frame, equality test, applies to all methods for asserting equality.

(2) Frame nest rule. The Frame A, which is used as a property in Frame B, can be stored in Frame B as a property. In other words, a frame can include frames as their properties.

For example, the histogram is a statistical method; it can also be used as a graphic method in such statistical methods as normality test, and variation comparison.

(3) Daemon activation rule. A frame can include a set of IF-NEEDED daemons in its slots. A daemon can be either a rule or a procedure. A daemon is activated when the value in the slot is needed. For instance, the daemon of assumption satisfaction check is activated when the value of the slot of the frame parametric methods for asserting mean equality between two samples is required.

5.3.2.2 RULE REPRESENTATION FOR ADVICE-ORIENTED KNOWLEDGE

Nine types of advice will be derived from the advice-oriented knowledge. Each of them will be dealt with by a set of rules. Using the derivation of the best statistical method as an example, Figure 6 shows a set of rules used for the selection of statistical methods. The deductive logic conforms to the deductive principle of statistical method selections (Section 4.3.2.2). Thus, following the principles of deductions described in Section 4.3.2.2, as well as the rule format shown by the example, it will be possible to formulate deductive rules for other advisory issues.

Fig. 6 A rule representation example for advice-oriented knowledge

-- rules for selecting a statistical method

IF statistical methods need to be selected,
and statistical issue is SI,
THEN retrieve all statistical methods with SI to form statistical method
hypothesis SMH_i ,
and ask users for analysis purpose.

IF analysis purpose is AP,
THEN retrieve statistical function SF of the hypothetical statistical method
 SMH_i ,
and compare AP with SF.

IF $AP = SF$,
 and analytical themes are $\{THM_i\}$,
 THEN retrieve measurement scale MS required by SMH_i ,
 and retrieve measurement scales $\{MS_i\}$ of $\{THM_i\}$,
 and compare $\{MS_i\}$ with MS .

IF measurement scale $MS = MS_i$,
 THEN check if SMH_i has assumptions.

IF SMH_i has an assumption,
 THEN conduct assumption satisfaction check,
 and check if SMH_i has uncertainty descriptions.

IF SMH_i has uncertainty descriptions
 THEN conduct uncertainty judgement.

IF uncertainty judgement comes to positive conclusions,
 THEN the hypothesis SMH_i is appropriate with certainty factors.

IF uncertainty judgement comes to negative conclusions,
 or uncertainty judgement comes to no conclusion,
 THEN the hypothesis SMH_i is rejected.

IF SMH_i has no uncertainty description,
 and no negative result in the assumption satisfaction check,
 THEN the hypothesis SMH_i is accepted (the method is appropriate).

IF SMH_i has no uncertainty description,
 and there is negative result in assumption satisfaction check,
 THEN the hypothesis SMH_i is rejected.

IF SMH_i 's measurement scale is not equal to the measurement scale of each
 analytical theme ($MS \neq MS_i$),
 THEN the hypothesis SMH_i is rejected.

IF analysis purpose is not equal to SMH_i 's statistical function ($AP \neq SF$),
 THEN the hypothesis SMH_i is rejected.

Two points must be stressed. First, the set of rules in statistical method selections by EASSA are universal. This means that for all the statistical issues, and for all the statistical methods selected, the inferential mechanism is the same; the difference lies in the different instantiations of variables involved in the rules. Second, the rules seem to be addressing one hypothetical statistical method SMH_i , however, logic programming languages or expert system shells usually provide the function of exhausted instantiation (non-

deterministic predicate), which allows the rules to be applied to all candidates. Therefore, this set of rules will be able to accomplish statistical method selections, on the basis of the domain knowledge. Essentially, they are an abstraction of inferential logic applied by human experts in statistical method selections.

5.4 SUMMARY

Knowledge formalism for EASSA has been presented in this chapter. This was carried out first at a conceptual level, and then at a logical level. The conceptual formalism has produced a set of knowledge models which provide EASSA's three layer knowledge with mathematical descriptions. The logical formalism has dealt with the design of knowledge representation methods. It is intended that the domain knowledge, except for sampling schemes, be represented in frames; the sampling schemes, advice-oriented knowledge, and task-oriented knowledge, be represented in rules or Horn clauses; and meta knowledge be represented in either frames or rules/Horn clauses, depending upon whether it is descriptive or deductive.

6. SYSTEM DESIGN AND BEHAVIOR DEMONSTRATION

6.1 INTRODUCTION

The descriptions of the system's functionality and the knowledge schemas, as well as their structure, have presented a conceptual design for the advisory system. This chapter provides a detailed design of the system at the logical level. Firstly, the logical structure of the system is illustrated. Then, the design is described from the perspective of logical reasoning. Although the physical implementation of the system will depend on the programming tool to be used (e.g. Prolog, or an expert system shell), the logical design is relatively independent, hence applicable to any language. Moreover, in order to present a clear picture of the advisory system, the system's behavior is demonstrated through two cases.

6.2 SYSTEM STRUCTURE OF EASSA

The logical structure of the system is discussed at two levels: the top level and the second level. Several components involved in the top level structure are detailed in the second level illustrations.

6.2.1 TOP LEVEL STRUCTURE

The advisory system will consist of five basic components: the human machine interface, the inference engine, the knowledge base, the knowledge acquisition mechanism, and the management module for frames, as shown in Figure 7.

Fig. 7 Top level structure of EASSA

The first three components are the basis of an expert system. The *human machine interface* is responsible for the coordination tasks needed between users and the system. It essentially refers to the task-oriented knowledge described in the conceptual design (Section 4.3.3). Accepting the users' requirements, the system then invokes the *inference engine*, in order to carry out particular tasks. It makes judgements on the possibility of conducting an analysis, and then derives the various types of advice necessary. The derivations correspond to the layer of advice-oriented knowledge discussed in the conceptual system (Section 4.3.2). Further, the inference engine conducts reasoning by appropriately retrieving the needed knowledge in the knowledge base. Thus, the *knowledge base* refers to the layer of domain knowledge (Section 4.3.1).

The last two components are included in order to meet the functional requirements of the conceptual system. The *knowledge acquisition module* is devoted to dynamic acquisition of new knowledge when the system is in operation. As indicated in Section 3.7, the initial

system is intended to automatically acquire two types of knowledge: Atlas contexts and analysis issues. The acquisition mechanism is separated from the inference engine, since the functionality of knowledge auto-acquisition may be expanded in further developments of the system. The *frame management module* is needed since frames are an important knowledge representation method for the advisory system.

6.2.2 SECOND LEVEL STRUCTURE

6.2.2.1 INTERFACE MODULE

The *human machine interface* module, as shown in Figure 8, will consist of three components: an advisory mode server, an explanation server, and a knowledge acquisition server.

The *advisory mode server* is intended to incorporate users' information and the system's reasoning behavior in various advisory modes (as discussed in Section 3.4 and 4.3.3). This is carried out by utilizing the analysis issue generator, the map context generator, and the advice server. The *analysis issue generator* involves several processes in relation to the generation of analysis issues. Firstly, it generates the list of statistical issues, and displays it for the users. After accepting the users' choice of statistical issues, it proceeds with two alternatives. The first one, the issue-dependent mode, is to generate the list of analytical themes, display it to users, and incorporate the users selected themes into a final analysis issue. The other, referred to as the issue-independent mode, prompts users for analytical themes, and uses the themes provided to finalize an analysis issue. The *map context generator* deals with the generation of analytical themes, as well as the contexts of analytical maps. The Atlas-dependent mode generates map contexts, including the measurement scales of variables, and the spatial dimensions and spatial distributions of analytical themes,

by using the inference engine of the system. For the Atlas-independent mode, the map context generator prompts the users for related map information, and accepts it for further usage. The *advice server* uses the inference engine to derive various types of advice, and displays them to users.

Fig. 8 Interface design of EASSA

The *explanation server* is responsible for giving explanations upon user request. As discussed in Section 3.6, explanations will be made for such questions as 'WHAT', 'WHY' and 'HOW'. The explanation server checks if the user is requesting explanations (*user request receiver*), recognizes which type of explanations are needed, uses the inference engine to generate the explanations (*explanation generator*), and displays it to the user (*explanation displayer*).

The *knowledge acquisition server* deals with issues related to automatic knowledge acquisition at the interface. Specifically, it checks which analysis has potential for knowledge acquisition, and if the user intends to store it into the knowledge base. The two types of knowledge, analysis issues, and map contexts, are dealt with in the acquisition process, within the *analysis issue server* and the *Atlas context server* respectively.

6.2.2.2 INFERENCE ENGINE MODULE

Generally, the *inference engine* of an expert system carries out such tasks as forward or backward chaining, uncertainty handling, and giving explanations. Considering EASSA's particularities, the inference engine of the system will take over several processes. First, it will provide a forward chaining mechanism based on the backward chaining capability, since the potential programming tools (Prolog or expert system shells) to be considered for the system often have backward chaining mechanisms but not forward chaining mechanisms. Secondly, the inference engine is responsible for making use of the knowledge in the knowledge base, in terms of the inferential requirements made by the user interface. This actually corresponds to the inferential knowledge discussed in the knowledge structure in the previous chapter. Thirdly, it will deal with the uncertainty problems encountered in statistical method selections, through the synthesis of uncertainty factors and priority factors of statistical methods. Fourthly, it will make decisions on whether a bivariate or multivariate analysis can be conducted based on the temporal differences of the surveyed data. Finally, a multi-level explanation mechanism will be incorporated in it, in order to have explanations in different levels of detail. The last three processes refer essentially to meta knowledge (Section 4.3.4).

Thus, the inference engine module will consist of five components: the *forward chaining* sub-module, the *advice-oriented reasoning* module, the *uncertainty & priority solver*, the

temporal difference solver, and the *explanation mechanism*. Figure 9 shows this basic structure.

Fig. 9 Components of EASSA's inference engine

6.2.2.3 KNOWLEDGE BASE MODULE

The *knowledge base* of the system includes the domain knowledge, referred to as the domain knowledge layer in the knowledge structure. It consists of four sets of domain knowledge, namely, the *analysis issues*, the *statistical knowledge*, the *sampling knowledge*, and the *Atlas contexts* (as presented in Figure 10). The analysis issues, statistical knowledge, and Atlas contexts are represented in frames, while the sampling knowledge is in rules.

6.2.2.4 FRAME MANAGEMENT MODULE

The *frame management module* deals with such frame operations as frame-retrieving, frame-removing, frame-adding, and frame-updating (cf. Figure 11). The operation of *frame-retrieving* finds a frame in terms of the frame name, and obtains all attributes of the

Fig. 10 Components of EASSA's knowledge module

frame through inheritances built in the frame. The *frame-removing* operation deletes a frame according to the frame name, and removes all attributes related to the frame through inheritances. The *frame-adding* operation adds a frame upon request. The operation involves the instantiation of a frame, i.e. filling a frame, and establishing inheritances in relation to the frame. This is essential for the automatic acquisition of frame instances, and

Fig. 11 Components of frame management module

thus, of importance to the advisory system. The *frame-updating* operation is used to replace original frame instances with new frame instances. This can be accomplished by combining the frame-removing and the frame-adding operations.

6.3 REASONING PROCEDURES OF EASSA

The reasoning procedures are discussed in an attempt to present the design from the perspective of inferential flow. The top level reasoning shows the whole picture of the inferential procedures of the system. The major steps involved in the top level processes are discussed in detail in the second level reasoning.

6.3.1 TOP LEVEL REASONING

Figure 12 illustrates the reasoning flow of the system at the top level. All components included in the system structure are involved in the processes. The second column from the left represents the knowledge base, which includes four types of domain knowledge, as discussed above. The third column from the left represents the inference mechanism of the system. It is presented with links to the various types of advice, and, the chaining mechanism is implied in the inferences. The temporal difference solver, the uncertainty & priority solver, and the explanation mechanism also belong to the inference engine. The right column refers to the user interface, where communications between users and the advisor are carried out. The knowledge acquisition module is utilized in acquiring map contexts and analysis issues, while the frame management module supports frame operations.

Thus, once being invoked, the system starts to deal with the derivation of analysis issues that can be analysed based on either Atlas maps or user provided maps, and either existing

Fig. 12 EASSA's top level reasoning chart

analysis issues or user proposed issues. This is carried out by using analysis issue knowledge, map context knowledge, and statistical knowledge. If the analytical maps and/or analysis issues are selected from the Electronic Atlas and/or the EASSA advisor, the inference engine retrieves the corresponding knowledge of Atlas contexts and analysis issues, as well as statistical knowledge, in order to derive advice. Otherwise, the advisor incorporates user provided information into the reasoning, and, afterwards, carries out a learning procedure, upon request, by using the knowledge acquisition module to store the user provided knowledge into the corresponding frame bases. If the analysis involves more than one variable, an additional problem is solved during the process, namely, the temporal differences of the surveyed data sets. This is supported by the Temporal Difference Solver. The reasoning result is displayed for the user at the interface.

If an analysis is recommended, the system then determines the statistical method suitable for the analysis. The analysis issue, the map context, and the statistical knowledge are taken into account in the inference. The uncertainty problem, which may be encountered in the method selection, is handled through the uncertainty and the priority solver. The derived statistical method is then displayed on the screen.

The next step is to select a set of sampling modes (one for each analytical theme) that is appropriate for the analysis. This inferential process is carried out by considering all the four types of knowledge, namely, the analysis issue, the map contexts, the statistical method being used, and the sampling knowledge. The result is recommended to the user through the interface.

The last two steps deal with the inferences for other types of advice, that is the hypothesis, the assumptions, the result interpretation, the graphic methods, the limitations of the analysis, and especially, the analysis procedure. The map contexts and statistical

knowledge are considered for the first five issues, while these two types of knowledge, as well as the selected sampling approach, are taken into account in the derivation of analysis procedures. Again, the results are displayed on the screen.

A whole consultation is thus completed. In all these stages, explanations can be requested either for WHY, HOW, or WHAT (Section 4.3.4.3). This is supported by the explanation mechanism. This mechanism is partially supported by the knowledge base.

6.3.2 SECOND LEVEL REASONING

6.3.2.1 REASONING FOR ANALYSIS ISSUES

The data driven flow for reasoning analysis issues is shown in Figure 13. As mentioned in the previous section, this inference is carried out through the inference mechanism, the three types of knowledge (analysis issues, maps contexts, statistical knowledge), as well as the knowledge acquisition module. Information may also be needed from users. For simplicity, only reasoning flows are presented, while the functional components are left out (this also applies to the other reasoning charts).

Upon receiving a request for analysis issue recommendation from the user interface, the inference engine attempts to generate an analysis issue, either by retrieving it from EASSA's analysis issue frame base, or by accepting it from users. This is followed by a test, completed through three separate channels, to determine whether the hypothesis is true.

The first channel A-D (Atlas-Dependent), as shown on the left side of the figure, deals with the case where the user attempts to conduct an analysis based on the maps in the

Fig. 13 Reasoning for analysis issues

Electronic Atlas. Following this channel, the analysis issue is separated into a statistical issue and a set of analytical themes. The system, on the one hand, retrieves measurement scales {MSj} required by the statistical methods related to the statistical issue. On the other hand, it searches the analytical themes from the Atlas context frame base, and checks if the information in the EA is sufficient for the analysis. If yes, it retrieves the measurement scales {MSi} of the themes from the Atlas context frame base, and compares {MSi} with {MSj}. If there exists an MSj which is the same as or is more generalized than the measurement scales of {MSi}, then the inference turns to temporal difference reasoning. If more than one variable is involved in the analysis, the reasoning mechanism makes judgement on whether the differences in the surveyed dates of the data are tolerable for the analysis. If yes, the analysis hypothesis is proven to be true, and the system recommends it to the users. If the analysis issue is proposed by the users and the system is requested to store the issue into EASSA's knowledge base, then an automatic knowledge acquisition procedure is carried out. If the temporal reasoning or the measurement scale comparison comes to a negative conclusion, the inference engine exits the current channel, and asks the user whether one is going to try other channels or give up this hypothesis. The reasoning is turned to channel 2 or channel 3 upon the first option. If the hypothesis is given up, the system tries to generate another hypothesis, and the procedure is repeated.

The second channel A-S (Atlas-Semi-dependent), shown in the central portion of the diagram, deals with the case where the user is going to use maps found in the EA, and provide other map characteristics through the interface. The major difference between this channel and the first channel is that the system searches the analytical themes from the Atlas context frame base first, and then prompts the user to input the themes and data characteristics of the other maps through the interface. If the themes provided are not sufficient, the analysis is not feasible, and the system attempts to find another hypothesis. Another difference of the A-S channel from the A-D channel is that, in the case where an

analysis hypothesis proves to be true, knowledge acquisition of map contexts will be carried out upon request.

The right side of the figure shows the third channel A-I (Atlas-Independent), in which all maps for analysis are supplied by the user. The difference between this channel and the A-S channel is that the system does not access the Atlas context frame base, but, rather, it accepts all analytical themes, as well as the map characteristics from the user.

6.3.2.2 REASONING FOR STATISTICAL METHODS

Once an analysis issue has been selected, the system then considers the statistical method. Figure 14 illustrates the inferential procedure. Firstly, following the separation of the analysis issue into a statistical issue and a set of analytical themes made in the previous reasoning, the system generates a statistical method hypothesis with the statistical function SF, from the statistical knowledge frame base. It then prompts the user for an analysis purpose AP. This is followed by a comparison between AP and SF. If they are equal, the system retrieves the measurement scale MS of the statistical method, as well as {MS_j} of the analytical themes. If MS is the same as or is more generalized than {MS_j}, then the inference proceeds with the next step; otherwise, it returns to the generation of another hypothesis. The next inference deals with the assumption check, and the uncertainty judgement based on the assumption satisfactions. Moreover, if both certainty factors and priority factors are present, they should be taken into account, in order to derive an optimal solution. The reasoning terminates with a suggestion on the statistical method hypothesis.

Fig. 14 Reasoning for statistical methods

6.3.2.3 REASONING FOR SAMPLING MODES

After the statistical method is selected, the system turns to determination of sampling modes suitable for analysis. Figure 15 presents the basic reasoning process. First of all, the system checks which advisory mode is adopted in the analysis. Along the Atlas-dependent (A-D) channel, it obtains the spatial dimensions and spatial distributions (Section 4.3.1.3) of the analytical themes from the Atlas context frame base. Yet, the A-S and A-I channels obtain this information either by partially retrieving from the Atlas context frame base and partially accepting from users, or by completely acquiring from users. On the basis of this information, the system then calls the rules from the sampling knowledge base, and determines the sampling base for the analysis. Moreover, it utilizes the related rules in the sampling knowledge base, as well as the selected statistical method, to infer the sampling mode(s) for the analytical theme(s). Finally, it recommends a set of inferred sampling mode(s) to the user.

6.3.2.4 REASONING FOR OTHER ADVISORY ISSUES

The inferences for the remaining advisory issues are straightforward, as shown in Figure 16. The statistical method used and the analytical themes involved are the two major knowledge sources in the generation of such advice as hypotheses, assumptions, analysis procedures, result interpretations, graphic methods, and limitations of the analysis. Based on these sources, the system retrieves the knowledge, with respect to each of the advisory issues, from the statistical knowledge frame base, and combines it with the analytical themes. For the analysis procedure, one more factor needs to be considered, namely, the sampling mode(s) recommended for the analysis. Various types of advice are thus generated for displays.

Fig. 15 Reasoning for sampling modes

Fig. 16 Reasoning for hypothesis, assumptions, analysis procedures, result interpretations, graphic methods, and limitations

6.4 DEMONSTRATION OF SYSTEM'S BEHAVIOR

The design of the system's structure and reasoning procedure provides a clear picture of the internal mechanism of the system. Its external behaviors are demonstrated by emulating consultations. Two cases have been selected for this purpose, as shown in Appendix 3.

The first case tests if life expectancy is spatially autocorrelated. The Issue-Atlas-dependent advisory mode is selected for this test; both the analysis issue and the analytical themes are provided by the advisor in this analysis. The presentation of the analysis issue is actually carried out through two steps: the presentation of the statistical issue, and the presentation of the analytical themes that are suitable for the analysis issue, as shown in screens [3] - [6]. The user makes choices following every prompt, and the advisor gives step by step advice. Whenever a decision is made, the system is ready to answer HOW it is done, as shown in screens [8] and [8'], as well as screens [9] and [9']. The question WHAT can be answered by the advisor for the concepts that may not be understandable for general users; the explanation of measurement scales (screen [8'']) is an example.

