

Copyright Undertaking

This thesis is protected by copyright, with all rights reserved.

By reading and using the thesis, the reader understands and agrees to the following terms:

1. The reader will abide by the rules and legal ordinances governing copyright regarding the use of the thesis.
2. The reader will use the thesis for the purpose of research or private study only and not for distribution or further reproduction or any other purpose.
3. The reader agrees to indemnify and hold the University harmless from and against any loss, damage, cost, liability or expenses arising from copyright infringement or unauthorized usage.

IMPORTANT

If you have reasons to believe that any materials in this thesis are deemed not suitable to be distributed in this form, or a copyright owner having difficulty with the material being included in our database, please contact lbsys@polyu.edu.hk providing details. The Library will look into your claim and consider taking remedial action upon receipt of the written requests.

UNCERTAINTY-BASED SPATIAL ASSOCIATION RULE MINING

ZHANG ANSHU

Ph.D

The Hong Kong Polytechnic University

2017

The Hong Kong Polytechnic University
Department of Land Surveying and Geo-Informatics

Uncertainty-based Spatial Association Rule Mining

ZHANG Anshu

A thesis submitted in partial fulfilment of the requirements
for the degree of Doctor of Philosophy

February 2017

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

_____ (Signed)

ZHANG Anshu (Name of student)

Abstract

Spatial association rule mining (SARM) is the discovery of implicit ‘antecedent → consequence’ rules from spatial databases. SARM is an emerging topic in geographical information science (GISc) and a powerful tool in research and practice. The key to usefulness of SARM results is their reliability: the abundance of authentic rules, control over the risk of spurious rules, and goodness of rule interestingness measure (RIM) values. Such reliability, however, faces great challenges from uncertainties of various types and sources.

Uncertainty-based SARM, proposed in this thesis, aims at enhancing the reliability of SARM results on all three aforesaid aspects via novel and improved uncertainty handling methods. In response to three critical uncertainty issues in SARM: data error, gradual/vague spatial concept, and uncertain concept modelling, this thesis realises the following three interrelated objectives:

Mining significant spatial association rules from uncertain data: a new statistical test on the rules is developed to correct existing statistically sound test, which is indispensable for strict control over spurious rules, for distortions due to data error. The new test combines original data error propagation modelling as well as simulative processes. The new method can averagely compensate 50% loss of true rules due to data error, thus markedly enrich authentic results. Such efficacy is also largely robust to inaccurate data error information and dependent error probabilities in practical imperfect data.

Gaussian-curve-based fuzzy data discretization and crisp-fuzzy SARM: a Gaussian-curve-based model is presented to strengthen spatial semantics in fuzzy data discretization. Also, crisp-fuzzy SARM is originated to synthesise statistically sound testing based on ordinary (crisp) SARM, and RIM evaluation based on fuzzy rules. The techniques can discover at least twice as many authentic rules as conventional fuzzy SARM; avoid large overestimations of RIM values, usually by more than 50%, in ordinary SARM; and keep minimal risk of spurious rules.

Genetic algorithm (GA) for crisp-fuzzy SARM: the new GA integrates the merits of statistical evaluation, new Gaussian-curve-based data discretization and crisp-fuzzy SARM. Experimentwise and generationwise adjusted statistical tests are innovated for the GA to satisfy different user needs. The proposed GA can produce several times as

many rules, and as high RIM values as non-GA SARM. The risk of spurious rules is below low user specified levels for both testing approaches.

The developments for the three objectives are proven effective and robust, through synthetic and real-world data experiments of various experimental settings and data conditions. Case studies for these developments on urbanization-socioeconomic changes, wildfire risks, and hotel room price determinants inject new findings in corresponding research topics.

In sum, methods developed in this thesis can alleviate manifold uncertainty issues in SARM, thereby significantly improving the reliability of SARM results in all its three aforesaid aspects.

As a systematic study on uncertainty handling in SARM, this thesis would enrich GISc theories and methodologies. Particularly, it answers the increasingly pressing need for quality and reliability studies in GISc. The thesis work is also practically useful in improving decision making and user services in various domains involving spatial data, as exemplified by the case studies.

Keywords: spatial association rule, uncertainty, quality issues, uncertain data, fuzzy sets and logic, genetic algorithm, statistical evaluation, pattern discovery, spatial data mining

Publications Arising from the Thesis

Articles

Zhang, A. and Shi, W. (under review). Mining significant fuzzy association rules with genetic algorithm. Submitted to *Applied Soft Computing*.

Shi, W., Zhang, A., and Webb, G.I. (under review). Mining significant crisp-fuzzy spatial association rules. Submitted to *International Journal of Geographical Information Science*.

Zhang, A., Shi, W., and Webb, G.I., 2016. Mining significant association rules from uncertain data. *Data Mining and Knowledge Discovery*, 30(4), 928–963.

Zhang, A. and Shi, W., 2014. Mining significant geographical association rules from uncertain data (Abstract). In: *The 2014 Institute of Australian Geographers Conference*, 30 June–2 July 2014 Melbourne.

Shi, W., Zhang, A. and Ho, O., 2013. Spatial analysis of water mains failure clusters and factors: A Hong Kong case study. *Annals of GIS*, 19(2), 89–97.

Patent

Shi, W. and Zhang, A., *Method and apparatus for significance test on association rules with consideration of data uncertainties*. P.R. China Patent application 201510076329.0 (entered into substantive examination).

Acknowledgements

I wish to thank all people supporting and encouraging me in my Ph.D. study.

First, I would like to represent my sincere honour and gratitude to Prof. Shi Wenzhong in Hong Kong Polytechnic University (HKPU), for his throughout support and guidance. As my chief supervisor in both B.Sc. and Ph.D. studies, Prof. Shi is and will ever be the most important teacher in my lifetime.

I also thank very much to Prof. G.I. Webb in Monash University for his cordial and insightful guidance and help during my visits to the university and all later time. These experiences greatly improved my research horizon and skills.

Thanks to Dr. Chen Jiangping in Wuhan University for her academic suggestions and help. Thanks to fellow students and colleagues in HKPU and HKPU Shenzhen Base, especially Ke Linghong, Miao Zelang, Tian Yumiao, Li Wenyu, Yu Shiwei, Wang Bin, Yang Cheng, Zhou Xiaolin, Deng Susu, Xu Qianxiang, Xu Rui, Wang Chisheng, Zhu Wu, Yan Xing and Wang Qunming for their kind discussions, research inspirations, encouragement and caring.

Gratefully acknowledged are the financial supports from Hong Kong PhD Fellowship Scheme by the Research Grants Council of Hong Kong, HKPU, and project ‘Dynamic Characterization of Smart City (1-ZE24)’.

Finally, I wish to express my deepest thanks to my parents, for their unconditional support and love.

Table of Contents

CERTIFICATE OF ORIGINALITY	v
Abstract	i
Publications Arising from the Thesis	iii
Table of Contents	v
List of Figures	vii
List of Tables	viii
List of Abbreviations	ix
Chapter 1 Introduction	1-1
1.1 Research background	1-1
1.2 Objective, methodology and scope	1-6
1.3 Main contributions	1-7
1.4 Outline of Thesis	1-10
Chapter 2 Prior works on ARM/SARM with uncertainties	2-12
2.1 SARM and RIMs	2-12
2.2 Avoidance of spurious rules and statistical test approach	2-14
2.3 ARM/SARM with uncertain data	2-18
2.4 Fuzzy ARM/SARM	2-19
2.5 GAs for ARM/SARM	2-22
2.6 Other uncertainty handling techniques in SARM	2-24
Chapter 3 Mining significant SARs from uncertain data	3-27
3.1 Statistical test for association rules with uncertain data	3-27
3.1.1 Overview of corrected test	3-27
3.1.2 Modelling error propagation	3-29
3.1.3 Recovering test parameters	3-31
3.1.4 Controlling spurious rules	3-37
3.2 Synthetic data experiment	3-40
3.2.1 Data and methods	3-41
3.2.2 Results	3-47
3.3 Real-world data experiment: Mining spatio-temporal associations between land uses and socioeconomics	3-57
3.3.1 Data and methods	3-57
3.3.2 Results	3-62
3.4 Accuracy of error probability information and its practical implication to corrected test	3-66
3.5 Summary	3-68
Chapter 4 Mining significant crisp-fuzzy SARs	4-70
4.1 Proposed techniques	4-70

4.1.1	Gaussian-curve-based fuzzy data discretization	4-70
4.1.2	Crisp-fuzzy SARM for mining authentic and accurate rules	4-73
4.2	Experiment with synthetic data	4-75
4.2.1	Methods	4-75
4.2.2	Results: true and spurious rules	4-80
4.2.3	Results: RIM accuracy	4-81
4.3	Experiment with real-world data	4-83
4.3.1	Data and methods	4-83
4.3.2	Results	4-86
4.4	Summary	4-92
Chapter 5	Genetic algorithm for mining significant crisp-fuzzy SARs	5-93
5.1	Methods	5-93
5.1.1	Chromosome encoding	5-93
5.1.2	Fitness assignment with statistically sound tests	5-96
5.1.3	Evolutionary model	5-100
5.2	Experiments: Hotel room price determinants and wildfire risk factors	5-104
5.2.1	Data collection and preprocessing	5-104
5.2.2	Experiment specifications	5-108
5.2.3	Accessing control over spurious rules	5-110
5.2.4	Evaluating ability of discovering true rules	5-116
5.2.5	Analysing practical implications for Hotel experiment	5-122
5.3	Summary	5-135
Chapter 6	Conclusions and future work	6-136
6.1	Research summary and significance	6-136
6.2	Future work	6-140
Appendix	Evaluating discrepancy between exact and approximate $\hat{s}_0(c_i)$	
values		144
References		146

List of Figures

Figure 1.1 Research framework	1-10
Figure 2.1 A composite region with a broad boundary	2-25
Figure 3.1 Using $\sigma(s(c_j))$ and z to control probability of overestimating $E(s(c_j))$ at arbitrary user specified value	3-32
Figure 3.2 Synthetic data experiment results	3-48
Figure 3.3 Overview of Massachusetts and land uses of a small locality	3-59
Figure 3.4 Recovery of true rules by corrected test in real-world data experiment	3-64
Figure 3.5 Recovery of true rules involving land use changes in real data experiment	3-65
Figure 4.1 Illustration of proposed fuzzy data discretization model	4-73
Figure 4.2 Synthetic data experiment results on abundance of true rules and avoidance of spurious rules	4-80
Figure 4.3 Synthetic data experiment results on RIM accuracy	4-82
Figure 5.1 Encoding of attribute a using proposed chromosome encoding scheme	5-95
Figure 5.2 Overall procedures of proposed GA for crisp-fuzzy SARM	5-100
Figure 5.3 Relation between membership function and fuzziness for linear and Gaussian-curve-based discretization model	5-103
Figure 5.4 Hong Kong map with experimented hotel locations and selective resources	5-107
Figure 5.5 Result in discovering true rules: Hotel experiment, generationwise approach	5-117
Figure 5.6 Evolutions in numbers of rules and total leverages in GA for Hotel and Fire experiments	5-122

List of Tables

Table 2.1 Common forms of membership functions for fuzzy data discretization	2-20
Table 2.2 Representative GA-based ARM studies	2-23
Table 3.1 Estimated true values of test parameters $\hat{a}_0 - \hat{d}_0$ with derivations	3-36
Table 3.2 Conditional probabilities of att_3 values in synthetic data	3-42
Table 3.3 Numbers of true rules from ‘ideal’ data and remarks	3-44
Table 3.4 Numerical synthetic data experiment results for E and R treatments	3-49
Table 3.5 Summary of synthetic data experiment results	3-50
Table 3.6 True rule increases and recovery rates by error level	3-52
Table 3.7 Synthetic data experiment results with inaccurate error specifications or dependent data error	3-55
Table 3.8 Land use classes of study area	3-58
Table 4.1 RIM value exaggerations due to crisp data discretization in a miniature database of four records	4-74
Table 4.2 Dependence of $\mu_1(outcome)$ on other attributes	4-77
Table 4.3 Various experiment settings in synthetic data experiment	4-79
Table 4.4 Data attributes in real data experiment	4-85
Table 4.5 Real data experiment result on number of significant rules and RIM accuracy	4-87
Table 4.6 Real data experiment result on single wildfire risk factors	4-88
Table 4.7 Numbers of rule pairs like $X \rightarrow Y$ and $X \setminus \{x\} \rightarrow Y$	4-90
Table 5.1 Accessibility attributes in Hotel experiment	5-105
Table 5.2 Specifications for Hotel and Fire experiments	5-109
Table 5.3 Result on control over spurious rules: Hotel experiment	5-112
Table 5.4 Result on control over spurious rules: Fire experiment	5-115
Table 5.5 Result in discovering true rules: Fire experiment	5-117
Table 5.6 Past hedonic hotel room price modelling studies that involved hotel accessibilities	5-124
Table 5.7 Interpreted resultant rules of Hotel experiment	5-126
Table 5.8 Hotel room price distribution by star rating	5-130

List of Abbreviations

ARM	Association rule mining
FWER	Familywise error rate
GA	Genetic algorithm
GISc	Geographical information science
RIM	Rule interestingness measure
SAR(M)	Spatial association rule (mining)
SDM	Spatial data mining

Chapter 1 Introduction

1.1 Research background

Spatial data mining (SDM) is the process of discovering implicit patterns from spatial data, with the ultimate goal of gaining knowledge from the patterns for research and practice, particularly for decision support. Contemporary geographical information science (GISc) has been experiencing explosive growth in spatial data volume and complexity. Resultantly, it has been increasingly difficult to look beyond and find interesting patterns and knowledge from explicitly stored spatial data by human observations and labours. Computer-aided SDM methods have been of rising popularity and essentiality under such circumstances.

As an important type of targeted pattern in SDM, *spatial association rule (SAR)* is an implicit ‘antecedent \rightarrow consequence’ pattern, or if-then rule, that involve spatial element(s). For example, the SAR ‘(house) near river \rightarrow expensive’ can be linguistically represented as ‘if a house is near river, then it is expensive’. *Spatial association rule mining (SARM)* seeks for SARs that meet constraints on certain *rule interestingness measures* (RIMs). With the superiority in revealing and prioritizing enormous numbers of multiway interactions between numerous spatial entities, SARM has become powerful and valuable for investigating interactions in complex data and supporting user decisions. This has been proven in its wide applications in, for example, environment and socioeconomic condition relations (Mennis and Liu 2005, Rodman *et al.* 2006), vegetation-climate change (Shu *et al.* 2008), soil contamination (Sun *et al.*), transportation demands (Lisi and Malerba 2004), urban accessibility (Appice *et al.* 2003) and related profitability analysis (Feng *et al.* 2010), object mobility pattern extraction (Verhein and Chawla 2008), mobile navigation (Baralis *et al.* 2012), image database learning (Lee *et al.* 2007) and tourist attraction visit principles (Versichele *et al.* 2014).

The usefulness of SARM results highly depends on their reliability, which is a balance between:

- ***Abundance of authentic rules***: this is the base on which adequate knowledge can be derived and presented to users.
- ***Control over the risk of spurious rules***: spurious rules are rules that convey non-existent associations and mislead users into poor decisions. Since SARM must explore enormous number of potential rules for high dimensional data, it faces a very high risk of falsely ‘discovering’ numerous spurious rules. As modern voluminous data sources and SARM methods produce ever greater numbers of rules, this problem becomes a critical barrier for reliability of SARM results.
- ***Goodness of RIM values*** for resultant rules: this includes their *accuracy* and *fitness*. Accuracy measures the degree to which RIM values deviates from their true values, or most probable values if true values are unknown in reality. Fitness refers to if the RIM values are favourable for specific user needs and data mining tasks. For example, in SARM for store site selection, target RIM values indicating the profit gains by selecting appropriate sites are desired to be large.

The reliability of SARM results, like that of SDM results in general, has been seriously challenged by uncertainties. Uncertainties for SDM and SARM are in various types, including but not limited to positional uncertainty, attribute uncertainty, topological uncertainty, temporal uncertainty, inconsistency, incompleteness and knowledge uncertainty (Shi *et al.* 2003). Uncertainties can occur in source data, as well as each stage of SARM: data pre-processing, rule mining and knowledge representation, and will propagate to all subsequent stages once being generated. Uncertainties may lead to defective SARM results, questionable quality of knowledge discovered, and ultimately misuses of and poor decisions based on the knowledge.

Theories and methods for uncertainty modelling and reliability enhancement of spatial data and spatial analyses have been developed and systematically summarized by Shi (2010). This line of research has also been extended to SDM. Shi *et al.* (2003) proposed uncertainty-based SDM and pointed out that SDM researches then lacked integrations with uncertainty handling. They also pointed out that relevant studies mostly focused on uncertainties in source spatial data, instead of those rising from particular SDM methods or uncertainty propagations in SDM process. These gaps still exist to a considerable degree over 10 years after the proposal of uncertainty-based SDM.

Specific in SARM, three of the various uncertainty issues stand out regarding the severity of influences on reliability of SARM results and pending research needs:

- (1) **Data error**: this largely refers to random error in surveying discipline (as is tied with systematic error and blunders) or noise in computer science. The error is inevitably generated by numerous unknown factors in data acquisition. While the error can influence the reliability of SARM results on all three aforesaid aspects, it most directly impacts the abundance of true rules. The error adds to a random component into data that has no associations with the rest of data, thus it is expected to weaken associations between data elements, thereby causing true rules lost from resultant rules. Random error can hardly be reduced once data is acquired, as its value and distribution are unpredictable for individual spatial entities. A more feasible approach would be statistically modelling the error propagation in SARM and developing corresponding methods to alleviate its impact on SARM results. However, current association rule mining (ARM) with uncertain data (Chui *et al.* 2007) mostly focuses on probabilistic data structures rather than behaviours of random error, let alone SARM.

- (2) ***Gradual/vague spatial concepts***: ARM/SARM calls for *data discretization* which generate concepts as sets of raw numerical data values, so that the concepts can be used in rule mining, and rules in linguistic forms can be presented to users. Gradual or vague concepts are prevalent in spatial data, which can be mathematically interpreted as *fuzziness* (Shi *et al.* 2003). Hard divisions for such concepts incur bias, inaccurate semantic representations and finally unreliability in SARM results (Hüllermeier 2009, Farzanyar and Kangavari 2012). *Fuzzy SARM* (Ladner *et al.* 2003) can largely relieve this problem by modelling fuzzy spatial concepts as fuzzy sets, or intervals of raw numerical data values with fuzzy boundaries. Fuzzy data discretization models have been proposed specifically for spatial relations such as proximity (Ladner *et al.* 2003, Laube *et al.* 2008). However, the common belief that fuzzy SARM results are more reliable than that of ordinary (crisp) SARM, which is rooted in higher RIM accuracy of fuzzy rules, has seldom been examined by quantitative studies. Also, using fuzzy sets might make rules less significant and incur the risk of reducing authentic rules.
- (3) ***Uncertain concept modelling***: this primarily concerns the data discretization stage of SARM. Experts often do not have adequate knowledge for providing proper discretization schemes, which include the number of concepts for each attribute, and raw data intervals (crisp or fuzzy) corresponding to each concept. Moreover, the appropriateness of a data discretization scheme is subject to specific user need and SARM task (Kaya 2006). Good discretization schemes for individual attributes may be achieved by existing classification or clustering methods, but they are not equivalent to, and usually quite different from, a combination of all concepts in data that leads to good resultant rules. Genetic algorithms (GAs) are promising to address this problem. By progressively optimizing given objective(s) in an evolutionary approach, GA-based SARM can find near-optimal data discretization schemes for specific user demands (Fazzolari *et al.* 2013). This can markedly enhance the fitness

of RIM values, and also mostly increase the number of authentic rules with user specified characteristics.

Besides, a key approach for reducing spurious rules in ARM is the statistical hypothesis testing (Megiddo and Srikant 1998, Liu *et al.* 1999, Bay and Pazzani 2001, Zhang *et al.* 2004). Both sampled and population data are finite representations of associations between studied objectives which can potentially repeat for infinite times in the real world. Rules might fulfil specified RIM constraints in data by chance rather than due to real associations of the objectives. Such spurious rules can account for a high percentage or even the majority of resultant rules (Webb 2007, Zhang *et al.* 2016), thereby making the results unusable. Statistical tests are designed to filter out spurious rules, and only *significant rules* accepted by the tests will enter final SARM results.

Statistically sound evaluation (Webb 2007) is a particularly effective statistical testing technique and can control the familywise error rate (FWER), the chance that entire result includes any spurious rule, upon a low user specified level, for example 5%. Albeit highly effective, this technique has not been systematically integrated with data error treatment, fuzzy techniques or GA-based method for ARM/SARM. Without integrating these techniques to address multiple uncertainties which are typically concurrent in SARM, the efficacy of each individual technique for producing reliable SARM results can be considerably impaired. Furthermore, statistically sound tests are conservative and also reject many authentic rules. In conventional ARM, the tests may reserve thousands of rules for medium sized data, which is adequate for practical use (Webb 2007). However, when the tests are applied with uncertainties in data or SARM methods, or in conjunction with other uncertainty handling techniques, SARM results may suffer from severe loss of authentic rules. This has been proven true for all three above mentioned uncertainty issues in Chapters 3 to 5 of this thesis.

1.2 Objective, methodology and scope

This thesis presents ***uncertainty-based SARM*** which develops new techniques, and enhances existing techniques for handling uncertainties in SARM, with the ultimate goal of improving the reliability of SARM results regarding abundance of authentic rules, risk of spurious rules, and goodness of RIM values.

In response to the three key uncertainty issues for SARM stated in Section 1.2, the study takes the following approaches (also as subgoals) for enhancing the reliability of SARM results:

- (1) To mathematically model (random) data error propagation through statistically sound tests on SARs; based on the error model, to design a method for alleviating impacts of data error on results of mining significant SARs (those accepted by statistical tests), particularly the loss of authentic rules.
- (2) To present a fuzzy data discretization model for SARM with enhanced spatial semantics; to conduct comparative studies on reliability of fuzzy and ordinary SARM results; and to originate a method for mining significant SARs that combines advantages of fuzzy and ordinary SARM, thereby obtaining both higher RIM accuracy and abundant authentic rules.
- (3) To investigate a feasible solution for statistically sound tests during evolutionary process in GAs, and then to develop a GA for mining significant SARs. The new GA-based method is for obtaining near-optimal data discretization schemes, abundant rules and high RIM fitness for specific user requirements.

These approaches of the thesis adopt statistically sound tests for strict control over spurious rules, which has been successful in conventional ARM (Webb 2007). In presence of uncertain SARM data and methods and integrated uncertainty handling techniques, the work revalidates the efficacy of statistically sound tests, and makes

efforts on overcoming increased scarification of abundant authentic rules and RIM goodness due to the tests.

The above new developments are evaluated with both synthetic and real-world data. Synthetic data experiments possess predesigned rules and RIM values, thus the reliability of SARM results can be confidently and quantitatively measured. Such experiments were rare in prior SARM studies, and largely resolve the difficulty of past research in evaluating the reliability of SARM results due to unknown true results. Real-world case studies on geographical and socioeconomic topics are also conducted to examine the practical value of the thesis work.

It is notable that the three aforementioned reliability issues present for both spatial (geometrical) and non-spatial attributes in SARM data. If methods exclusively for spatial attributes are first applied for handling these issues, their efficacies will likely to be hindered by reliability deficiencies caused by non-spatial attributes and hence difficult to evaluate. Therefore, this thesis attempts at methods that are also applicable for non-spatial data, while it emphasizes GISc theories and spatial data characteristics in method development and evaluation. For this purpose, the thesis focuses on the SARM stage where raw numerical values of spatial attributes have been computed from the geographic objects and tabulated in attribute-value data tables. As for spatial relations, the thesis concentrates on nearness instead of directional or topological relations, as the former is more readily transformed to numerical values. The case studies show that the new developments are particularly effective on spatial data and inspire new insights into GISc research topics.

1.3 Main contributions

The main contributions of this thesis are:

- (1) ***Mining significant SARs from uncertain data***: an original mathematical model is established for data error propagation in statistically sound tests on SARs. Based on the error model, a method combining analytic and simulative processes is designed to correct the statistical test for distortions caused by data error. Experiments show that this corrected test method significantly recovers the loss in authentic rules due to data error, averagely by 50%. The corrected test basically maintains superior control over the FWER of original statistically sound tests. The corrected test is robust against inaccurate data error probability information and dependent error probabilities, which increases its usefulness in practical SARM with imperfect data.
- (2) ***Gaussian-curve-based fuzzy data discretization and crisp-fuzzy SARM***: first, a fuzzy data discretization model based on Gaussian curves is presented. This model extended past works by strengthening spatial semantics and multi-concept relations. Second, the crisp-fuzzy SARM is originated. Crisp-fuzzy SARM explores crisp rules and prunes dubious ones using statistically sound tests, and then evaluates RIMs of accepted rules using fuzzy measures. The combination of this two techniques can at least double the number of resultant true rules, compared with using pure fuzzy SARM; and avoid large positive errors in RIM values committed by crisp SARM, which typically exceeded 50% for representative RIMs. The use of statistically sound evaluation guarantees minimal risk of spurious rules.
- (3) ***GA for mining significant crisp-fuzzy SARs***: the newly developed GA produces near-optimal SARM results for user specifications, including more abundant rules and RIM values of higher fitness than conventional SARM results, while strictly controlling spurious rules by integrating statistically sound tests. Two statistical testing approaches exclusively for the GA, the experimentwise and generationwise adjustment approach, are developed to control the FWER and percentage of spurious rules, respectively. The new GA adopts and thus combined the advantages of an efficient and flexible approach

for encoding candidate resultant rules, the newly proposed Gaussian-curve-based data discretization model and crisp-fuzzy SARM. The proposed GA is experimentally proven to produce several times as many rules as using data discretization based on standard classifications, and effectively keeps the FWER or percentage of spurious rules under user specified level (5% in experiment).

- (4) Practical usefulness of the above three contributions is demonstrated in multiple case studies, the topics of which include associations between land use and socioeconomic changes, wildfire risk factors, and accessibilities as hotel room price determinants. Improved insights of each topic are obtained due to merits of the newly developed methods.

The framework of the thesis work and its relation to research backgrounds are illustrated in Figure 1.1.

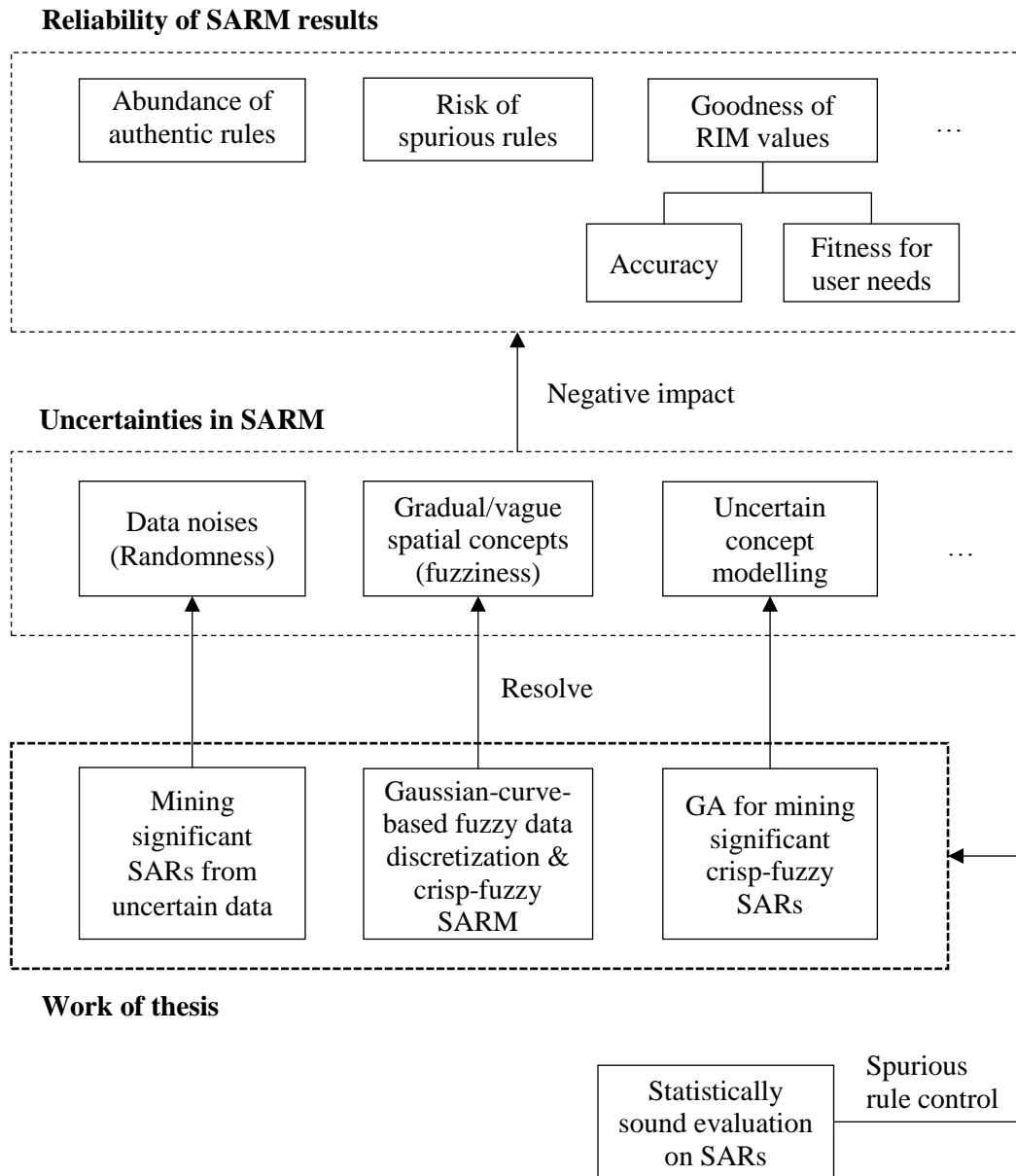


Figure 1.1 Research framework

1.4 Outline of Thesis

The subsequent chapters of this thesis are organized as follows:

Chapter 2 briefly introduces prior works for ARM/SARM with uncertainties, as well as relevant existing techniques as the bases of the thesis work, including essentials of SARM and RIMs; statistical tests on rules; ARM/SARM with uncertain data, fuzzy concepts and GAs; and other uncertainty handling techniques for SARM.

Chapter 3 constructs the error model and corrected test for mining significant SARs with uncertain data, with a synthetic data experiment and a real data experiment on land use and socioeconomic changes.

Chapter 4 presents the Gaussian-curve-based fuzzy data discretization model and crisp-fuzzy SARM, and evaluated them by a second synthetic data experiment and the wildfire risk factor case study.

Chapter 5 is dedicated to GA for mining significant crisp-fuzzy SARs. The proposed GA are comparatively examined by two case studies, one again for wildfire risk factors, and the other in depth for hotel room price determinants.

Chapter 6 concludes the thesis and suggests further researches for uncertainty-based SARM.

Chapter 2 Prior works on ARM/SARM with uncertainties

* Sections 2.1–2.3 in this chapter is partially based on the published research article of the thesis author entitled “Mining significant association rules from uncertain data”.

2.1 SARM and RIMs

The thesis work hereafter describes the preprocessed data ready for rule mining in terms of attribute-value data, one of the two most used data types in ARM/SARM. For an attribute-value dataset D , each record $R \in D$ is a set of *items* in the form ‘attribute = value’. A representative type of attribute-value data is the categorical data, where each value is a class label. Transactional data, the other most used data types, may be handled as binary valued data, where values 0 and 1 represent nonexistence and existence of an entry in the record. Numerical data is typically transformed into attribute-value one via data discretization before being explored for rules. Geometrical information of spatial entities can generally be first computed into numerical data.

An association rule is a pattern $X \rightarrow Y$, where the *antecedent* $X = \{x_1 \dots x_p\}$ and *consequent* $Y = \{y_1 \dots y_q\}$ are itemsets consisting of items in D , $X \cup Y$ contains at most one item for each attribute. $X \rightarrow Y$ is a SAR (spatial association rule) if it includes item(s) for spatial attribute(s). This thesis limits Y to single-item consequent y , which is common for SARM tasks, and most of its developments can be extended straightforwardly to multi-item consequents.

SARM was introduced by Koperski and Han (1995) as the spatial extension of general ARM (Agrawal et al. 1993). SARM aims at finding all spatial association rules that meet designated criteria, mostly being above specified minimum values of certain RIMs. It generally includes two tasks: first, to compute necessary spatial attributes from geometries of spatial entities, either before or during rule exploration via spatial query; second, to explore rules from data including the computed spatial attributes,

using algorithms similar to those for general association rule mining, such as the popular Apriori (Agrawal and Srikant 1994) and its improved versions. Numerous RIMs have been proposed for general association rules, most of which are applicable to SARM. Tew *et al.* (2014) reviewed and analysed 61 well-known RIMs using clustering technique, and revealed that many RIMs actually have very similar rule-ranking behaviours. That is, when rules are ranked by their RIM values, many RIMs result in very similar ranks. Some of the commonest RIMs are:

- *support* (Agrawal *et al.* 1993):

$$supp(X \rightarrow y) = supp(X \cup y) = freq(X \cup \{y\}) / |D|;$$

- *confidence* (Agrawal *et al.* 1993):

$$conf(X \rightarrow y) = supp(X \rightarrow y) / supp(X);$$

- *improvement* (Bayardo *et al.* 2000):

$$imp(X \rightarrow y) = conf(X \rightarrow y) - \max_{Z \subset X} (conf(Z \rightarrow y));$$

- *leverage* (Piatetsky-Shapiro 1991):

$$lev(X \rightarrow y) = supp(X \rightarrow y) - supp(X)supp(y) / |D|;$$

- *lift* (International Business Machines 1996):

$$lift(X \rightarrow y) = supp(X \rightarrow y) / (supp(X)supp(y));$$

- *interestingness weighting dependency* (Gray and Orłowska 1998):

$$IWD(X \rightarrow y) = \left(\left(supp(X \rightarrow y) / (supp(X)supp(y)) \right)^l - 1 \right) supp(X \rightarrow y)^m,$$

To be neutral, set $l = 1$, $m = 1$ (Tew *et al.* 2013)

Where $freq(S)$ is the number of records in D containing all items in the set of items S . $X \rightarrow y$ is *productive* if $imp(X \rightarrow y) > 0$, that is, every item in X improves the confidence of the rule (Webb 2007). Unproductive rules include redundant items in X that are irrelevant to y , and are generally regarded as uninteresting and removed from final result.

RIMs exclusively for SARM have also been proposed. Laube *et al.* (2008) advised spatial support and spatial confidence which are quantified by various proximity measures other than Euclidean distance between geographic objects. The proximity measures included point-to-point, point-to-polyline, point-to-polygon, polygon-to-multipoint, polygon-to-polygon and more possible cases.

2.2 Avoidance of spurious rules and statistical test approach

As this thesis augments statistical tests in ARM for controlling false rules, the discussion of this section is largely based on general ARM, and equally applies to SARM.

RIMs are the most direct measures to avoid spurious rules; rules with RIM values unsatisfying specified thresholds are considered uninteresting and removed from the result. While this can be very useful, empirically generic solutions to setting RIM thresholds for interesting rules are often unavailable. The thresholds are usually up to subjective user specifications and bear high risk of being inappropriate and leading to questionable reliability of selected rules.

There have been also quantitative criteria for pruning uninteresting rules, which are often operations on RIMs. For instance, the productive rule criterion in Section 2.1 is equivalent to $imp(X \rightarrow y) > 0$. The non-redundant rule criterion (Zaki 2000) rejects rules whose antecedents contain items that are implied by other items in the antecedents, such as ‘ $type = waterbody \wedge type = river \rightarrow elevation = low$ ’, where $waterbody$ is implied by $river$. As the implied items cannot improve rule confidences, this criterion is entailed by $improvement > 0$. The actionable rule criterion (Liu *et al.* 2001) accepts rules with $improvement > 0$ and higher confidences than $\emptyset \rightarrow y$ even if the data records conforming to their specializations are removed. Specializations

refer to rules containing all items of the studied rules and also extra items in the antecedents.

Although these criteria appear more objective, many spurious rules can still fulfil these criteria in data by chance instead of due to real associations, thereby threatening the reliability of rule mining results. Thus statistical hypothesis tests have been used to avoid such spurious rules. The test result is the probability p that a rule $X \rightarrow y$ has observed RIM value even if $X \rightarrow y$ association is nonexistent in reality. This equates to the risk that $X \rightarrow y$ is spurious. Only rules with p values below designated significance level α , say 0.05, are significant and accepted, while others are rejected and removed. Techniques in this line include correlation rules (Brin *et al.* 1997), association pruning (Liu *et al.* 1999), significant statistical quantitative rules (Zhang *et al.* 2004) and so on.

Hereafter the statistical tests are exemplified by the test for productive rules, a typical test for pruning redundant rules. The same approach can be right applied to other tests. According to the formulation of improvement in Section 2.1, the productivity of a rule $X \rightarrow y$, or whether $imp(X \rightarrow y) > 0$, can be tested by:

$$\forall Z \subseteq X, \Pr(y | X) > \Pr(y | X \setminus Z), \quad (2.1)$$

where ‘\’ denotes set difference. This thesis follows accepted practice (Webb 2007) to conduct a more computationally economic test, the result of which is quite similar to that of testing Equation (2.1), on

$$\forall m = 1 \dots p, \Pr(y | X) > \Pr(y | X \setminus \{x_m\}). \quad (2.2)$$

The null hypothesis for the test is $\exists m = 1 \dots p, \Pr(y | X) \leq \Pr(y | X \setminus \{x_m\})$, suggesting that $X \rightarrow y$ has a higher confidence in data by chance rather than due to real association between x_m and the remaining items.

Chi-square is commonly used for testing conditions like Equation (2.2), yet it is criticized as inaccurate for small data of sizes up to hundreds (McDonald 2014), due to the approximation of integral supports by continuous χ^2 distribution. Also, the chi-square test is two-tailed, while Equation (2.2) is a one-tailed condition. Thus, for ordinary ARM/SARM with integral crisp support of patterns, the Fisher exact test (Agresti 1992) is more appropriate for Equation (2.2). This test is reliable for any sample size, and can be one-tailed or two-tailed. Let a, b, c, d be numbers of records in data D containing the following patterns:

$$\begin{aligned} a &= \text{supp}(X \cup \{y\}) \\ b &= \text{supp}(X \cup \neg\{y\}) \\ c &= \text{supp}((X \setminus \{x_m\}) \cup \neg\{x_m\} \cup \{y\}) \\ d &= \text{supp}((X \setminus \{x_m\}) \cup \neg\{x_m\} \cup \neg\{y\}) \end{aligned} \quad , \quad (2.3)$$

where \neg refers to that the record must not contain the item. The p value of the test with respect to x_m is

$$p_m = \sum_{i=0}^{\min(b,c)} \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{(a+b+c+d)!(a+i)!(b-i)!(c-i)!(d+i)!} . \quad (2.4)$$

The final result of the test is equal to $\max(p_m)$, and is equivalent to the risk that $X \rightarrow Y$ is spurious. $X \rightarrow Y$ is accepted and included in resultant rules only if p is below the significance level.

In fuzzy ARM/SARM, as patterns evaluated have fuzzy fractional supports, the Fisher exact test is inapplicable, as it can only handle integral supports. In this case, chi-square test should still be used, and actually its inaccuracy for integral supports in small datasets no longer holds for the continuous fuzzy supports. The testing statistic is

$$\chi^2 = \frac{(ad - bc)(a + b + c + d)}{(a+b)(c+d)(a+c)(b+d)} , \quad (2.5)$$

and p_m is looked up from the χ^2 table for the computed χ^2 value with one degree of freedom.

The multiple testing problem occurs when statistical tests on rules are applied many times. If a test is applied with a significance level α , say 0.05, then there is no more than 0.05 probability that the null hypothesis will be rejected even though it is true. In the association rule context this means a rule is accepted even though there is no association. If many potential rules are tested, then the statistical test should pass 5% of the ones that should be rejected. When large numbers of potential rules are explored, this can even mean that more of the accepted rules are spurious than true. This problem may be resolved with a Bonferroni correction to the significance level (Shaffer 1995). A previous solution to control the familywise error rate (FWER) below α is to set the significance level $\kappa = \alpha/n$, where n is the number of rules tested. Yet this does not really work, as the tested rules usually have passed other interestingness measures such as the minimum confidence, and tend more to pass the test than arbitrary rules.

Webb (2007) suggests an approach which is statistically sound, meaning that it can place a strict upper limit on the FWER. The approach sets $\kappa = \alpha/s$, where s is the total number of potential rules as combinations of all data items. Suppose the data constitutes i attributes $att_1 \dots att_i$, and the numbers of values in corresponding attributes are $n_1 \dots n_i$. Denote the number of different combinations of up to j items coming from $att_1 \dots att_k$, $k \leq i$ as $c_{att,j,k}$; $c_{att,j,k}$ can include at most one value from each attribute. Then

$$c_{att,j,k} = \begin{cases} n_k, & j = 1, k = 1 \\ 0, & j > 1, k = 1 \\ c_{att,1,k-1} + n_k, & j = 1, k > 1 \\ c_{att,j,k-1} + n_k \times c_{att,j-1,k-1}, & \text{otherwise} \end{cases} \quad (2.6)$$

The number of potential rules with att_m as the consequent and up to $maxL$ items in the antecedents from all other attributes is equal to

$$n_m \times \sum_{j=1}^{maxL} c_{att-m,j,i} , \quad (2.7)$$

where $att-m,i$ means the set of attributes $\{att_1 \dots att_i\}$ excluding att_m . Finally,

$$s = \sum_{m=1}^i \left(n_m \times \sum_{j=1}^{maxL} c_{att-m,j,i} \right) \quad (2.8)$$

if all attributes can be in the rule antecedents and consequents. The s value when only specified attributes are in the rule antecedents and consequents can be similarly derived. With only a modest number of items, s can reach tens of thousands or even billions. While the κ value is then extremely small, experiments show that such κ value usually allows a substantial percentage of true rules to past the test and thus be discovered.

The statistically sound approach can achieve an FWER below 1% with $\alpha = 0.05$. Such high efficacy, however, also suggests that it is more conservative than users need. Although this is not problematic with accurate data and ordinary rule mining, it can lead to major loss of true rules with noisy data or fuzzy SARM, and will be shown in Chapter 3 and 4.

2.3 ARM/SARM with uncertain data

Frequent itemset mining (similar to ARM except for not arranging antecedents and consequents) and ARM with uncertain data has also attract much research effort. The studies mostly employed a probabilistic data structure, where a probability value is associated with each record or attribute value to present the degree of uncertainty. Chui *et al.* (2007) found that mining frequent itemsets in uncertain probabilistic transactional data was either inapplicable or very inefficient by simple extensions of traditional algorithms, and presented a data trimming framework to improve the efficiency. Chui and Kao (2008) developed a decremental pruning technique for

itemset mining in uncertain data that was more efficient and robust than data trimming. Aggarwal *et al.* (2009) examined a wider variety of conventional itemset mining algorithms, and pointed out that they have very different performances and degrees of suitability for being extended to uncertain probabilistic data. Zhu *et al.* (2010) presented the data sampling technique for ARM with uncertain data, which was much faster and maintained relatively high accuracy as compared with mining from all data. Gonzales and Zettsu (2012) developed an ARM method using genetic network programming for uncertain data that was efficient for large databases.

The studies are of great value, yet hard to be used for resolving data error impact in statistical tests on association rules. Even these studies commonly list data error as a major source of uncertainty, they have yet addressed random error behaviours which are far from single probabilities associated with data entries. According to an exhaustive review by Carvalho and Ruiz (2013), all past research articles in several major indexing databases about uncertain ARM algorithms employed the probabilistic data structure, and none was for random data error.

SARM studies usually focus on specific uncertainties in geographic objects, and handle such uncertainties using fuzzy sets or other soft computing models, as will be detailed in Sections 2.4 and 2.6, instead of probabilistic data structure. But if SARM adopts the probabilistic structure after spatial attributes are computed and tabulated, it will share the above problem of ARM in random error handling.

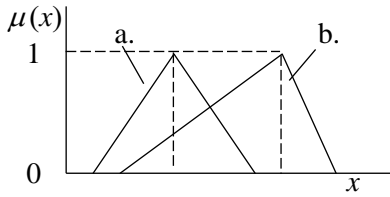
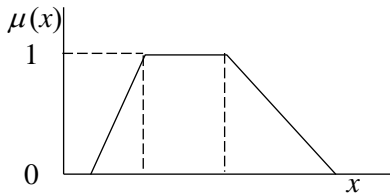
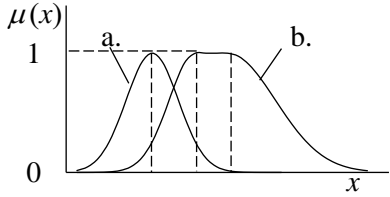
2.4 Fuzzy ARM/SARM

As said in Section 1.1, a key data preprocessing procedure in ARM/SARM is to discretize, or transform, raw numerical data into attribute-value ones, where the values are linguistic concepts, so that the data are ready for generating linguistic rules. In ordinary rule mining, each concept corresponds to a crisp raw data value interval,

which can cause bias and inaccurate representations of gradual or vague concepts. Fuzzy rule mining may relieve this problem, thereby improving the reliability of resultant rules, by modelling each concept as a fuzzy set of raw data values.

Consider a numerical attribute x with a generic value indiscriminate with the attribute name. A fuzzy data discretization model defines a *membership function*, μ_l : domain of $x \rightarrow [0,1]$, for each concept l . The *membership degree* of x in μ_l , $\mu_l(x)$, represents the degree to which x belongs to l . The *core* and *support* of μ_l is respectively $core(\mu_l) = \{x \in U \mid \mu_l(x) = 1\}$ and $supp(\mu_l) = \{x \in U \mid \mu_l(x) > 0\}$ (Bosc *et al.* 2007). Table 2.1 lists common types of previously proposed membership functions.

Table 2.1 Common forms of membership functions for fuzzy data discretization

Form	Graph	Reference(s)
Triangular		<p>a. Bilaterally symmetry: Herrera and Martinez (2000), Chen <i>et al.</i> (2008), Carmona <i>et al.</i> (2010)</p> <p>b. Bilaterally asymmetry: Alhajj and Kaya (2008)</p>
Trapezoidal		<p>Ladner <i>et al.</i> (2003), Patrick <i>et al.</i> (2007)</p>
Gaussian-curve		<p>a. Core for a single x value: Bordogna and Pasi (1993)</p> <p>b. Core for an x value range: Burda <i>et al.</i> (2014)</p>

Conjunctions of multiple membership degrees are evaluated by t-norm, an associative, commutative and monotone function $\otimes: [0, 1] \times [0, 1] \rightarrow [0, 1]$, $\alpha \otimes 1 = \alpha$ and $\alpha \otimes 0 = 0$ for each $\alpha \in [0, 1]$. SARM mostly adopts minimum t-norm: $\alpha \otimes_{\min} \beta =$

$\min(\alpha, \beta)$ and product t-norm: $\alpha \otimes_{\text{prod}} \beta = \alpha\beta$ (Laube *et al.* 2008). The *fuzzy support* of an itemset $V = \{x_1 = v_1 \dots x_m = v_m\}$ is

$$\text{supp}(V) = \sum_{R \in D} \mu_{v_1}(r_1) \otimes \dots \otimes \mu_{v_m}(r_m). \quad (2.9)$$

Fuzzy rule mining can then be conducted by using fuzzy instead of crisp supports in all patterns involved in RIMs. Replacing all membership degrees in fuzzy rules with binary memberships 0/1 reduces the task to ordinary rule mining.

Fuzzy SARM is of particularly value, as spatial data is rich of fuzzy concepts. Some obvious examples are nearness between spatial entities, and land covers with gradual transitions between land parcels of different covers. Fuzzy SARM was first formally presented by Ladner *et al.* (2003), and aforementioned spatial support and spatial confidence (Laube *et al.* 2008) are also based on fuzzy rules. Shu *et al.* (2008) explored association rules from vegetation and climate change data, where they used fuzzy *c*-means clustering to discretize numerical weather measurements and NDVI into vegetation and climate concepts. Fuzzy SARM has also been employed to find weights of risk degrees in site selection of emergency response centres (Fan 2014), and for investigating impact of air pollutant distributions on allergic asthma occurrences (Sadat *et al.* 2015).

These useful researches takes the premise that fuzzy SARM results are more reliable than the ordinary one, due to improved RIM accuracy for gradual or vague concepts in the former. However, it has rarely been empirically evaluated how much crisp RIM values actually deviate from fuzzy ones. The risk that fuzzy memberships reduce significance of rules and hence the number of true rules also need to be assessed, as said in Section 1.1. Besides, existing data discretization models for fuzzy SARM calls for more comprehensive mathematical justifications, for example, whether it is more reasonable to assign linear or curved membership functions for transitions of the fuzzy

sets, even existing spatial semantic studies have relevant outcomes to be extended to and examined in SARM.

2.5 GAs for ARM/SARM

As stated in Section 1.1, GAs may be used to find appropriate data discretization schemes for ARM/SARM, which addresses the prevalent inadequacy of expert knowledge to do so and variant user preferences on the schemes. GAs are metaheuristics that mimic natural selection and usually used to solve optimization and search problems (Mitchell 1996). The basic unit of evolution in a GA is a *chromosome*, or *individual*, which is a candidate of the entire or part of solution to an optimization problem. The GA starts with an initial *population* of individuals. In each successive generation, three *genetic operators* are applied to evolve the population toward better solutions:

- *Selection*: to distribute chances of surviving to the next generation or giving offspring among individuals. Individuals with better *fitness values*, computed according to one or more *objective functions*, have larger chances. Individuals with the best fitness values may become *elites* and have 100% chance to survive.
- *Crossover*: to recombine two chromosomes into offspring ones. A common approach is to select one or more crossover points on the parent chromosomes, and to swap parent genes between adjacent crossover points to produce two children. One or both of the children may be passed to the next generation.
- *Mutation*: to diversify the population genes by altering one chromosome. Locations where the values mutate are often randomly selected from the chromosome with a certain probability.

In GAs for ARM, individuals can be candidate rules or data discretization schemes. If candidate rules are encoded, membership functions for items in the rules may either

be predefined or encoded and optimized together. In any case, membership functions may have predefined shapes and other constraints. The objective functions are measures on goodness of the rules, including but not limited to RIMs.

Table 2.2 lists some representative GA-based ARM studies. The entire data discretization scheme includes the number of concepts for each attribute and membership functions of each concept. A *main rule* is a collection of candidate rules with the same attributes in the antecedents and same attributes in the consequents (to be detailed in Section 5.1.1). A DNF-type fuzzy rule is another collection with the same attributes in the antecedents and same item as the consequent.