The second case demonstrates the dialog between the user and the advisor when the user suggests analysis issues, and/or use his/her own maps for analysis. The equality test on the average income per capita between two regions is used as an example. The analysis issue is proposed by the user in the following manners: first, the user selects a statistical issue following request by the advisor, and then the user inputs the themes for analysis, the spatial dimensions and distributions of the themes, the measurement scales used to measure the themes in the maps, and the temporal attributes. The advisor presents serial advice (screens [6] to [14]), after the judgement has been made on the feasibility of the analysis. Some preprocessing (screen [7]), such as normality test and autocorrelation test, is needed for this case, according to the requirements of equality test.

The structure of the screens presented in the demonstration are simplified for convenience. In the working system, it should be constructed using multiple windows, with respect to advice, messages, explanations, questions, etc. Pop-up or pull down menus can be used to facilitate users accessing various types of advice.

6.5 SUMMARY

This chapter has presented the design of the advisory system, from system structures, reasoning procedures, and probable behavior. The system will be composed of three basic components: the user interface, the inference engine, and the knowledge base. Additionally, it will include a frame manager to deal with frame base management, and a knowledge acquisition module to carry out the acquisition of analysis issues and map contexts from users. The reasoning of the system is advice oriented. According to specified analysis issues, the system derives all related advice by using appropriate domain knowledge. The emulated consultation cases demonstrate the behaviors of the system in consultations.

7. SUMMARY, CONCLUSIONS, AND FURTHER DEVELOPMENT

7.1 SUMMARY

This research has accomplished major tasks of knowledge engineering in the development of EASSA, the expert advisory system for statistical analysis for electronic mapping and GISs, in particular for the Electronic Atlas of Canada. Specifically, EASSA's framework has been developed; the core knowledge required for the system has been acquired; major types of knowledge have been formalized in models; knowledge representation methods have been described; and system design has been accomplished.

System Framework. EASSA will act as an expert advisor to give users advice on the utilization of statistical tools in electronic mapping. The advice involves significant and feasible analysis issues, the statistical methods suitable for the analyses, the sampling approach to data preparation for the analysis, the analytical procedure of the analyses, the interpretations of the analytical results, and other relevant issues such as hypotheses, assumptions, graphic representation methods, and limitations of the analyses. Multiple advisory modes will be available for users, including the Issue-Atlas-dependent mode and the Issue-Atlas-independent mode. The Issue-Atlas-dependent mode performs a consultation based on the analysis issues suggested by the system and the maps found in the EA, while for the Issue-Atlas-independent mode, both analysis issues and themes of maps are provided by the users. Between these two modes are four other advisory modes that deal with the combinations of different information sources (users or EASSA) for analysis issues and map themes. Thus, the system not only serves as an advisor for statistical analysis in the EA environment, but can also help users who are working on general GIS systems, or on other data analysis tasks. To be an expert advisor, the system

will not only give these types of advice, but will also have self-explanation capabilities, justifying, upon user request, the manner in which advice is derived, the reason why an input is necessary, and the meaning of a concept. In addition, it will have basic learning capability, dealing with the automatic acquisition of analysis issues and map contexts during consultation. This capability allows these two types of domain knowledge to be acquired after the system is developed.

Knowledge Structure. In order to flexibly support the derivation of advice, a three layer knowledge structure is designed. The three layers are structured in a hierarchy, with meta knowledge being used in any layer. The bottom layer, referred to as the domain knowledge, includes statistical knowledge, analysis issues, sampling knowledge, and Atlas contexts. The intermediate layer, called the advice-oriented knowledge layer, deals with the inferences for various types of advice, by appropriately using the domain knowledge. On the top layer is the task-oriented knowledge, which incorporates user information and the system's behavior in order to fulfil specific advisory tasks. The meta knowledge deals with such problems as temporal difference between multiple variables, optimal selections of statistical methods, and explanations.

Knowledge Acquisition. Knowledge acquisition is the key issue of this thesis. Since a large amount of knowledge is needed for the intended system, the focus of this thesis was on the generation of statistical knowledge, which is the core knowledge needed by the system. A basic set of statistical knowledge has been formulated, covering statistics ranging from descriptive to inferential, and from univariate and bivariate to multivariate. The emphasis is on conventional statistics, but spatial contexts have been taken into account, and several primary spatial statistics have been included. In total, about forty statistical methods are involved. For each of these methods, fourteen items of knowledge are presented, indicating the principles of statistics, such as functions, assumptions,

analysis procedures, and limitations. Effort was also made to generate analysis issues and sampling knowledge. The former aims to provide the users with analysis issues that are of common interest, and make sense in geography. The latter deals with the selection of spatial sampling techniques in a GIS environment. They both remain to be refined.

Knowledge Formalism. The three layer knowledge has been modeled in the thesis, in order to provide precise, concise, and formal descriptions for the knowledge. The models also provide clear semantics for the automatic acquisition of knowledge. In addition to the conceptual level descriptions, logical representation methods have also been designed. Basically, the frame method will be primarily used to represent such types of knowledge as statistics, analysis issues, and map contexts. The sampling knowledge, as well as the advice-oriented and task-oriented knowledge will be represented in rules (if using an expert system shell for implementation) or Horn clauses (if using Prolog). The meta knowledge will be represented partially in frames and partially in rules, depending on the nature of the knowledge.

System Design. A design of the advisory system has been presented in the thesis. This is presented from three facets: the logical structure, the reasoning flows, and the system behavior in consultation. Logically, the system will be composed of three major components (a user interface, an inference engine, and a knowledge base), and two additional elements (a frame management module, and a knowledge acquisition mechanism). The second level design for the major components has been further illustrated. The reasoning flows of the system, a data-driven design, have presented the inferential processes, at the top level and the second level, for various types of advice to be given by the advisor. Both the structure and the reasoning flows address the internal mechanism of the system; the system behavior, however, has demonstrated the system from an external view. Such a multiple facet design will make it easier to implement the

advisory system.

Methodology for Knowledge Engineering. The way to accomplish a knowledge engineering task, to a large extent, depends on the nature of the knowledge domain (Kidd, 1987). A systematic knowledge engineering methodology was designed for the development of EASSA. First, the domain problems were analyzed and identified, so as to formulate the outline of the expert system. Then, the concept of problem-solving space was proposed, and used to decompose the complex problems into individual ones. Based on the decomposition, the knowledge schemas for the system was formulated. The schemas were then followed to extract and develop the expertise involved in the system. The expertise was further validated by interviewing domain experts, through case studies, and by extended reviews of the literature.

7.2 CONCLUSIONS

This research is an exploration of the development of an expert advisor for guiding statistical analysis in a spatial information system. Several conclusions can be drawn from the work thus far.

1. *Applying the expert system technology to the spatial statistics domain is innovative but feasible.* With the development of GIS technology, statistical analysis and expert system techniques are drawing increasing interest from GIS developers and users. Nevertheless, no system has been developed that incorporates these two approaches in facilitating users' utilization of statistics in a spatial information system environment. This research has conceptually proven that this integration is feasible and worthwhile.

2. *The research has provided a theoretical basis for the strategic development of a statistical*

advisory system.

Methodology for Knowledge Engineering. The way to accomplish a knowledge engineering task, to a large extent, depends on the nature of the knowledge domain (Kidd, 1987). A systematic knowledge engineering methodology was designed for the development of EASSA. First, the domain problems were analyzed and identified, so as to formulate the outline of the expert system. Then, the concept of problem-solving space was proposed, and used to decompose the complex problems into individual ones. Based on the decomposition, the knowledge schemas for the system was formulated. The schemas were then followed to extract and develop the expertise involved in the system. The expertise was further validated by interviewing domain experts, through case studies, and by extended reviews of the literature.

7.2 CONCLUSIONS

This research is an exploration of the development of an expert advisor for guiding statistical analysis in a spatial information system. Several conclusions can be drawn from the work thus far.

1. *Applying the expert system technology to the spatial statistics domain is innovative but feasible.* With the development of GIS technology, statistical analysis and expert system techniques are drawing increasing interest from GIS developers and users. Nevertheless, no system has been developed that incorporates these two approaches in facilitating users' utilization of statistics in a spatial information system environment. This research has conceptually proven that this integration is feasible and worthwhile.

2. *The research has provided a theoretical basis for the strategic development of a statistical*

expert system in GIS and electronic mapping environments, in particular in the Electronic Atlas of Canada. The development of such a system can start with a small system, and then consider expansion, or it can begin with a thorough design, and then have incremental development. The latter approach is being used in the development of EASSA, and related issues have been fully investigated. This will allow the development of a strategic plan for system development, and the development of a system that possesses the potential and flexibility needed for systematic expansion of the system's capabilities.

3. The hierarchical knowledge structure provides a clear reasoning framework for the advisory system, and allows the advisor to access knowledge flexibly. Statistical analysis in a spatial context is thought to be complex, since a range of spatially related factors have to be taken into account. The hierarchical knowledge structure presents the conceptual framework involved in human experts' reasoning behavior. This can enhance the flexibility of the system's accessing of domain knowledge, since the inferential knowledge is separated from the domain knowledge, and the various types of domain knowledge are organized independently. Essentially, it is this characteristic of the system that allows for the easy addition of the knowledge acquisition capability to the system.

4. The partial knowledge acquisition mechanism will facilitate the development and the use of the advisory system. The acquisition of analysis issues and Atlas contexts requires a large amount of work. Putting this task after the implementation of the system can significantly speed up system development, and this is made possible by the knowledge acquisition mechanism designed for the system. The mechanism also makes the use of the system more flexible, since users can develop their own knowledge base for specific uses.

5. The statistical knowledge set developed for EASSA can also be used in other related consultations. Many statistical text books are available. However, none of them covers the

whole range of commonly used statistical methods. It is especially difficult to find a text book that provides thorough discussion on spatially-related issues and limitations, and covers a full range of basic statistical techniques. This research has produced concise and thorough collections of commonly used statistical methods in a spatial context. Emphasis was placed on the integration of the techniques that are recently developed to deal with spatial issues into classical statistics. Other issues such as limitations and certainty of method suitabilities have also been addressed.

6. The sampling knowledge developed in this research provides a preliminary guide for the development of a spatial query tool box in a GIS or electronic mapping environment. Geographical phenomena have different types of spatial distributions. They are modeled to different spatial dimensions (or cells) in GISs or EA systems. Appropriate spatial query techniques are necessary, when an analysis involves several variables which have different spatial distributions in the real world and varied spatial dimensions in the computer system. This research has summarized the query functions needed for conducting various statistical analyses in a digital environment. It will be helpful for users to select suitable approaches for statistical analysis. It will also be useful for GIS developers to develop thorough spatial query functions.

7.3 FURTHER DEVELOPMENT

The development of EASSA involves a large amount of work. This thesis has addressed a major part of the knowledge engineering task; further development is necessary in order to develop a practical system.

System Implementation. The first task of further development will be to implement a prototype system of EASSA, by using the knowledge already acquired in this research,

following the logical design of the system presented in the thesis, and selecting a proper programming tool. Before prototype development, however, an investigation needs to be carried out, in order to define the priorities of various statistical methods to national mapping and to other EA users. Based on this, an incremental development approach should be applied. Namely, the methods ranked with high priorities for national mapping should be included in the first prototype system, so that the system can quickly be used in national mapping. This can be followed by an expansion of the system to one that contains statistical methods suitable for other EA users. A prototype validation should also be carried out after the prototype is used in application, and this should be followed by the development of a production system.

Further Knowledge Engineering Issues. This thesis focuses on statistical knowledge. A prototype with this knowledge, as well as the preliminary sampling knowledge, will be able to work in the Issue-Atlas-independent mode. Nevertheless, for EA users, the contexts of EA maps are especially useful in consultations, since, in most cases, the users should not be expected to provide maps for analysis. Thus, the first knowledge engineering task, after prototyping, will be the context formulation for the maps that have been or will be included in the Electronic Atlas. The analysis issues, which have been formulated in this research, are important information for EA map interpreters. A refinement on this knowledge should be carried out, and the knowledge should be input into the prototype, so as to enhance the system's capability. Another important issue of knowledge engineering will be the development of meta knowledge (cf. Section 4.3.4). In addition, knowledge should be updated, and, if desired, expanded.

Analysis Environment. EASSA will give advice on statistical analysis; however, an analysis environment is needed to carry out the analysis, as indicated in Section 3.2. Statistical functions constitute the first component of the environment, which can be either

an existing package, adapted for the EA analysis environment, or a module included in the EA system. Sampling functions are the second component of the analysis environment. This component should be implemented in the Electronic Atlas system as a set of advanced spatial query functions. Only after these query functions have been made available can the sampling knowledge of the system be validated, and the statistical analysis in the EA environment become possible. Graphic functions for displaying data or analyzed results are the last component that is required for EASSA related analysis. It is usually included in statistical software; therefore, how it will be implemented in the EA system largely depends on the way the statistical functions are developed.

Future evolution of EASSA. After preliminary development of the advisor, efforts can be made in two directions. One is to develop a system that can act as an statistical expert directly. That is, the system will accept users' requirement for analysis, as well as related information, and then take over the analysis task, with a minimum requirement of interactions by users. The other direction is to include more spatial statistical methods, with the development of the spatial statistics domain itself. The final evolution of the system should be towards an advanced decision support system, oriented to particular decision making domains, that could make wise recommendations for complex problem-solving tasks, on the basis of spatial statistical analysis.

7.4 FINAL COMMENT

Spatially based statistical analysis is one of the inevitable evolutions in the development of geographical information systems. This research is an exploration of this evolution. Although the work done so far has shown the feasibility of using statistics in electronic mapping, a large amount of work is needed to develop a practical expert advisory system which will bring the statistical methods closer to the user community.

REFERENCES USED IN STATISTICAL KNOWLEDGE FORMULATION:

Cliff, A. D. & J. K. Ord, 1981, Spatial Processes - Models and Applications. Pion Ltd.

Dent, Borden D., 1985, Principles of Thematic Map Design. Addison-Wesley Publishing Company.

Goodchild, M. F., 1986, Spatial Autocorrelation. Geo Books, Norwich.

Griffith, D. A., 1987, Spatial Autocorrelation - A Primer. Commercial Printing Inc.

Hammond R. and P. S. McCullagh, 1980, Quantitative Techniques in Geography. Second edition, Clarendon Press-Oxford.

Devore, J, and R. Peck, 1986, Statistics. West Publication Company.

Lewis, P., 1977, Maps and Statistics. Methuen & Co Ltd.

Mather, P. M., 1976, Computational Methods of Multivariate Analysis in Physical Geography. John Wiley & Sons.

Myers, R. H., 1990, Classical and Modern Regression with Applications. 2nd Edition, PWS-KENT Publishing Company.

Norusis, M. J., 1985, SPSS - Advanced Statistics Guide. McGraw-Hill Book Company.

Ripley, B. D., 1981, Spatial Statistics. John Wiley & Sons.

Velleman P. F., 1989, Data Desk: Statistics Guide. ODESTA Corporation.

Williams R. B. G., 1984, Introduction to Statistics for Geography and Earth Scientists. Macmillan Publishers Ltd.

Zhang K. and R. Guo, 1986, Mathematical Methods for Thematic Mapping (in Chinese). Wuhan Technical University of Surveying and Mapping, Wuhan, China.

REFERENCES CITED

- Aiello, Luigia, C. Cecchi, and D. Sartini, 1986, "Representation and Use of Meta knowledge". *Proceedings of the IEEE*, Vol. 74, No. 10, pp. 1304-1321.
- Arnberg, W. and B. Justusson, 1989, Concepts of the PC-version of the Atlas of Sweden. Stockholm.
- Bailey D. and Thompson D., 1990, "How to Develop Neural-Network Applications". *AI Expert*, Vol. 5, No. 6, pp. 38-47.
- Bailey, T. C., 1990, "A Geographical Spreadsheet for Visual Interactive Spatial Analysis". *Proceedings of the First European Conference on Geographical Information Systems*, Amsterdam, The Netherlands, pp. 21-29.
- Burrough, P. A., 1990, "Preface". *Int. J. of Geographical Information systems*, Vol. 4, No. 3, pp. 221-223.
- Cardinal Consulting, 1990, Electronic Atlas Mark III - GIS functions. Prepared for the Geographical Services Division, Energy, Mines, and Resources Canada.
- Chung, C.F., A. G. Fabbri, and R. Sinding-Larsen (ed.), 1988., "Quantitative Analysis of Mineral and Energy Resources". NATO ASI Series, Series C: *Mathematical and Physical Sciences*, Vol. 223, D. Reidel Publishing Company.
- Cliff A. D., and J. K. Ord, 1973, Spatial Autocorrelation. Pion.
- Cook, D. G. and S. J. Pocock, 1983, "Multiple Regression in Geographical Mortality Studies, with Allowance for Spatial Correlated Errors". *Biometrics* 39, pp. 361-371.
- Currie, S., 1989, "Data Classifications for Choropleth Mapping". M.A. Thesis, Department of Geography, University of Ottawa, Ottawa, Canada.
- Dambroise, E. and P. Massotte, 1986, "MUSE: An Expert System in Statistics". *COMPSTAT 86*, Rizzi ed. Heidelberg: Physica.

- Darwish, A. M. and A. K. Jain, 1988, "A Rule Based Approach for Visual Pattern Inspection". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 10, No. 1, pp. 56-68.
- Eastman, J. R., 1990, IDRISI: A Grid-based Geographic Analysis System. Clark University, Graduate School of Geography.
- Egenhofer, M. J. and A. U. Frank, 1990, "LOBSTER: Combining AI and Database Techniques for GIS". *PE & RS*, Vol. LVI, No. 6, pp 919 - 926.
- Engle, S. W. 1985, "CERBERUS release notes", Version 1.0, NASA Ames Research Centre, California.
- Freeman, H. and J. Ann, 1984, "AUTONAP - An Expert System for Automatic Map Name Placement". *Proceedings of First Int. Symp. on Spatial Data Handling*, pp. 544-571.
- Frost, R. A., 1986, Introduction to Knowledge Base Systems. William Collins Sons & Co. Ltd.
- Gatrell, A. and C. Dunn, 1990, "GIS in Epidemiological Research: Analysis of Cancer of Larynx in North West England". *Proceedings of the First European Conference on Geographical Information Systems*, Amsterdam, The Netherlands, pp. 346-355.
- Getis, A. and B. Boots, 1978, Models of Spatial Processes. Cambridge University Press.
- Goodchild, M. F., 1987, "A Spatial Analytical Perspective on Geographical Information Systems". *Int. J. of Geographical Information Systems*, Vol. 1, No. 4, pp. 327-334.
- Goodchild, M. F., 1990, "Geographic Information Systems and Cartography". *Cartography*, Vol. 19, No. 1, pp. 1-13.
- Goodenough, D. G., G. W. Plunkett, and J. Zelek, 1987, "An Expert System for Remote Sensing". *IEEE Transaction on Geoscience and Remote Sensing*, Vol. Ge-25, No. 3, pp. 349-358.

- Goldberg, M. M. Alvo, and G. Karam, 1984, "The analysis of Landsat Imagery Using an Expert System: Forestry Applications". *Proceedings of International Symposium on Computer-Assisted Cartography, AUTOCARTO 6*, Ottawa, pp. 493-503.
- Grimaldi, R. P., 1985, *Discrete and Combinational Mathematics*. Addison-Wesley Publishing Company.
- Groot, R., 1990, "Diversification of the Delivery of Geographic Information About Canada". *Proceedings, GIS for 1990s*, Ottawa, Canada, CISM, pp. 549-556.
- Haslett, J., G. Wills, and A. Unwin, 1990, "SPIDER - an Interactive Statistical Tool for the Analysis of Spatial Distributed Data". *Int. J. Geographical Information System*, Vol.4, No.3, pp. 285-292.
- Havens, W. S., 1985, "A Theory of Schema Labelling". *Computational Intelligence*, Vol. 1, No. 3-4, pp. 127-139.
- Huang, S., 1991, "An Overview and Perspectives on Expert Systems in GIS and Remote Sensing". Graduate seminar paper, Dept. of Geography, Univ. of Ottawa, Ottawa, Canada.
- Jankowski, P. and T. Nyerges, 1989, "A Knowledge Base for Map Projection Selection". *The American Cartographer*, Vol. 16, No. 1, pp. 29-38.
- Kehris, E., 1990, "Geographical Modeling Environment Built Around ARC/INFO". *Proceedings of the First European Conference on Geographical Information Systems*, Amsterdam, The Netherlands, pp. 556-564.
- Kidd, A. (ed.), 1987, *Knowledge Acquisition for Expert System*. Plenum Press.
- Kilchenmann, A., 1990, "Geographical Information Systems and Map Analysis using Factor Analysis and Character Analysis". *Proceedings of the First European Conference on Geographical Information Systems*, Amsterdam, The Netherlands, pp. 585-594.
- Levine, A. S., 1977, *Canonical Analysis and Factor Comparison*. Sage Publications.

- Marble, D. F. and D. J. Peuquet (author-editors). 1986. "Geographical Information Systems and Remote Sensing". *Manual of Remote Sensing*, Vol. 2. pp. 923-957.
- Matheron, G., 1971. *The Theory of Regionalised Variables*. Centre de Morphologie Mathématique, Fontainebleau.
- McKeown, D. M., W. A. Harvey, and L. E. Wixson, 1989. "Automating Knowledge Acquisition for Areal Image Interpretation". *Computer Vision, Graphics, and Image Processing* 46, pp. 37-81.
- Molenaar, M., 1989, "Towards a Geographic Information Theory". *ITC Journal*. 1989-1, pp. 5-11.
- Morse, B. 1987, "Expert interface to a GIS". *Proceedings of International Symposium on Computer-Assisted Cartography, AUTOCARTO 8*, Baltimore, pp.535-541.
- Mounsey, H. and D. Briggs, 1988, "Corine: An Environmental Database for the European Community". *Proceedings of the third Int. Symp. on Spatial Data Handling*, Sydney, pp. 387-406.
- Mulder, J.A., A. K. Mackworth, and W. S. Havens, 1988, "Knowledge Structuring and Constraint Satisfaction: The Mapsee Approach". *IEEE Transaction on Pattern Analysis and Machine Intelligence*, Vol. 10, No. 6, pp. 866-879.
- Muller, J. C., et al., 1986, "A Knowledge Based Approach for Developing Cartographic Expertise". *Proceedings of 2nd Int. Symp. on Spatial Data Handling*, Seattle, USA, pp. 557-571.
- Muller, J. C. and Z. Wang, 1990. "A Knowledge Based System for Cartographic Symbol Design". *The Cartographic Journal*, Vol. 27, pp. 24-30.
- Muzard, J., E. Falardeau, and M. G. Strobel, forthcoming, 1991, "A Blackboard Architecture for an Expert Assistant in Statistical Analysis".

Niemann, H., G. F. Sagerer, S. Schroder, and F. Kummert, 1990, "ETNEST: A Semantic Network System for Pattern Understanding", *IEEE Transaction on Pattern Analysis and Machine Intelligence*, Vol. 12, No. 9, pp.883-905.

Nickerson, B. G. and H. Freeman, 1986, "Development of a Rule-based System for Automatic Map Generalization". *Proceedings of 2nd Int. Symp. on Spatial Data Handling*, Seattle, USA, pp. 537-556.

Nyerges, T. and P. Jankowski, 1989, "Design Considerations for MaPKBS-Map Projection Knowledge-based System". *The American Cartographer*, Vol. 16, No. 2, pp. 85-95.

Oliver, M. A. and R. Webster, 1990, "Kriging: A Method of Interpolation for Geographical Information Systems". *Int. J. of Geographical Information Systems*, Vol. 4, No. 3, pp. 313-332.

Openshaw, S. and H. Mounsey, 1987, "Geographic Information Systems and the BBC's Domesday Interactive Videodisc", *Int. J. of Geographical Information Systems*, Vol. 1, No. 2, pp. 173-179.

Openshaw, S., M. Charlton, C. Wymer, and A. Craft, 1987, "A Mark I Geographical Analysis Machine for the Automated Analysis of Point Pattern Data". *Int. J. of Geographical Information Systems*, Vol.1, No.4, pp.335-358.

Openshaw S., A. Cross, and M. Charlton, 1990, "Building a Prototype Geographical Correlates Exploration Machine". *Int.J. Geographical Information System*, Vol.4, No.3, pp. 297-311.

Okx, J. P., T. Gorter, L. A. A. Ignatius, and J. H. Schalkwijk, 1990, "Integrated Use of GIS and (Geo)statistics in Soil Pollution Research". *Proceedings of the First European Conference on Geographical Information Systems*, Amsterdam, The Netherlands, pp. 812-820.

Ord, J. K., 1972, *Families of Frequency Distributions*. Griffin, London.

Payne, E. C. and R. C. McArthur, 1990, *Developing Expert System - A Knowledge Engineer's Handbook for Rules and Objects*. John Wiley & Sons Inc.

Pereira, L.M., P. Sabatier, and E. de Oliveira 1982, "ORBI- an expert system for environmental resource evaluation through natural language". Report FCT/DI-3/82, Departamento de Informatica, Universidade Nova de Lisboa.

Pfefferkorn, C., D. Burr, D. Harrison, B. Hechman, C. Oresky, and J. Rothermel, 1985, "ACES: A Cartographic Expert System", *Proceedings of International Symposium on Computer-Assisted Cartography, AUTOCARTO 7*, Washington D. C., pp.399-407.

Qian, J., R. W. Enrich, and J. B. Campbell, 1990, "DNESYS - An expert system for automatic extraction of drainage networks from digital elevation data". *IEEE Transactions on Geoscience & Remote Sensing*, Vol. 28, No. 1, pp.29-45.

Raveneau, J. L., M. Miller, Y. Brousseau, and C. Dufour, 1991, "Micro-Atlases and the Diffusion of Geographic Information: An Experiment with Hypercard", in D. R. F. Taylor, (ed.) *Geographical Information Systems: The Microcomputer and Contemporary Cartography*, Oxford: Pergammon.

Robinson, G. and M. Jackson, 1985, "Expert Systems in Map Design". *Proceedings of International Symposium on Computer-Assisted Cartography, AUTOCARTO 7*, Washington D. C., pp. 430-439.

Robinson, V. B. and A. U. Frank, 1987, "Expert Systems for Geographic Information Systems". *PE & RS*, Vol. 111, No. 10, pp. 1435-1441.

Robinson, V. B. and R. N. Wong, 1987, "Acquiring approximate representations of some spatial relations", *Proceedings of International Symposium on Computer-Assisted Cartography, AUTOCARTO 8*, Baltimore, pp. 604-622.

Rogers, A., 1974, *Statistical Analysis of Spatial Dispersion: The Quadrat Method*. Pion, London.

Siekierska, E. M., 1983, "Towards An Electronic Atlas", *Proceedings of International Symposium on Computer-Assisted Cartography, AUTOCARTO 6*, Ottawa, pp. 464-474.

Siekierska, E. M. and S. Palko, 1986, "Development of the Electronic Atlas of Canada", in P. S. Glaeser, (ed.) *Computer Handling and Dissemination of Data*, North Holland: CODATA, 1987, Elsevier Science Publishers B. V., pp. 308-312.

Siekierska, E. and D.R.F. Taylor, 1990, "Electronic Mapping and Electronic Atlases: New Cartographic Products for the information Era - the Electronic Atlas of Canada". *CISM Journal ACSGC*, Vol. 45, No. 1, Spring 1991, pp. 11 - 21.

Smith, R., 1987, "Electronic Atlas of Arkansas: Design and Operational Considerations". *Proceedings of the 13th Int. Cartographic Conf.*, Morelia, Mexico, pp. 161-166.

Smith, T. R., D. Peuquet, S. Menon, and P. Agarwal, 1987, "KBGIS II: A Knowledge-based Geographic Information System". *Int. J. of Geographical Information System*, Vol.1, No.2, pp.149-172.

Stanley, J., 1989, *Introduction to Neural Networks*. California Scientific Software.

Tanic, L. 1986, "Urban planning and AI: the URBYS system". *Computers, Environment, and Urban System*. 01(3/4), pp. 135-146.

Taylor, D. R. F., 1987, "New Methods and Technologies in Cartography: The Digital World Atlas". *Int. Yearbook of Cartography*, pp. 219-224.

Thompson, J. J. S., 1989, "The National Atlas Information System: An Overview". *Proceedings of Canadian Conference Challenges for 1990's GIS*, CISM, Ottawa, Canada, pp. 62-74 .

Tobler, W. R., 1970, "A Computer Movie Simulating Urban Growth in the Detroit Region". *Economic Geography*, Supplement 46, pp. 234-240.

TYDAC Technologies Inc., 1990, "SPANS Learning System". Unpublished document, Ottawa, Canada.

Upton, G. J. G. and B. Fingleton, 1985, *Spatial Data Analysis by Examples*. Vol I: Point Pattern and Quantitative Data. Wiley, Toronto.

Wang, Zheng, L. Lu, and L. Fang, 1990, "PURSIS in Regional Planning". *Proceedings of Canadian Conference GIS for 1990s*, CISM, Ottawa, Canada, pp. 456-466.

Warren, W. G. and C. L. Batcheler, 1979, "The Density of Spatial Patterns: Robust Estimation through Distance Methods". *Spatial and Temporal Processes in Ecology*, Edited by R. M. Connack and J. K. Ord, pp. 247-270.

Waterman, D. A., 1986, *A Guide to Expert Systems*. Addison-Wesley Publishing Company.

Wilkinson, R. K., and E. M. Sigsworth, 1972, "Attitudes to the Housing Environment. An Analysis of Private and Local Authority Housholds in Batley, Leeds, York". *Urban Studies*, 9, pp193-209.

Zhang, A., 1990, "Knowledge Acquisition for EASSA: An Expert Advisory System for Statistical Analysis for the Electronic Atlas of Canada". COSEP Project Report, prepared for the Geographical Services Division, Energy, Mines, and Resources Canada.

REFERENCES CONSULTED

Barabba, V. P., 1990, "Through A Glass Less Darkly". Presidential Address at *the Annual Meeting of the American Statistical Association*.

Blalock, H. M., 1972, *Social Statistics*. McGraw-Hill, Inc.

Chandra, N. and W. Goran, 1986, "Steps toward a knowledge-based geographical data analysis system", in B. Optitz (ed.), *Geographical Information Systems in Government*, A. Deepak: Hampton, Virginia, pp. 749-763.

Dutton, G., 1991, "Improving Spatial Analysis in GIS Environments". *Proceedings of International Symposium on Computer-Assisted Cartography, AUTOCARTO 10*, Baltimore, pp. 168-185.

3

PM-1 3 1/2" x 4" PHOTOGRAPHIC MICROCOPY TARGET
NBS 1010a ANSI/ISO #2 EQUIVALENT

Energy, Mines, and Resources Canada, "The National Atlas of Canada", Fourth edition, 1974, the Macmillan Company of Canada Ltd.; Fifth edition, in publishing.

Huang, W. and R. Lippmann, 1987, "Comparisons between Neural Net and Conventional Classifiers". *IEEE First Int. Conf. on Neural Network*, Vol. IV, San Diego, California, pp. 485-494.

Lavigne, G. J. W., 1990, "Integrated Systems or Combined Products for Spatial Analysis". *Proceedings of the First European Conference on Geographical Information Systems*, Amsterdam, The Netherlands, pp. 661-663.

Manola, F., 1990, "Object-oriented Knowledge Bases". *AI Expert*, Vol. 5, No. 3, pp. 26-37.

National Atlas Information Service, Geographical Services Division, Energy, Mines and Resources Canada, 1989, "The National Atlas of Canada-5th Edition: Final Content". Unpublished document.

Peuquet, D. J. and D. F. Marble, 1990, *Introductory Readings in Geographic Information System*. Taylor & Francis Inc.

Reiter, R. and A. K. Mackworth, 1989, "A Logical Framework for Depiction and Image Interpretation". *Artificial Intelligence*, Vol. 41, No. 2, pp. 125-156.

Rhind, D., 1988, "A GIS research Agenda". *Int. J. of Geographical Information Systems*, Vol. 2, No. 1, pp. 23-28.

Taylor, D. R. F., 1991, *Geographical Information Systems: The Microcomputer and Contemporary Cartography*, Oxford: Pergammon.

Wood, M., 1989, "Expert Systems, Expert Tutors and Training in Elementary Statistics", *Proceedings of Expert System 89, the Ninth Annual Tech. Conf. of the British computer Society Specialists Group on Expert Systems*, London, pp. 195-206.

Xiu, J., 1982, *Introduction to Discrete Mathematics*. Advanced Educational Press, in Chinese.

Zhan, F., 1990, "A Knowledge-based System for Urban Land Classification". *Proceedings of Canadian Conference GIS for 1990s*, CISM, Ottawa, Canada, pp. 1507-1527.

Zhang A., E. M. Siekierska, and D. R. F. Taylor, 1991, "EASSA: An Expert Advisory System for Statistical Analysis for the Electronic Atlas of Canada". *Proceedings of Canadian Conference on GISs*, CISM, Ottawa, Canada, pp. 981-993.

Zhang, A., 1991, "Report on the EASSA contract". Prepared for the Geographical Services Division, Energy, Mines, and Resources Canada.

APPENDIX 1

STATISTICAL KNOWLEDGE FOR EASSA

LIST OF TABLES

1. STATISTICAL SUMMARY

TABLE 1: CENTRAL TENDENCY	183
TABLE 2: DATA VARIABILITY AND SKEWNESS	184
TABLE 3: HISTOGRAM, FREQUENCY TABLE AND CONTINGENCY TABLE	186

2. EQUALITY TEST

TABLE 4: COMPARING THE MEANS OF TWO GROUPS.....	187
TABLE 5: TESTING THE DIFFERENCE BETWEEN TWO GROUPS (NON- PARAMETRIC).....	189
TABLE 6: SINGLE-FACTOR ANOVA.....	191
TABLE 7: COMPARING THE MEANS AMONG MULTIPLE GROUPS (NON- PARAMETRIC).....	193
TABLE 8: COMPARING CATEGORICAL POPULATIONS.....	195
TABLE 9: MULTI-FACTOR ANOVA	197
TABLE 10: COMPARE MULTI-COMPONENT OBJECTS (NON-PARAMETRIC).....	199
TABLE 11: NORMALITY TEST.....	201
TABLE 12: COMPARING VARIATIONS BETWEEN TWO GROUPS.....	203

3. BIVARIATE RELATIONSHIP ANALYSIS

TABLE 13: INDEPENDENCE TEST.....	204
TABLE 14: PEARSON'S PRODUCT-MOMENT CORRELATION.....	205
TABLE 15: SPEARMAN'S RANK CORRELATION.....	207
TABLE 16: KENDAL'S Tau.....	208
TABLE 17: BIVARIATE REGRESSION.....	210

4. MULTIVARIATE RELATIONSHIP ANALYSIS

TABLE 18: MULTIPLE REGRESSION.....	213
TABLE 19: FACTOR ANALYSIS.....	216

5. UNIARIATE CLASSIFICATION

TABLE 20: CLASSIFICATION BY DISTRIBUTION.....	220
TABLE 21: CLASSIFICATION BY PROCEDURES	222

6. MULTIVARIATE CLASSIFICATION

TABLE 22: CLUSTER ANALYSIS	224
TABLE 23: DISCRIMINANT ANALYSIS	225

7. RANDOMNESS TEST

TABLE 24: AUTOCORRELATION ANALYSIS - JOINT STATISTIC'	230
TABLE 25: AUTOCORRELATION ANALYSIS - MORAN'S I STATISTIC'	232
TABLE 26: POINT PATTERN RANDOMNESS TEST	234

8. SPATIAL SURFACE ANALYSIS

TABLE 27: KRIGING	235
TABLE 28: TREND SURFACE ANALYSIS	237

STATISTICAL SUMMARY

TABLE 1: STATISTICAL KNOWLEDGE -- CENTRAL TENDENCY

ITEM	KNOWLEDGE
Statistical Issue	Statistical Summary
Statistical Method	Mean, Median, Mode
Goals	Find central tendency of a data set
Examples	. Mean of population age . Mode of soil type
How It's Done	. Mean -- calculate the arithmetic average of a data set. . Median -- find the middle value in a list of observations. . Mode -- find the value that most frequently occurs in a data set.
Measurement Scales	. Mean -- Interval, Ratio . Median -- Ordinal, Interval, Ratio . Mode -- Nominal, Ordinal, Interval, Ratio
Assumptions	None
Hypothesis	None
Graphic Representation	Box plot for Ordinal, Interval, and Ratio scales
Analysis Processes	. Organize a data set. . Use 'Central tendency' to calculate the value. . Use 'Box plot' to show the data.
Result Interpretation	See the explanations for 'How It's Done' of the methods.
Priority Factors	Mean - 1; Median - 2; Mode - 2.
Limitations	The three methods are complementary: Mean: good in a mathematical sense; it can be used as an estimation in inference, but is sensitive to errors and outliers. Median: not inordinately affected by errors and outliers; in a skewed distribution, it is more nearly like the value of the data set than is the value of mean and mode; the sum of the absolute deviation around a median is a minimum, and is thus the most representative or typical value of the data; it cannot be used in inference. Mode: the most easily understood concept of average; it is the only measure of central tendency for nominal data.

STATISTICAL SUMMARY**TABLE 2: DATA VARIABILITY AND SKEWNESS**

ITEM	KNOWLEDGE
Statistical Issue	Statistical Summary
Statistical Method	Maximum, Minimum, Dispersion measures, Skewness coefficient
Goals	<ul style="list-style-type: none"> . To express the degree of variability or dispersion in a set of data. . To show the skewness of the distribution of a data set.
Examples	<ul style="list-style-type: none"> . To express the variability of life expectancy over Canada. . To show the skewness of the distribution of income per capita in Canada.
How it's done	<ul style="list-style-type: none"> . Maximum and minimum values are the extreme values of a data set . Standard deviation, the square root of variance, is the typical measure used to express dispersion. To calculate variance, first, calculate the mean of the data set; second, sum the square of the differences between the mean and the observations; then, divide the sum by the number of the observations. . Skewness coefficient is a ratio. Its numerator is the average cubed deviation from the mean. The denominator is the cube of the standard deviation.
Measurement Scale	Maximum, Minimum -- Interval, Ratio Dispersion -- Variation ratio -- Nominal Percentiles -- Ordinal Standard deviation -- Interval Coefficient of variation -- Ratio Skewness -- Ordinal, Interval, Ratio
Assumptions	None
Hypothesis	None
Analysis Processes	<ul style="list-style-type: none"> . Organize a data set; . Choose 'Data variability' to calculate the dispersion measure with indications of maximum and minimum values; . Choose 'Skewness' to calculate the skewness coefficient.

Result Interpretations	<p>. The larger the dispersion measure is, the more the values of the data set vary: standard deviation is actually an average dispersion of a data set. If all the values are the same, the standard deviation is zero.</p> <p>. The larger the skewness coefficient, the larger the skewness of the distribution. Left skew results in a negative coefficient, and right skew in a positive coefficient. A normal distribution has a coefficient of zero.</p>
Graphic Representation	Box plot, Histogram
Limitations	No discussion on limitations is found in literature.