Table 2.2 Representative GA-based ARM studies. \uparrow and \downarrow respectively indicates to maximize and minimize

Reference	Objective(s)	Chromosome	Membership functions (MFs)
Kaya (2006)	\uparrow support, \uparrow confidence, \downarrow No. of attributes	Main rule + MFs	Fuzzy, triangular
Salleb-Aouissi <i>et al.</i> (2007)	\uparrow gain-based measure ^a	Itemset + MFs	Ordinary
Chen <i>et al.</i> (2008)	\uparrow No. of large items ^b , \uparrow suitability ^c	Entire discretization scheme	Fuzzy, triangular
Alcalá-Fdez <i>et al.</i> (2009)	$\frac{\sum support(\text{large items})}{suitability}$	Entire discretization scheme, with fixed No. of concepts and shape of MFs; displacements of MF centres evolve in GA	Fuzzy, symmetrical triangular
Casillas and Martínez-López (2009)	\downarrow approximation error, \downarrow No. of DNF-type fuzzy rules/equivalent Mamdani fuzzy rules ^d	DNF-type fuzzy rule; MFs are predefined	Fuzzy, triangular

^a this measure favours high-support rules with low-support antecedent items

^b large items: items with support $>$ predefined min support

^c this measure favours MFs with small mutual overlaps and covering larger ranges

^d these measures favour more concise and interpretable rules

In the SARM context, after spatial attribute values are computed and tabulated together with non-spatial data, it is straightforward to employ GA for optimizing data discretization schemes and candidate rules. For example, Barb and Kilicay-Ergin (2013) utilized GA to explore association rules for more accurate semantic ranking of satellite images according to using low level features such as colours and shapes.

A pending issue for GA-based ARM/SARM is that the resultant rules are optimal takes the premise that these rules are authentic, or with minimal risk of being spurious. However, as said in Section 1.1 and will be reconfirmed in Chapter 3–5, spurious rules can take a large portion in the result. The issue may be resolved by integrating statistically sound tests on SARs. This calls for new methods to adjust significant levels of the tests, as the total number of potential rules in GA-based rule mining methods is different from that in conventional ones. Besides, not all chromosome encoding approaches for GA are suitable for the integration with statistical tests; in particular, integrating the encoding of entire data discretization schemes can be infeasible in terms of time consumption, as will be elaborated in Chapter 5.

2.6 Other uncertainty handling techniques in SARM

While this thesis focuses on nearness spatial relations, this section also presents exemplary SARM studies mainly for uncertain topologies, for a more wholesome overview on uncertainty handling in SARM.

(1) Pruning rules by known geographic dependencies (Bogorny *et al.* 2008)

This approach aims at pruning uninteresting SARs containing known geographic dependencies that are explicit in spatial database schemas, such as ‘gas stations must be on roadsides’. Such rules may not be spurious, but indeed downgrade the quality of SARM results and adds to the difficulty for users to interpret the results and make good decisions. The study utilized a two-step pruning strategy, the first step in data

preprocessing and the second in generation of itemsets, to prevent occurrences of well-known geographic dependencies in resultant rules before these rules are generated. This proved much more efficient than filtering out rules containing such dependences after all rules were generated indiscriminately. As this strategy can work after the computation and tabulation of spatial attributes, it may be readily incorporated into the methods for enhancing reliability of SARM results developed in this thesis.

(2) Spatial objects with broad boundaries

Studies have proposed models like the RCC-8 (Randell *et al.* 1992) and egg-yolk model (Cohn and Gotts 1996) for representing vague regions around areal spatial objects without crisp boundaries. Clementini *et al.* (2000) extended vague region modelling to multi-level SARM by proposing ‘composite regions with broad boundaries’. Such a composite region includes an inner certain region $A_1 \subseteq A_2$ and an outer region A_2 including its certain and uncertain parts. A_1 and A_2 need not to be single polygons, which allows for more general cases of uncertain regions. The broad boundary of the region is $\Delta A = A_1 \setminus A_2$ (Figure 2.2). Two areas with broad boundaries may constitute 56 topological relations. The study took these relations as bottom-level ones, and grouped them into 14 mid-level clusters and further to 4 top-level ones. The study also developed an efficient method for determining the topology between two regions across and within each of the taxonomical levels during SARM process.

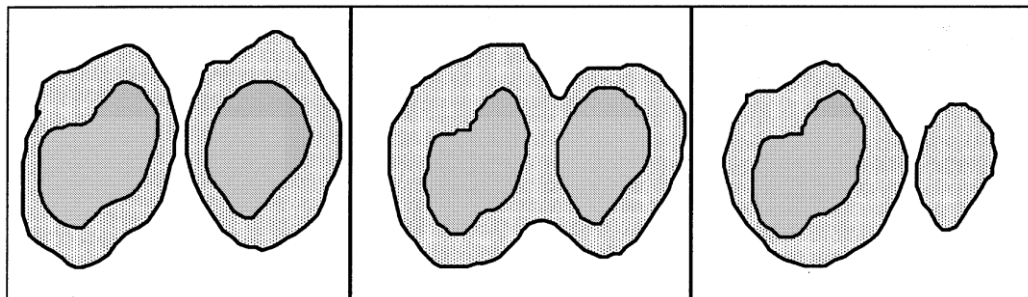


Figure 2.1 A composite region with a broad boundary (Clementini *et al.* 2000, p254)

(3) Rough set approach

Rough set theory is another extension to traditional set theory which can handles spatial data uncertainties. A rough set represents a set X by its lower approximation: $\underline{R}(X) = \{x \in U \mid [x]_R \subseteq X\}$ and upper approximation: $\overline{R}(X) = \{x \in U \mid [x]_R \cap X \neq \emptyset\}$, where U is the universe of all data; R is the equivalence relation defining the property held by X ; and $[x]_R$ is the set of x with the designated property. In SARM, $\underline{R}(X)$ and $\overline{R}(X)$ can model the uncertain inner and outer boundary of an area, or raw data value intervals that fully or at least partially support concept X in the data discretization for X . Each record may have weight 1 in relevant pattern supports if its raw numerical coordinates or value are in $\underline{R}(X)$, 0 if beyond $\overline{R}(X)$, and a , $0 < a < 1$ if in the boundary $\underline{R}(X) - \overline{R}(X)$ (Beaubouef *et al.* 2004).

While the weight is somehow like membership degrees in fuzzy SARM, fuzzy and rough sets depict different types of uncertainties in spatial data. Fuzzy sets are for vague or gradual concepts, such as ‘high’ with respect to elevation. In contrast, rough sets approximate concepts that are not vague, but cannot be precisely defined due to inadequate information available (Bai *et al.* 2014).

Taking highness as an example, if elevations over 2000m are definitely high while those below 1500m are definitely not, then elevations from 1500m to 2000m have increasing fuzzy membership degrees for ‘high’ in the fuzzy set approach, while they all have the same weight a for ‘high’ in the rough set one. Obviously the former is more reasonable. A case where the rough set is suitable is about a lake with clear boundary. In a low-resolution satellite image, all pixels the boundary falls in, and probably some nearby pixels, are uncertain to be in or out of the lake. As the exact location of the boundary is unknown, all areas in these pixels have equal probability to be in the lake. It is appropriate to represent these pixels as a rough boundary with equal weight a for ‘in the lake’.

Chapter 3 Mining significant SARs from uncertain data

* This chapter is primarily based on the published research article of the thesis author entitled “Mining significant association rules from uncertain data”.

The chapter presents a new ARM/SARM method for uncertain, or erroneous data. This method concerns the statistical testing stage in rule mining, and is based on the existing statistically sound test on rules reviewed in Section 2.2. Compared with existing statistical tests, the new method can discover more true rules, by making original mathematical corrections for impacts of random data error, and recovering rules lost due to the error. The method can also limit the risk of spurious rules upon a low user specified level. Hence, the new method will be referred to as the *corrected test*, and the existing statistically sound test will be called the *original test*.

While the corrected test is applicable for both spatial and non-spatial attributes, the case study in Section 3.3 demonstrates its particular usefulness for spatial data and mining spatio-temporal associations.

In this chapter, Section 3.1 presents the methodologies of the corrected test. More detailed outline of the methodologies is presented in Section 3.1.1. Sections 3.2 and 3.3 respectively illustrate methods and results of, and discuss about the synthetic and real-world data experiments on the new test. The real-world case study was for mining spatio-temporal associations between land uses and socioeconomics. Section 3.4 discusses the accuracy of data error information in practice and its implication to the practical value of the corrected test. Section 3.5 is a summary of this chapter.

3.1 Statistical test for association rules with uncertain data

3.1.1 Overview of corrected test

As said in Section 1.1 and will be confirmed by experiment results in Sections 3.2 and

3.3, data error mostly causes reduction of true rules in ARM/SARM results, since the error is of random nature and irrelevant to real associations in data. Data error can distort computational parameter values in statistical tests on the rules. The parameters are variables in the tests that involve supports of relevant itemsets, for example, $a-d$ for the Fisher's exact test in Equation (2.5). Distorted parameter values then result in distorted p values of the tests, mostly larger than their true values, as the rules are weakened. Finally, true rules may be rejected by the tests if the true p values resulted from the test on them are below the significant level, but the distorted p values go beyond.

In response of this problem, the corrected test models and corrects the distortions of the test parameter values due to data error. The corrected parameter values become more accurate, or closer to their true values, which in turn lead to more accurate p values, recovery of true rules lost due to the error, and finally the discovery of more true rules.

To correct the distortions in the test parameter values, the distortions must be quantified first. For this purpose, the coming Section 3.1.2 is devoted to a mathematical model describing the error propagation from source data to distortions of the test parameters. The model is originated in this thesis, as past uncertain ARM studies did not provide a model exclusively for random data error behaviours (see Section 2.3). Once the parameter value distortions are quantified by the error propagation model, the amount of correction to be made on the distorted parameters can be derived, and then the corrected test can be formally formulated. Such work is done in Section 3.1.3.

The corrected test can strictly control the risk of spurious rules since it makes the corrections based on the statistically sound evaluation. As explained in Section 2.2, existing statistically sound tests can limit the FWER to below a low user specified

level, for example 5%, and an even much lower percentage of spurious rules. Still, the corrections need to be carefully moderated to ensure that the corrected test can inherit the distinct advantage in spurious rule control from statistically sound tests. Section 3.1.4 discusses the technique to do such moderation.

This thesis exemplifies the original test by the statistically sound Fisher's exact test for productive rules (reviewed in Section 2.2). The method of modelling and correcting the data error is applicable to other statistical tests.

It is worth noticing that the losses of true rules due to data error cannot be alleviated by simply increasing the significance level of the original test: doing so can result in substantial increases of spurious rules, as only a small to moderate percentage of potential rules are authentic, according to Webb (2007) and experiment results in this chapter. It may be acceptable to increase true rules at the price of slightly more spurious ones. Yet since spurious rules can be very harmful, increase in true rules needs to be many times of that in spurious ones. This calls for targeted corrections to the original test for reducing impacts of data errors according to their statistical behaviours, as is done by the corrected test.

3.1.2 *Modelling error propagation*

To model errors in the test parameters, we start from the error on a single item in data. Consider an attribute a with values $1, \dots, k$ and any data record containing a . For $i, j=1 \dots k$, denote by p_{ij} the probability that the value of a in the record is i on condition that the true value of a is j . That is, $p_{ij} = \Pr(\text{value in data} = i \mid \text{true value} = j)$. Then there is

$$\mathbf{P} = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1k} \\ p_{21} & p_{22} & \cdots & p_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ p_{k1} & p_{k2} & \cdots & p_{kk} \end{pmatrix}.$$

Probabilities on the principal diagonal of \mathbf{P} corresponds to cases where $i = j$, or that the attribute values are correctly recorded. Other elements all represent the probabilities of error between different true and recorded value pairs.

This study adopts a common simplifying assumption in past researches on mining association rules from uncertain data: the independence between uncertain probability behaviors of different data items (Aggarwal *et al.* 2009). Under this assumption, the probability of each case that the error occurs in a , as recorded by each p_{ij} in \mathbf{P} , is invariant regardless of any other attribute values in each record. Thus, \mathbf{P} can provide all information about chance of error occurrence in the entire data that is needed for modeling propagation of error in a during the statistical test. We call \mathbf{P} the *proportional error matrix* of a . \mathbf{P} can be seen as a standardized form of the population error matrix, or confusion matrix (Ting 2011), where the standardization makes $\sum_i p_{ij} = 1$ for $j=1 \cdots k$.

Let c_i be the item representing value i in a . The *observed support* of c_i , $s(c_i)$, is the number of records containing c_i . Due to the data error, $s(c_i)$ is typically different from the unknown *true support* of c_i , $s_0(c_i)$. For $j \in [1, k]$, there are $s_0(c_j)$ records where the true value of a is j . In each of these records, recording the value of a as i can be seen as a Bernoulli experiment with the probability of success equal to p_{ij} . Then the number of records with true value j and recorded value i , $s(c_j \rightarrow c_i)$, is the number of successes in $s_0(c_j)$ such independent Bernoulli experiments, and follows a binomial distribution: $s(c_j \rightarrow c_i) \sim B(s_0(c_j), p_{ij})$. In ARM/SARM, normally $s_0(c_j) \gg 30$, $s(c_j)p_{ij} \gg 5$ and $s(c_j)(1-p_{ij}) \gg 5$, so the distribution of $s(c_j \rightarrow c_i)$ can be approximated by a normal distribution: $s(c_j \rightarrow c_i) \sim N(s_0(c_j)p_{ij}, s_0(c_j)p_{ij}(1-p_{ij}))$.

$s(c_i)$ is the number of records with any true values and recorded value i of a , that is,
 $s(c_i) = \sum_{j=1}^k s(c_j \rightarrow c_i)$. As $s(c_1 \rightarrow c_i) \dots s(c_k \rightarrow c_i)$ are mutually independent,

$$s(c_i) \sim N\left(\sum_{j=1}^k p_{ij}s_0(c_j), \sum_{j=1}^k p_{ij}(1-p_{ij})s_0(c_j)\right). \quad (3.1)$$

The expectation and variance of $s(c_i)$ are

$$E(s(c_i)) = \sum_{j=1}^k p_{ij}s_0(c_j), \quad (3.2)$$

$$\sigma^2(s(c_i)) = \sum_{j=1}^k p_{ij}(1-p_{ij})s_0(c_j). \quad (3.3)$$

Distributions for observed supports of all classes $1, \dots, k$ can be written in a matrix form:

$$\begin{pmatrix} E(s(c_1)) \\ \vdots \\ E(s(c_k)) \end{pmatrix} = \begin{pmatrix} p_{11} & \dots & p_{1k} \\ \vdots & \ddots & \vdots \\ p_{k1} & \dots & p_{kk} \end{pmatrix} \begin{pmatrix} s_0(c_1) \\ \vdots \\ s_0(c_k) \end{pmatrix}, \quad (3.4)$$

$$\mathbf{E}(\mathbf{S}(a)) = \mathbf{P}\mathbf{S}_0(a)$$

$$\mathbf{\Sigma}(\mathbf{S}(a)) = \begin{pmatrix} \sigma(s(c_1)) \\ \vdots \\ \sigma(s(c_k)) \end{pmatrix} = \begin{pmatrix} ((p_{11}(1-p_{11}))(s_0(c_1)) + \dots + (p_{1k}(1-p_{1k}))(s_0(c_k)))^{1/2} \\ \vdots \\ ((p_{k1}(1-p_{k1}))(s_0(c_1)) + \dots + (p_{kk}(1-p_{kk}))(s_0(c_k)))^{1/2} \end{pmatrix}. \quad (3.5)$$

3.1.3 Recovering test parameters

Equation (3.4) is equivalent to $\mathbf{S}_0(a) = \mathbf{P}^{-1}\mathbf{E}(\mathbf{S}(a))$. $\mathbf{E}(\mathbf{S}(a))$ is determined by \mathbf{P} and $\mathbf{S}_0(a)$, the latter being a vector of true supports and unknown in reality, thus $\mathbf{E}(\mathbf{S}(a))$ is also unknown and needs to be estimated. Once an estimation of $\mathbf{E}(\mathbf{S}(a))$, denoted by $\hat{\mathbf{E}}(\mathbf{S}(a))$, is determined, the estimation of $\mathbf{S}_0(a)$, $\hat{\mathbf{S}}_0(a)$, can then be solved:

$$\hat{\mathbf{S}}_0(a) = \mathbf{P}^{-1}\hat{\mathbf{E}}(\mathbf{S}(a)). \quad (3.6)$$

When expanded, Equation (3.6) is a matrix of k equations, each for one value in a .

The i th row of the matrix form shows the *estimated true support* for value i :

$$\hat{s}_0(c_i) = \sum_{j=1}^k p_{ij}^{-1} \hat{E}(s(c_j)), \quad (3.7)$$

where p_{ij}^{-1} is the element at position (i, j) of \mathbf{P}^{-1} .

As $\hat{E}(s(c_j))$ is the most probable value of the observed support $s(c_j)$, it is straightforward to take $\hat{E}(s(c_j)) = s(c_j)$. The probabilities that $s(c_j) > E(s(c_j))$ and $s(c_j) < E(s(c_j))$, or that $E(s(c_j))$ is overestimated and underestimated, are both 0.5. This “neutral” estimation is not always best with respect to the purpose of estimating $s_0(c_i)$; a more generic solution should be controlling the probability that $\hat{E}(s(c_j)) > E(s(c_j))$, or that $E(s(c_j))$ is overestimated, at any user specified value between (0,1). This can be achieved by incorporating the variance of $s(c_j)$ and a constant z . By perceiving $s(c_j)$ as $E(s(c_j)) + z\sigma(s(c_j))$, we take $\hat{E}(s(c_j)) = s(c_j) - z\sigma(s(c_j))$. The probability that $s(c_j) > E(s(c_j)) + z\sigma(s(c_j))$ is $1 - \Phi(z)$, where Φ is the cumulative distribution function of the standard normal distribution. The probability that $\hat{E}(s(c_j)) > E(s(c_j))$, equivalent to $s(c_j) > E(s(c_j)) + z\sigma(s(c_j))$ by this estimation, is also $1 - \Phi(z)$ (Figure 3.1).

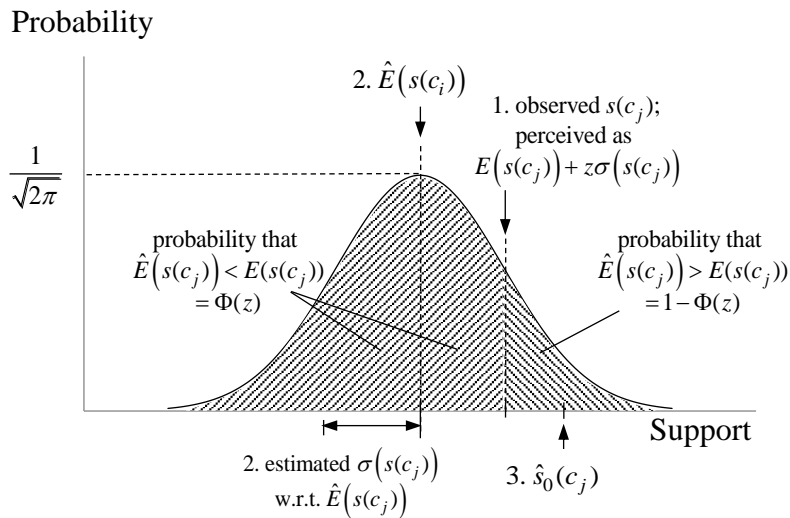


Figure 3.1 Using $\sigma(s(c_j))$ and z to control probability of overestimating $E(s(c_j))$ at arbitrary user specified value

For Equation (3.7), substitute $\hat{E}(s(c_j))$ by $s(c_j) - z\sigma(s(c_j))$, and $\sigma(s(c_j))$ by its expression in Equation (3.3):

$$\hat{s}_0(c_i) = \sum_{j=1}^k \left(p_{ij}^{-1} \left(s(c_j) - z \left(\sum_{l=1}^k p_{jl}(1-p_{jl})s_0(c_l) \right)^{1/2} \right) \right). \quad (3.8)$$

$s_0(c_l)$ is an unknown true support, so it should also take the estimated value $\hat{s}_0(c_l)$:

$$\hat{s}_0(c_i) = \sum_{j=1}^k \left(p_{ij}^{-1} \left(s(c_j) - z \left(\sum_{l=1}^k p_{jl}(1-p_{jl})\hat{s}_0(c_l) \right)^{1/2} \right) \right). \quad (3.9)$$

List all equations like Equation (3.9) for $\hat{s}_0(c_l) = \hat{s}_0(c_1) \dots \hat{s}_0(c_k)$ and combine them into a matrix:

$$\begin{pmatrix} \hat{s}_0(c_1) \\ \vdots \\ \hat{s}_0(c_k) \end{pmatrix} = \mathbf{P}^{-1} \begin{pmatrix} s(c_1) \\ \vdots \\ s(c_k) \end{pmatrix} - z \begin{pmatrix} ((p_{11}(1-p_{11}))(\hat{s}_0(c_1)) + \dots + (p_{1k}(1-p_{1k}))(\hat{s}_0(c_k)))^{1/2} \\ \vdots \\ ((p_{k1}(1-p_{k1}))(\hat{s}_0(c_1)) + \dots + (p_{kk}(1-p_{kk}))(\hat{s}_0(c_k)))^{1/2} \end{pmatrix}. \quad (3.10)$$

Equation (3.10) includes k equations and should have a unique solution for its k unknowns $\hat{s}_0(c_1) \dots \hat{s}_0(c_k)$. However, an exact solution of Equation (3.10) is complicated and computationally uneconomic. When only one $\hat{s}_0(c_i)$ is needed, all equations in Equation (3.10) have to be solved, and all of $\hat{s}_0(c_1) \dots \hat{s}_0(c_k)$ will be obtained. In the real operation, $\hat{s}_0(c_l)$ on the right side of Equation (3.9) can be approximated by the observed support $s(c_l)$:

$$\hat{s}_0(c_i) = \sum_{j=1}^k \left(p_{ij}^{-1} \left(s(c_j) - z \left(\sum_{l=1}^k p_{jl}(1-p_{jl})s(c_l) \right)^{1/2} \right) \right). \quad (3.11)$$

An analytic evaluation of the discrepancy between the $\hat{s}_0(c_l)$ values solved from Equations (3.11) and (3.10) is provided in Appendix 1. Also shown is that such discrepancy has minimal effect on the corrected test.

Let I be a set of items other than c_i . We first consider I as error free; if not, other erroneous items can in turn take the place of c_i and have their errors addressed. The “true” support of $I \cup \{c_i\}$ without the impact of error in c_i is $s_0(I \cup \{c_i\})$, and the observed support is $s(I \cup \{c_i\})$. Under the assumption in Section 3.1.2 that items are independent in their chances of error occurrence, Equation (3.11) still holds if c_i is substituted by $I \cup \{c_i\}$. Denote the estimated true value of $s(I \cup \{c_i\})$ with respect to \mathbf{P} and z by $\hat{E}(c_i, I, \mathbf{P}, z)$:

$$\hat{E}(c_i, I, \mathbf{P}, z) = \hat{s}_0(I \cup \{c_i\}) = \sum_{j=1}^k \left(p_{ij}^{-1} \left(s(I \cup \{c_j\}) - z \left(\sum_{l=1}^k p_{jl}(1 - p_{jl}) s(I \cup \{c_l\}) \right)^{1/2} \right) \right). \quad (3.12)$$

Equations (3.6)–(3.12) are applicable as long as \mathbf{P} is nonsingular, which is always the case when \mathbf{P} is diagonal dominant (Taussky 1949). This is equal to that $p_{ii} > 0.5$ for all $i = 1 \dots k$, as $\sum_{j=1}^k p_{ji} = 1$ (see Section 3.1.2). According to Section 3.1.2, if $p_{ii} \leq 0.5$, then the accuracy of value i is no more than 50%, and $s(c_i)$ will be distorted by at least 50%. In this case, any rules containing c_i would be so unreliable that it is recommended to remove i from \mathbf{P} and discard the rules containing c_i , instead of repairing $s(c_i)$ by the corrected test. Yet if one would like to anyway preserve c_i in resultant rules, $\hat{E}(c_i, I, \mathbf{P}, z)$ can still be solved by replacing \mathbf{P}^{-1} in Equations (3.6)–(3.12) by the Moore-Penrose inverse (Penrose 1955) of \mathbf{P} . The Moore-Penrose inverse is existent for any matrix and can be computed by well-established methods such as the one presented by Ben-Israel and Greville (2003). The resultant $\hat{E}(c_i, I, \mathbf{P}, z)$ is actually a minimum norm least squares solution, which is proven a minimum bias one (Rao and Mitra 1972).

Consider the Fisher’s exact test for productivity of a rule $X \rightarrow y$ (Equation 2.5) on one of its items $x_m \in X$. Parameters a, b, c and d for the test, as defined in Equation (2.3), can be rewritten as

$$\begin{aligned}
a &= s(X \cup \{y\}) \\
b &= s(X) - s(X \cup \{y\}) \\
c &= s((X - \{x_m\}) \cup \{y\}) - s(X \cup \{y\}) \\
d &= s(X - \{x_m\}) - s(X) - s((X - \{x_m\}) \cup \{y\}) + s(X \cup \{y\})
\end{aligned} \tag{3.13}$$

Where s denotes observed support for the itemset with data error impact. Let a_0, b_0, c_0 and d_0 be unknown true values of $a-d$. Applying Equation (3.12) to $a-d$ by altering contents of I and c_i will produce $\hat{a}_0, \hat{b}_0, \hat{c}_0$ and \hat{d}_0 , the estimations of a_0, b_0, c_0 and d_0 . Values of $\hat{a}_0 - \hat{d}_0$ should be less distorted than $a-d$. Therefore, when conducting a statistically sound Fisher's exact test following Equation (2.5) and using the significance level determined by Equation (2.6) – (2.8), replacing $a-d$ with $\hat{a}_0 - \hat{d}_0$ in may lead to more accurate p value, recover true rules lost due to data error, and finally increase the number of true rules discovered.

According to Equation (2.4), increasing a and d values and decreasing b and c values will reduce the p value, which makes both true and false rules more likely to pass the test. To guarantee that using the parameter z does not add to the risk of spurious rules, the z value should neither make a or d values increase nor b or c values decrease. Thus a non-negative z value should be used with $\hat{E}(c_i, I, \mathbf{P}, z)$ for correcting a and d , and $\hat{E}(c_i, I, \mathbf{P}, -z)$ for b and c .

In a rule $X \rightarrow y$, the erroneous item c_i may be x_m, y , or an item $x_e \in X$ other than x_m . The three conditions result in three different formulations of $\hat{a}_0 - \hat{d}_0$ values, as listed in Table 3.1. Values of $\hat{a}_0 - \hat{d}_0$ need to be rounded to the closest integers for the use in the Fisher exact test.

Table 3.1 Estimated true values of test parameters $\hat{a}_0 - \hat{d}_0$ with derivations

(a) Case 1: $c_i = x_m$

Derivation	$a = s(X \cup \{y\})$ $= s((X - \{x_m\}) \cup \{x_m\} \cup \{y\})$ $= s((X - \{x_m\}) \cup \{y\} \cup \{c_i\})$ $b = s(X) - s(X \cup \{y\})$ $= s((X - \{x_m\}) \cup \{x_m\}) - s((X - \{x_m\}) \cup \{x_m\} \cup \{y\})$ $= s((X - \{x_m\}) \cup \{c_i\}) - s((X - \{x_m\}) \cup \{y\} \cup \{c_i\})$ $a + c = s((X - \{x_m\}) \cup \{y\})$ $b + d = (a + b + c + d) - (a + c)$ $= s(X - \{x_m\}) - s((X - \{x_m\}) \cup \{y\})$
$\hat{a}_0 - \hat{d}_0$	$\hat{a}_0 = \hat{E}(c_i, (X - \{x_m\}) \cup \{y\}, \mathbf{P}, z)$ $\hat{b}_0 = \hat{E}(c_i, X - \{x_m\}, \mathbf{P}, -z) - \hat{E}(c_i, (X - \{x_m\}) \cup \{y\}, \mathbf{P}, -z)$ $\hat{c}_0 = a + c - \hat{a}_0$ $\hat{d}_0 = b + d - \hat{b}_0$

(b) Case 2: $c_i = y$

Derivation	$a = s(X \cup \{y\})$ $= s(X \cup \{c_i\})$ $a + b = s(X)$ $c = s((X - \{x_m\}) \cup \{y\}) - s(X \cup \{y\})$ $= s((X - \{x_m\}) \cup \{c_i\}) - s(X \cup \{c_i\})$ $c + d = (a + b + c + d) - (a + b)$ $= s(X - \{x_m\}) - s(X)$
$\hat{a}_0 - \hat{d}_0$	$\hat{a}_0 = \hat{E}(c_i, X, \mathbf{P}, z)$ $\hat{b}_0 = a + b - \hat{a}_0$ $\hat{c}_0 = \hat{E}(c_i, X - \{x_m\}, \mathbf{P}, -z) - \hat{E}(c_i, X, \mathbf{P}, -z)$ $\hat{d}_0 = c + d - \hat{c}_0$

(c) Case 3: $c_i = x_e \in X - \{x_m\}$

Derivation	$a = s(X \cup \{y\})$ $= s((X - \{x_e\}) \cup \{x_e\} \cup \{y\})$ $= s((X - \{x_e\}) \cup \{y\} \cup \{c_i\})$
------------	---

	$ \begin{aligned} b &= s(X) - s(X \cup \{y\}) \\ &= s((X - \{x_e\}) \cup \{x_e\}) - s((X - \{x_e\}) \cup \{x_e\} \cup \{y\}) \\ &= s((X - \{x_e\}) \cup \{c_i\}) - s((X - \{x_e\}) \cup \{y\} \cup \{c_i\}) \end{aligned} $
	$ \begin{aligned} c &= s((X - \{x_m\}) \cup \{y\}) - s(X \cup \{y\}) \\ &= s((X - \{x_m\} - \{x_e\}) \cup \{x_e\} \cup \{y\}) - s((X - \{x_e\}) \cup \{x_e\} \cup \{y\}) \\ &= s((X - \{x_m\} - \{x_e\}) \cup \{y\} \cup \{c_i\}) - s((X - \{x_e\}) \cup \{y\} \cup \{c_i\}) \end{aligned} $
	$ \begin{aligned} d &= s(X - \{x_m\}) - s(X) - s((X - \{x_m\}) \cup \{y\}) + s(X \cup \{y\}) \\ &= s((X - \{x_m\} - \{x_e\}) \cup \{x_e\}) - s((X - \{x_e\}) \cup \{x_e\}) \\ &\quad - s((X - \{x_m\} - \{x_e\}) \cup \{x_e\} \cup \{y\}) + s((X - \{x_e\}) \cup \{x_e\} \cup \{y\}) \\ &= s((X - \{x_m\} - \{x_e\}) \cup \{c_i\}) - s((X - \{x_e\}) \cup \{c_i\}) \\ &\quad - s((X - \{x_m\} - \{x_e\}) \cup \{y\} \cup \{c_i\}) + s((X - \{x_e\}) \cup \{y\} \cup \{c_i\}) \end{aligned} $
$\hat{a}_0 - \hat{d}_0$	$ \begin{aligned} \hat{a}_0 &= \hat{E}(c_i, (X - \{x_e\}) \cup \{y\}, \mathbf{P}, z) \\ \hat{b}_0 &= \hat{E}(c_i, X - \{x_e\}, \mathbf{P}, -z) - \hat{E}(c_i, (X - \{x_e\}) \cup \{y\}, \mathbf{P}, -z) \\ \hat{c}_0 &= \hat{E}(c_i, (X - \{x_m\} - \{x_e\}) \cup \{y\}, \mathbf{P}, -z) - \hat{E}(c_i, (X - \{x_e\}) \cup \{y\}, \mathbf{P}, -z) \\ \hat{d}_0 &= \hat{E}(c_i, X - \{x_m\} - \{x_e\}, \mathbf{P}, z) - \hat{E}(c_i, X - \{x_e\}, \mathbf{P}, z) \\ &\quad - \hat{E}(c_i, (X - \{x_m\} - \{x_e\}) \cup \{y\}, \mathbf{P}, z) + \hat{E}(c_i, (X - \{x_e\}) \cup \{y\}, \mathbf{P}, z) \end{aligned} $

As has been stated, the corrected test applies to other statistical tests than the exemplified Fisher's exact test. For this purpose, users need to derive the estimated true values of each parameter involving supports of relevant itemsets in corresponding tests based on Equation (3.12). Given $z \geq 0$, $\hat{E}(c_i, I, \mathbf{P}, z)$ should be used for parameters whose increase in values will reduce the p value of the test, and $\hat{E}(c_i, I, \mathbf{P}, -z)$ be used for parameters whose increase in values will enlarge the p value. The estimated true values should be derived considering the three conditions for the location of erroneous items in the rules, as has been done in Table 3.1.

3.1.4 Controlling spurious rules

For the method in Section 3.1.3, the z value is the key to the increase of true rules as well as the risk of spurious rules. A smaller z value leads to larger corrections to the Fisher exact test parameters, higher potential to recover true rules lost due to data error, yet also higher risk of overcorrecting these parameters and eventually spurious rules.

Ideally, a quantitative relation shall be established between the z value and the risk of spurious rules, particularly the FWER, thus z can be determined for a user specified maximum FWER. However, such an analytical solution is very difficult to achieve. As explained in Section 3.1.3, the z value only directly relates to the probability of overcorrecting each Fisher exact test parameter. This probability then links to the probability of overcorrecting p value of the test, the risk of individual spurious rules, and finally the FWER. There are numerous uncertain factors in this multi-step relation, for example, the value of each element in \mathbf{P} , original p value of the test before the correction for error impact, the significance level κ and the data size. It appears impossible to clearly quantify all the impacts of these uncertainties on the relation between z value and the FWER. If any factor is modelled very inaccurately, the entire quantitative relation will not work.

An alternative solution takes a simulation approach. The simulation skips the links in the above multi-step relation and directly identifies the z value that are expected to result in an FWER of up to a user specified maximum, denoted by r_{\max} . The simulation includes three steps:

- (1) For each column in the data table representing a certain attribute, reorder all values in the column in random sequence;
- (2) Generate association rules from the above randomized data, and apply the corrected test on generated rules. Starting from $z = 0$, increase the z value until all rules are rejected by the test. Record this smallest z value that makes all rules rejected;
- (3) Repeat steps 1 and 2 for n times. Find the largest z value recorded in n loops.

The largest z value recorded is then used in the statistical test on actually mining the erroneous data. The randomization in step 1 creates a new dataset where the support of each item is equal to that in the erroneous data, but all items are independent from each other. Any rules discovered from such randomized data must be spurious rules.

The randomized data maintains all features of the erroneous data except for the associations, thus it may simulate the numerous affecting factors to the relation between the z value and FWER.

There is a factor to this relation beyond the simulation, due to the fact that the p value of the Fisher exact test is more sensitive to a certain amount of change in a test parameter a , b , c , or d if the parameter is smaller. In the simulation with randomized data, items are expected to be independent, and even spurious productive rules occur, the rules are mostly weaker and thus have smaller supports than rules in actual data to be explored. As a is equal to the rule support, there are more large a values when the test is conducted on the actual data than in the simulation. Similarly, rules in the actual data have more small b and c values and large d values, as inferred by definitions of these parameters in Equation (2.3). Corrections in similar magnitudes using the same z value to rules in actual and randomized data can then lead to different influences on the resultant p value. On a small number of rules with very small b and c values, the influence of the correction may exceed the maximum influence encountered in the simulation. This may cause higher risk of over-correction and more spurious rules.

To address this issue, the simulation also recorded a *correctable range* for each test parameter, defined as the range of percentage changes, which can be positive or negative, in the parameter due to the correction in the simulation. When exploring the actual data, the four parameters in one test are corrected only if the correction to every parameter is within its correctable range. Otherwise, the correction is discarded, leaving the parameter values the same as in the original test. The correctable range limit is not posed when the simulation generates no false rules even at $z = 0$. In that situation, there will be false rules only if $z < 0$, which leads to larger corrections to the test parameters than $z = 0$. Still, the zero instead of negative z value is used, as z is for controlling the FWER and should not cause larger corrections than not using z or making it zero. Thus the potential of increasing true rules by the corrected test is

underutilized, and the range of corrections in the simulation is not the widest range that can control the FWER under r_{\max} .

The number of necessary loops n is determined by r_{\max} . Each loop is like a random sample from an infinite number of data randomizations that could be realized. If each time the randomized data has a chance of r_{\max} to accept any false rules, then the probability of obtaining up to one false discovery in each loop is

$$\begin{aligned}\Pr(K \leq 1) &= \Pr(K = 0) + \Pr(K = 1) \\ &= C_n^0 r_{\max}^0 (1 - r_{\max})^{n-0} + C_n^1 r_{\max}^1 (1 - r_{\max})^{n-1} \\ &= (1 - r_{\max})^n + n r_{\max} (1 - r_{\max})^{n-1}.\end{aligned}\quad (3.14)$$

As reducing z value by a minimum enumeration step will lead to spurious rules, to be on the safe side, $\Pr(K = 1)$ should be included. The n value is the smallest one making $\Pr(K \leq 1) \leq 0.5$. That is, when the data error shows average effect on the test in the simulation, the FWER cannot exceed r_{\max} . When $r_{\max} = 0.05$, the number of necessary loops is $n = 34$.

It needs to be noticed that through the simulation, the maximum FWER depends on r_{\max} rather than the significance level κ of the statistical test. Yet optimally the statistically sound test also takes $\kappa = \alpha/s$, where s is the total number of potential rules and $\alpha = r_{\max}$. This is because both the simulation and the statistically sound test control the FWER instead of individual spurious rules. Also, the two techniques should aim at achieving the same user specified maximum FWER (α or r_{\max}).

3.2 Synthetic data experiment

The corrected statistical test on association rules was experimented with both synthetic and real data. Potential rules to be evaluated were generated by the test using the K -Optimal Rule Discovery (KORD) algorithm (Webb and Zhang 2005). For unbiased comparison between results of the original and corrected test, the rules should undergo

minimal filtering by minimum support and confidence prior to the test. In this case, KORD is much more efficient than the popular Apriori typed algorithms in both Webb and Zhang (2005) and pilot experiments of this study. The experiments and all experiments in subsequent chapters were mainly implemented in MATLAB® R2012a for Microsoft Windows operating system.

The synthetic data experiment firstly examined the impact of data error on the statistical test, especially the loss of true rules, hence confirming the need for the corrected test; and secondly evaluated the corrected test in terms of recovering true rules and controlling spurious rules. The corrected test was also examined for its robustness against inaccurate error probability specifications and dependence between error probabilities of different data items. The synthetic data was generated with predesigned true rules, so true and spurious rules can be correctly judged when evaluating rules accepted by the statistical tests. Thus the synthetic data experiment serves a strong support to the later real data experiment: the latter can show practical value of the corrected test, but has less confidence in evaluating the correctness of resultant rules, as true rules behind real data are rarely known.

3.2.1 Data and methods

The data was generated as a set of records, each containing 8 attributes: att_0 , att_1 , att_2 , att_3 , x_0 , x_1 , x_2 and x_3 . att_0 included five values from 0 to 4. The other seven were binary attributes. In every record, the value of each attribute was assigned at random following a predesigned probability distribution. Thus the support of each value followed a binomial distribution: in n records with attribute att , if the probability of $att = 0$ is equal to p in each record, then $s(att = 0) \sim B(n, p)$. Following likewise distributions, the supports of all patterns have fluctuations. This is exactly the cause of spurious rules.

Value assignments of all attributes were equiprobable and independent, except that the probabilities of att_3 values depended on att_0 , att_1 and att_2 values, and such dependences are summarized in Table 3.2. Consider $att_0 = 0$ as the basic case. Then conditions $att_0 = 1$ or $att_0 = 2$ alone increased the probability that $att_3 = 1$, while conditions $att_0 = 3$ or $att_0 = 4$ increased the probability that $att_3 = 1$ only if $att_1 = 1$ and $att_2 = 1$. This was to simulate real-world data associations: sometimes a factor alone is associated to other factors, while sometimes only the concurrence of several factors establishes a new association with other factors. x_0 – x_3 were not in any predesigned associations, and they simulated the numerous ‘noise’ attributes irrelevant to interested associations in practical rule mining.

Table 3.2 Conditional probabilities of att_3 values in synthetic data

att_0	att_1	att_2	Probability of att_3 values	
			$att_3 = 0$	$att_3 = 1$
0	Any values		0.5	0.5
1	Any values		0.1	0.9
2	Any values		0.3	0.7
3	1	1	0.1	0.9
	Otherwise		0.5	0.5
4	1	1	0.3	0.7
	Otherwise		0.5	0.5

The predesigned associations led to 61 productive rules:

- $att_0 = 1$ or $att_0 = 2 \rightarrow att_3 = 1$ (2 rules);
- zero or more of $att_1 = 1$ or $att_2 = 1$, with or without $att_0 = 3 \rightarrow att_3 = 1$ (6 rules);
- $att_0 = 0 \rightarrow att_3 = 0$ (1 rule);
- zero or one of $att_1 = 0$ and $att_2 = 0$, and zero or one of $att_0 = 3$ and $att_0 = 4 \rightarrow att_3 = 0$ (8 rules);
- zero or one of $att_0 = 3$ and $att_0 = 4$, and $att_3 = 1$, with or without $att_2 = 1 \rightarrow att_1 = 1$ (6 rules);

- zero or one of $att_0 = 3$ and $att_0 = 4$, and $att_3 = 1$, with or without $att_1 = 1 \rightarrow att_2 = 1$ (6 rules);
- zero or one of $a_0 = 3$ and $att_0 = 4$, and $att_3 = 0$, with or without $att_2 = 0 \rightarrow att_1 = 0$ (6 rules);
- zero or one of $att_0 = 3$ and $att_0 = 4$, and $att_3 = 0$, with or without $att_1 = 0 \rightarrow att_2 = 0$ (6 rules);
- zero or one of $att_1 = 0$ and $att_2 = 0$, and $att_3 = 0 \rightarrow att_0 = 3$ or $att_0 = 4$ (6 rules);
- zero or one of $att_1 = 0$ and $att_2 = 0$, and $att_3 = 1 \rightarrow att_0 = 1$ or $att_0 = 2$ (6 rules);
- zero or one of $att_1 = 1$ and $att_2 = 1$, and $att_3 = 0 \rightarrow att_0 = 0$ (3 rules);
- one or more of $att_1 = 1$ and $att_2 = 1$, and $att_3 = 1 \rightarrow att_0 = 3$ (3 rules);
- $att_1 = 1$, $att_2 = 1$ and $att_3 = 1 \rightarrow att_0 = 0$ or $att_0 = 2$ (2 rules).

These rules varied a lot in their strength, or productivity. Rules with lower strength are more sensitive and likely to be lost when the data has error.

The predesigned productive rules included up to 3 items in their antecedents. In practice, the number of items involved in associations is usually unknown. Thus the experiment used variable largest number of items allowed in the antecedent, denoted by $maxL$. Each dataset was explored using $maxL = 3$ and $maxL = 4$. When $maxL = 3$, the total number of potential rules was $s = 10248$, and the statistically sound significance level was $\kappa = 4.88 \times 10^{-6}$. When $maxL = 4$, $s = 27608$ and $\kappa = 1.81 \times 10^{-6}$. For each dataset, data of five sizes comprising 4000, 8000, 16,000, 32,000 and 64,000 records were generated. The experimented statistical tests were not expected to discover all predesigned rules; the larger the data size, the more likely the rules would pass the test. For statistical hypothesis tests in general, increasing amount of data can provide more evidences to the hypotheses, thus making the hypotheses more statistically significant. For SARM this translates to that more rules become significant and pass the test. An example is given in Table 3.3a for a Fisher Exact test on the rule “(house) near river \rightarrow expensive”. In both the small and large datasets, 60%

houses near river and 40% houses not near river are expensive, and exactly half of the houses are near river. However, the test produces much smaller p value for the large dataset, showing that with more data, nearness to river is more certain to contribute to house price premiums. The rules are then more likely to have p values lower than the significance level and thus pass the test.

Table 3.3 Numbers of true rules from ‘ideal’ data and remarks

(a) Example of higher statistical significance of rules with increasing dataset size; a , b , c and d are Fisher’s exact test parameters in Equation (2.4)

No. of records	Small dataset		Large dataset	
	Near river	Not near river	Near river	Not near river
Expensive	$a = 6$	$c = 4$	$a = 60$	$c = 40$
Not expensive	$b = 4$	$d = 6$	$b = 40$	$d = 60$
Test result	$p = 0.3281$		$p = 0.0035$	

(b) Numbers of true rules from ‘ideal’ data

	$maxL$	Data size				
		4000	8000	16,000	32,000	64,000
No. of true rules	3	12	32	40	42	49
	4	12	30	38	42	49

For each data size, such ‘ideal’ data was generated that all items and itemsets in data had their expected supports. For instance, in the example earlier in this subsection, $s(att = 0)$ would be equal to np . The predesigned rules were examined using the ‘ideal’ data with the statistically sound test. The numbers of rules accepted for each data size were listed in Table 3.3b.

The original and corrected tests were also applied to data with artificial errors added into the original error-free data. The erroneous attributes were set to att_0 and att_3 , which were keys to the predesigned rules. For each attribute, a designated percentage of records containing each possible value were randomly selected to include the error and have the attribute value changed. For att_0 the value was assigned with

equiprobability to one of the remaining values, and for att_3 the value was swapped to the other of the two possible values. Records of all the attribute values were equiprobable to include the error. Denote the total error by e , then the error matrices for att_0 and att_3 are

$$\mathbf{P}(att_0) = \begin{pmatrix} 1-e & e/4 & e/4 & e/4 & e/4 \\ e/4 & 1-e & e/4 & e/4 & e/4 \\ e/4 & e/4 & 1-e & e/4 & e/4 \\ e/4 & e/4 & e/4 & 1-e & e/4 \\ e/4 & e/4 & e/4 & e/4 & 1-e \end{pmatrix}; \quad \mathbf{P}(att_3) = \begin{pmatrix} 1-e & e \\ e & 1-e \end{pmatrix}.$$

Data in four error levels were generated, and made nine experiment groups together with the original data:

- Original: the error-free data and original test was used;
- E20, E10, E05 and E02: for each of att_0 and att_3 , 20%, 10%, 5% and 2% of the records contained error. The selection of erroneous records was in random and independent. The original test was used;
- R20, R10, R05 and R02: the same data as their ‘E’ counterparts but the corrected test was used.

In E and R groups, elements in $\mathbf{P}(att_0)$ and $\mathbf{P}(att_3)$ were equal to actual error probabilities between corresponding attribute value pairs. That is, the error probability information was completely accurate. In practice, however, the error probability is also subject to inaccuracy (see details in Section 3.4). To evaluate the robustness of the corrected test to inaccurate error probability specifications, four more experiment groups using the corrected test were added:

- R20−/−: the data in E20/R20 with 20% actual error level for both att_0 and att_3 was used, while the perceived error level for the corrected test (the total e in \mathbf{P}) was 10% for both attributes;
- R10+/+: the data in E10/R10 with 10% actual error level was used, while the perceived error level was 20% for both attributes;

- R20+/-: the data in E20/R20 was used, while the perceived error level was 30% for att_0 and 10% for att_3 ;
- R10+/-: the data in E10/R10 was used, while the perceived error level was 15% for att_0 and 5% for att_3 .

‘+’ and ‘-’ refer to overestimation and underestimation of data error in the corrected test, respectively. These groups contained large inaccuracies in error specifications and focused on the highest two error levels, which pose the highest risk in affecting the corrected test. If the corrected test is robust then, so should it be with smaller data error or inaccuracy in error specification.

For reinforced practical value, the corrected test was further examined for its robustness against the breakage of the widely accepted assumption on independent error probability behaviours between data items (see Section 3.1.2). Four groups were designed to include two types of dependence between error probabilities:

- E10_ErrDep, R10_ErrDep: the data was the same as that in E10/R10, except that the error level of att_3 depended on that of att_0 : the error level of att_3 was 25% and 8.33% for records with erroneous and true att_0 values, respectively;
- E10_ValDep, R10_ValDep: the data was the same as that in E10/R10, except that the error level of att_3 depended on the value of att_0 : the error level of att_3 was 6% and 16% for records with $att_0 = 0-2$ and $att_0 = 3-4$, respectively.

In these four groups, att_3 had an aggregate error level of 10%, and both attributes had perceived error levels of 10%. Groups began with ‘E’ and ‘R’ respectively employed the original and corrected test.

The above 17 experiment groups, five data sizes and two $maxL$ values made up 170 combinations, each called a *treatment*. 50 datasets were generated for each treatment. The application of each treatment to a dataset is a *run*. There were 8500 runs in total.

For each dataset, the Original treatments produced a set of *reference rules* for each data size. In the corrected test, the simulation looped 34 times as required by a 5% maximum FWER.

3.2.2 Results

Among resultant rules accepted by the statistical tests, a rule was true if it was also in the 61 predesigned productive rules, and was false if not. Figure 3.2 plotted true rules, spurious rules and FWER for the E and R treatments. The corresponding numerical results are listed in Table 3.4. Each point in the figure and number in the table are the aggregation for 50 datasets. Results for treatments with inaccurate error probability specifications or dependent error probabilities will be listed later. In the results, the number of rules counted only rules containing att_0 and/or att_3 . The numbers of other rules being discovered were irrelevant to data error or the evaluation to the tests.

As described in Section 3.2.1, the Original treatments produced a reference rule set for each data size in every dataset. The numbers of true rules containing att_0 and/or att_3 in each reference rule set, also reported in the ‘Original’ rows in Table 3.4, were taken as 100% for computing percentages of rules in Figure 3.2 and hereafter for corresponding E and R treatments. This was because the error-free data and original test for Original treatments were control conditions against the erroneous data and corrected test.

As shown in Table 3.4, results for $maxL=3$ and $maxL=4$ were similar and shown almost identical trend of variation in relation to various conditions. This suggests the robustness of the statistical tests against variable $maxL$ values, which is desirable, as in practice the numbers of associated data items are usually unclear. Starting from Figure 3.2, all results refer to averages for the two $maxL$ values. In Figure 3.2, percentages of true rules varied significantly with data sizes and thus are plotted for

each data size; the spurious rules in different data sizes were relatively stable and thus aggregated. The essential numerical results are also summarized in Table 3.5.