STATISTICAL SUMMARY**TABLE 3: HISTOGRAM, FREQUENCY TABLE AND CONTINGENCY TABLE**

ITEM	KNOWLEDGE
Statistical Issue	Summary
Statistical Method	Histogram, Frequency table, Contingency table.
Goals	<ul style="list-style-type: none"> . To display the frequency distribution of a data set. . To report how many cases fall into each category. Frequency table -- for one variable category; Contingency table -- for two factor categories with rows for one and columns for the other.
Lxamples	<ul style="list-style-type: none"> . Histogram of population ages . Contingency table of incomes by educational levels
How it's done	Specify classes for a data set, and count frequencies in each class.
Measurement Scale	Any if grouped
Assumptions	None
Hypothesis	None
Analysis Processes	<ul style="list-style-type: none"> . Organize a data set with one or two factors; . Choose a histogram to display the distribution for one variable; . Choose a frequency table to display a one factor data set; or . Choose 'Contingency table' to display a two factor data set.
Result Interpretation	<ul style="list-style-type: none"> . Histogram shows the data frequency distribution of a data set graphically. . Tables -- display counts and percentages of cases to different categories.
Graphic Representation	. Histogram
Limitations	It is not easy to decide class intervals without the help of other methods (refer to Table 20, 21 for univariate classification)

*EQUALITY TEST***TABLE 4: COMPARING THE MEANS OF TWO GROUPS**

ITEM	KNOWLEDGE
Statistical Issue	Equality test
Statistical Method	Parametric methods for comparing means of two groups, including <ul style="list-style-type: none"> . Student's t-test . Modified t-test (1) for dealing with different sample sizes . Modified t-test (2) for dealing with autocorrelation
Goals	Assert equality or non-equality between two groups.
Examples	Do people in Ontario and Quebec differ in mean income?
How it's done	For the basic t-test: Assume two samples are drawn from two independent populations. Calculate the samples' means as well as their differences. The sampling distribution of the differences will be normal either when the populations have normal distribution, or when the populations are infinite. With the knowledge of sampling distribution, we can then conduct a significance test. Assume the distributions of populations are normal, t-test can be used for the test directly.
Measurement Scale	Interval, Ratio
Assumptions	<ul style="list-style-type: none"> . Random sample . Independent observations (for Student's t-test and Modified t-test (1)) . Normal or roughly normal distributions
Certainty factors on applicability	<ul style="list-style-type: none"> . If no autocorrelation and (normal or roughly normal) and $n_1=n_2$ and $s_1=s_2$, then t-test with CF (Certainty Factor) =95; . If no autocorrelation and very approximately normal and $n_1=n_1$ and sample size is very large (infinite), then t-test with CF=90; . If no autocorrelation and very approximately normal and $n_1=n_2$ and $s_1=s_2$ and $n_1>30$, then t-test with CF=80; . If no autocorrelation and normal and $n_1=n_2$ and $s_1\neq s_2$, then t-test with CF=70; . If no autocorrelation and non-normal and $n_1\neq n_2$ then t-test with CF=-70; . If no autocorrelation and $n_1\neq n_2$ and $s_1\neq s_2$, then t-test with CF=-90;

Certainty factors on applicability (continued)	<ul style="list-style-type: none"> . If no autocorrelation and $n1 \neq n2$ and $s1 \neq s2$ and $n1 \rightarrow \text{infinite}$ and $n2 \rightarrow \text{infinite}$, then t-test with CF = 70; . If normal or roughly normal and $n1 \neq n2$ and $s1 \neq s2$, then Modified t-test (1) with CF = 95; . If normal or roughly normal and autocorrelation, then Modified t-test (2) with CF = 95.
Hypothesis	Ho: there is no significant difference in means between the two populations.
Analysis Processes	<ul style="list-style-type: none"> . Take a sample of each population. . Use 'Normality test' to check if the data sets have normal or roughly normal distributions; if not normal, use transformation techniques to transform it into a normal or roughly normal distribution. . Calculate standard deviation for each of the samples. . Use 'Autocorrelation test' to analyze if the observations are independent; choose Normality Assumption if the data sets have normal distributions; otherwise, use Randomization Assumption (see Table 24). . Set Ho, and significant level α (0.01 – 0.1) at which you can reject the null hypothesis Ho and accept the alternative hypothesis Ha. . Use the suggested type of t-test to get p-value
Result Interpretation	P-value is the smallest value to reject Ho. If it is smaller than α , then Ho is rejected, meaning that the two populations have significantly different means.
Graphic Representation	Boxplot, Distribution curves for each of the groups.
Priority Factor	1
Limitations	<ul style="list-style-type: none"> . Type I or Type II error exists. . No appropriate method has been found when data are autocorrelated but have a large departure from normal distribution, or the standard deviation of the two groups significantly differ.

*EQUALITY TEST***TABLE 5: TESTING THE DIFFERENCE BETWEEN TWO GROUPS (NON-PARAMETRIC)**

ITEM	KNOWLEDGE
Statistical Issue	Equality test
Statistical Method	Mann-Whitney's U-test
Goals	Assert equality between two groups' means
Examples	Does mean life expectancy differ between Eastern Canada and Southern Canada?
How It's Done	If there is no difference in means between two groups, the sums of ranks for the two samples are similar when they are ranked together. By randomization, the probability for one of the sums to be a certain value has been calculated. Counting the rank sums for our studied samples, one can then find the probability with which the rank sum occurs when the two samples actually have the same means. When the samples' sizes are greater than 20, z-score is used for the test as approximation.
Measurement Scale	Ordinal, Interval, Ratio
Assumptions	<ul style="list-style-type: none"> . Random sample . Independent observations . Identical distributions of the two populations
Hypothesis	Ho: there is no significant difference in means between the two populations.
Analysis Processes	<ul style="list-style-type: none"> . Take a sample of each population. . Use 'Autocorrelation test' to confirm there is no significant autocorrelation between the observations, under Randomization Assumption (see Table 24); if they are autocorrelated, it is questionable to use the U-test. . Set H_0 and α (0.01 - 0.1) at which you are going to accept the hypothesis. . Choose 'Mann-Whitney U-test' to get p-value.
Result Interpretation	If p is smaller than the given α , H_0 is rejected, meaning that the two groups have significantly different means.
Graphic Representation	<ul style="list-style-type: none"> . Boxplot . Distribution curves for each of the groups

Priority Factor	2
Limitation	<ul style="list-style-type: none">. Type I or Type II error exists.. No appropriate method has been found if data are far from normality, and not identically distributed or autocorrelated.

EQUALITY TEST**TABLE 6: SINGLE-FACTOR ANOVA**

ITEM	KNOWLEDGE
Statistical Issue	Equality test
Statistical Method	Parametric methods for multiple samples comparison -- . Analysis of Variance for single-factor (ANOVA); . Modified ANOVA for dealing with autocorrelation.
Goals	Comparing means among multiple populations sharing a single factor.
Examples	Do different types of agriculture (e.g. grain, vegetable, etc.) have different mean productivities (\$ return per hour)?
How It's Done	For the basic ANOVA: Calculate measure of discrepancy between samples, and divide it by their combined variance, in order to measure the degree of sample variations. When the normality assumption is met, the sampling distribution of the ratio follows F distribution. Tests can hence be conducted based on the knowledge of the distribution.
Measurement Scale	Interval, Ratio
Assumptions	. Random sample . Independent observations (for ANOVA) . Each population is normal or roughly normal . All populations have the same standard deviation
Certainty factors on applicability	. If distributions are roughly normal, but standard deviations are the same, CF=90; . If distributions are roughly normal, and $3s_{min} > s_{max}$, CF=70; . If standard deviations differ greatly, CF=-70 . If distributions are far from normality, CF=-70.
Hypothesis	. General comparison -- H_0 : there is no significant difference in means among the studied groups. . Multiple comparisons -- H_0 : there is no significant difference in means between pairs of groups.

Analysis Processes	<ul style="list-style-type: none"> . Take a sample for each population. . Use 'Normality test' to check if the distributions are normal or roughly normal; if not normal, transform each of the distributions into normal or roughly normal ones if possible; if far from normality, use of this method is questionable. . Use 'Autocorrelation test' to analyze if the independence assumption is violated, choosing Normality Assumption (see Table 24). . Set H_0, and significant level α, at which you can accept the hypothesis. . Use suggested type of method to obtain an ANOVA table.
Result Interpretation	<ul style="list-style-type: none"> . If p-value for F-ratio is smaller than α, H_0 is rejected. This means that the difference in means among the multiple groups is not due to chance; in other words, these populations have significantly different means. . If p-value for t-test is smaller than α, H_0 for multiple comparison is rejected, meaning that the two groups followed by this p-value differ in means significantly. A p-value that is greater than α suggests that the corresponding two groups do not have significantly different means.
Graphic Representation	Box plot
Priority Factor	1
Limitations	<ul style="list-style-type: none"> . If the normality assumption is not well met, type I error may not be controlled at the desired level of significance α, and type II error probabilities may be rather large so that substantial departures from H_0 are likely undetected. . No good approach has been found when standard deviations are greatly different.

*EQUALITY TEST***TABLE 7: COMPARING THE MEANS AMONG MULTIPLE GROUPS (NON-PARAMETRIC)**

ITEM	KNOWLEDGE
Statistical Issue	Equality test
Statistical Method	Non-parametric equivalence of ANOVA -- Kruskal-Wallis test
Goals	Assert equality among multiple populations sharing a single factor.
Examples	Do violent crime rates differ in means among provinces of Canada?
How It's Done	For groups with the same means, the rank sum for each of them should be similar. The Kruskal-Wallis test statistic is constructed. The probabilities for randomized occurrences have been calculated. It approximately follows χ^2 distribution. Based on this distribution, inferences can be made about the groups studied.
Measurement Scale	Ordinal, Interval, Ratio
Assumptions	<ul style="list-style-type: none"> . Random samples. . Independent observations. . If the number of populations is k, and the number of objects for each population is n_i, then when $k = 3$, $n_i > 6$; when $k > 3$, $n_i > 5$. . The studied k-population distributions have the same shape and spread.
Hypothesis	H_0 : there is no significant difference in means among the multiple populations.
Analysis Processes	<ul style="list-style-type: none"> . Draw a random sample for each population. . Use 'Autocorrelation test' to confirm that the observations are independent, choosing Randomization Assumption; if there is significant dependency, use of this method is questionable. . Choose 'Distribution comparison' to check if the k-populations have similar skewness coefficients and similar kurtoses; if not, it is questionable to use this method. . Set H_0, and significant level α. . Choose K-W equality test to get p-value.
Result Interpretation	If p is smaller than the given α , H_0 is rejected, meaning the difference in means among the k -populations is significant.
Graphic Representation	Box plot
Priority Factor	2

Limitations	<ul style="list-style-type: none">. Type I or Type II error exists.. No appropriate approach has been found in the reviewed texts when the shape or spread of the populations' distributions are greatly different.. No good non-parametric approach has been found for autocorrelated data sets.
-------------	---

*EQUALITY TEST***TABLE 8: COMPARING CATEGORICAL POPULATIONS**

ITEM	KNOWLEDGE
Statistical Issue	Equality test
Statistical Method	Chi-square test (Goodness-of-fit)
Goals	To test if two or more populations are homogeneous with respect to some defined categories.
Examples	Comparing the distribution of population ages between men and women.
How It's Done	Form a contingency table with frequency counts for each category. The expected frequencies for each cell when the distributions are the same can be computed. Based on these and the counts for the studied populations, the Chi-square statistic can be calculated. According to the Chi-square curve distribution, it can be inferred whether the differences between two distributions are due to chance or not.
Measurement Scale	Any if grouped
Assumptions	<ul style="list-style-type: none"> . Random sample. . The items in the sampled population must be classified independently of each other. . The classes or cells must be mutually exclusive and exhaustive. . Each cell frequency > 5, or 80% of cells' frequencies > 5 and no cell frequency < 1.
Hypothesis	H_0 : there is no significant difference in frequencies between the studied populations.
Analysis Processes	<ul style="list-style-type: none"> . Draw a random sample for each population under study; percentage data should be converted into frequency data. . Specify categories if data are not categorical, and use 'Contingency table' to display the frequency distributions. . Check the cell sizes, if the assumption about frequency is violated, increase sample size, or change categories, or take risk, which means that the result may be erroneous. . Set H_0 and significant level α (0.01 - 0.1). . Choose 'Chi-square test' to get p-value.
Result Interpretation	If p is smaller than the given α , H_0 is rejected, meaning that the studied populations have significantly different category proportions.

Graphic Representation	Histogram
Priority Factor	3
Limitations	<ul style="list-style-type: none">. If the assumptions are not well met, the test results in inaccurate p-value.. Ordering information is neglected in the test. For this, Kolmogorov-Smirnov is an alternative, but Type II error increases when using this test.

EQUALITY TEST**TABLE 9: MULTI-FACTOR ANOVA**

ITEM	KNOWLEDGE
Statistical Issue	Equality test
Statistical Method	Multi-factor ANOVA
Goals	<ul style="list-style-type: none"> . To compare means of multiple populations under multiple factor categories. . To assess the effects of multiple factors on a response variable.
Examples	Do energy consumptions differ in means by income per capita in different regions?
How It's Done	The variability in the dependent variables is subdivided into two components: that attributable to differences between groups (the hypothesis sums of squares matrix), and that attributable to variability within groups (the error sums of squares matrix). Various F-ratio statistics that compare the magnitudes of these two matrices are computed.
Measurement Scale	Interval, Ratio
Assumptions	<ul style="list-style-type: none"> . Random samples . Independent observations . Multi-normal or roughly normal distributions . Variance and covariance matrices are equal for all groups.
Hypothesis	<ul style="list-style-type: none"> . General hypothesis -- H_0: there is no significant difference in means among the multiple populations based on the multiple factors. . Hypothesis for each of factors -- H_0: the factor being analyzed has no significant effect on the variable; or, the variable response does not differ with the variation of a certain factor.
Analysis Processes	<ul style="list-style-type: none"> . Take samples for each treatment. . Use 'Normality test' to check if each of the groups is normally or roughly normally distributed; if there is great departure, try to transform it into a normal or roughly normal distribution; otherwise, the use of this method is questionable. . Use 'Autocorrelation test' to confirm that the observations are independent, choosing Normality Assumption; if they are significantly autocorrelated, it is questionable to use this method. . Set H_0 and α. . Choose 'Multi-factor ANOVA' to get various F-ratio's and p-values

Results	ANOVA table with . P for the overall test; . {Pi} for each of the studied factors.
Result Interpretation	. If P is smaller than the given α , H_0 is rejected, meaning there is interaction among the factors; . If P_i is smaller than α , the null hypothesis that there is no main effect of the i th factor on the variable is rejected, meaning that this factor has main effects on the variable; it also means that the variable response differs in means with the variation of the i th factor.
Graphic Representation	Box plot
Priority Factor	1
Limitations	. If the assumptions are not well met, type I error may not be controlled at the desired level of significance α , and type II error probabilities may be rather larger so that substantial departures from H_0 are likely undetected. . No non-parametric equivalent method found. . No approach found to deal with the autocorrelation problem.

*EQUALITY TEST***TABLE 10: COMPARING MULTI-COMPONENT OBJECTS (NON-PARAMETRIC)**

ITEM	KNOWLEDGE
Statistical Issue	Equality test
Statistical Method	Non-parametric equivalent ANOVA (randomized block design) -- Friedman's test
Goals	To assert that the mean value does not depend on which treatment is applied.
Examples	<ul style="list-style-type: none"> . Do Canadian cities have different employment growth rates in different years? . Do population age structures vary with provinces (or cities)?
How It's Done	k observations in each block (or object) are first ranked separately from 1 to k. Then the rank average for each treatment is computed. (For example, for k cities, k proportional values in each population age group are first ranked from 1 to k; then the rank average for each city is computed). If the mean value does not depend on which treatment is applied, the rank averages will be similar to one another. The test statistic, Friedman's statistic based on the ranks, has approximately a Chi-square distribution. With the sampling distribution knowledge, we can test the significance of the difference in means for a specific case.
Measurement Scale	Ordinal, Interval, Ratio
Assumptions	<ul style="list-style-type: none"> . Random sample . Independent observations . Observations are drawn from populations having the same shape and spread.
Hypothesis	H ₀ : (e.g.) the proportional structures of population ages do not significantly differ between cities.

Analysis Processes	<ul style="list-style-type: none"> . Take a sample (e.g. a set of objects, like cities) with responses for each treatment (e.g. population proportional values for each age group). . If the objects are not spatially or temporally discrete, use 'Autocorrelation test' to confirm that there is no significant autocorrelation between the objects, choosing Randomization Assumption; if yes, it is questionable to use this method. . Use 'Friedman's test' to get p-value.
Result Interpretation	If p is smaller than the given α , H_0 is rejected, meaning, e.g. the proportional structures of population ages are significantly different in different cities.
Graphic Representation	<ul style="list-style-type: none"> . Stacked bar . Stacked population pyramid . Pie
Priority Factor	2
Limitations	<ul style="list-style-type: none"> . No alternative found when distribution shapes are not the same, or data are autocorrelated. . Type I or II error exists.

EQUALITY TEST**TABLE 11: NORMALITY TEST**

ITEM	KNOWLEDGE
Statistical Issue	Equality test
Statistical Method	Lilliefors' test (Goodness-of-fit)
Goals	To test if sample measurements closely approximate a normal distribution.
Examples	<ul style="list-style-type: none"> . Does income per capita follow a normal distribution? . Test whether a distribution is normal, so as to decide if a non-parametric method should be used.
How It's Done	<ul style="list-style-type: none"> . For each value of X in the sample, count the number of items having this value or a smaller value. Divide every number by the sample size, to obtain the cumulative probability distribution of the X values. . Calculate the cumulative probability distribution of the normal curve with the same mean and standard deviation as the sample. . Use the largest difference between the two sets of cumulative probabilities as a test statistic, to infer whether the departure is due to chance.
Measurement Scale	Interval, Ratio
Assumptions	<ul style="list-style-type: none"> . Random sample drawn from a large population . Independent observations
Certainty factors on applicability	<ul style="list-style-type: none"> . If the population is infinite and no autocorrelation and continuous variable, then CF=95; . If the population is finite but not extremely small and no autocorrelation and continuous variable, then CF=70; . If the population is extremely small, CF=-90; . If autocorrelation, CF=-80; . If discontinuous variable, CF=60.
Hypothesis	Ho: the sample has a normal distribution
Analysis Processes	<ul style="list-style-type: none"> . Draw a random sample from the population. . Use 'Autocorrelation test' to analyze if the data are autocorrelated, choosing Randomization Assumption; if seriously autocorrelated, it is questionable to use this method. . Use 'Lilliefors' test' to get p-value.

Result Interpretation	If p is smaller than the given α , H_0 is rejected, meaning that the studied population has significant departure from normality.
Graphic Representation	<ul style="list-style-type: none">. Histogram. Cumulative probability plot
Limitations	<ul style="list-style-type: none">. Can not deal with autocorrelation.. Can not handle extremely small samples.. Tends to be somewhat biased in favour of the null hypothesis when the variable is discontinuous.

EQUALITY TEST**TABLE 12: COMPARING VARIATIONS BETWEEN TWO GROUPS**

ITEM	KNOWLEDGE
Statistical Issue	Equality test
Statistical Method	F-test, Modified F-test
Goals	To test whether two populations differ significantly in variance.
Examples	Does income per capita differ significantly in variance between Ontario and Quebec?
How It's Done	For the basic F-test: Calculate variance for each of the samples. F-ratio, the ratio of the larger variance to the smaller variance, is used as a statistic. F-curve is used for inference. When observations are autocorrelated, Modified F-test is used.
Measurement Scale	Interval, Ratio
Assumptions	. Random sample . Independent observations (for F-test) . Populations have normal distributions.
Hypothesis	H ₀ : there is no significant difference in variance between the two populations.
Analysis Processes	. Draw a random sample for each population under study. . Use 'Normality test' to check if the distributions are normal; if far from normality, it is questionable to use this method. . Use 'Autocorrelation test' to analyze whether the observations are autocorrelated, under Normality Assumption. . Set H ₀ and α (0.01 - 0.1). . Use suggested type of F-test to get p-value.
Result Interpretation	If p is smaller than the given α , H ₀ is rejected, meaning that the studied populations are significantly different in variance.
Graphic Representation	. Histogram . Box plot
Limitations	. No alternative found for dealing with Non-normality. . Type I or Type II error exists.

BIVARIATE RELATIONSHIP ANALYSIS

TABLE 13: INDEPENDENCE TEST

ITEM	KNOWLEDGE
Statistical Issue	Bivariate relationship analysis -- Independence analysis
Statistical Method	Chi-square test
Goals	Test if two classification schemes are independent.
Examples	Do immigrants from different origins differ in educational levels? In other words, are educational levels of immigrants dependent on country of origin?
How It's Done	Form a contingency table with cell values for the counted number against the classification schemes. The expected values, when the two schemes are independent, can be calculated, and the Chi-square statistic can then be computed. It follows χ^2 distribution, so the inference can be made with confidence.
Measurement Scale	Any if grouped
Assumptions	<ul style="list-style-type: none"> . Random sample . The items in the sampled population must be classified independently. . The classes or cells must be mutually exclusive and exhaustive. . Each cell frequency > 5, or, 80% of cells' frequencies > 5 and no frequency < 1.
Hypothesis	Ho: the two classification schemes are independent.
Analysis Processes	<ul style="list-style-type: none"> . Draw a sample against the theme categories under study. . Use 'Contingency table' to display the cell counts as well as the marginal totals. . Set Ho and significant level α (0.01 - 0.1). . Choose the 'Chi-square test'.
Result Interpretation	. If p is smaller than the given α , Ho is rejected, meaning that the two classification schemes are significantly dependent.
Graphic Representation	Histogram for each category
Priority Factor	3
Limitations	<ul style="list-style-type: none"> . If the assumptions are not well met, the test produces an inaccurate p-value. . Ordering information is neglected in the test. For this, Kolmogorov-Smirnov is an alternative, but Type II error increases when using this test.