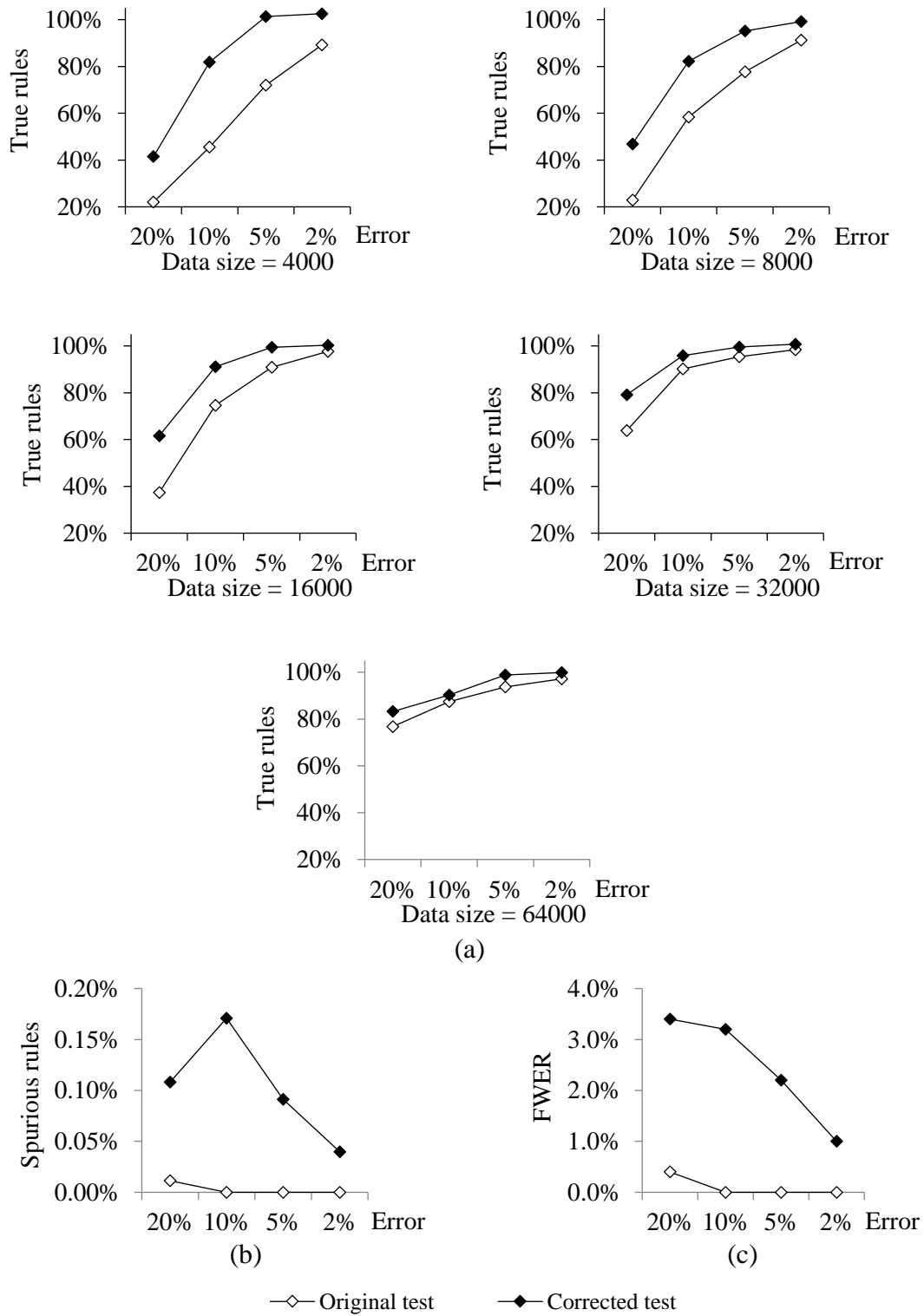


Figure 3.2 Synthetic data experiment results. (a) True rules (b) Spurious rules (c) FWER with respect to total number of runs

Table 3.4 Numerical synthetic data experiment results for E and R treatments

(a) True rules

Data size <i>maxL</i>	4000		8000		16,000		32,000		64,000	
	3	4	3	4	3	4	3	4	3	4
Original	15.62	14.20	29.36	27.32	39.44	38.66	44.04	43.38	49.90	49.18
E20	3.40	3.16	6.96	5.98	15.32	13.84	28.54	27.24	38.40	37.70
E10	7.20	6.38	17.36	15.74	29.96	28.32	39.72	39.08	43.72	42.96
E05	11.42	10.06	22.74	21.32	36.10	34.90	41.94	41.40	46.86	46.00
E02	13.94	12.68	26.90	24.80	38.50	37.76	43.38	42.64	48.58	47.70
R20	6.38	5.98	13.80	12.78	24.98	23.14	35.12	34.04	41.52	40.94
R10	12.88	11.52	24.06	22.56	36.24	34.94	42.22	41.60	45.14	44.32
R05	15.90	14.32	27.98	25.96	39.26	38.42	43.96	43.08	49.48	48.40
R02	16.30	14.28	29.08	27.12	39.52	38.82	44.30	43.78	49.92	49.04

(b) Serendipitous discoveries

Data size <i>maxL</i>	4000		8000		16,000		32,000		64,000	
	3	4	3	4	3	4	3	4	3	4
E20	0	0	0.04	0.04	0.02	0.02	0	0	0	0
E10	0	0	0.08	0.12	0.02	0.06	0.02	0	0.04	0.04
E05	0.06	0.14	0.12	0.20	0.02	0.06	0.10	0.02	0.20	0.18
E02	0.18	0.30	0.26	0.42	0.12	0.16	0.24	0.14	0.14	0.16
R20	0.50	0.42	0.28	0.28	0.14	0.14	0.08	0.06	0.18	0.08
R10	1.54	1.20	1.02	1.28	0.56	0.64	0.50	0.46	0.38	0.38
R05	1.66	1.60	1.38	1.50	0.78	0.76	0.78	0.66	0.82	0.78
R02	1.24	0.84	1.00	1.28	0.52	0.52	0.60	0.72	0.56	0.54

(c) Spurious rules

Data size <i>maxL</i>	4000		8000		16,000		32,000		64,000	
	3	4	3	4	3	4	3	4	3	4
Original	0	0	0	0	0	0	0	0	0	0
E20	0	0	0	0	0	0	0	0	0.02	0.02
E10	0	0	0	0	0	0	0	0	0	0
E05	0	0	0	0	0	0	0	0	0	0
E02	0	0	0	0	0	0	0	0	0	0
R20	0.08	0.08	0	0.02	0.02	0.02	0.06	0.02	0.06	0.02
R10	0.02	0	0.04	0.04	0.08	0.24	0.08	0.06	0.02	0.02
R05	0	0	0	0.02	0.08	0.02	0.04	0.04	0.06	0.06
R02	0.06	0.02	0	0	0	0	0	0	0.04	0.02

Table 3.5 Summary of synthetic data experiment results

	Original data ^a	Original test (E treatments)	Corrected test (R treatments)
No. of true rules	1404.2	1100.6	1253.1
No. of spurious rules	0	0.04	1.44
% of spurious rules	0%	0.01%	0.10%
FWER	0%	0.10%	2.50%

^a Values are (result in Original treatments) $\times 4$, corresponding to E and R treatments of 4 error levels.

(1) Original test

When the statistical test was applied to the original data with a significance level of 0.05, averagely over 140 rules were accepted. Most of these rules must be false since there were only 61 predesigned true rules. With the statistically sound significance levels, the test generated zero false rules, though a very small number of false rules could be generated provided more runs in the experiment (Webb 2007). The numerous spurious rules in absence of, and the minimal spurious rules in presence of the statistically sound test, were consistent with the previous study of Webb (2007). This confirmed the necessity and effectiveness of the statistically sound test as the basis of this chapter.

When applied to erroneous data, the original test maintained strict control on spurious rules. With the maximum FWER set at 5%, the actual FWER was only 0.1%, and the percentage of spurious rules was 0.01% (Table 3.5). That is, in terms of controlling spurious rules, the statistically sound test was robust to distortions posed by the data error to all patterns containing att_0 and att_3 . Apparently, RIM values computed from distorted pattern supports should also distort and cause the spurious rules to increase. However, under the assumption that the probability of error occurrence on each attribute was independent from values of other attributes (see Section 3.1.2), such error mostly distorts in proportion the support of a rule and its sub-patterns containing erroneous attributes. The proportional distortions largely cancel out each other when

these support values are used together to compute RIM values. For a rule $X \rightarrow y$, data error in an attribute value $x \in X$ occurs by a constant probability, whether the record contains y or not. Thus $support(X \cup y)$ and $support(X)$ tend to distort in proportion, and their distortions largely cancel out in $confidence(X \rightarrow y) = support(X \cup y)/support(X)$. Also, error in y changes $confidence(X \rightarrow y)$ and $confidence(X \setminus \{x\} \rightarrow y)$, $x \in X$ in rough proportion, while maintaining their differences. In both cases, the productivity of $X \rightarrow y$ with respect to its generalization $X \setminus \{x\} \rightarrow y$ would not be much affected, nor would relevant spurious rules be generated.

However, the data error did cause marked loss of true rules. The loss worsened with higher error levels and smaller data sizes. In E20 with data sizes 4000 and 8000, almost 80% true rules lost and only 3–6 rules were preserved (Figure 3.2a, Table 3.4a). Such few true rules hardly made up a meaningful resultant rule set. While 90% data accuracy is satisfactory for many applications, the true rule loss was still prominent in E10 and up to 50% with small data sizes. Thus the original test may not obtain enough true rules for practical uses. This poses the need for the corrected test.

(2) Corrected test with accurate error specifications

The corrected test in R treatments obtained more true rules than the original test in E treatments for all error levels and data sizes (Figure 3.2a). The true rule increase was more significant when the true rule loss in the original test was severer. For medium error levels and data sizes, true rule rates raise from 60%~70% with the original test to 80%~90% with the corrected test.

The increase of true rules can be standardized by their loss in the original test into:

$$recovery\ rate = \frac{\text{No. of SR/DR true rules} - \text{No. of SE/DE true rules}}{\text{No. of reference rules} - \text{No. of SE/DE true rules}} \times 100\% . \quad (3.15)$$

The true rule increases and recovery rates for various error levels are listed in Table 3.6. With smaller data error, the true rule increase dropped due to decreased room of improvement, yet the recovery rate became significantly higher. The average recovery rate for all error levels was 50.2%, suggesting that the corrected test made up around half of the loss in true rules, or the loss in value of resultant rules.

Table 3.6 True rule increases and recovery rates by error level

	R20	R10	R05	R02
z	0.78	0.27	0.06	0.01
True rule increase	16.6%	12.8%	9.7%	4.4%
Serendipitous discovery increase	0.6%	2.1%	2.7%	1.6%
Recovery rate	34.1%	55.8%	88.7%	107.5%
Average recovery rate	50.2%			

While the data error mostly led to loss of true rules, productive rules that were not discovered in Original treatments were in fact occasionally discovered in E and R treatments. We call such true rules ‘gained’ from erroneous data *serendipitous discoveries*. These rules are favourable but contrary to the expectation that random data error should cause loss of true rules. Still, serendipitous discoveries do result from the random nature of data error. As explained in Section 3.1.2, the observed support of a pattern S in erroneous data, $s(S)$, roughly follows a normal distribution with expectation $E(s(S))$ and variance $\sigma^2(s(S))$. When S contains associated items, the error tends to make $E(s(S))$ smaller than the true support $s_0(S)$. Thus, usually $s(S) < s_0(S)$, and rules like $X \rightarrow y, X \cup \{y\} = S$ become less significant. However, there is a small probability equal to $\Phi((E(s(S)) - s_0(S)) / \sigma(s(S)))$ that $s(S) > s_0(S)$, or that $X \rightarrow y$ might become more significant. Thus some rules originally rejected by the statistical test may now pass the test and become serendipitous discoveries.

To reassure that serendipitous discoveries happened purely by chance, and were not artefacts resulted from specific predesigned rules, an auxiliary experiment was conducted. Data error was added to the ‘ideal’ data used in Section 3.2.1 where all data patterns had their expected supports, and uniformly distributed among all attribute value combinations. Then the 61 predesigned rules were evaluated by the Fisher exact test using this ‘ideal erroneous’ data. This was similar to setting $s(S) = E(s(S))$ for each pattern S tested. All rules turned out to have larger p values and become less significant.

Serendipitous discoveries were small in numbers, and their increases from E to R treatments took small percentages relative to numbers of reference rules (Table 3.6). However, the increases were actually sharp and around 2–10 times of the number in E treatments (Table 3.4b). This was because serendipitous discoveries often have borderline p values barely exceeding the significance level but still lowest among all rejected rules. Such borderline rules were much more likely to get p values decreased below the significance level and accepted by the corrected test than arbitrary rules.

As shown in Table 3.6, serendipitous discovery increase was largely stable, with some decrease at the highest and lowest error levels (R20 and R02). The number of serendipitous discoveries in the corrected test itself changed likewise, as the corrected test had several times more serendipitous discoveries than the original test. At low data error levels, the increase of true rules dropped due to reduced room of improvement, thus serendipitous discoveries became more important, and took as many as 50% of the increase of true rules in R02. This is also the reason that the recovery rate rise at low data error levels. In R02, the recovery rate exceeded 100%, suggesting that the corrected test discovered even more true rules than the original test with error-free data. This was reasonable as serendipitous discoveries were not recovered from the lost true rules.

The number of serendipitous discoveries was relatively stable because their space of improvement included all productive rules that were rejected with the original data. This space did not change with variable error levels. Serendipitous discoveries decreased at very high and low error levels due to the z values determined by the simulation process. Recall that larger z values meant smaller correction to the test parameters. At very high error level, the same z value would lead to larger corrections and higher risk of spurious rules. Then z value had to be larger to control the FWER, as can be seen in Table 3.6. The large z value also limited the true rules, especially the serendipitous discoveries with borderline significances. In R02, the average z value were only 0.01, and z values in most runs were actually zero. As explained in Section 3.1.4, when $z = 0$ the potential of the corrected test in increasing true rules was not fully utilized, and the ability to recover serendipitous discoveries with borderline significances again became the most affected.

At all error levels, the corrected test controlled the FWER below 5% and spurious rules below 0.2% (Figure 3.2b, c). The much higher error rates in the corrected test than those in the original test seems inevitable, since the former must bear some risk of overcorrection. However, the FWER in the corrected test was still quite low. The average FWER was 2.5% (Table 3.5), indicating that on 97.5% occasions there were not any spurious rules. Computed from Table 3.5, the ratio between true and false discovery increases was about 109:1. Users would obtain 109 more true rules at the risk of one more false discovery. As shown in Webb (2007) and result part (1) of this section, statistically unsound tests on the rules usually had 100% FWER and high percentages of spurious rules. Thus the corrected test can be regarded to have essentially equal advantage in spurious rule control to the original statistically sound test.

(3) Corrected test with inaccurate error specifications or dependent data error

Experiment results with inaccurate error specifications or dependent data error were summarized in Table 3.7a. Results of R20 and R10 treatments with accurate error matrices and independent data error are also listed for direct comparison. The corrected test turned out largely maintained its efficacy for increasing true rules, compared with corresponding R20 or R10 treatments. The recovery rates were sometimes lower than that in R20 or R10, yet sometimes even higher. The largest recovery rate reduction occurred in R20+/- where the recovery rate was 20.3%, or 60% of that in R20. Considering the major loss of true rules in E20, the 20.3% increase of true rules was still significant.

Table 3.7 Synthetic data experiment results with inaccurate error specifications or dependent data error

(a) Results of corrected test

	R20	R20 -/-	R20 +/-	R10	R10 +/+	R10 +/-	R10_ ErrDep	R10_ ValDep
Recovery rate	34.1%	26.5%	20.3%	55.8%	70.4%	42.8%	65.3%	38.3%
% of spurious rules	0.11%	0.07%	0.09%	0.17%	0.10%	0.17%	0.23%	0.25%
FWER	3.4%	2.6%	2.4%	3.2%	2.8%	3.8%	3.4%	3.0%
z	0.78	0.24	0.77	0.27	0.83	0.23	0.27	0.27

(b) True rules in original test

	E20	E10	E10_ErrDep	E10_ValDep
True rule rate	51.4%	77.0%	79.3%	70.5%

R20-/- shows less decrease in the recovery rate than R20+/-, thanks to the dynamic determination of the z value by the simulation. While the underestimation of error levels reduced the corrections to the Fisher exact test parameters and consequently the chance of recovering true rules, it also decreased the chance that the simulation accepted any false rules at a certain z value. Then the z value determined was much smaller, only 0.24 for R20-/-, compared with 0.78 for R20 (Table 3.5a). This again enlarged the corrections to the parameters, as explained in 3.3. R20+/- lost more

recovery rate than R20-- as its z value had no obvious reduce compared to R20 (Table 3.5a). It seemed that the possibility of accepting false rules in the simulation, and thus the limitation on z values, mainly depended on the most overestimated error. R10++ exhibited an even higher recovery rate than R10, but it is not recommended to intentionally overestimate the error in order to obtain a higher recovery rate. As the true error probabilities are usually unknown, such practice may result in mixed overestimations and underestimations on error levels and decrease in the recovery rate, as with the case of R20+/-.

As specified in Section 3.2.1, in R10_ErrDep the error levels in att_0 and att_3 were positively correlated. Although contracting to the assumption of independent uncertain probability behaviours, this hardly disturbed the corrections to the test parameters according to error matrices of the attributes. Thus the recovery rate in R10_ErrDep was expected to be unaffected, and it actually increased to 65.3% as compared with 55.8% in R10. Yet this was unlikely to suggest that the corrected test worked better with correlated error probabilities. Rather, the positive correlation between error probabilities in att_0 and att_3 made the error concentrate on a smaller number of erroneous records than when the error probabilities were independent. Thus E10_ErrDep actually used less noisy data and also preserved more true rules than E10 (Table 3.7b). R10_ErrDep seemed to simply follow the previous revealed trend of higher recovery rates at lower data error levels.

On the other hand, the dependence of error probabilities in att_3 on att_0 values in R10_ValDep indeed disturbed the corrected test. The dependence actually made the error probabilities of att_3 beyond the representation of a single error matrix. Hence the recovery rate decrease in R10_ValDep should be due to the limit of the mathematical model for the corrected test, but only partially. Another reason should be the noisier data in R10_ValDep than that in R10. This can be inferred from the lower true rule rate in the original test (E10_ValDep) than E10 (Table 3.7b). According to Table 3.3,

att_0 values of 3 and 4 were more involved in data associations than other values. In R10_ValDep, att_3 had a higher error probability when $att_0 = 3-4$, making more information lost compared with the data in R10.

For all the treatments, the FWER was similar to that of the corresponding R treatments using accurate data error specifications, though there seemed a slight increase of spurious rules in R10_ErrDep and R10_ValDep. The robustness of the corrected test in controlling spurious rules is also expected, as the FWER was controlled by the simulation which worked regardless of the efficacy in increasing true rules.

3.3 Real-world data experiment: Mining spatio-temporal associations between land uses and socioeconomics

This experiment investigates how the corrected statistical test improves the value of real-world SARM results. The case study targeted at spatio-temporal association rules between land use and socioeconomic changes in Massachusetts, the US, in 1985 to 1999, using data in a geographical information system (GIS). Previous data analyses, including some SARM studies, drew some inconsistent conclusions on relations between land use and socioeconomic developments. It is difficult to convincingly judge whether the land use transformations were authentically, or significantly, relevant to socioeconomic developments. For example, in the representative SARM study on this topic of Mennis and Liu (2005), relevancy of land use changes was judged only with human perception, by whether confidences of rules with land use changes were significantly higher than those without. The statistical test appears a promising solution to this longstanding research difficulty.

3.3.1 Data and methods

Raw data was mainly collected from Office of Geographic Information (MassGIS),

Commonwealth of Massachusetts (2012) through online open access. The data was in ESRI® vector shapefile format, and primarily preprocessed with ESRI® ArcGIS Desktop. The data format and preprocessing platform combination was a popular mainstream in GISc studies and projects.

The land use data consisted of about 263,000 land parcel polygons, each having a land use attribute value in 1985 and another for 1999. The land uses included 21 classes, as defined by MassGIS and listed in Table 3.8. Figure 3.3 shows an overview of the study area and land uses in 1985 and 1999 of a small locality within it. Clearly, the locality experienced land use transformations towards urban ones. Such a geographic region is suitable for investigating associations between urbanization and socioeconomic changes.

Table 3.8 Land use classes of study area

Urban (12 classes)		Non-urban (9 classes)
Participation recreation	Urban open area	Forest
Spectator recreation	Transportation	Cropland
Water based recreation	Waste disposal	Pasture
Residential, multi-family		Wetland
Residential, < 1/4 acre lots		Mining
Residential, 1/4–1/2 acre lots		Rural open area
Residential, >1/2 acre lots		Salt wetland
Commercial		Water
Industrial		Woody perennial

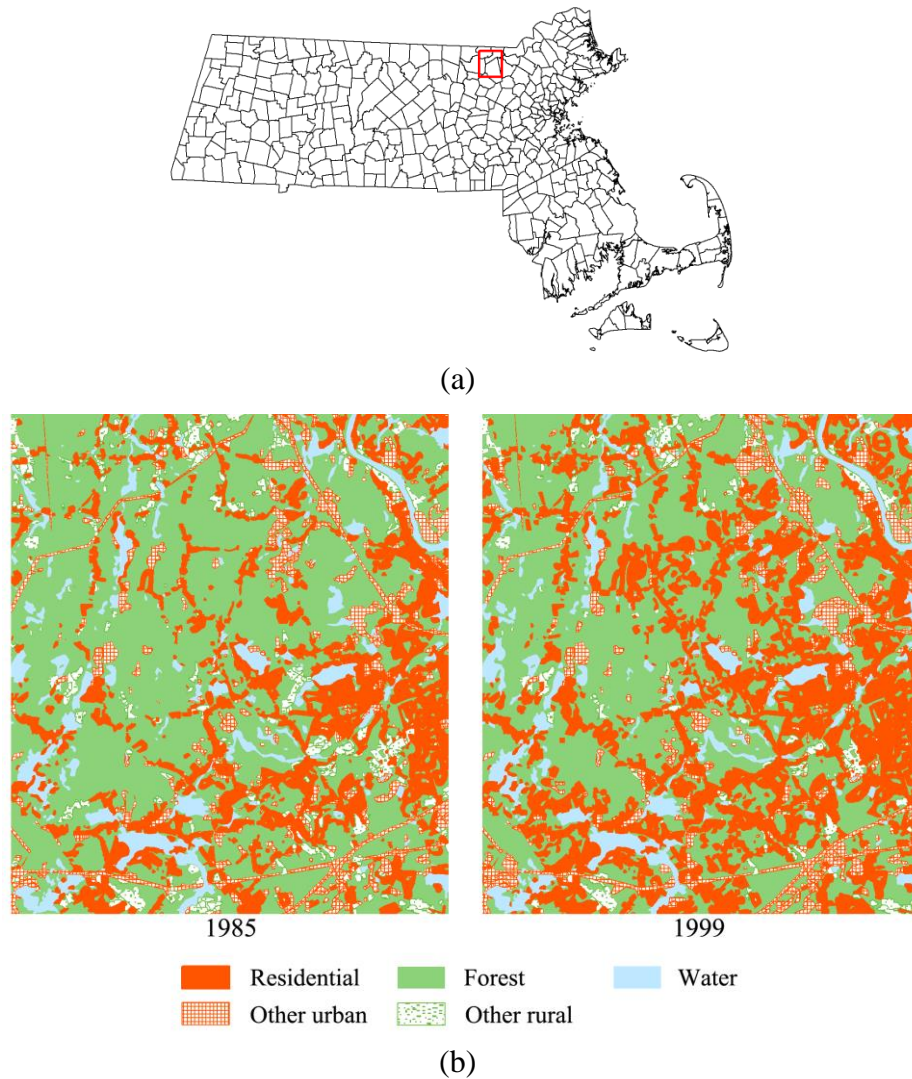


Figure 3.3 (a) Overview of Massachusetts with town boundaries. (b) Land uses of a small locality. The locality is marked by rectangular box in (a)

The socioeconomic data came from 1990 and 2000 census showing statistics in 1989 and 1999, respectively. It was the available GIS data closest in time to the land use data used. The socioeconomic data provided one record for each of the about 6000 building block groups in the study area. Four key socioeconomic measures were selected and computed into percent changes in 1990~2000:

- population,
- percentage of non-whites out of total population,
- house value median, and
- family income median.

Each of the four percent change attributes was discretized into five ordinal classes to be used as items of association rules. The five classes, labelled as Class 0–4, represented the lowest to highest percent increment (or decrement as negative increment) of a socioeconomic measure. The classification took the natural breaks scheme. The scheme could well reflect natural socioeconomic groupings across different areas, and it produced more reasonable results than quantile and equal interval schemes in both this experiment and representative past study on similar topic (Mennis and Liu 2005).

Land uses and socioeconomic data were collected using different geographic units, thus the two datasets were overlaid into a new layer of land parcel polygons. Each polygon was homogenous in all attribute values. For adding artificial data error in later process, each land parcel of area polygon within $[a \times 10000 - 5000, a \times 10000 + 5000)$ m² was further divided into a smaller ones. The land parcels were averagely divided into around 8 small polygons. Thus, most small polygons were around 10000 m², and none had more than 5000 m² discrepancy from that. Given the large number of polygons, discrepancies of individual polygon areas from 10000 m² were mostly cancelled out in later addition of artificial land use errors, and made the errors cover almost exactly the designed percentages of land areas.

The final data consisted of around 2,044,000 split land polygons, each linked to a record with six attributes, namely land uses in the two years and the four socioeconomic changes. This was the ‘original’ data regarded as error-free in this experiment.

To reconfirm the need for statistically sound control on spurious rules, an Original treatment was first applied to rules extracted from the original data with the original statistically sound test. To maintain feasible computation time and number of rules, a minimum support that generated only 10000 productive rules with at most four items

in the antecedents in the Original treatment was determined. The value was $7.06 \times 10^6 \text{m}^2$, or 0.035% of study area, and applied to all subsequent treatments.

Unlike the synthetic data, there were no predesigned rules behind real data for evaluating the efficacy of the statistical test in preventing spurious rules. An alternative evaluation was made by adding two artificial attributes to the data. Each artificial attribute contained five classes like the socioeconomic attributes, but the class values were randomly generated, equiprobable, and independent from values of other attributes. These two attributes then had no association with the rest of data, and any rules involving them must be false. Five datasets with artificial attributes were generated and experimented with the original test.

For evaluating the corrected test, 20%, 10% and 5% artificial error was added to each land use attribute in original data. The error levels simulated the quality of real-world data: land use data from automatic satellite image classification typically contain 10–15% error, and 20% error was a common threshold for acceptance (Olson 2008). This produced six treatments:

- E20, E10 and E05: used data with 20%, 10% and 5% error, respectively, and the original statistical test;
- R20, R10 and R05: used the above data and the corrected test.

Among all the land uses, the dominant Forest class covered 60% of the study area. This class tends to have much higher classification accuracy than terrestrial non-Forest classes, according to studies on primarily Massachusetts (Hollister *et al.* 2004) and eastern United States which covers Massachusetts (Yang *et al.* 2001). In the study area, Forest is in larger patches, or continuous areas of single land uses, than non-Forest classes. Land uses in large patches can be more accurately classified than those in small, fragmental patches (Smith *et al.* 2003). In order to include this essential realistic condition in the experiment while maintaining its simplicity, the error was

assigned equiprobable for non-Forest classes, but decreased for Forest to such a degree that the area of Forest remained unchanged after the error was added. The aggregated error for all classes was 5–20% as designated. The resultant error probability for Forest was about 2/3 of that for non-Forest classes. At each error level, five erroneous datasets were generated and experimented.