BIVARIATE RELATIONSHIP ANALYSIS**TABLE 14: PEARSON'S PRODUCT-MOMENT CORRELATION**

ITEM	KNOWLEDGE
Statistical Issue	Bivariate relationship -- Correlation analysis
Statistical Method	<ul style="list-style-type: none"> . Pearson's product-moment correlation coefficient r . Modified Pearson's correlation coefficient r'
Goals	Analyze how strongly two variables in a sample are linearly associated
Examples	Does violent crime rate correlate with employment rate? To what degree are they correlated?
How It's Done	<p>For r:</p> <p>For a bivariate sample with variables X and Y, calculate the covariance and standard deviations of X and Y. The ratio of the covariance to the product of the standard deviations is the Pearson's product-moment coefficient r, which measures the degree of linear correlation.</p>
Measurement Scale	Interval, Ratio.
Assumptions	<p>For test:</p> <ul style="list-style-type: none"> . Sampling with replacement, or infinite population/sample constitutes less than about 10% of the population. . If sample size is small, Y value should be normally distributed on each value of X. . Independent X, Independent Y on each X. No serious autocorrelation (for r).
Hypothesis	H_0 : there is no correlation between the two variables in the population.
Analysis Processes	<ul style="list-style-type: none"> . Draw a bivariate sample from a population of pairs . Use 'Scatter plot' to check if the distribution appears to be a linear trend. . Use 'Pearson's correlation' to calculate the coefficient r. . Examine the variables under study; if there is a large possibility to have autocorrelation, then use 'Autocorrelation test' to check if data are seriously autocorrelated. If yes, use r' to recalculate the coefficient. . If sample size < 20, use 'Normality test' to check if Y has a normal distribution on each X; if far from normality, it is better to transform Y distribution to a normal or basically normal one. . Set H_0 and significant level α (0.01 - 0.1) . Use suggested 't-test' to obtain the p-value.

Result Interpretation	<p>If $r = 0$, no correlation</p> <p>If $0 < r \leq 0.5$, weak linear correlation</p> <p>If $0.5 < r \leq 0.8$, moderate linear correlation</p> <p>If $0.8 < r \leq 1.0$, strong linear correlation</p> <p>Sign of the coefficient indicates the direction of the correlation (positive or negative)</p> <p>If p is smaller than α, H_0 is rejected, meaning the fact that X and Y share information.</p>
Graphic Representation	Scatter plot
Priority Factor	1
Limitations	<p>On the coefficient interpretations:</p> <ul style="list-style-type: none"> - r is a measure of linear correlation. r may be very small, even though the variables are strongly curvilinearly correlated. - No-sense autocorrelation: there may be no evidence suggesting that the r is meaningful, even though it is high. This means that an empirical analysis is always necessary in interpreting the coefficient. - Confounding: the relationship may be systematically affected by the unexamined variables, which may cause a linear relationship to be nonlinear, or a nonlinear to be linear; when this is the case, two approaches can be used: <ul style="list-style-type: none"> a) stratify sampling so that X and Y are collected for selected Z; b) use partial or multiple correlation analysis method. - Heterogeneity population and samples: if the population or the samples includes multiple groups, the coefficient may be misleading, e.g. the amount of positive correlation may exactly balance the amount of negative correlation. However it is difficult to define the 'heterogeneity'. - Spurious correlation: if X and Y share a common element, e.g. population size, their correlation may be partially spurious and partially real. An approach to this is to calculate the spurious correlation, and subtract it from r; the remaining value is the real correlation. - The value of r may be greatly affected by the presence of even one or two outlying pairs that are far from the center of the scatter plot. - For spatial analysis, one r may not be representative. <p>On the test:- When the normality assumption is seriously violated, the test is doubtful.</p>

BIVARIATE RELATIONSHIP ANALYSIS

TABLE 15: SPEARMAN'S RANK CORRELATION

ITEM	KNOWLEDGE
Statistical Issue	Bivariate relationship analysis -- Correlation analysis
Statistical Method	Spearman's rank correlation r_s
Goals	<ul style="list-style-type: none"> . Analyze how strongly the ranks of two variables are related. . Identify both linear and nonlinear relationships. . Deal with samples with outliers.
Examples	Does infant mortality significantly correlate to income per capita? To what degree?
How It's Done	Rank X and Y separately. Calculate the covariance and standard deviations of X and Y ranks. The r_s is the ratio of the covariance to the product of the standard deviations.
Measurement Scale	Ordinal (Interval, Ratio data can be transferred to ordinal)
Assumptions	<ul style="list-style-type: none"> . Random sampling . Sample size > 10 if conducting test.
Hypothesis	H_0 : there is no significant correlation between the two variables.
Analysis Processes	<ul style="list-style-type: none"> . Take a bivariate sample X, Y from a population of pairs. . Choose Spearman's correlation to get r_s. . Set H_0 and significant level α (0.01 - 0.1). . Use suggested test (Hotelling-Pabst test for small sample, and z-test for large sample) to obtain p-value.
Result Interpretation	<ul style="list-style-type: none"> . If $r_s = 0$, no correlation . If $0 < r_s \leq 0.5$, weak monotone correlation . If $0.5 < r_s \leq 0.8$, moderate monotone correlation . If $0.8 < r_s \leq 1.0$, strong monotone correlation . Sign of the coefficient indicates the direction of the correlation (positive or negative) . If p smaller than α, H_0 is rejected, meaning that the variables are significantly correlated.
Graphic Representation	Scatter plot
Priority Factor	2
Limitations	<ul style="list-style-type: none"> . Too many ties will cause results to be less rigorous. . Not suitable for small samples. . If far from normality, z-test is questionable.

BIVARIATE RELATIONSHIP ANALYSIS

TABLE 16: KENDAL'S Tau

ITEM	KNOWLEDGE
Statistical Issue	Bivariate relationship analysis -- Correlation analysis
Statistical Method	Kendal's <i>Tau</i>
Goals	<ul style="list-style-type: none"> . Analyze how strongly the ranks of two variables in a sample are related. . Identify both linear and nonlinear relationships. . Deal with sample outliers. . Deal with small samples. . Deal with samples with many ties.
Examples	Does life expectancy significantly correlate to income per capita? To what degree?
How It's Done	Rank X and Y separately. Count the concordant pairs and discordant pairs (it is called concordant if $x_1 > x_2$, then $y_1 > y_2$; else it is called discordant). Tau statistic is constructed based on the counts so that when all pairs are concordant, the coefficient is +1, meaning perfect positive correlation; when all pairs are discordant, it is -1, meaning perfect negative correlation; and when they are the same, it is zero, meaning no correlation.
Measurement Scale	Ordinal (or ranked to ordinal)
Assumptions	Random bivariate sample
Hypothesis	H_0 : there is no significant correlation between the two variables.
Analysis Processes	<ul style="list-style-type: none"> . Take a bivariate sample from an entire population of pairs. . Choose 'Kendal's correlation' to get <i>tau</i>. . Set H_0 and significant level α (0.01 – 0.1). . Use suggested test (Kendal's test for sample size ≤ 10; z-test for sample size > 10) to get p-value.
Result Interpretation	<ul style="list-style-type: none"> . If $Tau = 0$, no correlation . If $0 < Tau \leq 0.5$, weak monotone correlation . If $0.5 < Tau \leq 0.8$, moderate monotone correlation . If $0.8 < Tau \leq 1.0$, strong monotone correlation . Sign indicates the direction of the correlation (positive or negative). . If p is smaller than α, H_0 is rejected, meaning that the two variables are significantly correlated.
Graphic Representation	Scatter plot

Priority Factor	2
Limitations	It deals with limitations of Pearson's r at the price of losing sensitivity.

BIVARIATE RELATIONSHIP ANALYSIS**TABLE 17: BIVARIATE REGRESSION**

ITEM	KNOWLEDGE
Statistical Issue	. Bivariate relationship analysis -- dependence analysis
Statistical Method	. Simple linear regression . Curvilinear regression . Poisson regression
Goals	. To develop an equation that summarizes the relationship between a dependent variable and an independent variable; . To predict values for a dependent variable from the values of the independent variable.
Examples	How is mean annual temperature related to annual precipitation?
How It's Done	For simple linear regression: Formulate a model, e.g. $Y = a + bX$, where Y is a dependent variable and X is an independent variable. The coefficients a and b are chosen so that the sum of the squared differences between the observed and predicted values of the dependent variable based on the model is as small as possible. Various diagnostic plots and statistics are used to identify departures from the model and 'unusual points'.
Measurement Scale	Simple linear regression, curvilinear regression, and modified regression - Interval, Ratio Poisson regression - Binary variable.
Assumptions	To draw inference about population values based on sample results: . For any X , the distribution of Y is normal, with a constant variance. . The Y 's are statistically independent of each other (no autocorrelation; for simple linear regression and curvilinear regression).
Hypothesis	$H_0: r = 0$, i.e. the regression is meaningless.

Analysis Processes

- . Take a random sample Y for certain X's.
- . Make a scatter plot to confirm the relationship between two variables is linear; if not, transform it into linear (if positively skewed, use the log transformation on X or Y; if negatively skewed, use the square transformation) or use curvilinear regression.
- . Use suggested 'Regression analysis' to compute the regression model.
- . Inference from sample to population:
 - Set H_0 and α (0.01 - 0.1);
 - Use 'F-test' to test if $R^2 = 0$ is true;
 - Calculate standard errors for predicted values or mean predicted values, then construct their confidence intervals if required.
- . Searching for violations of assumptions using residuals:
 - Use 'Residual plot' to show the distribution of the residuals, then check the equality of variance: if the spread of the residuals increases or decreases with the values of the independent variables or with the predicted values, the assumption of constant variance is questionable;
 - Use 'Autocorrelation test' to check if the residuals are autocorrelated; if yes, the assumption of no autocorrelation on Y is violated;
 - Use 'Normality check' to confirm if the residual distribution is normal; if not, the normality assumption is violated.
- . Approaches to the violations:
 - Stabilizing the variance: when the variance is proportional to the mean of Y for a given X, use the square root of Y if all Y's are positive; when the standard deviation is proportional to the mean, try logarithmic transformation; when the standard deviation is proportional to the square of the mean, use the reciprocal of Y; when Y is a proportion or rate, the arc sine transformation may stabilize the variance.
 - On autocorrelation: there are three possibilities which cause autocorrelation to be detected: a) incorrect model; b) omission of variables; and c) autocorrelation.
 - a) check the model effectiveness by looking at the residual plot or X-Y scatter plot to see how the line fits the distribution;
 - b) if the model is correct, try to add some variables (use the multiple regression); if systematic variation occurs, the variable should be added to the model;

Analysis Processes (continued)	<p>c) If still autocorrelated, use 'Cochran-Orcutt iterative procedure' to eliminate it.</p> <p>. Other diagnostics:</p> <ul style="list-style-type: none"> - Locating outliers by looking at the residual plot: the cases with very large positive or negative residuals are outliers, which should be checked to determine whether or not they are caused by errors; - Use 'Mahalanobis' distance' to identify the cases having an unusual X value; - Use 'Cook's distance' to identify the case as an influential point. <p>Analyze all these cases to decide if they should be removed.</p>
Result Interpretation	<p>. R^2 -- coefficient of determination; an overall measure of the success of the regression in predicting Y from X; its square root is the correlation coefficient.</p> <p>. Adjusted R^2 -- corrected R^2 more closely reflecting the goodness of fit of the model in the population.</p> <p>. s -- standard deviation of residuals.</p> <p>. Coefficients of regression -- intercept coefficient a; slope coefficient of the regression line b.</p> <p>. Standard errors of coefficients -- the average errors of the estimated coefficients.</p> <p>. If p is smaller than the given α, H_0 is rejected, meaning the regression is significant.</p> <p>. You can predict the Y values from X values if all assumptions are met.</p>
Graphic Representation	<p>. Scatter plot;</p> <p>. Residual plot;</p> <p>. Simple line;</p> <p>. Simple projection line.</p>
Limitations	<p>. The checks for the assumptions are mostly empirical, therefore more statistical knowledge is required for one to conduct the analysis.</p> <p>. It is not easy to select all related variables to add to the regression equation when autocorrelation is detected.</p>

MULTIVARIATE RELATIONSHIP ANALYSIS

TABLE 18: MULTIPLE REGRESSION

ITEM	KNOWLEDGE
Statistical Issue	Multivariate relationship analysis -- dependence analysis
Statistical Method	<ul style="list-style-type: none"> . Multiple linear regression . Poisson regression
Goals	<ul style="list-style-type: none"> . To develop an equation that summarizes the relationship between a dependent variable and a set of independent variables. . To identify the subset of independent variables that are most useful for predicting the dependent variable. . To predict values for a dependent variable from the values of the independent variables.
Examples	How is the area of interior living space per resident related to: income per capita, the proportion of income spent on housing, population density, size of household, educational level, dominant occupation, etc.? Which variables are most useful for predicting area per resident?
How It's Done	<p>For Multiple linear regression:</p> <p>Formulate a model, e.g. $Y = b_0 + b_1X_1 + \dots + b_nX_n$, where Y is a dependent variable and X_i, $i = 1, 2, \dots, n$, are independent variables. The coefficients b_i's, $i = 0, 1, \dots, n$, are chosen so that the sum of the squared differences between the observed and predicted values of the dependent variable based on the model is as small as possible. Various diagnostic plots and statistics are used to identify departures from the model and 'unusual points'.</p>
Measurement Scale	<p>Multiple linear regression and Modified regression: Interval, Ratio</p> <p>Poisson regression: Binary variable.</p>
Assumptions	<p>To draw inference about population values based on sample results:</p> <ul style="list-style-type: none"> . For any X, the distribution of Y is normal, with a constant variance. . The Y's are statistically independent of each other (no autocorrelation; for Multiple regression and Modified regression). . Independent variables are not strongly correlated linearly. . Linear relationship between Y and X's for the multiple linear regression.
Hypothesis	<ul style="list-style-type: none"> . H_0: any one of the regression coefficients b_i is zero; . H_0: $R^2 = 0$, i.e. the regression equation is meaningless.

Analysis Processes

- . Draw a multivariate sample from a population with certain X's and random Y.
- . Compute the correlation matrix.
- . Use 'Stepwise method' to select variables and build the model.
- . Seeking for the violations of assumptions:
 - Use 'Partial Regression Plot (PRP) of residuals' for each of the independent variables to check the linearity relationship between Y and X_i .
 - Use PRP to check influential points: a point far from the other points is influential.
 - Use 'Normality check' to check if Y's are normally distributed on each of the X values.
 - Use PRP to check the equality of variance; if the spread of the residuals increases or decreases with values of the independent variables or with the predicted values, the equality of variance assumption is questionable.
 - Use 'Autocorrelation test' to check if the Y's are autocorrelated.
 - Check multicollinearity: in each step of stepwise regression, check if there is any tolerance of variables in the equation that is less than the given tolerance (default tolerance: 0.01); if yes, the new variable should not be added.
- . Approaches to the violations:
 - The same as the approaches used in bivariate regression;
 - The Ridge regression method can be used to deal with the multicollinearity problem.
- . Other diagnostics: the same as those used in the bivariate regression.
- . Draw inferences from sample to population:
 - Use 'F-test' to test the significance of the regression.
 - Use 't-test' to test the significance of each of the independent variables in the regression.
 - Calculate the standard deviation of the residuals, and construct the confidence interval if required.

Result Interpretation	<ul style="list-style-type: none"> . Correlation matrix: shows the correlation between all pairs of variables under study. . R^2 -- coefficient of determination; an overall measure of the success of the regression in predicting Y from Xs; its square root is the multiple correlation coefficient; it is also the proportion of the variation in the dependent variable 'explained' by the model. . Adjusted R^2 -- corrected R^2 more closely reflecting the goodness of fit of the model in the population. . s -- standard deviation of residuals. . Partial regression coefficients -- each is a slope of the regression line for the independent variable, but does not reveal anything about the relative importance of the independent variables to the predicted variable, since they are not comparable in scales, also, they are affected by the correlations between the independent variables. . Standardized partial regression coefficients -- they have the same meaning as the regression coefficients, but are standardized to dimensionless values that are comparable (they do not suggest the relative importance since the effects of correlations between the independent variables exist). . Standard errors of coefficients -- the average errors of the estimated coefficients. . If p is smaller than the given α, H_0 is rejected, meaning that the regression is significant, or a certain coefficient is significant in the regression. . You can predict the Y values from X's values, if all assumptions are well met.
Graphic Representation	<ul style="list-style-type: none"> . Partial scatter plot; . Partial residual plot; . Simple line, Multiple line, Area line graph; . Simple projection line, Multiple projection line graph.
Limitations	<p>A fairly amount of knowledge concerned with both statistics and the variables under study is required to achieve the correct solutions.</p>

MULTIVARIATE RELATIONSHIP ANALYSIS**TABLE 19: FACTOR ANALYSIS**

ITEM	KNOWLEDGE
Statistical Issue	Multivariate analysis -- interdependence analysis
Statistical Method	Factor analysis
Goals	<ul style="list-style-type: none"> . To identify a set of descriptive concepts which summarize the relationships among the components of a system of interactive variables; . To determine the number of dimensions required to represent a set of variables.
Examples	Determine the characteristics of cities, from variables such as population size, income per capita, proportions of different occupations, educational levels, area per resident, proportion of income spent on housing, transportation network density, telecommunication facilities, etc.
How It's Done	Each variable is expressed as a linear combination of a small number of common factors, which are shared by all variables, and a unique factor that is specific to that variable. The correlations between the variables arise from the 'sharing' of the common factors. The common factors in turn are estimated as linear combinations of the original variables. To improve the interpretability of the factors, solutions can be rotated. A good factor solution is both simple and meaningful.
Measurement Scale	Interval, Ratio.
Assumptions	<ul style="list-style-type: none"> . Each variable is correlated to at least one other, and the observed correlations are due to the sharing of common factors; . The relationships are linear; . Residuals are non-systematic and independent of one another; . Multivariate normal distribution is required when conducting statistical tests.
Hypothesis	H ₀ : K is the true number of descriptive concepts (for maximum likelihood and generalized least square factor extraction methods).

Analysis Processes	<ul style="list-style-type: none"> . Specify a number of variables and organize a multivariate data set. . If the observations make up a population and the purpose of the analysis is to infer the common factors from selected variables to the universe of variables, then use 'Non-statistical estimation method'; if the observations make up a sample, and the purpose of the analyses is to infer the common factors from the sample to the population, then use 'Statistical method'. . Use 'Correlation matrix' to compute the correlations between all pairs of variables. . Check the applicability of factor analysis with the data: <ul style="list-style-type: none"> - Use 'Bartlett's test of sphericity to test the hypothesis that the correlation matrix is an identity matrix; if $p\text{-value} > \alpha$, the use of the factor analysis should be reconsidered. - Use 'KMO (Kaiser-Mayer-Okin) calculation' to measure the sampling adequacy: <ul style="list-style-type: none"> If $KMO \geq 0.9$ ---- marvelous for using FA If $0.8 < KMO < 0.9$ ---- meritorious If $0.7 < KMO < 0.8$ ---- middling If $0.6 < KMO < 0.7$ ---- mediocre If $0.5 < KMO < 0.6$ ---- miserable If $KMO < 0.5$ ---- cannot use FA - Use 'MSA (Measure Sampling Adequacy) calculation' to measure sampling adequacy for each variable; then use the same criteria as above to determine the elimination of a variable. . Factor extraction <ul style="list-style-type: none"> - Use 'PCA (Principle Component Analysis)' to get PCA table. - Decide how many factors to retain by using 'Eigen value 1 criterion' or by looking at the cumulative percentages of variables explained by factors on the PCA table. - Use 'Factor loading matrix' to compute factor loadings on each variable. - Compute 'Final statistics table'. - Compute reproduced correlation matrix and residuals from the original correlations. - You can also use one of the advanced extraction methods: Maximum Likelihood Method(MLM); Principal axis factoring; and Generalized Least Square Method(GLSM); Alpha method, etc.
--------------------	---

Analysis Processes (continued)	<ul style="list-style-type: none"> - If using MLM or GLSM, Chi-square test can be used to test the goodness of fit for the adequacy of a k-factor model; in this case, normality should be tested by using 'Normality test'. . Factor rotation <ul style="list-style-type: none"> - Use one of 'Orthogonal rotation' methods or oblique rotation methods to achieve a simple structure of the factors. - Use 'factor loading plot' to judge the effectiveness of the rotation: if the rotation has achieved a simple structure, clusters of the variables should occur near the ends of the axes and at their intersection. . Compute 'Factor pattern matrix' on which variables with high loadings on the same factor appear together, and small loadings are eliminated. . Compute 'Factor scores'.
Results	<ul style="list-style-type: none"> . K -- number of common factors . L -- Factor loading matrix . F -- Factor score matrix . Reproduced Correlation Matrix (RCM) and their residuals from the original correlations . Final statistical table
Result Interpretation	<ul style="list-style-type: none"> . K defines the number of descriptive concepts summarized from the studied variables of the system. . The factor loading matrix defines the K-dimensional space of the common factors found by the analysis. Each column in the matrix represents a common factor, and each cell in the column refers to one of the original variables. The larger absolute values suggest the corresponding variables are the main concept of the common factor. . The factor score matrix gives coordinates of the observations in the K-dimensional space defined by L. . RCM and residuals: RCM consists of the correlations between variables that are reproduced by using the estimated correlations between common factors and variables. Their residuals indicate how well the factor model fits the data. If they are large, the model should be reconsidered. . The final statistical table shows the communality of variables and the percentages of variance accounted for by each of the retained common factors. The communality means the proportion of variance explained by the common factors.

Graphic Representation	<ul style="list-style-type: none">. Scree plot;. Factor loading plot;. Retation plot.
Limitations	<ul style="list-style-type: none">. Variables used in the analysis should cover the domain of interest as much as possible. This requires a large amount of data capture.. The criteria for deciding on the retention of a number of factors should differ from applications, which requires more knowledge; the suggested criterion can usually work well, but sometimes may delete important factors.. Interpretation of factors needs cautious analysis.

UNIVARIATE CLASSIFICATION**TABLE 20: CLASSIFICATION BY DISTRIBUTION**

ITEM	KNOWLEDGE
Statistical Issue	Univariate classification
Statistical Method	Equal step, equal frequency, histogram.
Goals	To classify univariate data into several classes according to the data distribution.
Examples	Group population density into several classes.
How It's Done	<p>. Analyze the distribution of the data set using goodness of fit techniques. For data sets that follow uniform or normal distributions, equal step or equal frequency methods are appropriate and efficient. The data set is separated evenly either on class frequencies or on divisions.</p> <p>. Histogram method is a graphic approach, based on the frequency distribution.</p>
Measurement Scale	Interval, Ratio.
Assumptions	None
Hypothesis	Ho: there is no significant difference between the given distribution and the classified frequency distribution.
Analysis Processes	<p>. Take a data set to be classified.</p> <p>. Set Ho and significant level α (0.05 - 0.1).</p> <p>. Use 'Normality test' to check if the distribution is normal; or use 'Uniform test' (Kolmogorov-Smirnov test) to check if it follows uniform distribution. If the p-value is smaller than the given α, Ho is rejected, meaning it is not normal or not uniform.</p> <p>. Specify the number of classes.</p> <p>. For data sets neither normal nor uniform, use 'Histogram' to classify the data visually, or use procedural methods as indicated in the next table. Histogram should be used so that low frequency points are the candidates of divisions.</p> <p>. For a data set with a normal or uniform distribution, use 'Equal step method', if you want a classification with equal steps; use 'Equal frequency method' if you want every class to have the same observations.</p> <p>. Adjust or round the classes as required.</p>

Result Interpretation	<ul style="list-style-type: none"> . If the distribution is uniform, both the equal step and equal frequency methods generate a set of uniform distribution classes. . If the distribution is normal, the equal step method generates a set of normal distribution classes; while the equal frequency method results in a set of uniform distribution classes.
Graphic Representation	<ul style="list-style-type: none"> . Histogram . Choropleth map . Classified point symbol map . Prism map
Priority Factor	1
Limitations	<ul style="list-style-type: none"> . Data are rarely uniform or normal, therefore the 'equal' methods have their limitations, though they are efficient and straightforward. . Histogram method is subjective.

*UNIVARIATE CLASSIFICATION***TABLE 21: CLASSIFICATION BY PROCEDURES**

ITEM	KNOWLEDGE
Statistical Issue	Univariate classification
Statistical Method	Jenks' procedure, Monmonier's procedure
Goals	To classify univariate data into several classes so that data in the same class are the most similar and those in different classes are the least similar.
Examples	Group housing costs into several classes.
How It's Done	First, conduct reiterative cycling: arbitrarily arrange data into the desired number of groups; calculate an index that measures the accuracy of the classification (e.g. Tabular Accuracy Index (TAI), Overview Accuracy Index (OAI), variance, etc.; cf. Currie, 1989); group the data with the nearest means, and recalculate the measure; compare the index with the original one in order to select the best classification; repeat the procedure until the index and the classes repeat themselves. Then, carry out force cycling: force the first observation in the second class into the first class, calculate the index, compare it with the original one, and keep the best one as well as the classes; if it is better than the original, force the second value of the class into the first class, until the index gets worse; conduct this process in all classes until the end; then, proceed the same way from the last group to the first group, until the best index is obtained.
Measurement Scale	Interval, Ratio
Assumptions	None
Hypothesis	None
Analysis Processes	<ul style="list-style-type: none"> . Select a data set to be classified. . Choose Jenks' or Monmonie's procedure. . Specify the number of classes. . Adjust or round the divisions as required.
Result Interpretation	The classes have the characteristic that data in the same class are more similar than those in different classes.

Graphic Representation	<ul style="list-style-type: none">. Histogram. Choropleth map. Classified point symbol map. Prism map
Priority Factor	2
Limitations	The iterative methods can work well in most situations, and are hence regarded as the best approaches for univariate data classification. However, they usually take a longer time to complete than other methods.

MULTIVARIATE CLASSIFICATION**TABLE 22: CLUSTER ANALYSIS**

ITEM	KNOWLEDGE
Statistical Issue	Classification
Statistical Method	Agglomerative Cluster analysis
Goals	<p>. To identify relatively homogeneous groups of cases based on selected attributes (case clustering).</p> <p>To classify variables into relatively homogeneous groups (variable clustering).</p>
Examples	<p>. Group cities into homogeneous clusters so that comparable cities can be selected to test various marketing strategies.</p> <p>. Group air quality indices into groups so that air quality can be described based on its major characteristics.</p>
How it's done	<p>Measures that compute the similarity or distance between all pairs of objects are computed. For case clustering, the objects are cases; for variable clustering, they are variables. Based on these, similar objects are grouped into clusters using one of several criteria for cluster formation: single linkage, complete linkage, or average linkage. Single linkage means that the minimum distance between two clusters is used as the clustering criterion; complete linkage deals with the maximum distance between two clusters; average linkage computes the distance between the average values of clusters. The combination proceeds step by step until all objects are merged into a single group.</p>
Measurement scale	Ordinal, Interval, Ratio, Binary
Assumptions	There is significant difference among means of groups
Hypotheses	Ho: there is no significant difference among means of groups
Analysis processes	<p>. Use appropriate spatial sampling schemes to get a set of multivariate data.</p> <p>. Choose a 'Standardization' function to standardize the variables.</p> <p>. Specify clustering requirement: case clustering or variable clustering.</p> <p>. Choose a 'Proximity' function to compute a distance matrix or a similarity measure matrix.</p> <p>. Choose 'Clustering procedure', specify a clustering criterion, and obtain clustering results</p> <p>. Set Ho and significant level α, use 'ANOVA' to test the hypothesis.</p>

<p>Result Interpretations</p>	<ul style="list-style-type: none"> . Clustering tree (Dendrogram) -- a step by step grouped graph. . Icicle plot -- showing the hierarchical clustering procedure. . Cluster membership at different stages -- showing the memberships of the clustered cases at several potentially acceptable clustering stages. . Cluster characteristics -- showing the means and standard deviations for various variables when a certain number of clusters is chosen. . The results can be similarly interpreted for variable clustering. . If p-value in MANOVA table is smaller than the given α, H_0 is rejected and the classification makes sense; otherwise, it is questionable.
<p>Graphic representations</p>	<ul style="list-style-type: none"> . Dendrogram . Icicle plot . Scatter plot for 2 or 3 dimensional cases . Nominal map . Choropleth map, Prism map
<p>Limitations</p>	<ul style="list-style-type: none"> . Cluster methods are application sensitive (Mather 1976, p310) . but no literature has been found which gives detailed illustrations on the suitability of the various clustering methods. . If the analysis is based on overlaid polygons, there may be error polygons and the classifications for these polygons would be meaningless.

MULTIVARIATE CLASSIFICATION**TABLE 23: DISCRIMINANT ANALYSIS**

ITEM	KNOWLEDGE
Statistical Issue	Multivariate classification
Statistical Method	Discriminant analysis
Goals	<ul style="list-style-type: none"> . To classify cases into one of several mutually exclusive groups on the basis of various characteristics. . To establish which characteristics are important in distinguishing the groups. . To evaluate the accuracy of the classification.
Examples	<ul style="list-style-type: none"> . Classify cities into exclusive groups according to variables concerned with housing: area per resident, proportion of income spent on housing, income per capita, housing costs, renovation schemes, age of house structure, size of household, etc. . Identify which variables are important for distinguishing the groups. . Evaluate the accuracy of the classification according to preclassified cases.
How It's Done	Based on a set of cases for which group membership is known, linear (or other) combinations of the variables are formed and serve as the basis for assigning cases to groups. The coefficients for the combinations are so chosen that they result in the 'best' separation among the groups. The accuracy of the classification can be estimated by applying the model to cases for whom group membership is known and comparing predicted group membership to actual membership.
Measurement Scale	<ul style="list-style-type: none"> . Nominal or ordinal for grouping variable. . Linear discriminant function requires independent variables be interval or ratio, or dichotomous. . Nonparametric procedure can be used when independent variables are binary, or a mixture of continuous and discrete variables.
Assumptions	<p>Linear discriminant function assumes:</p> <ul style="list-style-type: none"> . The variables in each group are from multivariate normal distributions; . The covariance matrices for all groups are equal.
Hypothesis	Ho: there is no significant difference in the classified group means.

Analysis Processes	<ul style="list-style-type: none"> . Specify grouping variables (housing class in the above example) and independent variables (all other variables in the example). . Draw a multivariate sample. . Select an appropriate discriminant procedure to build the discriminant function(s): <ul style="list-style-type: none"> 1) check the measure of independent variables: <ul style="list-style-type: none"> if the independent variables are all binary variables, or a mixture of continuous and discrete variables, use non-parametric procedure and go to Step 3); in all other cases, go to Step 2) 2) consider the distributions: <ul style="list-style-type: none"> - use 'Normal probability plot' for each independent variable to inspect normality; if anyone shows great departure from linear distribution (non-normal), then there is a reason to suspect the multivariate normality. - use 'Box'M' to test the equality of the group covariance matrix; a small significance probability suggests rejecting the null hypothesis that the covariance matrices are equal. - if the covariance matrices are unequal, the joint distribution of the variables is multivariate normal, and the sample size is large (>30), then the optimum classification rule is the quadratic discriminant function; if the matrices are not too dissimilar, the normality is met, and the sample size is small (<30), then the linear discriminant function is appropriate (CF 80). 3) choose a variable selection method (one of stepwise methods). . Classify the cases. . Evaluate classification accuracy: <ul style="list-style-type: none"> - check the correctly classified rate shown in the 'Classification table'; - use 'Histogram' to show the discriminant score distribution for two-group classification analysis; use 'Territorial map' to show the discriminant score distribution for three-group classification analysis; - use additional statistics: <ul style="list-style-type: none"> ANOVA for discriminant function(s) Eigenvalues Canonical correlation coefficients Wilks' Lambda test for discriminant function(s)
--------------------	---

Graphic Representations	<ul style="list-style-type: none"> . Histogram, scatter plot, normality probability plot, territorial map . Nominal map . Choropleth map, prism map
Result Interpretations	<ul style="list-style-type: none"> . Summary table for variable selection: <ul style="list-style-type: none"> - shows the selected variables that are most important for distinguishing the cases. - a smaller Wilks' Lambda value suggests a larger difference between group means; the values in the table mean that when the corresponding variable is taken out, how different the group means of the discriminant function will be, so the larger the value is, the more important the variable is. - the significance probability is the smallest value to reject the null hypothesis that the group means are the same; in the table, the larger the value is, the more important the corresponding variable is. . Discriminant function coefficients: <ul style="list-style-type: none"> - can be used as indicators of the relative importance of variables in the discriminant function when the coefficients are standardized; but note that they do not show the direct relationship between each of them and the function because the variables are correlated. - the sign of coefficients do not mean positive or negative correlation. . Classification output: <ul style="list-style-type: none"> - the sequence number is the case number. - the actual group is the group the case belongs to (prior knowledge). - the highest group is the most likely group that the case is assigned to by the function. - $P(G_i/D)$ is the probability that the case with a discriminant score of D belongs to group i; a larger value suggests higher certainty of the assignment. - the second group as well as the rest groups are probable classifications of the cases with less and less certainty. - the discriminant score is the predicted value using the discriminant function for the corresponding case. . Classification table: <ul style="list-style-type: none"> - shows how many cases are correctly or incorrectly classified for each of the groups.

<p>Result Interpretations (continued)</p>	<ul style="list-style-type: none"> - shows the correctly classified rate for each of the groups and for the whole cases (percentages). . ANOVA table: <ul style="list-style-type: none"> - to test the null hypothesis that the function means for all groups are the same; if the significance probability is smaller than your acceptable significant level α, the hypothesis is rejected, indicating the effectiveness of the discriminant function(s). Wilks' Lambda test: <ul style="list-style-type: none"> - Wilks' Lambda test: another way to test the hypothesis about the effectiveness of the discriminant function(s); if the significance probability is smaller than the acceptable α, H_0 is rejected, indicating the effectiveness of the function(s) - this test does not require a normal distribution of the discriminant scores, and is thus more useful. . Histogram for two-group classifications; discriminant function scatter plot for three(four)-group classification: <ul style="list-style-type: none"> - graphically shows the distributions of the discriminant scores for all studied cases. Territorial maps for three-group classification: <ul style="list-style-type: none"> - shows the ranges of different groups in the discriminant function space.
<p>Limitations</p>	<ul style="list-style-type: none"> . If the analysis is based on overlaid polygons, there may be error polygons and the classifications for these polygons would be meaningless. . In covariance matrices equality test, the Box's M test is sensitive to non-normality, and when the sample size is large, the significance probability may be small even if the matrices are not too dissimilar. . The selection of an appropriate method is somewhat ambiguous because the extent to which the assumptions can be violated is not always certain.

RANDOMNESS TEST**TABLE 24: AUTOCORRELATION ANALYSIS -- JOINT-COUNT STATISTIC**

ITEM	KNOWLEDGE
Statistical Issue	Randomness test
Statistical Method	Autocorrelation analysis -- joint-count statistic
Goals	To analyze whether a spatial pattern is randomly distributed, or spatially autocorrelated.
Examples	Are soil types spatially correlated or not.
How it's done	A joint-count statistic is constructed. The basic idea is to assign a distinct color to each of the types, and to count the total number of joins with the same colors. The count statistics are asymptotically normally distributed. Thus, z-score can be used to test the null hypothesis that there is no autocorrelation.
Measurement scale	Nominal
Assumptions	<ul style="list-style-type: none"> . Free sampling assumption: assume the spatial objects are independently coded; or . Non-free sampling assumption: each spatial object has the same probability of having a certain code.
Hypotheses	Ho: there is no autocorrelation in the spatial pattern.
Analysis processes	<ul style="list-style-type: none"> . Organize a data set as a sample. . Set Ho and significant α (0.01 - 0.1). . Use 'Joint-count statistic' under either free-sampling assumption or non-free sampling assumption. . Choose 'Binary matrix' as weight matrix, or define a weight matrix according to your prior knowledge.
Result Interpretations	<ul style="list-style-type: none"> . P-value -- the smallest probability to reject Ho; If $p < \alpha$, Ho is rejected, meaning that there is significant autocorrelation in the spatial pattern. . z -- standard score; if z is negative, a larger absolute value of z suggests a larger positive autocorrelation (less variability); if z is positive, a larger z value suggests a stronger negative autocorrelation (high diversity).
Graphic representations	If a series of spatial lags are applied in the construction of the proximity weight matrix, correlogram can be used to inspect the spatial process.

Limitations	<ul style="list-style-type: none">. When sample size < 20, the approximation of z-test is dubious.. When weight matrix is used, care must be used in the choice of weights if spurious correlations are to be avoided; sufficient prior knowledge on the analysis is desirable. Otherwise, the results may be misleading.. Type I or Type II error exists.
-------------	--

RANDOMNESS TEST**TABLE 25: AUTOCORRELATION ANALYSIS -- MORAN'S I STATISTIC**

ITEM	KNOWLEDGE
Statistical Issue	Randomness test
Statistical Method	Autocorrelation analysis - Moran's I statistic
Goals	To analyze whether there is spatial autocorrelation in a spatial distribution.
Examples	Is income per capita spatially correlated?
How it's done	Similar to Pearson's product moment correlation coefficient, Moran's I statistic also contains covariance in its numerator, and the measure of variance in its denominator. The difference is that Moran's I statistic additionally takes into account the spatial relationships between objects, which is expressed as a matrix representing a certain kind of proximity, such as connection, sharing a boundary, transportation convenience, etc. This statistic is also asymptotically normally distributed as sample size increases, therefore, the normal standard score z is applied to test the significance of the correlation.
Measurement scale	Ordinal, Interval, Ratio
Assumptions	<ul style="list-style-type: none"> . R assumption (assumption of randomization): the given data set is a random sample of a population, which consists of all possible permutations of the data in spatial arrangement; or . N assumption (assumption of normality): consider a map pattern as being the joint realization of n random variables, and the values of the variables are assumed to be multivariate normal and identically distributed.
Hypotheses	H_0 : there is no significant spatial autocorrelation in the spatial distribution.
Analysis processes	<ul style="list-style-type: none"> . Organize a data set as a sample. . Use 'Normality test' to check if the data are normally distributed. . Set H_0 and α (0.01 - 0.1). . Use 'Moran's I statistic': if normal, choose 'Normality Assumption'; if non-normal, choose 'Randomization Assumption'. . Choose 'Binary matrix' as a weight matrix, or define weight matrix according to your prior knowledge.

Result Interpretations	<p>. P-value: if $p < \alpha$, H_0 is rejected, meaning that there is significant autocorrelation in the spatial distribution.</p> <p>. z -- if z is negative, the larger the absolute value of z is, the stronger the positive autocorrelation the distribution has (less spatial variability); if z is positive, a larger z value suggests a stronger negative autocorrelation (high diversity).</p> <p>. I -- I $\rightarrow$ 1 suggests positive autocorrelation; I $\rightarrow$ -1 suggests negative autocorrelation.</p>
Graphic representations	<p>If I statistic is evaluated for serial weights which are assigned for varied spatial lags, then a Spatial Correlogram can be made for inspection of the variation in the spatial pattern.</p>
Limitations	<p>. If a weight matrix is to be defined, care must be taken in the choice of weights, and sufficient prior knowledge is desirable; otherwise, it may be misleading.</p> <p>. Type I or Type II error exists.</p>

RANDOMNESS TEST**TABLE 26: POINT PATTERN RANDOMNESS TEST**

ITEM	KNOWLEDGE
Statistical Issue	Randomness test
Statistical Method	Nearest neighbor distance method
Goals	To test if a mapped point pattern is uniform (Poisson process)
Examples	To analyze the spatial distribution of weather stations in Southern Canada
How it's done	Calculate the distances for selected objects from their nearest neighbors. Estimate the intensity of the point pattern. Use the intensity and the sum of the nearest neighbour distances to construct a test statistic whose sampling distribution is approximately normal. Use z deviate to test the randomness hypothesis.
Measurement Scale	Any scale, but treated as nominal.
Assumptions	A random sample should be drawn for the estimation of the intensity.
Hypothesis	Ho: the points are randomly distributed, following the spatial Poisson process
Analysis Processes	<ul style="list-style-type: none"> . Organize a set of points. . Draw a random sample from the data set. . Use 'Intensity estimation' to get the estimate of the intensity of the data set. . Use 'Distance method' to calculate the nearest neighbor distances. . Set Ho and α (0.01 - 0.1). . Use 'Clark-Evans test'.
Result Interpretation	If the p-value is smaller than α , Ho is rejected, meaning that the point pattern does not significantly follow a Poisson distribution.
Graphic Representation	None has been found.
Limitations	<ul style="list-style-type: none"> . The point pattern randomness test is still a research topic. A variety of test statistics are available, but every one has its drawbacks. The major difficulty is that a mapped point pattern is not a random sample; considering only a random sample may cause the loss of information on spatial structures. The Clark-Evans test is used here, since it is recommended in Cliff and Ord (1980, p112). . Type I error or Type II error exists.