The six attributes in data made up 988,360 potential rules with up to four items in the antecedents. The statistically sound significance level κ was 5.06×10^{-8} with respect to a 5% maximum FWER. For the experiment with two artificial attributes, there were 5,619,990 potential rules, and κ was equal to 8.90×10^{-9} . The corrected test used the same increment step for z value and number of loops as the synthetic data experiment.

3.3.2 Results

From each dataset with the two artificial attributes, about 50,000 productive rules containing artificial attributes were generated besides the 10,000 rules involving only the six original attributes. None of the rules containing artificial attributes were accepted by the statistically sound test. The result from only five datasets was not enough for computing an FWER, yet it should demonstrate the effectiveness of the test in pruning spurious rules, and imply that rules accepted by the test were indeed likely to be true.

Out of the 10,000 productive rules from original data, about 3800 rules were rejected by the test. The large number of rules with dubious reliability coincided with the study by Webb (2007), and reconfirmed the essentiality of the statistical sound test to protect users against harmful spurious rules. Below is an example of rejected rules:

Land use changed from Forest to Residential, >1/2 acre lots →
 Percentage of non-whites increase = 4 (highest)
support = 0.289%, *confidence* = 0.164, *p* = 0.0220

‘Residential, >1/2 acre lots’ is the dominating residential category. Without the statistically sound test, this rule would be presented to users and deliver likely fake information that large increase in non-whites was related to development of residential area. Policy makers might be misled to concentrate facilities for ethnic minorities on new residential areas in former woodlands. This could waste resources and hinder allocation of the facilities to actual needy places.

As commonly in practical data mining, besides the statistical test, more RIMs were needed to trim resultant rules to the amount that human users could consider. Here the leverage measure as listed in Section 2.1 was used. Leverage is very suitable for evaluating productivity of rules, as it directly measures the number of additional records containing the association between the antecedent and consequent of a rule more than that if the antecedent and consequent are unrelated (Webb and Zhang 2005). Another filtering measure was to only include ‘non-Forest’ rules which contained at least one non-Forest land uses. Rules involving only the dominant Forest and socioeconomic changes were of dubious value, as they were often artefacts due to associations between other land uses and opposite socioeconomic changes. For instance, rules for associations between several urban land uses and high population increase were likely to result in another rule between Forest and low population increase.

1000, 500 and 200 significant ‘non-Forest’ rules with the highest leverages from original data were used as reference results. Rules accepted in E and R treatments were regarded as true rules if they were also among the reference results. According to Section 3.2.2, some rules in E and R treatments but not in reference results could be serendipitous discoveries and still true, but this could not be evaluated, as the real data held no predesigned true rules. Yet Section 3.2.2 also suggests that serendipitous discoveries take small fractions of true rules, and their absence should only cause slight underestimation on the merit of the corrected test, rather than overestimation.

Figure 3.4 shows true rules in different treatments, in terms of percentages relative to sizes of reference results (1000, 500 or 200). With all error levels and sizes of reference results, the corrected test improved the numbers of true rules, and recovered 20–50% of true rule loss in the original test. Compared with the synthetic data experiment, the real data experiment had less significant true rule loss and less improvement of the corrected test. This seemed mainly attributable to the much larger volume of the real data than synthetic data. As suggested by Webb (2007) and the synthetic data experiment in Section 3.2.2, more sufficient data leads to more true rules in the test. Recovery rate of the corrected test might also be underestimated due to the exclusion of serendipitous discoveries.

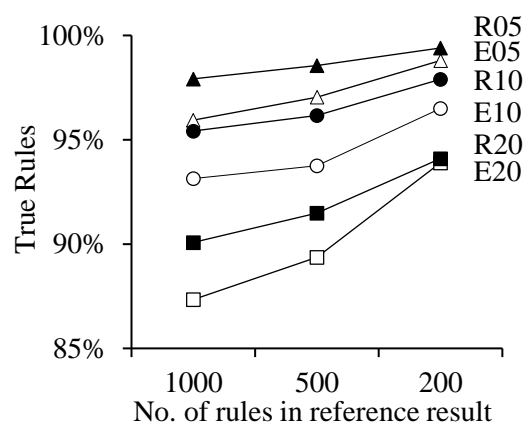


Figure 3.4 Recovery of true rules by corrected test in real-world data experiment

It may be argued that over 85% true rules discovered by the original test (Figure 3.4) are already informative for users, thus the corrected test is unnecessary. However, the situation became very different for rules that included land use changes. Such rules are of the highest interest among resultant rules, as they reveal relations between land use transformations, mostly towards urbanisation, and socioeconomic developments. Out of over 6000 significant rules from original data, only 99 contained different land uses in 1985 and 1999. At the scale of the entire state, even significant land use changes usually involved only small parts of total land area, while most places

maintained their land uses. Thus these valuable land use change rules typically had small supports and leverages, and were rare in the result.

The same reason made the land use change rules highly sensitive to data error. The true rules involving land use changes are plotted in Figure 3.5. What were taken as 100% in the figure were not the 99 significant rules as said above, but parts of them that were at least productive in corresponding treatments. The land use change rules were so sensitive that some significant rules in original data became even unproductive in erroneous data. Those rules were excluded because they were lost before the statistical test and beyond the study scope. The original test lost half of true rules at even 5% error level, and preserved few true rules at 20% error level. Meanwhile, the corrected test resulted in 2–4 times as many true rules as the original test.

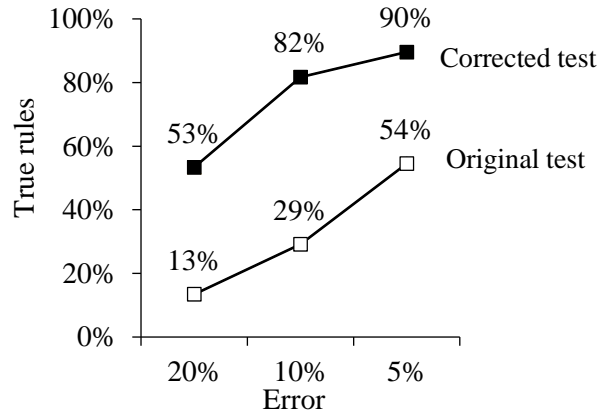


Figure 3.5 Recovery of true rules involving land use changes in real data experiment

The following is an example of recovered land use change rules:

Land use changed from Forest to Residential, >1/2 acre lots \wedge
House value increase = 4 (highest)
→ Income increase = 4 (highest)
support=0.188%, *confidence* = 0.410, $p=1.00 \times 10^{-55}$

This rule suggests that the Forest to residential land use change and large house value increase had an additive association with large income increase, compared with their individual relations to the latter.

While the most meaningful rules suffered from severe loss in the statistical test, the corrected test exhibited remarkable ability in recovering the loss and improving user knowledge to associations between urban and socioeconomic developments. Cases in practice are usually similar: rules with large supports and leverages and robust to data error are usually trivial or do not contain attributes of high interest. This provides great potential of the corrected test in adding value to SARM results.

The corrected test is also promising in saving time and cost of practical data mining, by allowing for the use of cheaper or faster collected data of slightly lower quality, while obtaining a result as good as mining more accurate data with the original test. In this experiment, the corrected test achieved over 80% true rules at 10% error level, far above the 54% true rules of the original test at 5% error level. While land use data with 5% error generally requires manual interpretation, data with 10% error could be achieved by automatic computerized classification which consumes only a small fraction of time and cost taken by the manual process.

3.4 Accuracy of error probability information and its practical implication to corrected test

With explosive accumulation of data in the contemporary world, increasing efforts have also been invested to the data quality assessment by scientific and industrial communities. For categorical data, quality measurement based on the confusion matrix (Ting 2011) is very popular and very often a standard approach. The assessment requires a set of reference data covering a sample of instances in the data under the quality assessment. Usually the reference data is from a more accurate

source and perceived as error-free. The confusion matrix can be obtained by counting the number of records with each pair of “true” value in the reference data and the value in the assessed data, and then standardized into the error matrix **P** used in the corrected test. Sometimes the reference data is not regarded as error-free but provides strong clues for estimating the element values in **P**.

When assessing the quality of remote sensing image classification, the reference data is usually the true classification through field surveys, or from a more accurate remote sensing data source, for example, human interpreted aerial photos as the reference for assessing automatically classified satellite images (Foody 2002, Stehman *et al.* 2008). For evaluating the quality of business and social statistics, quality surveys are conducted for producing the reference data with higher accuracy, such as face-to-face interviews for assessing statistics from self-completion questionnaires (Office for National Statistics 2014), and re-interviews by experienced staff for assessing statistics from interviews (Jones and Lewis 2003). Reference data initially collected for other purposes have also been used, such as detailed demographic registration data (Fosu 2001) and Census Dress Rehearsal data (Bishop 2009) for assessing census data.

Albeit widely available, the error matrix is seldom completely accurate. As the data quality assessment evaluates only a sample of the assessed data, the resultant **P** is subject to sampling error (Office for National Statistics 2014). Moreover, the bias in sampling the assessed data may sometimes be inevitable. For example, field surveys for accessing remote sensing image classification accuracy are limited to human accessible places. Different collection times of the assessed data and reference data can also add to the discrepancy between them (Hollister *et al.* 2004). Such discrepancy would be attributed to the error in the assessed data and lead to overestimation of the error probability. One of the rare cases of obtaining perfect **P** is when the data is deliberately perturbed with error, for purposes like privacy protection. Therefore, the

robustness of the corrected test to inaccurate error probability specifications, as demonstrated in Section 3.2.2, is crucial and advantageous for its practical usefulness.

Appropriate reference data may be unavailable for assessing the quality of, for example, historical data or data for rapidly changing phenomena. In this case, machining learning methods like the one presented by Zhu *et al.* (2004) can be used to detect the data error solely using the assessed data. These methods usually work by identifying the instances that most disturb inherent characteristics in data as erroneous, so they construct minimum estimations of the error probabilities instead of accurate error matrices. The corrected test would be still effective using error matrices filled with such minimum estimations, as it largely maintains the ability of increasing true rules with underestimated error probabilities (see Section 3.2.2). Using error-aware data mining methods like the corrected test is usually preferable to removing or trying to correct the erroneous records, as the latter can incur information loss or introduce new errors (Zhu and Wu 2006).

3.5 Summary

This chapter presents a novel method for ARM/SARM with uncertain data. The method improves the reliability of rule mining results by recovering true resultant rules lost due to random error in data, while controlling the risk of spurious rules at a low user specified level. A mathematical model was originated to describe the propagation of data error in the statistical test computations. Based on this model, techniques were developed to recover true rules via correcting the test for impacts of data error, as well as to control the risk of spurious rules.

When assessed with synthetic data, the new method recovered averagely 50% true rules lost due to data error with accurate error probability information. Its ability for recovering true rules is also robust against inaccurate error information and

dependences among the error and attribute values. The new method maintained superior control on spurious rules by existing statistically sound technique, and achieved a spurious rule rate below 0.2% and a FWER below 5%. In the case study on spatio-temporal rule mining from land use and socioeconomic data, the new method discovered several times as many those most practically useful but sensitive rules containing land use changes as by the existing technique.

Chapter 4 Mining significant crisp-fuzzy SARs

This chapter presents new techniques for improving reliability of fuzzy SARM results. The first technique, presented in Section 4.1.1, is a *Gaussian-curve-based fuzzy data discretization model*. While studies have proposed fuzzy data discretization based on Gaussian curves for individual concepts (Table 2.1), the newly proposed model more comprehensively integrates spatial semantics and multi-concept relations. The second technique, presented in Section 4.1.2, is *crisp-fuzzy SARM* that integrates statistically sound tests on crisp rules and evaluation of RIMs based on fuzzy supports. This method is designed to combine more abundant true rules in crisp SARM, higher RIM accuracy of fuzzy SARM and minimal spurious rules attained by statistically sound tests.

The techniques are experimented with synthetic data in Section 4.2 and real-world wildfire factor data in Section 4.3. The synthetic data experiment also seeks to prove the superior RIM accuracy of fuzzy SARM to the ordinary one, which was previously widely believed by rarely experimentally examined. Finally, a summary is presented in Section 4.4.

4.1 Proposed techniques

4.1.1 Gaussian-curve-based fuzzy data discretization

The proposed model transforms a numerical spatial attribute x to ordinal concepts $l_1 \dots l_k$ that are gradual or vague in nature. For instance, x is distance and $(l_1, l_2, l_3) = (\text{near}, \text{medium}, \text{far})$. By saying ordinal concepts, we mean that the concepts cannot fully or partially imply one another, though their corresponding x value ranges can overlap. Therefore, $(l_1, l_2, l_3) = (\text{near}, \text{medium}, \text{medium to far})$ is invalid. For concepts with well-defined boundaries (not gradual or vague), such as ‘above/below sea level’ for numerical attribute ‘elevation’, the model shall be

specialized to place crisp boundaries between the concepts.

The model has the following characteristics:

- (1) For each membership function $\mu_{l_j}, 1 \leq j \leq k$, the sections with $0 < \mu_{l_j}(x) < 1$, named *transitions* after ‘transitions between concepts’, are Gaussian curves and can be symmetric or not. $core(\mu_{l_j})$ is non-empty and in arbitrary size.

Gaussian curves have been widely used for characterizing degrees to which numerical spatial attribute values belong to linguistic concepts, especially for proximity measures. Gaussian weighting function is most commonly used in fixed-kernel geographically weighted regression (Wu *et al.* 2014). The weight represents the impact factor due to nearness between spatial objects, and is equivalent to μ_{near} in fuzzy SARM. Robinson (2000) used Gaussian functions to learn fuzzy spatial relations via human-machine interaction. Worboys (2001) proved by empirical study that the degrees people perceive two places as ‘near’ or ‘not near’, translating to μ_{near} and μ_{far} in fuzzy SARM, exhibit S-curve trend against Euclidean distances. The S-curve (sigmoid function) trend was visually interpreted and thus indistinguishable from a Gaussian-curve trend. Actually, the precise curve form is not that essential; the essence is, in contrast to triangular or trapezoidal membership functions, transition curves shall have larger slopes in the middle and smaller towards the endpoints. This reflects the fact that for x values in the middle of transitions, people have more uncertainty (modelled by curved slopes) in judging to which concept x should belong. Gaussian-curve transitions are also robust to uncertain and usually non-ideal value intervals of $core(\mu_{l_j})$, since the curves smoothly connect to core endpoints with reducing slopes towards zero (Bordogna *et al.* 1991).

Asymmetric transitions are critical for spatial concepts, as geographical data prevalently have rank-size relations in heavy-tailed distributions. That is, the data includes a ‘head’ containing a minority of extra-large sized objects, and a ‘tail’

containing the majority of small-sized objects; ‘size’ may be city population, road connectivity, or other impact measures (Jiang 2013). The heavy-tailed distribution recursively happens within the ‘head’ of data, thus raw numerical data value intervals for low-impact to high-impact concepts should increase exponentially instead of linearly. For example, populations of small, medium and large towns are more likely 1:5:25 than 1:5:9. Then the left transition of ‘medium town’ concept towards ‘small town’ shall be much narrower than the right one towards ‘large town’.

- (2) For relations between $l_1 \dots l_k$, each $supp(\mu_{l_{j-1}})$ touches $core(\mu_{l_j})$, $1 < j \leq k$, neither overlap nor disjoint.

This characteristic follows a common approach in past studies (Herrera and Martinez 2000, Alhajj and Kaya 2008, Carmona *et al.* 2010). Then $\mu_{l_j}(x) = 1 \Leftrightarrow \mu_{i \neq j}(x) = 0$, or each x value completely belongs to l_j if and only if it does not at all belong to any other concepts. The proposed $\mu_{l_1} \dots \mu_{l_k}$ are as below and illustrated in Figure 4.1:

$$\begin{aligned}
 \mu_{l_1}(x) &= \begin{cases} 1, & x \leq cr_{1_R} \\ \exp\left[-(x - c_{1_L})^2 / (2\sigma_{1_L}^2)\right], & cr_{1_R} < x < cr_{2_L} \\ 0, & x \geq cr_{2_L} \end{cases} \\
 \mu_{l_j}(x) &= \begin{cases} 0, & x \leq cr_{(j-1)_R} \text{ or } x \geq cr_{(j+1)_L} \\ \exp\left[-(x - c_{j_L})^2 / (2\sigma_{j_L}^2)\right], & cr_{(j-1)_R} < x < cr_{j_L} \\ 1, & cr_{j_L} \leq x \leq cr_{j_R} \\ \exp\left[-(x - c_{j_R})^2 / (2\sigma_{j_R}^2)\right], & cr_{j_R} < x < cr_{(j+1)_L} \end{cases}, \quad 1 < j < k, \\
 \mu_{l_n}(x) &= \begin{cases} 0, & x \leq cr_{(n-1)_R} \\ \exp\left[-(x - c_{n_L})^2 / (2\sigma_{n_L}^2)\right], & cr_{(n-1)_R} < x < cr_{n_L} \\ 1, & x \geq cr_{n_L} \end{cases}
 \end{aligned} \tag{4.1}$$

where $core(\mu_{l_j}) = [cr_{j_L}, cr_{j_R}]$, and σ_{j_L} and σ_{j_R} are standard deviations of left and right transitions in μ_{l_j} .

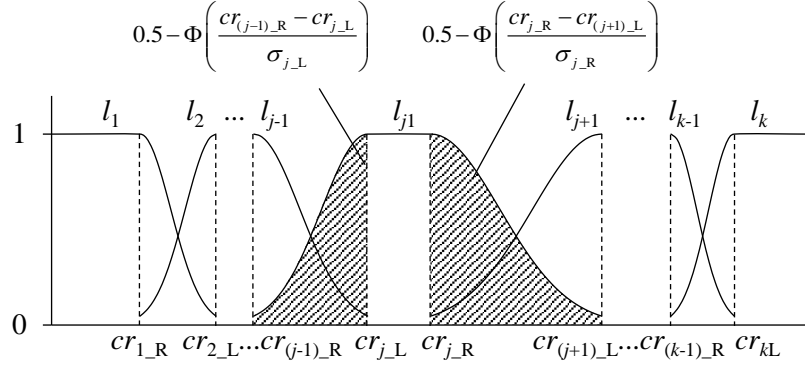


Figure 4.1 Illustration of proposed fuzzy data discretization model

- (3) Given no further pre-knowledge, there is an intuitively unbiased suggestion to set σ_{j_L} and σ_{j_R} as such that the cumulative $\mu_{l_j}(x)$ is half the size of transition ranges, or takes half of full memberships of l_j in the transitions.

Under this condition, $0.5 - \Phi\left(\frac{(cr_{(j-1)_R} - cr_{j_L})}{\sigma_{j_L}}\right) = 0.5 \times (cr_{j_L} - cr_{(j-1)_R})$, where Φ is the cumulative standard normal distribution function. Resultantly, $\sigma_{j_L} = (cr_{j_L} - cr_{(j-1)_R}) / 2.473$ and similarly $\sigma_{j_R} = (cr_{(j+1)_L} - cr_{j_R}) / 2.473$. Following Bordogna and Pasi (1993), who left out low-value tails of Gaussian-curve membership functions, it is set that $\mu_{l_j}(x) = 0$ for $x < cr_{(j-1)_R}$ and $x > cr_{(j+1)_L}$.

4.1.2 Crisp-fuzzy SARM for mining authentic and accurate rules

Effective control over spurious rules by statistically sound tests has been experimentally proven using predesigned associations between already categorized numerical attributes by Webb (2007) and Chapter 3 of this thesis. Section 4.2.1 in this chapter synthetizes more realistic data which contains associations of variant strength directly depending on raw numerical attribute values in a gradual manner, and some data disturbances. The statistically sound evaluation turns out to maintain very low FWER.

Meanwhile, artificial crisp representations of gradual or vague concepts are expected to distort, mostly exaggerate, RIM values. As exemplified in Table 4.1, supports of individual items like $supp(A)$ may not be exaggerated. More undesirably, positive

associations between data items are overstated via t-norm operations. Thus $supp(A \rightarrow B)$ is more exaggerated than $supp(A)$, and finally $imp(A \rightarrow B)$ and $lev(A \rightarrow B)$ become exaggerated. This holds for both product and minimum t-norms, and worsens when rules contain more items and thus RIM evaluations include more t-norm operations.

Table 4.1 RIM value exaggerations due to crisp data discretization in a miniature database of four records. Numerical attributes a and b are discretized into ordinal concepts respectively including l_A and l_B . Item A is $l_A = a$ and B is $l_B = b$

Record#	Fuzzy				Crisp		
	$\mu_{l_A}(a)$	$\mu_{l_B}(b)$	$\mu_{l_A}(a) \otimes_{\text{prod}} \mu_{l_B}(b)$	$\mu_{l_A}(a) \otimes_{\text{min}} \mu_{l_B}(b)$	$\mu_{l_A}(a)$	$\mu_{l_B}(b)$	$\mu_{l_A}(a) \otimes \mu_{l_B}(b)$
1	1	1	1	1	1	1	1
2	0.6	0.8	0.48	0.6	1	1	1
3	0.4	0.2	0.08	0.2	0	0	0
4	0	0	0	0	0	0	0
$supp(A) = \sum \mu_{l_A}(a)$			2	2	2		
$supp(B) = \sum \mu_{l_B}(b)$			2	2	2		
$supp(A \rightarrow B) = \sum \mu_{l_A}(a) \otimes \mu_{l_B}(b)$			1.56	1.8	2		
$imp(A \rightarrow B)$			0.28	0.4	0.5		
$= conf(A \rightarrow B) - conf(\emptyset \rightarrow B)$							
$= supp(A \rightarrow B) / supp(A) - supp(B) / 4$							
$lev(A \rightarrow B)$			0.56	0.8	1		
$= supp(A \rightarrow B) - supp(A) supp(B) / 4$							

Compared with support and confidence, RIMs for evaluating how different the associations between items in rules are from independence among the items, such as improvement and leverage, appear to be exaggerated more severely. These RIMs are ‘margins’ of itemset supports, thus small overestimations in supports can be much amplified in them. In Table 4.1, the crisp rule has 28% exaggeration on $supp(A \rightarrow B)$ but 79% on $imp(A \rightarrow B)$ and $lev(A \rightarrow B)$ with respect to fuzzy product t-norm results. For experiments in Sections 4.2–4.3, improvement and leverage are typically exaggerated by over 50%. Unfortunately, support (and sometimes confidence) mainly

serve as mechanisms for controlling the number of rules considered, while other RIMs that are more severely exaggerated are usually of higher interest for users.

While fuzzy rules have more accurate RIM values, they are also more moderate and less significant in the statistically sound evaluation. This significantly reduces the true rules accepted as SARM results compared with crisp SARM, typically by at least 50%, as experimentally shown in Sections 4.2–4.3. To combine the abundance of true rules using crisp supports and higher accuracy of fuzzy RIM values, the crisp-fuzzy SARM method is proposed:

- Firstly, perform statistically sound tests on rules, using crisp supports of itemsets involved;
- Then evaluate RIM values of the significant rules accepted by the tests using fuzzy supports.

As will be elaborated in Section 4.2.2, statistically sound tests are still indispensable to keep the FWER under control for fuzzy SARM. Thus the crisp-fuzzy SARM indeed achieves the greatest number of true rules without sacrificing strong control over the risk of spurious rules. The statistical test stage shall use a crisp discretization model matching the fuzzy one for RIM evaluation:

$$\mu_{l_j}(x) = \begin{cases} 1, & (c_{(j-1)\text{-R}} + c_{j\text{-L}})/2 \leq x \leq (c_{j\text{-R}} + c_{(j+1)\text{-L}})/2 \\ 0, & \text{otherwise} \end{cases} \quad (4.2)$$

Equation (4.2) ensures that each x value has the same concept of maximum membership degrees as in the fuzzy model.

4.2 Experiment with synthetic data

4.2.1 Methods

The experiment data included three spatial point sets: objective, resource1 and

resource2. Each point had two-dimensional coordinates (x, y). Each objective linked to a record containing nine attributes: *near1*, *near2*, *other1*, *other2*, *noise1*–*noise 4*, and *outcome*.

near1 and *near2* were respectively fuzzy attributes representing nearness, or accessibility of objective to resource1 and resource2. They each had two values 1 (near) and 0 (far). Following the proposed data discretization model in Section 4.1.1, $\mu_1(\text{near1})$ and $\mu_0(\text{near1})$ were computed from Euclidean distance *near1* (same as attribute name, see 3.1) between each objective and the nearest resource1:

$$\mu_1(\text{near1}) = \begin{cases} 1, & \text{near1} \leq d_{0.01} \\ \exp\left[-(\text{near1} - d_{0.01})^2 / \left(2 \times \left(\frac{d_{0.99} - d_{0.01}}{2.473}\right)^2\right)\right], & d_{0.01} < \text{near1} < d_{0.99} \\ 0, & \text{near1} \geq d_{0.99} \end{cases},$$

$$\mu_0(\text{near1}) = \begin{cases} 0, & \text{near1} \leq d_{0.01} \\ \exp\left[-(\text{near1} - d_{0.99})^2 / \left(2 \times \left(\frac{d_{0.99} - d_{0.01}}{2.473}\right)^2\right)\right], & d_{0.01} < \text{near1} < d_{0.99} \\ 1, & \text{near1} \geq d_{0.99} \end{cases} \quad (4.3)$$

where $d_{0.01}$ and $d_{0.99}$ are the first and 99th percentiles of *near1* values in data. Without pre-knowledge for more proper membership functions, partial membership degrees were assigned to nearly full range of *near1* to better differentiate accessibilities of different objectives to resources. Cutting data at $d_{0.01}$ and $d_{0.99}$ was for lessening sensitivities of c_{1_R} and c_{2_L} to extreme distances. This mimicked real-world SARM practice in handling extreme data. All above equally applied for *near2*.

other1, *other2* and *noise1* to *noise4* were categorical attributes; *other2* had 10 possible values (0–9) and the others had four (0–3). Values of these attributes were generated randomly, independently and equiprobably for every possible value.

outcome was also fuzzy, with values 0 (bad) and 1 (good). It was the only attribute with dependences on other attributes. The dependences were listed in Table 4.2: higher $\mu_1(\text{near1})$, or accessibility to resource1, unconditionally improved *outcome*, while higher *other2* values improved *outcome* only when *other1* = 0 or 2, and higher $\mu_1(\text{near2})$ did so only when *other1* = 0 or 1. Such unconditional and conditional associations are both common in practical SARM. *Noise1–noise4* had no association with *outcome* and were for examining the proposed techniques for tolerance to irrelevant data. As there were no raw numerical values for computing $\mu_0(\text{outcome})$, it was set that $\mu_0(\text{outcome}) = 1 - \mu_1(\text{outcome})$. Factors *fac1* and *fac2* were to adjust expectations of $(\mu_1(\text{near1}) + \mu_1(\text{near2})) / \text{fac}_1$ and $\mu_1(\text{near1}) / \text{fac}_2$ to 0.5, so as to cancel out data variations due to capping negative $\mu_1(\text{outcome})$ to 0 and $\mu_1(\text{outcome})$ above 1 to 1. These two adjustments were unlikely to affect the comparative evaluation of experiment results, since they proportionally changed RIM values for all experiment groups.

Table 4.2 Dependence of $\mu_1(\text{outcome})$ on other attributes

<i>other1</i>	<i>other2</i>	Expectation of $\mu_1(\text{outcome})$ (Standard deviation = 0.15)
0	0, 1, 2, 3, 4	$(\mu_1(\text{near1}) + \mu_1(\text{near2})) / \text{fac}_1^a - 0.35, -0.3, -0.25, -0.2, -0.15$
	5, 6, 7, 8, 9	$(\mu_1(\text{near1}) + \mu_1(\text{near2})) / \text{fac}_1 + 0.15, +0.2, +0.25, +0.3, +0.35$
1	Any	$(\mu_1(\text{near1}) + \mu_1(\text{near2})) / \text{fac}_1$
2	0, 1, 2, 3, 4	$\mu_1(\text{near1}) / \text{fac}_2 - 0.35, -0.3, -0.25, -0.2, -0.15$
	5, 6, 7, 8, 9	$\mu_1(\text{near1}) / \text{fac}_2 + 0.15, +0.2, +0.25, +0.3, +0.35$
3	Any	$\mu_1(\text{near1}) / \text{fac}_2$

^a *fac1*: mean value of all $(\text{near1} + \text{near2})$, *fac2*: mean value of all *near1*

The predesigned data associations generated 118 productive rules with *outcome* values as the consequents:

- $near1 = 1 \wedge \text{zero or one of } other1 = 2 \text{ or } 3 \rightarrow outcome = 1$ (3 rules);
- $near1 = 0 \wedge \text{zero or one of } other1 = 2 \text{ or } 3 \rightarrow outcome = 0$ (3 rules);
- Zero or one of $near1 = 1 \wedge near2 = 1 \wedge \text{zero or one of } other1 = 0 \text{ or } 1 \rightarrow outcome = 1$ (6 rules);
- Zero or one of $near1 = 0 \wedge near2 = 0 \wedge \text{zero or one of } other1 = 0 \text{ or } 1 \rightarrow outcome = 0$ (6 rules);
- Zero or one of $near1 = 1 \wedge other1 = 2 \wedge other2 = 5-9 \rightarrow outcome = 1$ (10 rules);
- Zero or one of $near1 = 0 \wedge other1 = 2 \wedge other2 = 0-4 \rightarrow outcome = 0$ (10 rules);
- Zero or more of $near1 = 1, near2 = 1$ and $other1 = 0 \wedge other2 = 5-9 \rightarrow outcome = 1$ (40 rules);
- Zero to three of $near1 = 0, near2 = 0$ and $other1 = 0 \wedge other2 = 0-4 \rightarrow outcome = 1$ (40 rules).

To evaluate the robustness of the proposed techniques, experiment groups were constructed by various alternations, as summarized in Table 4.3. The *extraneous factors* simulated numerous affecting factors to real-world imperfect distance data, including but not limited to measurement errors. For instance, actual distance between two places for all citizens could be longer than recorded shortest path, if barrier-free paths between the places are long detours. As stated in Section 4.1.1, previous work has suggested that Gaussian-curve relations between concept memberships and raw numerical data are more usual. Still, groups using ‘true’ and ‘perceived’ memberships with linear transitions were generated, for a more comprehensive comparison of the proposed data discretization model with triangular/trapezoidal ones.

Table 4.3 Various experiment settings in synthetic data experiment

Item	Variations	Remarks
1. Data size (No. of records)	5000, 20,000, 80,000	
2. Spatial patterns of objectives and resources	Clustered, random, dispersed	Point sets must pass nearest neighbour index test (Mitchell 2005) for designated spatial patterns with threshold $p = 0.05$
3. Extraneous factors	$\sigma = 0, 10\%, 20\%$	Added to data by multiplying raw <i>near1</i> or <i>near2</i> a random variable following normal distribution $N(1, \sigma^2)$
4. ‘True’ membership functions for <i>near1/near2</i> w.r.t. their raw values	Linear, Gaussian-curve	Linear: membership functions with linear transitions and endpoints coincided with Equation (4.3), e.g. $\mu_1(near1) = \begin{cases} 1, & near1 \leq d_{.01} \\ \frac{near1 - d_{.99}}{d_{.01} - d_{.99}}, & d_{.01} < near1 < d_{.99} \\ 0, & near1 \geq d_{.99} \end{cases}$
5. Memberships used in statistical tests	Crisp, linear, Gaussian-curve	Crisp functions matching Equation (4.3) were defined according to Equation (4.2). Both linear and Gaussian-curve groups used original $\mu_0(outcome)$ and $\mu_1(outcome)$ computed from Table 4.2
8. ‘Perceived’ memberships for evaluating RIMs	Crisp, linear, Gaussian-curve	For simulating how people perceive unknown ‘true’ relations

All these alternations multiplied into 486 unique experiment groups, or treatments as defined in Chapter 3. Each treatment was applied for 10 runs to 10 independently generated datasets to produce stable average results. In each run, association rules were first extracted from data using the KORD algorithm, and tested against both unadjusted (without Bonferroni correction) and statistically sound chi-square tests for productive rules at $\alpha = 0.05$. The statistically sound test adopted the direct adjustment approach, with the number of potential rules $s = 44,796$ and significance level $\kappa = 0.05 / s = 1.12 \times 10^{-6}$ computed according to Webb (2007). Pilot runs revealed that direct adjustment and holdout approaches discovered similar number of true rules

from experimented datasets. Then the direct adjustment was favourable, since the holdout results might be affected by random selection of holdout data from the dataset.

4.2.2 Results: true and spurious rules

Figure 4.2 illustrates the numbers of true rules, numbers of spurious rules, and FWER against variations in data sizes, statistical soundness of tests and perceived memberships for fuzzy attributes. True and spurious rules were those accepted by statistical tests that were within and beyond predefined productive rules, respectively. True rules only counted for 88 out of 118 predefined rules containing *near1* or *near2*, as the others were irrelevant to fuzzy nearness. Each plotted point was an aggregation for all spatial patterns of objectives and resources, true $\mu_1(\text{outcome})$ and extraneous factors, as the plots had quite similar patterns when these settings varied.

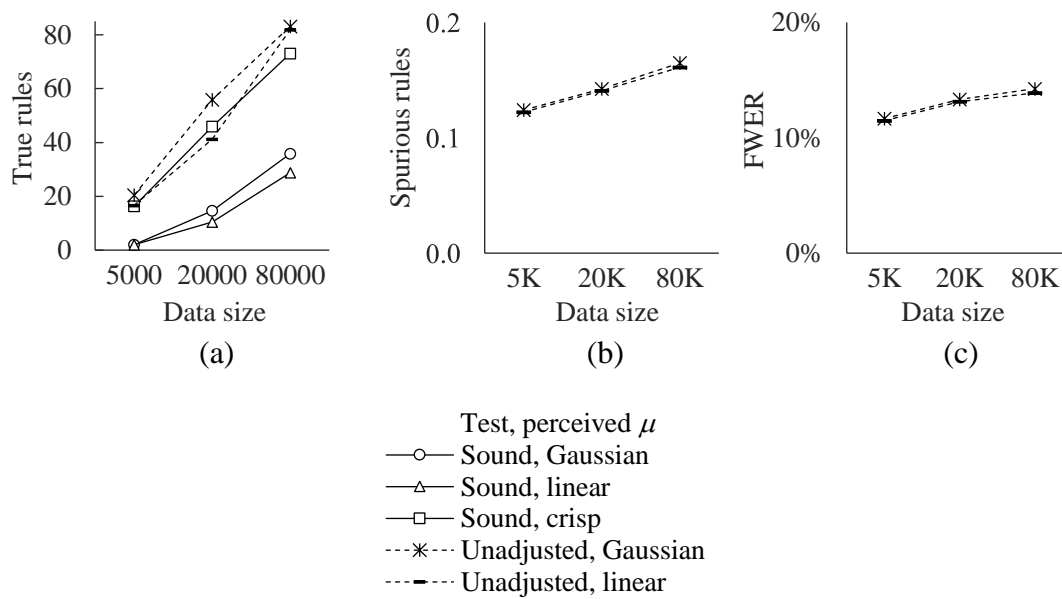


Figure 4.2 Synthetic data experiment results on abundance of true rules and avoidance of spurious rules

Statistically sound tests are proven indispensable for both crisp and fuzzy rules in order to strictly control spurious rules (Figure 4.2b, c). In crisp rule treatments, unadjusted tests resulted in dozens of spurious rules per run and 100% FWER, which were too large to be plotted. Such extreme risk of spurious rules in absence of

statistically sound tests agrees with Webb (2007) and Section 3.2.2. While this risk largely reduced in fuzzy rule treatments (both Gaussian-curve and linear), after unadjusted tests the FWER was still 10–15%, far above 5% as user specified by setting $\alpha = 0.05$. Moreover, though not shown in the figure, more than half of the runs produced rules with p values between 0.05 and 1. Thus, unadjusted tests with $\alpha = 0.1$ will produce over 50% FWER, while 10% was what user expected. This is likely to be unacceptable in practice, as users never know the maximum risk of spurious rules under the significance level they set. Meanwhile, statistically sound tests did not accept any spurious rules, and thus their plots are absent from Figure 4.2b and c. Webb (2007) and Section 3.2.2 reported higher 0.01–0.1% spurious rules and 0.1–1% FWER for statistically sound tests on crisp rules when $\alpha = 0.05$, which are still only several tenths of those for unadjusted tests with fuzzy memberships, and well fulfil user specified 5% maximum FWER. In this study, this approach produced even fewer spurious rules, actually zero rules in 1620 runs, probably due to the limitation of rule consequents to *outcome* which was most related to other attributes.

On the condition that statistically sound tests are necessary, using crisp memberships exhibited great superiority in discovering more true rules. Averaging results of all data sizes, statistically sound tests using crisp memberships discovered at least 2–4 times as many true rules as using fuzzy ones, and comparable number of true rules as unadjusted tests using fuzzy memberships (Figure 4.2a).

Overall, the results support the suggestion to conduct statistically sound tests with crisp memberships in SARM, as this is the best for finding abundant true rules while maintaining strict control over spurious rules.

4.2.3 Results: RIM accuracy

Figure 4.3 shows the accuracy of RIM values against the variation of extraneous

factors. Variations in other experiment settings made little difference in the changing trend of RIM accuracy. The plotted values are percentage errors of RIM values to their true values computed using ‘true’ memberships of fuzzy attributes (see Section 4.2.1) and 0% extraneous factors.

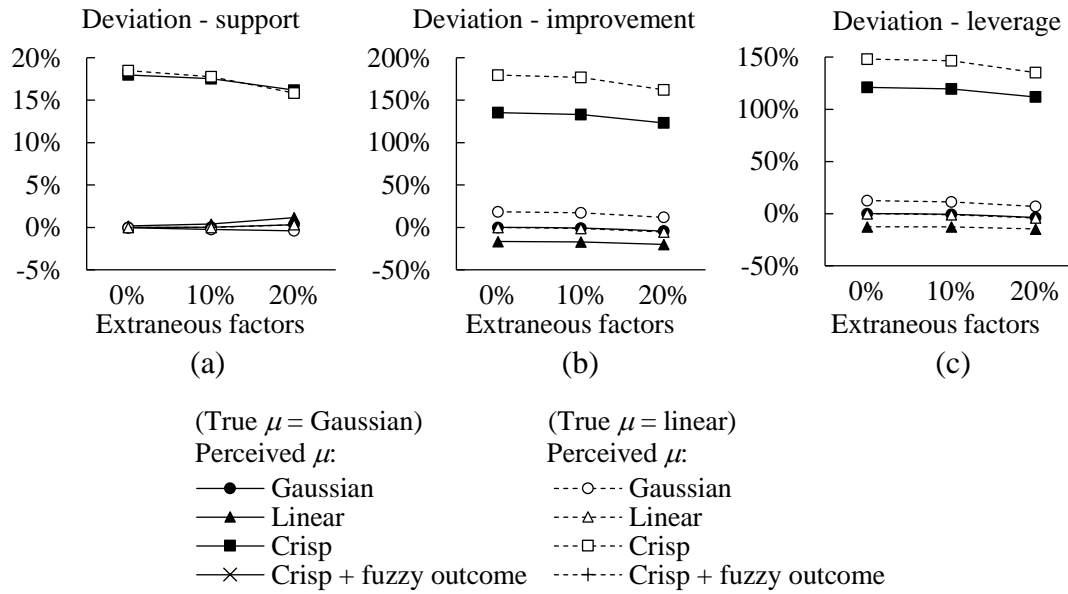


Figure 4.3 Synthetic data experiment results on RIM accuracy

Crisp rules committed more than 100% positive errors on improvements and leverages, while only 15–20% error on supports (Figure 4.3). This confirms the inference in 3.2 about large exaggerations on RIMs by the crisp membership, and that the exaggerations are worse on RIMs other than support which are usually more useful for decision support. A further investigation revealed that crisp *outcome* in rule consequents contributed about 2/3 of the RIM exaggerations. This is because itemset supports containing rule consequents dominate the computation of improvement and leverage (equations in Section 2.1).

As extraneous factors grew, RIM values generally decreased, except for minimal increases in fuzzy supports that had little effect on SARM results (Figure 4.3). This

conforms to the expectation that extraneous factors blur data associations and weaken rules.

Gaussian-curve and linear perceived memberships caused positive and negative RIM errors, respectively, when true memberships were the opposite. Yet such errors were only 10–20% of those in counterpart crisp treatments. Thus in RIM evaluation, by replacing crisp membership for gradual or vague concepts with fuzzy ones, there can be a major improvement in RIM accuracy, even if true fuzzy memberships might not be accurately defined due to inadequacy of expert knowledge.

When the true membership is unknown, the proposed Gaussian-curve model is still recommended for RIM evaluation, after identifying true rules using crisp membership and statistically sound tests. First, as suggested in Section 4.1.1, Gaussian-curve memberships are widely regarded better illustrating linguistic concepts. Second, Gaussian-curve memberships produced larger RIM values than linear ones (Figure 4.3), since linear membership degrees change more constantly (with constant slopes) across transitions between concepts. Then the reduction in RIM values, due to practically inevitable extraneous factors, may partially offset positive errors in RIM values caused by the Gaussian-curve model while enlarging negative errors committed by the linear model.

4.3 Experiment with real-world data

4.3.1 Data and methods

This case study investigated how the proposed techniques help improved practical SARM results and decision support, through an association analysis between wildfire risks and fire-inducing environmental factors. The raw data was the Covertypes dataset (Blackard 1998) from the UCI Machine Learning Repository. The data covered around 52,000 hectares of land in the Colorado Front Range, US, a longstanding

wildfire-prone region (Calkin *et al.* 2014, Sherriff *et al.* 2014). The data was based on a 30-m resolution raster of 581,012 cells. Each cell linked to a record of mixed numerical and categorical attributes, including distance to nearest past fire ignition point which indicated past wildfire risks and was highly weighted for predicting future risks (The Virginia Department of Forestry 2003, Lein and Stump 2009), and various environmental conditions serving as wildfire risk factors.

Numerical attributes in raw data were discretized into concepts, the result of which and the categorical attributes are listed in Table 4.4. Three treatments with different data discretization models were applied:

- AllFuzzy: all numerical attributes adopted fuzzy discretization models proposed in Section 4.1.1;
- FuzzyDist: distance attributes in Table 4.4 adopted the fuzzy discretization models. Other numerical attributes adopted matching crisp models specified following Equation (4.2);
- AllCrisp: all numerical attributes adopted matching crisp models.

The FuzzyDist treatment was for examining if the advantage of proposed techniques revealed with synthetic data were generalizable to non-distance and even non-spatial concepts.

Table 4.4 Data attributes in real data experiment

Type	Name	Description	No. of categories/ concepts
In rule antecedents: wildfire risk factors			
Numerical, distance	1. horz_dist_to_water	Horizontal distance to nearest water	2 (near, far)
	2. above_water	Absolute distance	1 (far)
	3. below_water	above/below nearest water; empty if cell is below/above nearest water, respectively	1 (far)
	4. horz_dist_to_road	Horizontal distance to nearest road	2 (near, far)
Numerical, non- distance	5. elevation	-	2 (low, high)
	6. aspect	-	8 (N to NW clockwise)
	7. slope	-	2 (low, high)
	8. hillshade_9am	Summer hillshade index at 9am/noon/3pm	2 (low, high)
	9. hillshade_12nn		
	10. hillshade_3pm		
Categorical	11. soil	Soil type	40
	12. cover	Forest cover type	7
In rule consequents: past wildfire risk			
Numerical, distance	13 horz_dist_to_fire	Horizontal distance to nearest past wildfire ignition point	1 (near)

For AllFuzzy, numerical attributes other than aspect adopted Gaussian-curve-based fuzzy discretization following Equation (4.1). ‘Low’ and ‘high’ for non-distance attributes were defined like ‘near’ and ‘far’. Most attributes had 1–5% cells of the same or indistinguishable lowest values. For example, 4% of cells had 0m distance to, or were dominated by water, while still partially forested and might be burnt. Also, 2% of cells had 2% slopes or below. Considering that the slopes were computed from 30-m DEM cells, detailed terrains within these cells were hardly discernible. To reasonably assign full membership of ‘near/low’ to all these cells, the lowest and highest 5% values had $\mu = 1$ or 0 for corresponding concepts, instead of 1% as in

Section 4.2.1. Aspect in degrees was discretized into eight common directions. $\mu_{dir_i}(aspect) = 1$ only when $aspect = 45i$, where dir_i , $i = 0$ to 7 was for direction N to NW clockwise. Transitions of dir_i had suggested standard deviations in 3.1 and 45° widths in both clockwise and counterclockwise directions.

In each treatment, rules like “fire risk factor value(s) \rightarrow horz_dist_to_fire = near ” with up to four items in antecedents were extracted. Extracted rules were passed to statistically sound test for productivity using the direct adjustment approach with $\kappa = 3.2024 \times 10^{-7}$ computed according to Webb (2007).

4.3.2 Results

Experiment results reconfirmed advantages of the proposed crisp-fuzzy SARM in obtaining abundant true rules and accurate RIM values (Table 4.5). AllCrisp resulted in more than twice as many significant rules as AllFuzzy. Synthetic data experiments in 4.2 and previous studies have revealed that statistically sound tests yield extremely few spurious rules. Thus, AllCrisp more than doubled the number of true rules. However, similar to Section 4.2.3, AllCrisp and AllFuzzy had large discrepancies in RIM values, especially improvements and leverages, which should be largely errors of AllCrisp due to crisp discretization for gradual concepts. The RIM errors and analyses hereafter are based on 1225 significant rules ($p < \kappa$) in AllCrisp which also had $p < 0.05$ in AllFuzzy. These rules generally had much larger leverages and thus should be more important than the remaining rules. FuzzyDist also caused RIM exaggerations, though only by 10–25% of those in AllCrisp. This suggests that the advantage in RIM accuracy of fuzzy data discretization also applies to gradual non-distance concepts, and it is recommended to evaluate RIMs using fuzzy memberships on all applicable attributes in SARM.

Table 4.5 Real data experiment result on number of significant rules and RIM accuracy

		AllCrisp	FuzzyDist	AllFuzzy
No. of significant rules		1952	1004	803
Among 1225 rules with $p < \kappa$ in AllCrisp and $p < 0.05$ in AllFuzzy				
Exaggeration w.r.t. AllFuzzy	Support	11%	1%	-
	Improvement	67%	18%	-
	Leverage	37%	7%	-

Table 4.6a lists the effects of single fire risk factors suggested by significant rules with single-item antecedents, like ‘ horz_dist_to_water = near \rightarrow horz_dist_to_fire = near ’, and compared them with those in empirical fire risk models. The results generally cohere with empirical models, except for the aspect effect. South slopes in northern hemisphere are often the riskiest, as they receive more direct sunlight and consequently have higher temperature and drier fuels. Yet direct sunlight should be less influential in the study area with relatively homogeneous dryness and cool summer afternoon of no more than 20°C (Colorado Climate Center 2016). Instead, north slopes can be riskier than south slopes due to considerably higher vegetation density on the former (Kaufmann *et al.* 2006, Chambers *et al.* 2016), as fuel mass is usually the most highly weighed fire-inducing factor (Noble *et al.* 1980, The Virginia Department of Forestry 2003, US Forest Service 2010). Riskier west slopes than east ones are likely due to the prevalent dry west wind (Colorado Climate Center 2016). Low morning/high afternoon hillshade seems also represent riskier west slopes, as hillshade and aspect in raw data was computed from the same DEM.

Table 4.6 Real data experiment result on single wildfire risk factors

(a) Fire-inducing effects and comparison with empirical fire risk models

Factor	Values suggesting high wildfire risk	
	In empirical models	In this study
1. horz_dist_to_water	- ^a	Horizontal proximity to water
2. above_water, below_water	-	Vertical farness to water, both above and below
3. horz_dist_to_road	Proximity to roads [1–4] ^b	Proximity to roads
4. elevation	Low elevation, for areas with elevation > 1600m - case of study area [1, 4, 5]	Low elevation
5. aspect	Depend on study area; south slopes in northern hemisphere are usually riskiest [1–7], but high risk on north slopes is also reported [7]	Northwest, west and southwest slopes
6. slope	Steep slopes [2, 3, 5, 6, 8]	Steep slopes
7. hillshade	-	Low value at 9am and high value at 3pm
8. forest type	Conifer/evergreen forest [3, 8]	Not meaningful, as study area is dominated by evergreen forest

^a: Factors are not common empirical model inputs.

^b: [1] Anchor Point Group (2010), [2] Gerdzheva (2014), [3] The Virginia Department of Forestry (2003), [4] Thompson *et al.* (2000), [5] Ghobadi *et al.* (2012) [6] US Forest Service (2010) [7] Yang *et al.* (2013) [8] Bradshaw *et al.* (1984).

(b) Sensitivity of RIM values to change in class boundaries (transition midpoints for AllFuzzy) by 4% of 5–95th percentile attribute value ranges

	Class boundary of attribute in antecedent	AllCrisp	AllFuzzy
<i>imp</i> (elevation = low → horz_dist_to_fire = near)	Original	0.0337	0.0383
	Increase by 4%	0.0288	0.0303
	Decrease by 4%	0.0536	0.0468
<i>imp</i> (below_water = far → horz_dist_to_fire = near)	Original	0.0574	0.0262
	Increase by 4%	0.0499	0.0239
	Decrease by 4%	0.0557	0.0279

Albeit seemingly unusual, the SARM result suggests increased fire risks in places near waters. The overall cool and dry locality could weaken the moistening, cooling, and thus fire-mitigating effect of waters. Moreover, proximity to streams is the most powerful predictor of forest density in the study area due to water stress (Krasnow *et al.* 2009), and can be thereby linked to higher fire risk.

In AllCrisp and AllFuzzy, elevation and below_water exhibited the most inconsistent ranks of rule improvements, or relative importance to fire risks. As shown in Table 4.6b, rule improvements in AllCrisp are more sensitive to slight variations in class boundaries. If the class boundaries lowered by 4% for elevation and raise by 4% for below_water, elevation would be more important than below_water, which contradicted to the current result. Meanwhile, AllFuzzy results consistently suggests higher importance of elevation. This suggests that the proposed crisp-fuzzy SARM are more robust against uncertainties in expert classifications for data discretization, and often more reliable than crisp SARM.

The following paragraphs will shift from single fire risk factors to interactions between multiple factors, the study on which is the distinctive advantage of SARM. Empirical models typically evaluate fire risks by equations of risk factors, with constantly valued coefficient for each factor. Actually, fire-inducing effects of many factors vary according to presences or values of other factors. Conditionally variable coefficients have been employed to improve model accuracy (Nobel *et al.* 1980, The Virginia Department of Forestry 2003), but only on occasion and not systematically. This is probability because mainstream fire risk modelling techniques, such as regression and laboratory test, are unsuitable for studying interactions between multiple risk factors. Meanwhile, SARM has strong ability to reveal such interactions from rules with multiple risk factors in the antecedents.

Define *semi-improvement* of a rule $X \rightarrow Y$:

$$semi-imp(X \rightarrow Y) \text{ w.r.t } x \in X, |X| > 1 = conf(X \rightarrow Y) - conf(X \setminus \{x\} \rightarrow Y). \quad (4.4)$$

$semi-imp(X \rightarrow Y)$ w.r.t x and $imp(x \rightarrow Y)$ are respectively conditional impacts of x in presence and absence of other items in X . For fire risk modelling, all $x \in X$ are risk factors. If $semi-imp(X \rightarrow Y)$ w.r.t $x/imp(x \rightarrow Y)$ significantly deviates from 1, the effect of x may highly depend on other factors and worth the consideration of conditional coefficients.