SPATIAL SURFACE ANALYSIS**TABLE 27: KRIGING**

ITEM	KNOWLEDGE
Statistical Issue	Spatial surface analysis
Statistical Method	Kriging
Goals	<p>. To infer values of a theme at particular spatial locations between sampling points.</p> <p>. To convert spatial point data into an isarithmic map.</p>
Examples	Infer the air index value for Ottawa
How it's done	<p>Given a spatial lag (distance) and direction, a variogram (semi-variances) is estimated, based on the assumption of a constant local mean and a stationary variance of differences between places separated by a given distance and direction. The variogram contains useful information on the spatial variation of a property. The sample variogram is modeled so as to construct a continuous variogram. Spatial interpolation is then carried out by means of kriging method, in which the modeled variogram is used to compute weights concerned with the interpolation. With further tracing of the spatial values, which are either interpolated or original, an isarithmic map can be created. Given a spatial location, the value of the property at the location can be inferred.</p>
Measurement scale	Interval, Ratio.
Assumptions	The means and variances of differences between places separated by a given distance and direction are independent of location.
Hypotheses	None

Analysis processes	<ul style="list-style-type: none"> . Draw a sample covering the studied area. . Use 'Variogram' to compute a variogram as well as a variogram model for the data set. . Examine the shape of the variogram: if it appears to be horizontal (wholly nugget), then the data totally lack spatial dependence, so there is no point in attempting interpolation. . Otherwise, use 'Kriging' to interpolate the surface and to compute the kriging variance: if the variogram appears to rise in a concave upward form from the origin, then there exists a trend or 'drift' in the data, so use 'Universal kriging'; otherwise use 'Ordinary kriging'. . If the purpose of the analysis is to infer a value at particular points, give the locations of the points, and obtain the values; if the purpose is to draw a map, use 'Isarithmic map' to obtain the map.
Result Interpretations	<ul style="list-style-type: none"> . The inferred values are the optimal estimates of the true values at those points. They are unbiased, and have known minimum variances. . The isarithmic map displays the spatial surface that is the optimal estimate of the true surface, with minimum variances and no bias.
Graphic representations	<ul style="list-style-type: none"> . Isarithmic map . Perspective diagram
Limitations	<ul style="list-style-type: none"> . The technique needs relatively sophisticated knowledge. . Variogram modeling is still a research topic.

*SPATIAL SURFACE ANALYSIS***TABLE 28: TREND SURFACE ANALYSIS**

ITEM	KNOWLEDGE
Statistical Issue	Spatial surface analysis
Statistical Method	Trend surface analysis
Goals	To inspect the spatial variation trend of a spatial phenomenon.
Examples	What is the spatial variation of air quality over Canada?
How it's done	Formulate a spatial trend model (e.g. a quadratic model). The coefficient of the models are chosen so that the sum of the squared differences between the observed and predicted values of the dependent variable based on the model is as small as possible.
Measurement scale	Interval, Ratio
Assumptions	. The data should cover the area of interest; . The variable should be spatially continuous.
Hypotheses	None
Analysis processes	. Draw a sample covering the whole area of interest. . 'Use trend surface analysis' to compute the surface model, and calculate the values by grids over the interested area. . 'Use isarithmic map' or 'Perspective diagram' to display the surface.
Results and Interpretations	. Trend surface model -- the best fit of the spatial surface based on the sample. . Isarithmic map -- shows the spatial variation of the phenomenon (with low accuracy, hence not suitable to be used as a formal map)
Graphic representations	. Isarithmic map . Perspective diagram
Limitations	The model can only be used for the inspection of the overall variation trend of the phenomenon. It is not suitable for inferring local values, nor for making a formal isarithmic map.

APPENDIX 2: PRELIMINARY SAMPLING KNOWLEDGE

- Mode (1) -** collect attribute values of point features in specified regions.
- Mode (2) -** calculate the distance from a selected point feature to the nearest point feature.
- Mode (3) -** calculate a binary matrix of point features, by firstly using the Thiessen polygon method (Griffith, 1987) to transform the points to polygons, and then using Mode (7) to get the matrix (see Mode (7)).
- Mode (4) -** calculate line density (length of a linear feature per unit area) in specified regions.
- Mode (5) -** obtain total lengths of a linear feature in specified regions.
- Mode (6) -** collect attribute values of areal features in specified regions.
- Mode (7) -** calculate a binary matrix of areal features. The binary matrix measures the spatial proximity; each element of the matrix has either value 1, if the corresponding pair of areas share the same boundary, or value 0, if they do not.
- Mode (8) -** collect attribute values of volume features from points, e.g. obtain temperature data by weather stations.
- Mode (9) -** collect attribute values of point features from the points, e.g. capture the employment growth rates of cities by the cities.
- Mode (10) -** calculate distances from the points of a principal theme to the points of minor themes, e.g. from cities to power stations.
- Mode (11) -** calculate distances from points to lines, e.g. from cities to railways.
- Mode (12) -** count the number of lines connected to a point feature, e.g. the number of airways of a city.
- Mode (13) -** collect attribute values of areal features from points, e.g. soil types by weather stations.
- Mode (14) -** calculate point densities (number of points per unit area) in specified regions.

Rule-1: IF the analysis is univariate,
and the spatial dimension is 0-cell,
and the statistical method is summary,
THEN the sampling base is the user specified region,
and the sampling scheme is Mode (1).

Rule-2: IF the analysis is univariate,

and the spatial dimension is 0-cell,
and the statistical method is the nearest neighbor distance method,
THEN the sampling base is the theme,
and the sampling scheme is Mode (2).

- Rule-3:** IF the analysis is univariate,
and the spatial dimension is 0-cell,
and the statistical method is autocorrelation analysis,
THEN the sampling base is the theme,
and the sampling scheme is Mode (3).
- Rule-4:** IF the analysis is univariate,
and the spatial dimension is 1-cell,
and the analytical measure is density,
THEN the sampling base is the user specified region,
and the sampling scheme is Mode (4).
- Rule-5:** IF the analysis is univariate,
and spatial dimension is 1-cell,
and the analytical measure is length,
THEN the sampling base is the user specified region,
and the sampling scheme is Mode (5).
- Rule-6:** IF the analysis is univariate,
and the spatial dimension is 2-cell,
and the statistical method is summary or equality test,
THEN the sampling base is the 2-cell regions,
and the sampling scheme is Mode (6).
- Rule-7:** IF the analysis is univariate,
and the spatial dimension is 2-cell,
and the statistical method is autocorrelation analysis,
THEN the sampling base is the 2-cell regions,
and the sampling scheme is Mode (7).

Rule-8: IF the analysis is bivariate,

and the spatial dimension case is 0-0 cells,
 and only one of the themes has 3-cell spatial distribution,
 THEN transfer the theme THM1 with 3-cell spatial distribution into 3-cell spatial dimension,
 and the other theme THM2 is used as the sampling base,
 and the sampling scheme for THM1 is Mode (8),
 and the sampling scheme for THM2 is Mode (9).

Rule-9: IF the analysis is bivariate,
 and the spatial dimension case is 0-0 cells,
 and both the analytical themes have 3-cell spatial distribution,
 and the two sets of points fit each other,
 THEN the sampling base is the point sets,
 and the sampling scheme for each theme is Mode (9)

Rule-10: IF the analysis is bivariate,
 and the spatial dimension case is 0-0 cells,
 and both themes have 3-cell spatial distribution,
 and the two sets of points do not fit each other,
 THEN transfer one of the theme THM1 into 3-cell spatial dimension,
 and use the other theme THM2 as the sampling base,
 and the sampling scheme for THM1 is Mode (8),
 and the sampling scheme for THM2 is Mode (9).

Rule-11: IF the analysis is bivariate,
 and the spatial dimension case is 0-0 cells,
 and none of the themes has 3-cell spatial distribution,
 and the principal analytical theme is THM1,
 and the measure of the minor analytical theme is distance,
 THEN the sampling base is THM1,
 and the sampling scheme for THM1 is Mode (9),
 and the sampling scheme for the minor analytical theme is Mode (10).

Rule-12: IF the analysis is bivariate,
 and the spatial dimension case is 0-1 cells,
 and the spatial distribution of the 0-cell theme is 0-cell,

and analytical measures include distance,
THEN the sampling base is the 0-cell theme,
and the sampling scheme for the 0-cell theme is Mode (9),
and the sampling scheme for the 1-cell theme is Mode (11).

Rule-13: IF the analysis is bivariate,
and spatial dimension case is 0-1 cells,
and the spatial distribution of the 0-cell theme is 0-cell,
and analytical measures include the number of lines connected to points,
THEN the sampling base is the 0-cell theme,
and the sampling scheme for the 0-cell theme is Mode (9),
and the sampling scheme for the 1-cell theme is Mode (12).

Rule-14: IF the analysis is bivariate,
and spatial dimension case is 0-2 cells,
and the 0-cell theme has 3-cell spatial distribution or no density is involved
in the analytical measures,
THEN the sampling base is the 0-cell theme,
and the sampling scheme for the 0-cell theme is Mode (9),
and the sampling scheme for the 2-cell theme is Mode (13).

Rule-15: IF the analysis is bivariate,
and the spatial dimension case is 0-2 cells,
and the 0-cell theme does not have 3-cell spatial distribution,
and analytical measures include density,
THEN the sampling base is the 2-cell theme,
and the sampling scheme for the 0-cell theme is Mode (14),
and the sampling scheme for the 2-cell theme is Mode (6).

Rule-16: IF the analysis is bivariate,
and the spatial dimension case is 1-1 cells,
and analytical measure is density,
THEN the sampling base is the user specified region,
and the sampling scheme for each of the themes is Mode (4).

Rule-17: IF the analysis is bivariate,

and the spatial dimension case is 1-1 cells,
 and analytical measure is length,
 THEN the sampling base is the user specified region,
 and the sampling scheme for each of the themes is Mode (5).

Rule-18: IF the analysis is bivariate,
 and the spatial dimension case is 1-2 cells,
 and analytical measures include density,
 THEN the sampling base is the 2-cell regions,
 and the sampling scheme for the 1-cell theme is Mode (4),
 and the sampling scheme for the 2-cell theme is Mode (6).

Rule-19: IF the analysis is bivariate,
 and the spatial dimension case is 1-2 cells,
 and analytical measures include length,
 THEN the sampling base is the 2-cell regions,
 and the sampling scheme for the 1-cell theme is Mode (5),
 and the sampling scheme for the 2-cell theme is Mode (6).

Rule-20: IF the analysis is bivariate,
 and the spatial dimension case is 2-2 cells,
 and the two sets of polygons fit each other,
 THEN the sampling base is the 2-cell regions,
 and the sampling scheme for each of the themes is Mode (6).

Rule-21: IF the analysis is bivariate,
 and the spatial dimension case is 2-2 cells,
 and the two sets of polygons do not fit each other,
 and the statistical issue is not classification,
 THEN generate a set of random spatial sampling points as the sampling base,
 and the sampling scheme for each of the themes is Mode (13).

Rule-22: IF the analysis is bivariate,
 and the spatial dimension case is 2-2 cells,

and the two sets of polygons do not fit each other,
 and the statistical issue is classification,
 THEN overlay the two maps,
 and use the overlaid polygons as the sampling base,
 and the sampling scheme for each of the themes is Mode (14).

Rule-23: IF the analysis is multivariate,
 and one and only one of the analytical themes has 0-cell spatial dimension,
 THEN the 0-cell theme is used as the sampling base,
 and the sampling scheme for the 0-cell theme is Mode (9),
 and Rule-12 and Rule-13 apply to the 1-cell theme(s) involved in the
 analysis,
 and the sampling scheme for each of the 2-cell theme(s) involved is Mode
 (13).

Rule-24: IF the analysis is multivariate,
 and n themes have 0-cell spatial dimension,
 and $n > 1$,
 and the n point sets fit each other,
 THEN the 0-cell points are used as the sampling base,
 and the sampling scheme for each of the 0-cell themes is Mode (9),
 and Rule-12 and Rule-13 apply to the 1-cell theme(s) involved in the
 analysis,
 and the sampling scheme for each of the 2-cell theme(s) involved is Mode
 (13).

Rule-25: IF the analysis is multivariate,
 and n themes have 0-cell spatial dimension,
 and m theme(s) have 3-cell spatial distribution,
 and $n > m$,
 and the $(n - m)$ sets of points fit each other,
 THEN transfer the m themes with 3-cell spatial distribution into 3-cell
 spatial dimension,
 and use the $(n - m)$ 0-cell spatially distributed themes as the sampling base,
 and the sampling scheme for each of the $(n - m)$ themes is Mode (9),
 and the sampling scheme for each of the m theme(s) is Mode (8),

and Rule-12 and Rule-13 apply to the selection of sampling schemes for the 1-cell spatial dimension themes involved in the analysis,
and the sampling scheme for each of the 2-cell theme(s) involved is Mode (13).

Rule-26: IF the analysis is multivariate,
and n themes have 0-cell spatial dimension,
and m theme(s) have 3-cell spatial distribution,
and $n > m$,
and the $(n - m)$ sets of points do not fit each other,
and distances are required as the analytical measures between the $(n - m)$ sets of points,
and Principal Analytical Theme (PAT) is included in the $(n - m)$ themes,
THEN transfer the m themes with 3-cell spatial distribution into 3-cell spatial dimension,
and use the PAT as the sampling base,
and Rule-11 applies to the selection of sampling schemes for the $(n - m)$ 0-cell themes,
and the sampling scheme for each of the m 0-cell theme(s) is Mode (8),
and Rule-12 and Rule-13 apply to the selection of sampling schemes for the 1-cell spatial dimension themes involved in the analysis,
and the sampling scheme for each of the 2-cell theme(s) involved is Mode (13).

Rule-27: IF the analysis is multivariate,
and n theme(s) have 0-cell spatial dimension and 3-cell spatial distribution,
THEN transfer $n-1$ themes with 0-cell spatial dimension into 3-cell spatial dimension,
and use the remaining 0-cell theme THM as the sampling base,
and the sampling scheme for THM is Mode (9),
and the sampling scheme for each of the 3-cell theme(s) transformed from 0-cell dimension is Mode (8),
and Rule-12 and Rule-13 apply to the selection of sampling schemes for the 1-cell spatial dimension themes involved in the analysis,
and the sampling scheme for each of the 2-cell theme(s) involved is Mode (13).

- Rule-28:** IF the analysis is multivariate,
and no theme involved has 0-cell spatial dimension,
and one and only one theme has 2-cell spatial dimension,
THEN use the 2-cell theme as the sampling base,
and the sampling scheme for the 2-cell theme is Mode (6),
and Rule-18 and Rule-19 apply to the selection of sampling schemes for the
1-cell themes involved in the analysis.
- Rule-29:** IF the analysis is multivariate,
and no analytical theme has 0-cell spatial dimension,
and n themes have 2-cell spatial dimension,
and the n sets of 2-cell polygons fit each other,
THEN the sampling base is the 2-cell polygons,
and the sampling scheme for each of the 2-cell themes is Mode (6),
and Rule-18 and Rule-19 apply to the selection of sampling schemes for the
1-cell theme(s) involved in the analysis.
- Rule-30:** IF the analysis is multivariate,
and no analytical theme has 0-cell spatial dimension,
and n themes have 2-cell spatial dimension,
and the n sets of 2-cell polygons do not fit each other,
THEN overlay all the themes,
and use the overlay polygon set as the sampling base,
and the sampling scheme for each of the 2-cell themes is Mode (14),
and Rule-18 and Rule-19 apply to the selection of sampling schemes for the
1-cell theme(s) involved in the analysis.

APPENDIX 3: CASE DEMONSTRATIONS

Invoke EASSA by command: EASSA.

Screen [1]:

Welcome to EASSA --

The Expert Advisory System for Statistical Analysis for Electronic Mapping

Developed at Geographical Services Division

Energy, Mines, and Resources Canada

Ottawa, Canada

© Copyright *, Version 1.0**

-
- 1 - System introduction
 - 2 - Consultation
 - 3 - Case demonstration I - EASSA recommends analysis issues
 - 4 - Case demonstration II - user proposes analysis issues

Enter a **number** to make selection

Press '**E**' to exit

Enter '1' to get System introduction

Screen [2]:

You have entered the system introduction.

Introduction to EASSA

EASSA is an expert system for statistical analysis. It is capable of giving advice on statistical analysis in the environment of the Electronic Atlas of Canada (EA). It can also serve in other geographical information systems and general statistical consultations.

Six advisory modes are available in EASSA. You can ask EASSA to recommend geographical analysis issues, or, you can propose your analysis issues; you can conduct an analysis based on the maps in the Electronic Atlas, or, you can use your own maps for the analysis.

The system's responsibilities are: (1) to recommend statistical analysis issues that can be analyzed in EASSA's environment, the statistical methods suitable for specific analyses, and the sampling modes applicable to the analyses; (2) to advise on analysis procedures for carrying out an analysis, and on the interpretations of analytical results; (3) to give other relevant advice on statistical hypotheses, assumptions, graphic representation methods, and limitation of statistical methods.

The following statistical issues are involved: data descriptions, equality tests, bivariate relationship analysis, multivariate relationship analysis, univariate classifications, multivariate classifications, randomness test, and spatial surface analysis.

Statistics may sound difficult, but by simply following EASSA's instructions, you will find them easy to use!

Press 'B' to return to Screen [1]

press 'E' for exit

Return to Screen [1] and enter '3' to get into Case demonstration 1

CASE 1: RANDOMNESS TEST FOR LIFE EXPECTANCY

Screen [2]:

You have chosen Case demonstration 1

Please choose an advisory mode

Advisory Modes

Mode-1: EASSA recommends analysis issues

Mode-1-1: using the maps of the Electronic Atlas (EA)

Mode-1-2: using your maps

Mode-1-3: using both EA's maps and your maps

Enter a **combined-number** to choose a mode (e.g. 1-1)

Press 'E' to return to Screen [1]

Select Mode-1-1 by entering '1-1'

Screen [3]:

You have chosen Mode 1-1

Please choose a statistical issue

Statistical Issue Category

1 -- Data descriptions

2 -- Equality tests

3 -- Bivariate relationship analysis

4 -- Multivariate relationship analysis

5 -- Univariate classification

6 -- Multivariate classification

7 -- Randomness test

8 -- Spatial surface analysis

Enter '**WHAT**' followed with a number for the explanation of the corresponding statistical issue

Enter a **number** to make your choice

Press '**B**' to return to the previous screen

Press '**E**' to return to Screen [1]

Enter '**WHAT7**' to get the explanation of Randomness test

Screen [3']:

[Explanation]

On randomness test

Randomness test is devoted to the inspection of spatial processes of geographical phenomena. It tests if a mapped point pattern is uniform (Poisson process), or, whether there is spatial autocorrelation in a spatial distribution.

Enter '**EXAMPLE**' for showing examples of the analysis issue

Press '**B**' to return to the previous screen

Press '**E**' to return to Screen [1]

Enter 'EXAMPLE' to get examples of the analysis issue

Screen [3']:

[Explanation]

Examples of the analysis issues of randomness test

- . Are weather stations in Southern Canada randomly distributed?
- . Are income per capita spatially correlated?

Press 'B' to return to the previous screen

Press 'E' to return to Screen [1]

Return to Screen 3, and enter 7 to select Randomness test

Screen [4]:

You have chosen Randomness test

Please choose a category of geography

Geography Category

- 1 -- Physical and Environmental Geography
- 2 -- Social and Cultural Geography
- 3 -- Economic and Resource Geography

Enter a **number** to get a corresponding realm list

Press 'B' to return to the previous screen

Press 'E' to return to Screen [1]

Enter '2' to get the list of Social and Cultural geographical realms

Screen [5]:

You have chosen Social and Cultural Geography

Please choose a geographical realm under this category

Realms of social and cultural geography

11 -- Exploration	12 -- Settlement
13 -- Political Geography	14 -- Population
15 -- Ethnography	16 -- Languages
17 -- Migrations	18 -- Vital Statistics
19 -- Health and Welfare	20 -- Justice and Law
21 -- Culture	22 -- Religion
23 -- Education	38 -- Leisure
39 -- Urban Canada	

Enter a **number** to get a list of analytical themes of the realm

Press '**B**' to return to the previous screen

Press '**E**' to return to Screen [1]

Enter '14' to get the list of analytical themes

Screen [6]:

You have chosen the realm of Population

Please choose a theme for analysis

Analytical Themes for Randomness Test

-- Social and Cultural Geography: Health and Welfare --

1 -- Infant mortality

2 -- Life expectancy

Enter a **number** to select a theme

Press '**B**' to return to the previous screen

Press '**E**' to return to Screen [1]

Enter '2' to select Life expectancy for randomness test

Screen [7]:

[Advice] Analysis issue: randomness test for life expectancy distribution

Please input your analysis purpose

Purpose of the Analysis

1 -- to examine the variation of the analytical theme over space

2 -- to examine the distribution pattern of the analytical points.

Enter a **number** to give your analysis purpose

Press '**B**' for back to the previous screen

Press '**E**' to return to Screen [1]

Enter '1' to choose the variation

Screen [8].

Analysis issue: randomness test of life expectancy distribution

[Advice] Suggested statistical method: autocorrelation analysis - Moran's I statistic

Enter '**HOW**' for explanation of the selection of the method

Enter '**WHAT**' for explanation of the statistical method

Press '**Enter**' for further advice

Press '**B**' to return to the previous screen

Press '**E**' to return to Screen [1]

Enter 'HOW' to get explanation.

Screen [8]:

[Explanation]

How is the statistical method selected?

There are two candidate methods for the randomness test statistical issue: the nearest neighbor distance method, and the autocorrelation analysis method. Based on the analysis purpose, it is determined to use autocorrelation analysis, since the nearest neighbor distance method is used to examine the distribution pattern of analytical points, while the autocorrelation analysis deals with the variation of the analytical theme over space.

The system considers two candidate methods for autocorrelation analysis: joint-count statistic, and Moran's I statistic. The measurement scale is the major factor considered in the selection. The former is suitable for nominal data, while the latter for interval or ratio data. Since the life expectancy is measured on a ratio scale, Moran's I statistic is recommended.

Enter '**Measurement**' for the explanation of the measurement scale

Press '**B**' to return to the previous screen

Enter 'Measurement' to get the explanation of measurement scale

Screen [8"]

[Explanation]

What is measurement scale?

There are four types of measurement scales usually used in measuring geographical phenomena:

Nominal: labeling or numbering objects for identification;

Ordinal: objects are numbered from least to most or vice versa;

Interval: objects are arrayed in order of relative rank with known distance between the ranks;

Ratio: objects are arrayed in order of absolute rank with known distance between the ranks.

For further information, please refer to Dent (1985): Principles of Thematic Map Design.

Press 'Enter' for further advice

Press 'B' to return to the previous screen

Press 'E' to return to Screen [1]

Return to Screen [8], and enter 'WHAT' to get explanation about what the method means
Screen [8"]:

[Explanation]

What is autocorrelation analysis?

Autocorrelation refers to correlation between pairs of observations made on a single variable. If there is systematic spatial variation in the variate, then the phenomenon is said to exhibit spatial autocorrelation; otherwise, the phenomenon exhibits spatial randomness. The autocorrelation analysis is to use an appropriate statistic to ascertain the characteristic of spatial variation of phenomena.

For further explanation, please refer to Cliff and Ord (1981) or Goodchild (1986).

Press 'WHAT' for explanation of the Moran's I statistic

Press 'Enter' for further advice

Press 'B' to return to the previous screen

Press 'E' to return to Screen [1]

Press 'WHAT' to get further explanation of Moran's I statistic

Screen [8''']

*****'

[Explanation]

Moran's I statistic

Similar to Pearson's product moment correlation coefficient, which is the ratio of covariance to the square product of standard deviations, Moran's I statistic also contains covariance in its numerator, and the measure of variance in its denominator. The difference is that Moran's I statistic additionally takes into account the spatial relationships between objects, which is expressed as a matrix representing a certain kind of proximity, such as connection, sharing a boundary, transportation convenience, etc. This statistic is also asymptotically normally distributed as sample size increases, therefore, the normal standard deviate z is applied to test the significance of the correlation.

Press 'Enter' for further advice

Press 'B' to return to the previous screen

Press 'E' to return to Screen [1]

Press 'Enter' for further advice

Screen [9]:

Analysis issue: autocorrelation analysis for life expectancy distribution

Statistical method: Moran's I statistic

[Advice] Suggested sampling scheme: Mode (7)

Enter 'HOW' for the explanation of the scheme selection

Enter 'WHAT' for the explanation for the sampling scheme

Press 'Enter' for further advice

Press 'B' to return to the previous screen

Press 'E' to return to Screen [1]

Enter 'HOW' to get the explanation of the scheme selection

Screen [9]:

[Explanation]

How is the sampling scheme selected?

According to Sampling Rule-7:

IF the analysis is univariate,
and the spatial dimension is 2-cell (also called 2-dimension),
and the statistical method is autocorrelation analysis,
THEN the sampling base is the 2-cell regions,
and the sampling scheme is Mode (7).

Because the current analysis is univariate, and the spatial dimension of life expectancy is 2-cell, and the statistical method is autocorrelation, Rule-7 is applied.

Press 'Enter' for further advice

Press 'B' to return to the previous screen

Press 'E' to return to Screen [1]

Return to Screen [9] and enter 'WHAT' to get explanation for the sampling scheme

Screen [9]:

[Explanation]

What is the sampling scheme Mode (7)?

Over ten sampling modes are used to sample various spatial data sets. Sampling Mode (7) serves the calculation of a binary matrix of areal features. The binary matrix measures the spatial proximity; each element of the matrix has either value 1, if the corresponding pair of areas share the same boundary, or value 0, if they do not.

Press 'Enter' for further advice

Press 'B' to return to the previous screen

Press 'E' to return to Screen [1]

Press 'Enter' for further advice

Screen [10]:

Analysis issue: autocorrelation analysis for life expectancy distribution

Statistical method: Moran's I statistic

Sampling scheme: Mode (7)

[Advice] Null hypothesis H_0 : There is no autocorrelation in life expectancy distribution.

Alternative hypothesis H_a : Life expectancy distribution is spatially correlated.

Enter 'WHAT' for the explanation of the hypotheses

Press 'Enter' for further advice

Press 'B' to return to the previous screen

Press 'E' to return to Screen [1]

Enter 'WHAT' to get explanation

Screen [10]:

[Explanation]

What do hypotheses mean?

The major preoccupation of modern statistics is inferential statistics, which use sample information to draw conclusions about a population or several populations. Hypotheses, which are used in the inferences, are assertions about the values of population characteristics. The null hypothesis H_0 is initially assumed to be true, while the alternative hypothesis H_a specifies a claim that is contradictory to H_0 and is judge the more plausible claim when H_0 is rejected.

Press 'Enter' for further advice

Press 'B' to return to the previous screen

Press 'E' to return Screen [1]

Press 'Enter' to get further advice

Screen [11]:

Analysis issue: autocorrelation analysis for life expectancy distribution

Statistical method: Moran's I statistic

Sampling scheme: Mode (7)

Null hypothesis H_0 : There is no autocorrelation in life expectancy distribution.

Alternative hypothesis H_a : Life expectancy distribution is spatially correlated.

[Advice] Assumptions of autocorrelation analysis:

N Assumption: consider the map pattern as being the joint realization of n random variables in n areas, and the values of the variables are assumed to be multivariate normal and identically distributed; or

R assumption: treat the order of the observations in spatial arrangement is important, and all permutations of the order constitutes the population. The data to be studied is a random sample of the population.

Enter 'WHAT' for explanation of the assumptions

Press 'B' to return to the previous screen

Press 'E' to return to Screen [1]

Press 'WHAT' for explanation

Screen [11]:

[Explanation]

What do the assumptions mean?

An important factor in spatial autocorrelation analysis is to make statements in a larger context, by generalizing from the sample to the larger population. However, because of the way in which joint counts are taken, the set of objects used as a sample in measuring spatial autocorrelation is the entire set found within the study area boundary at some given point in time. Thus, "Great imagination has gone into turning what appears to be a population into a sample" (Upton and Fingleton, 1985). Consequently, two popular perspectives have been employed, namely, N Assumption and R Assumption.

N Assumption refers to resampling (or normality). It essentially has to do with stratified random sampling, where the areal units of a map represent a stratification, and data for each unit are treated as a random sample from a population in that areal unit.

R Assumption refers to randomization. It treats the order of arrangement of the data over a geographical surface as important, and then assumes that the observed map is a single sample drawn from the set of all possible data permutations for the given map.

For further explanation, please refer to Goodchild (1986), or Griffith (1987).

Press 'Enter' for further advice

Press 'B' to return to the previous screen

Press 'E' to return to Screen [1]

Press 'Enter' to get further advice

Screen [12]:

Analysis issue: autocorrelation analysis for life expectancy distribution

Ho: There is no autocorrelation in life expectancy distribution

Ha: Life expectancy distribution is spatially correlated

[Advice] Please follow the following procedure to conduct the analysis.

- 1) Use sampling scheme, Mode (7), in the EA to obtain the Life expectancy data.
- 2) Use 'Normality test' to examine if the sample distribution is normal.
- 3) If the data is approximately normal, use Moran's I statistic under N Assumption; otherwise use Moran's I statistic under R Assumption.
- 4) Choose 'Moran's I statistic' function in statistical software, specify the chosen assumption.
- 5) When requested, choose 'Binary Matrix', or define the Weight Matrix according to your *a priori* knowledge.

Press 'Enter' for interpretation of the results

- Press 'B' to return to the previous screen

Press 'E' to return to Screen [1]

Press 'Enter' to get the interpretation

Screen [13]:

Analysis issue: autocorrelation analysis for life expectancy distribution

Ho: There is no autocorrelation in life expectancy distribution

Ha: Life expectancy distribution is spatially correlated

[Advice]

Results: p -- the smallest probability to reject Ho

z -- normal deviate

I value -- the statistic

Please select a significant level α (0.01 -- 0.1)

Result interpretation:

(1) Compare the p value resulting from the analysis with the significant level α . If $\alpha \geq p$, reject Ho in favour of Ha, which means that there is positive spatial autocorrelation in the life expectancy distribution at a significant level; if $\alpha < p$, Ho can not be rejected, which means that there is no significant spatial autocorrelation at α significant level.

(2) If z is negative, the larger the absolute value of z, the more positive the autocorrelation is (less spatial variability); if z is positive, a larger z value suggests a stronger negative autocorrelation (high diversity).

(3) I --> 1 suggests positive autocorrelation; I --> -1 suggests negative autocorrelation.

For further information, please refer to Griffith (1987).

Enter 'WHY' for explanation of the significant level

Press 'B' to return to the previous screen

Press 'E' to return to Screen [1]

Enter 'WHY' to get the explanation of significant level
Screen [13]:

Why is a significant level needed?

In hypothesis testing, two errors may occur:

- Type I error: the error of rejecting H_0 when H_0 is true;
- Type II error: the error of not rejecting H_0 when H_0 is false.

The significant level is related to these errors. When you are giving a significant level, you are deciding the error range that you can accept. The two types of errors are mutually conflicting. When you decrease Type I error, you are increasing Type II error. A way to decrease both of the errors is to use a larger sample size. 0.05 is usually chosen as the significant level.

For further information, please refer to Jay and Poxy (1986).

Press 'Enter' for further advice

Press 'B' to return to the previous screen

Press 'E' to return to Screen [1]

Press 'Enter' for further advice

Screen [14]:

Analysis issue: autocorrelation analysis for life expectancy distribution

[Advice] Suggested Graphic representation method:

Areal maps can provide an overview for the spatial variation of life expectancy

Press 'Enter' for further advice

Press 'B' to return to the previous screen

Press 'E' to return to Screen [1]

Press 'Enter' to get further advice

Screen [14]:

Analysis issue: Autocorrelation analysis for life expectancy distribution

[Advice] Limitations:

(1) If weight matrix is defined, attention must be paid to the choice of weights if spurious correlations are to be avoided, and sufficient *a priori* knowledge is desirable; otherwise, it may be misleading. (For further information, reference to Cliff and Ord, 1981, p17).

(2) Type I or Type II error exists.

Press 'Enter' to continue

Enter 'WHAT' for the explanation of Type I and Type II errors

Press 'B' to return to the previous screen

Press 'E' to return to Screen [1]

Press 'Enter' to continue

Screen 15:

Advice ends

Thank you for consulting EASSA!

I hope it is helpful

If you have any comment, please contact the author at 992-4710

Press 'B' to return to the previous screen

Press 'E' to return to Screen [1]

Press 'Q' to exit

Press 'E' to return to Screen [1] and enter '4' to get into Case demonstration II.

Note: This case is intended to show the difference between advisory Mode-1-1 and Mode-2-2. The way to generate analysis issue is emphasized. The other aspects which apply to Case I are simplified.

CASE II: TWO SAMPLE TEST FOR AVERAGE INCOME PER CAPITA

Screen [2]:

You have entered the channel Case demonstration II

Please choose an advisory mode

ADVISORY MODES

Mode-2: user proposes analysis issues

Mode-2-1: using the maps of the Electronic Atlas (EA)

Mode-2-2: using your maps

Mode-2-3: using both EA's maps and your maps

Enter a **combined-number** to choose an advisory mode (e.g. 2-2)

Press 'E' to return to Screen [1]

Select Mode-2-2 by entering '2-2'

Screen [3]:

You have chosen Mode-2-2

Please choose a statistical issue

Statistical Issue Category

1 -- Data descriptions

2 -- Equality tests

3 -- Bivariate relationship analysis

4 -- Multivariate relationship analysis

5 -- Univariate classification

6 -- Multivariate classification

7 -- Randomness test for spatial distributions

8 -- Spatial surface analysis

Enter a **number** to make your choice

Press '**B**' to return to the previous screen

Press '**E**' to return to Screen [1]

Enter '3' to choose Equality test

Screen [4]:

You have chosen Equality test

Please choose an analysis purpose

Analysis Purposes of Equality Test

1--test equality of variable means

1-1 -- between two levels of objects

1-2 -- between two regions

1-3 -- between more than two levels of objects

1-4 -- between more than two regions

1-5 -- between different levels of objects in different regions

2--test equality of data variability

2-1 -- between two regions

2-2 -- between two levels of objects

3--test equality of frequency distribution between two data sets

4--test equality of proportions between two data sets

Enter a **combined-number** or a **number** to select an analysis purpose (e.g 2-2; 3)

Press 'B' to return to the previous screen

Press 'E' to return to Screen [1]

Enter '1-1' to select the equality test of variable means between two regions

Screen [5] (the italic type represents user input):

You have chosen equality test of variable means between two regions

[Input] Please input theme related information:

(1) Analytical theme name: *income per capita*

(2) The measurement scale (nominal/ordinal/interval/ratio): *ratio*

(3) The spatial dimension of the theme (0/1/2/3-cell): *2*

(4) The spatial distribution of the theme (0/1/2/3-cell): *2*

(5) The date when the data was surveyed (date/month/year): *03-06-91*

Enter 'WHAT' followed with a number to get explanation on the corresponding item

Press 'Enter' for further advice

Press 'B' to return to the previous screen

Press 'E' to return to Screen [1]

Enter 'WHAT3' to get explanation of spatial dimension

Screen [5]:

[Explanation]

On spatial dimension

In a broad sense, spatial dimension refers to the spatial form in which geographical phenomena appear. Usually, the spatial forms are summarized into four types: point (0-cell), line (1-cell), area (2-cell), and volume (3-cell).

In the EASSA context, the spatial dimension refers to the spatial dimension of phenomena modeled in maps or system data bases. The four types of dimensions apply to the context.

The spatial dimension needs to be input because it is a factor used to derive advice.

Press 'Enter' for further advice

Press 'B' to return

Press 'E' to return to Screen [1]

Return to screen [5], and enter 'WHAT4' for explanation of spatial distribution
Screen [5]:

[Explanation]

On spatial distribution

Spatial distribution means the spatial form in which geographical phenomena are distributed over space. They are summarized into four types: point (0-cell), line (1-cell), area (2-cell), and volume (3-cell).

The difference between spatial distribution and spatial dimension referred to in EASSA is that the spatial distribution refers to the spatial form of phenomena in the real world, while spatial dimension refers to the spatial form of phenomena modeled in maps or in system data bases. It should be noticed that measures of spatial forms of phenomena in the real world are relative. The judgement on whether they are 0-cell or 1-cell or 2-cell should be made based on the spatial scale under study.

Press 'Enter' for further advice

Press 'B' to return to the previous screen

Press 'E' to return to Screen [1]

Press 'Enter' for further advice

Screen [6]:

Analysis issue: test mean equality of income per capita between two regions

[Advice] Suggested sampling scheme: Mode (6)

Enter 'HOW' for explanation of the sampling scheme selection

Enter 'WHAT' for explanation of the sampling scheme

Press 'Enter' for further advice

Press 'B' to return to the previous screen

Press 'E' to return to Screen [1]

Press 'Enter' to get further advice

Screen [7] (the italics represent user input):

Analysis issue: test mean equality of income per capita between two regions

[Advice] Please follow the procedure to conduct a preprocessing for statistical method selection:

- (1) use Mode (7) to draw a sample from each population under study;
- (2) use 'Standard deviation' to calculate the statistic SD for each sample;
- (3) use 'Normality test' to examine if the samples are normally distributed;
- (4) use 'Autocorrelation analysis' to test if income per capita is spatially dependent.

[Input] Input the following information:

Sample size: $SS1 = 60$; $SS2 = 60$

Standard deviation: $SD1 = 2.5$; $SD2 = 2.8$

Normality (normal/roughly normal/non-normal): *roughly normal*

Autocorrelation: *yes*

Enter a **number** for the explanation of each procedure in the preprocessing

Press '**Enter**' for further advice

Press '**B**' to return to the previous screen

Press '**E**' to return to Screen [1]

Press 'Enter' to get further advice

Screen [8]:

Analysis issue: test mean equality of income per capita between two regions

[Advice] Suggested statistical method: Modified t-test (2)

Enter 'WHAT' for explanation of the method

Enter 'HOW' for explanation of method selection

Press 'Enter' for further advice

Press 'B' to return to the previous screen

Press 'E' to return to Screen [1]

Press 'Enter' to get further advice

Screen [9]:

Analysis issue: test mean equality of income per capita between two regions

Suggested statistical method: Modified t-test (2)

Suggested sampling scheme: Mode (6)

[Advice] Null hypothesis H_0 : There is no significant difference in mean income per capita
between the two regions.

Alternative hypothesis H_a : The income per capita are significantly different
between the two regions.

Enter 'WHAT' for the explanation of hypotheses

Press 'Enter' for further advice

Press 'B' to return to the previous screen

Press 'E' to return to Screen [1]

Press 'Enter' to get further advice

Screen [10]:

Analysis issue: test mean equality of income per capita between two regions

Suggested statistical method: Modified t-test (2)

Suggested sampling scheme: Mode (6)

[Advice] Assumptions of the statistical method:

- 1) random samples
- 2) normal or roughly normal frequency distribution

Enter 'WHAT' for the explanation of the assumptions

Press 'Enter' for further advice

Press 'B' to return to the previous screen

Press 'E' to return to Screen [1]

Press 'Enter' for further advice

Screen [11]:

Analysis issue: test mean equality of income per capita between two regions

[Advice] Please use the following procedure to carry out the analysis:

- (1) Use the samples drawn for the preprocessing;
- (2) Set null hypothesis H_0 and significant level α ;
- (3) Use 'Modified t-test (2)' to conduct the analysis.
- (4) Draw conclusion from the result.

Press 'Enter' for further advice

Press 'B' to return to the previous screen

Press 'E' to return to Screen [1]

Press 'Enter' for further advice

Screen [12]:

Analysis issue: test mean equality of income per capita between two regions

Null hypothesis H_0 : there is no significant mean difference between the income per capita of the two regions.

Alternative hypothesis H_a : the income per capita are significantly different between the two regions.

[Advice] Results: p value

Please select a significant level α (0.1 - 0.01)

Result interpretation:

If $\alpha \geq p$, H_0 is rejected in favor of H_a , meaning the means of income per capita between these two regions are significantly different;

If $\alpha < p$, there is no significant evidence to reject H_0 , meaning the means of income per capita are not significantly different between these two regions.

Press 'Enter' for further advice

Press 'B' to return to the previous screen

Press 'E' to return to Screen [1]

Press 'Enter' for further advice

Screen [13]:

Analysis issue: test mean equality of income per capita between two regions

[Advice] Suggested Graphic Representation method:

Histogram for each of the data sets

Press 'Enter' for further advice

Press 'B' to return to the previous screen

Press 'E' to return to Screen [1]

Press 'Enter' for further advice

Screen [14]:

Analysis issue: test mean equality of income per capita between two regions

[Advice] Limitations:

Type I or Type II errors exists.

Enter 'WHAT' for the explanation of Type I and Type II errors

Press 'B' to return to the previous screen

Press 'E' to return to Screen [1]

4

of/de

4

**PM-1 3½"x4" PHOTOGRAPHIC MICROCOPY TARGET
NBS 1010a ANSI/ISO #2 EQUIVALENT**

Press 'Enter' to continue

Screen 15:

Advice ends

Thank you for consulting EASSA!

I hope it is helpful

If you have any comment, please contact the author at 992-4710

Press 'B' to return to the previous screen

Press 'E' to return to Screen [1]

Press 'Q' to exit

END

28/05/92

FIN