Table 4.7 lists the numbers of rule pairs like $X \rightarrow Y$ and $X \setminus \{x\} \rightarrow Y$ where $semi-imp(X \rightarrow Y)$ w.r.t $x/imp(x \rightarrow Y)$ exceeds different thresholds. ‘Omissions’ and ‘commissions’ refer to omitted and extra eligible rule pairs in AllCrisp with respect to AllFuzzy, since they should be mainly attributed to RIM errors of AllCrisp, as explained earlier. AllCrisp did not have many omissions, but suffered from severe commissions, sometimes over 50%. This reconfirms the inference in Section 4.1.2 that crisp data discretization exaggerates RIMs more severely for rules containing more items, here $X \rightarrow Y$ compared with $x \rightarrow Y$.

Table 4.7 Numbers of rule pairs like $X \rightarrow Y$ and $X \setminus \{x\} \rightarrow Y$, where $semi-imp(X \rightarrow Y)$ w.r.t $x/imp(x \rightarrow Y)$ exceeds specified thresholds. Rule pairs with $imp(x \rightarrow Y) < 0.01$ are excluded, as resultant ratios were sensitive to small divisors

	Criteria for $semi-imp(X \rightarrow Y)$ w.r.t $x/imp(x \rightarrow Y)$		
	>1.2	>1.5	>1.8
AllFuzzy	546	393	281
AllCrisp	621	504	421
Omission	41	22	24
Commission	116	133	164

The numerous commissions in AllCrisp appear to misrepresent the needs for conditional coefficients. This makes crisp RIMs unusable, or only fuzzy RIMs accurate enough, for investigating conditional fire risk factors. The point is further

exemplified as below. Let *item1* be ‘ horz_dist_to_water = near ’, consider two resultant rules:

$$\begin{aligned}
 & \text{rule1: } item1 \wedge \text{above_water} = \text{far} \rightarrow \text{horz_dist_to_fire} = \text{near} \\
 & \text{semi-imp}(\text{rule1}) \text{ w.r.t. } item1 / \text{imp}(item1 \rightarrow \text{horz_dist_to_fire} = \text{near}) \\
 & \quad = 2.88 \text{ -AllFuzzy, } 5.46 \text{ -AllCrisp} \\
 & \text{rule2: } item1 \wedge \text{below_water} = \text{far} \rightarrow \text{horz_dist_to_fire} = \text{near} \\
 & \text{semi-imp}(\text{rule2}) \text{ w.r.t. } item1 / \text{imp}(item1 \rightarrow \text{horz_dist_to_fire} = \text{near}) \\
 & \quad = 0.85 \text{ -AllFuzzy, } 1.74 \text{ -AllCrisp}
 \end{aligned}$$

In AllFuzzy, the fire-inducing effect of *item1* was boosted by 1.88 times in *rule1*, yet suppressed by 15% in *rule2*. No risk factors in Table 4.6a were likely reasons for such inconsistent behaviours of *item1*. Areas above and below the nearest waters had quite similar values on all factors other than slope. Also, if slope had influenced the effect of *item1*, the influence should have been consistently boosting or suppressing in *rule1* and *rule2*, since slope and *item1* values exhibited significant positive correlation in all areas.

An alternative explanation may be that waters can weaken fire-inducing hot and dry airflows which typically go upslope (The Virginia Department of Forestry 2003). This reduces the risk immediately above waters, yet has little effect on upwind below-water areas (Wang and Fu 1991, Saaroni and Ziv 2003). In *rule1*, the antecedent implies steeper slopes and associated higher wildfire risks, thus the water barrier effect looks more prominent where above_water = near. This is equivalently represented as boosted fire-inducing effect of *item1* where above_water = far. In *rule2*, below_water = far may suppress the effect of *item1*, as it implies nearness to waters at even lower elevations and associated fire risk reduction.

If this explanation can be validated by additional weather observation data, it will be recommended to assign conditional coefficients to *item1*, and also to strengthen the conservation of montane waters for wildfire mitigation. Such discovery is impossible

using AllCrisp results, where *item1* enhanced fire-inducing effect in both *rule1* and *rule2*.

4.4 Summary

This chapter presents two techniques for improving the reliability of fuzzy SARM results. First, it presents a Gaussian-curve-based fuzzy data discretization model for SARM. Compared with existing models, this model summarizes spatial semantics of Gaussian curves and their advantages for SARM, and is more complete by covering multi-concept relations. Second, it originates a crisp-fuzzy SARM method: first to conduct statistically sound tests on crisp SARs, and then to evaluate RIMs of accepted rules using matching fuzzy data discretization schemes. This method can relieve the conservativeness of statistically sound tests and reduce their rejection of true rules particularly for fuzzy SARM, thereby increasing authentic rules; and avoid large overestimations in RIM values caused by crisp data discretization for gradual or vague concepts, hence improving the RIM accuracy.

Experiments show that the proposed techniques can significantly increase authentic resultant rules, typically by at least 100% compared with conventional fuzzy SARM. The techniques also largely avoid large positive errors in RIM values incurring in ordinary SARM, which is usually more than 50% for representative RIMs. The FWER was below 1%. A case study on wildfire risk factors demonstrates the practical value of the proposed techniques, especially the higher robustness against data discretization scheme changes and discoveries of sensible rules due to more accurate RIM values.

Chapter 5 Genetic algorithm for mining significant crisp-fuzzy SARs

This chapter develops the GA-based method for mining significant crisp-fuzzy SARs. This method aims at achieving genetically optimized SARM which produces more abundant rules and RIM values of higher fitness, compared with existing methods for mining significant SARs of minimal risk to be spurious. Section 5.1.1 presents the chromosome encoding for the new GA. Section 5.1.2 illustrates the core for this method: two original statistical testing approaches, the experimentwise and generationwise adjustment approach, for strictly limiting spurious rules; as well as the integration of the GA with the Gaussian-curve-based model and crisp-fuzzy SARM proposed in Chapter 4, for further improving numbers of true rules and RIM goodness. Section 5.1.3 overviews the algorithm procedures, and details the design for genetic operators and specific operators for the proposed GA.

Section 5.2 presents two experiments for the proposed GA: Hotel experiment on smaller-sized data for investigating Hong Kong hotel accessibilities to tourism resources as determinants of hotel room prices, and Fire experiment revisiting larger-sized Colorado wildfire risk factor data used in Chapter 4. Sections 5.2.1 and 5.2.2 illustrates experiment data collection and preprocessing, and GA computational specifications. Sections 5.2.3 and 5.2.4 evaluates the capability of the GA in controlling spurious rules and discovering true rules, respectively, as compared with conventional SARM. Section 5.2.5 elaborates new insights into hotel room price determinant studies contributed by the more reliable Hotel experiment results using the proposed GA. Section 5.3 summarizes this chapter.

5.1 Methods

5.1.1 Chromosome encoding

Referring to Section 2.5, in GA-based SARM, each chromosome, or individual, can

be used to encode either the entire data discretization scheme or part of the solutions like a main rule (Kaya 2006). This chapter adopts the main rule encoding approach and adapts it for the Gaussian-curve-based fuzzy data discretization model presented in Chapter 4. Recall that a main rule is a collection of rules with the same attributes in the antecedents and the same in the consequents. That is, all rules like $a_1 = l_{a_1 i_1} \wedge \dots \wedge a_q = l_{a_q i_q} \rightarrow b = l_{b j}$ is under the main rule $M: a_1 \wedge \dots \wedge a_q \rightarrow b$, where $a_1 \dots a_q b$ are attributes with corresponding concepts $l_{a_1 i_1} \dots l_{a_q i_q} l_{b j}$. For each attribute a , three groups of variables are encoded to define a main rule:

- k_a : the number of concepts for a . The concepts are $l_{a1} \dots l_{ak_a}$ and $k_a \leq k_{\max}$, k_{\max} is the predefined maximum number of concepts for any attribute;
- cr_{ai_L} , cr_{ai_R} : left and right endpoints of $core(\mu_{l_{ai}})$, $i = 1 \dots k_{\max}$. These variables are specific to the new Gaussian-curve-based data discretization model. If other models are applied, other variables that can fully define $l_{a1} \dots l_{ak_a}$ should be used instead;
- loc_a : the location of items involving a in rules; $loc_a = 1, 2, 0$ if the items are in the rule antecedent, rule consequent, and neither, respectively.

The encoding of a is

$$k_a \ cr_{a1_R} \ cr_{a2_L} \ cr_{a2_R} \ \dots \ cr_{a(k_{\max}-1)_L} \ cr_{a(k_{\max}-1)_R} \ cr_{ak_{\max}_L} \ loc_a, \quad (5.1)$$

$cr_{ak_a_R} \dots cr_{ak_{\max}_L}$ are assigned empty values. Membership functions of the concepts defined by Equation (5.1) are illustrated in Figure 5.1. The entire chromosome for all n attributes in data, with a length of $n(k_{\max} + 2)$, is

$$k_1 \ cr_{11_R} \ cr_{12_L} \ cr_{12_R} \ \dots \ cr_{1k_{\max}_L} \ loc_1 \ \dots \ k_n \ cr_{n1_R} \ cr_{n2_L} \ cr_{n2_R} \ \dots \ cr_{nk_{\max}_L} \ \dots \ loc_n. \quad (5.2)$$

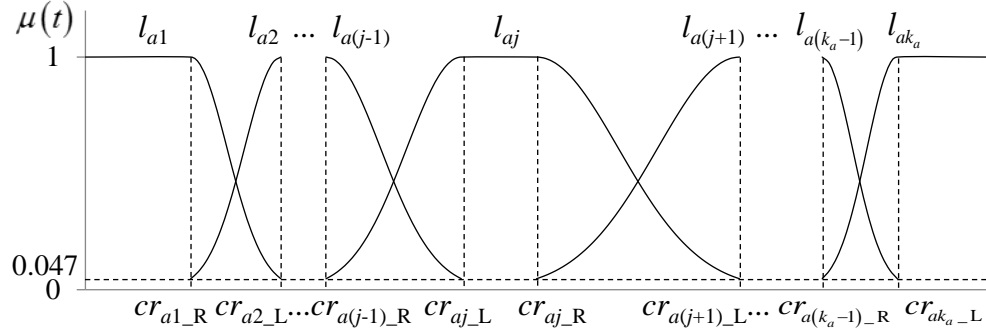


Figure 5.1 Encoding of attribute a using proposed chromosome encoding scheme

For the proposed GA, there are two main benefits for encoding a single main rule instead of the entire data discretization scheme in each chromosome. First, this enables resultant rules of higher flexibilities but still reasonable interpretability. Each main rule presents a group of associations between a unique combination of attributes and their presences in rule antecedents and consequents. In different groups of associations, an attribute may have different optimal raw value intervals for each concept that leads to rules of the highest user interest. Consider rules:

(hotel) near attractions ($< 800\text{m}$) \rightarrow price = high
 (hotel) near attractions ($< 1500\text{m}$) \wedge near subway stations \rightarrow price = high

The optimal intervals in parentheses refer to those with the largest membership degrees for the concepts among all concepts in corresponding attributes. These above two rules can suggest that hotels near subway stations are required to achieve only a looser degree of ‘near’ to attractions in order to obtain room price premium, compared with general hotels near attractions.

On the other hand, concepts for the same group of associations (main rule) should have consistent value intervals, otherwise the rules will be confusing. Consider rules:

(hotel) mid-far to attractions (300 – 600m) \rightarrow price = high
 (hotel) mid-near to attractions \rightarrow price = low

Looking into the numerical distance intervals, nearness to attractions are linked to higher hotel room prices. However, the semantic rules with inconsistent definitions

for nearness to attractions misleadingly suggest that nearness to attractions are linked to lower prices.

Second, encoding individual main rules is much more efficient when statistical tests are fully integrated, that is, rules are statistically tested in every generation of the GA. As seen in Chapters 3 and 4, even modest datasets typically have tens of thousands to billions of potential rules. The number of rules under a main rule of up to 4 items in the antecedent is $2^2-2^5 = 4-32$ when every attribute has two values, and $5^2-5^5 = 25-3125$ when every attribute has five. In the two experiments to be presented in this chapter, a main rule typically contained dozens to 300 rules. Encoding entire data discretization schemes would then require hundreds to tens of thousands times as many rules evaluated in each individual as encoding main rules. As the GA time consumption is roughly proportional to the number of rules evaluated, encoding entire data discretization schemes may even be infeasible in terms of time cost.

5.1.2 Fitness assignment with statistically sound tests

Subjects to user needs, the proposed method can use various RIMs as the objectives to be optimized, and conduct different statistical tests on the rules. For more abundant true rules, the statistical tests should follow the crisp-fuzzy approach: a fuzzy rule is accepted if its corresponding crisp rule passes the statistical test, and then its RIM value is evaluated using fuzzy membership. ‘Corresponding crisp rule’ refers to the rule with the same semantics as the fuzzy rule considered which follows the matching crisp data discretization scheme to that for the fuzzy rule, as defined in Equation (4.2).

The objective function $fval$ for computing the fitness value for each main rule M depends on the objective RIMs:

- For RIMs evaluating extra support of a rule or its subsets, compared with that if the items in the rule are unrelated, such as leverage (defined in Section 2.1): $fval(M)$ is equal to summed RIM value of all rules under M that meets constraint ϕ and have their corresponding crisp rules passing the statistical test. Such rules are hereafter called *eligible rules*;
- For RIMs evaluating higher occurrence probabilities of a rule or its subsets, compared with that if the items in the rule are unrelated, such as confidence and improvement: $fval(M)$ is equal to the average RIM value of all eligible rules. Other averaging measures than arithmetic mean may also be used.

ϕ is typically being more than a certain minimum value for the objective RIM. A loose and unbiased constraint is being more than the RIM value suggesting that not all items in the rule are associated, for example, leverage > 0 and improvement > 0 . Users may also set a stricter constraint like leverage > 0.1 under the consideration of specific SARM tasks. Other constraints such as minimum support can be jointly applied.

According to Section 5.1.1, all rules under a certain main rule should have consistent data discretization scheme, that is, they should come from the same individual. Therefore, if multiple individuals have the same $k_1 \dots k_n$ and $loc_1 \dots loc_n$ and thus encode the same main rule, only the one with the highest $fval$ remains unchanged. Other individuals are reset $fval = 0$, so that rules under them will not enter final SARM result, and they have the lowest chance to survive or produce offspring population.

To answer different user needs for balancing the abundance of true rules and risk of spurious rules, two approaches for rules in GA-based fuzzy SARM based on statistically sound tests are proposed. Let G be the number of generations in the GA, N be the population size, or the number of chromosomes evaluated in each generation, and rules under main rule $M: a_1 \wedge \dots \wedge a_q \rightarrow b$ with the number of concepts $k_{a_1} \dots k_{a_q} k_b$ are tested.

- (1) Under the *experimentwise adjustment* approach, the raw significance level α , say 0.05, is corrected for the number of potential rules throughout the GA, with the purpose of limiting the FWER in entire GA to no more than α . The significance level is adjusted to

$$\kappa = \alpha / \left(G \times N \times \prod_{i=1}^q k_{ai} \times k_b \right). \quad (5.3)$$

Equation (5.3) applies three-level Bonferroni corrections to α : first to limit the risk of having any spurious rules in each generation to at most α/G , then to limit such risk in each individual of a generation to no more than $\alpha/(G \times N)$, and finally share the risk in each individual among all rules under it. Alternatively, slightly more rules may be discovered by using Holm procedure (Holm 1979) to replace the last Bonferroni correction. That is, to rank p values of the tests on all eligible rules ascendingly from p_1 , and accept such rules corresponding to $p_1 \dots p_i$ that

$$\forall 1 \leq j \leq i, p_j \leq \alpha / \left(G \times N \times \left(\prod_{i=1}^q k_{ai} \times k_b - j + 1 \right) \right). \quad (5.4)$$

Equations (5.3) and (5.4) are multi-level extensions to the direct adjustment approach (Webb 2007) of correcting for the multiple test problem. As shown in Chapter 4, the direct adjustment approach can effectively control the FWER in fuzzy SARM. Thus, Equations (5.3) and (5.4) should also be able to strictly control the FWER to no more than α .

- (2) Under the *generationwise adjustment* approach, a correction to α is applied for the number of potential rules in each generation of the GA, with the aim of limiting the percentage of spurious rules among all resultant rules to no more than α . Using purely Bonferroni corrections, the adjusted significance level is

$$\kappa = \alpha / \left(N \times \prod_{i=1}^q k_{ai} \times k_b \right). \quad (5.5)$$

If the Holm procedure is adopted, j eligible rules with the smallest p values $p_1 \leq \dots \leq p_i$ in the tests will be accepted, if

$$\forall 1 \leq j \leq i, p_j \leq \alpha / \left(N \times \left(\prod_{i=1}^q k_{ai} \times k_b - j + 1 \right) \right). \quad (5.6)$$

Equation (5.5) or (5.6) restricts the probability of accepting any spurious rules in each generation to at most α , or makes that at most $\alpha \times 100\%$ generations generates spurious rules. As spurious rules are generated due to random data fluctuation, whether any spurious rule is produced in a generation should be independent from the total number of new rules discovered in this generation. That is, even all newly discovered rules in a generation are spurious if any of them are spurious, the expected percentage of spurious rules in final SARM results is still no more than α .

The generationwise adjusted test has a much higher significance level, about G times of that of the experimentwise approach. Thus, the former approach does not maintain a minimum FWER like the latter, but also enables considerably more rules than the latter. Experiments in Sections 5.2 will confirm this point, and also suggest that the percentage of spurious rules under the generationwise approach is often actually below 2% when $\alpha = 0.05$. Thus the generationwise approach still includes necessary correction to α . As has been experimentally proven in Webb (2007) and Section 3.2, when rules are tested at raw significance level of α , the resultant percentage of spurious rules is typically much higher than α . Users may choose the appropriate approach according to the benefit of extra rules discovered using the generationwise approach and the acceptable hazard of possible spurious rules, the balance between which are specific to each SARM task.

Similar to previous chapters, this chapter hereafter exemplifies the proposed technique using leverage as the RIM, $\varphi = \text{'leverage'} > 0$ and chi-square test for productive rules.

5.1.3 Evolutionary model

The proposed GA for crisp-fuzzy SARM is overviewed in Figure 5.2. Main considerations of the algorithm are detailed as follows.

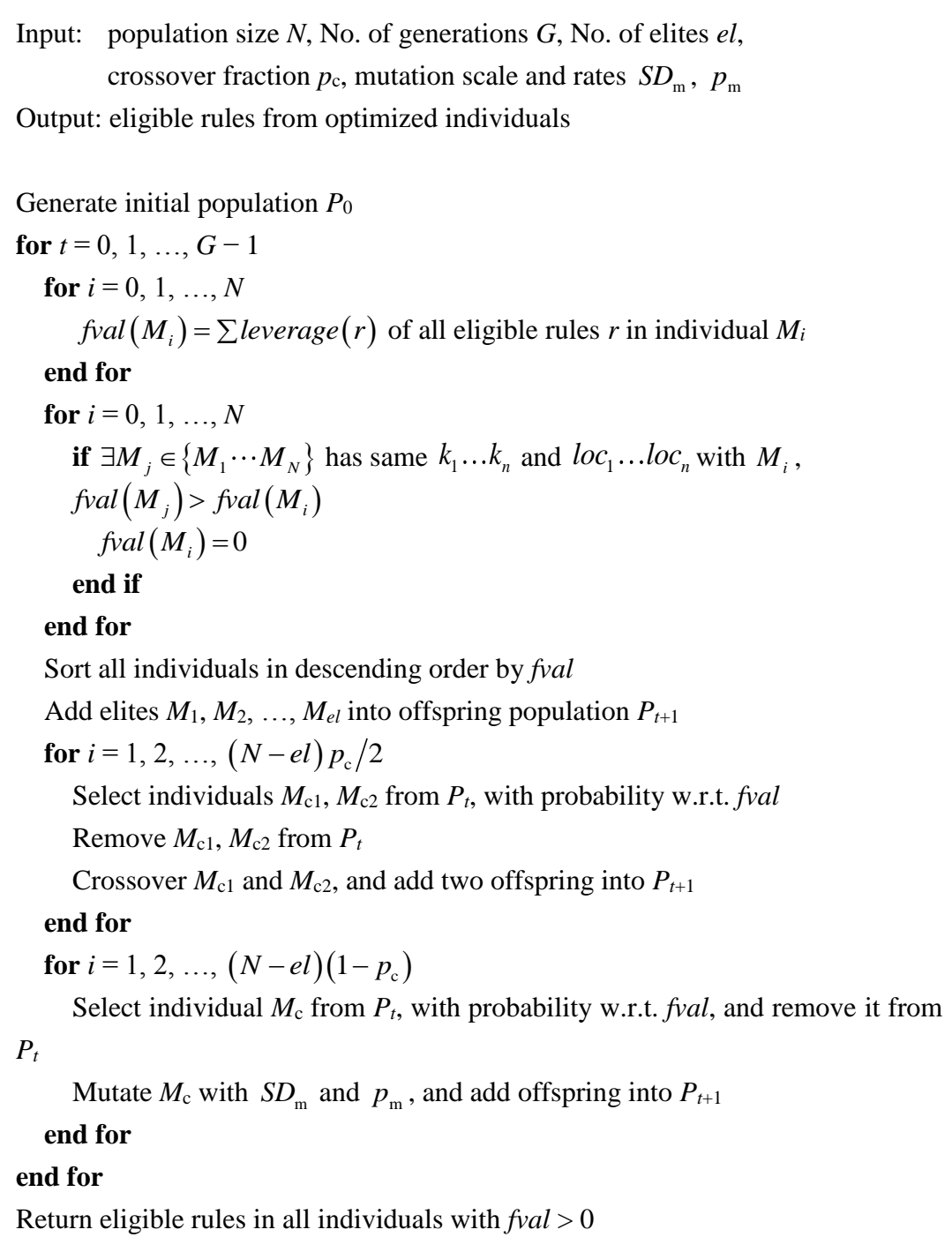


Figure 5.2 Overall procedures of proposed GA for crisp-fuzzy SARM

(1) Generating initial population

For each attribute a in individual M (undistinguished with the main rule it encodes), the number of concepts k_a , $2 \leq k_a \leq k_{\max}$ and number of items in antecedents of rules under M , L_M , are randomly generated. Then L_M attributes and one attribute are randomly selected for rule antecedents ($loc_a = 1$) and consequents ($loc_a = 2$), respectively. Concept core endpoints $cr_{a1_R} cr_{a2_L} cr_{a2_R} \dots cr_{ak_L}$ can also be generated in random. Alternatively, midpoints of each cr_{ai_R} and $cr_{a(i+1)_L}$, $1 \leq i < k_a$, can be decided by standard classification methods such as equisize classification plus random numbers. Then all core endpoints can be computed accordingly.

(2) Genetic operators

- Selection: under the main rule encoding approach, each individual stores only part of optimized data discretizing scheme. To avoid the loss of good data discretization genes in the evolution, a relatively large number of elite individuals with the highest $fval$ values need to survive to the next generation. The number of elites should be close to the number of main rules having eligible rules under them at the end of the GA. This number is specific to data and targeted rules of each SARM task and should be estimated in pilot studies. In crossover and mutation operations, each individual has a probability to be selected as the parent, with higher probabilities for individuals having higher fitness values.
- Crossover: one, two, or more crossover points are randomly selected from locations after cr_{ai_R} , $1 \leq i \leq k_a$ and before loc_a of all attributes a . To produce a crossover child, chromosome segment of one parent M_1 are broken at each crossover point, and the segment of the other parent M_2 starting from cr_{aj_L} will join after cr_{ai_R} of M_1 . j is the smallest such value that cr_{aj_L} in M_2 $> cr_{ai_R}$ in M_1 , or the smallest value that makes $k_a = k_{\max}$, whichever is larger.

- Mutation: for producing a mutation child, core endpoint values of the parent individual are mutated by adding a random number following normal distribution with mean 0 and standard deviation SD_m times the attribute range. loc_a is mutated to some alternative value with probability p_m .

(3) Examining and manipulating fuzziness of concepts

As has been revealed in Chapter 4, crisp discretization for gradual/vague concepts generally causes overestimations of RIM values representing positive data associations. The evolutionary process in GA continuously searches for core and transition intervals of the fuzzy concepts that lead to larger RIM values. As a result, the search is likely to end up with near-crisp concepts with very narrow transitions (defined in Section 4.1.1), which suffers from the RIM overestimation problem like crisp SARM. To avoid this situation, the *fraction of transition*, ft is defined to evaluate the fuzziness of concepts with core $[cr_L, cr_R]$ and base $[a, b]$:

$$ft = 1 - (cr_R - cr_L) / (b - a). \quad (5.7)$$

All crossover and mutation children are then required to have all concepts fulfilling a user specified minimum ft , ft_{\min} .

ft is in line with the classical and widely applied fuzziness measure of fuzzy sets proposed by Yager (1979): for a fuzzy set with base $[a, b]$ and continuous membership function μ ,

$$fuzziness = 1 - \frac{1}{(b-a)^{1/p}} \left[\int_a^b |2\mu(x) - 1|^p dx \right]^{1/p}. \quad (5.8)$$

As illustrated in Figure 5.3a, $\int_a^b |2\mu(x) - 1| dx$ is equal to the area under $|2\mu(x) - 1|$ curve. The left and right transitions have standard deviations $(cr_L - a)/2.473$ and

$(b - cr_R)/2.473$, as suggested in Section 4.1.1. Computed based on cumulative normal distribution function for Gaussian curves,

$$S_{\text{curve}} = (cr_R - cr_L) + 0.5905[(cr_L - a) + (b - cr_R)]. \quad (5.9)$$

Substituting Equation (5.9) into the simplest form of Equation (5.8) which takes $p=1$, $fuzziness = 0.4095ft$. For fuzzy concepts with linear transitions, which will be used in control experiment groups to compare with the proposed Gaussian-curve-based data discretization model in Sections 5.2 and 5.3, it can be similarly derived that $S_{\text{curve}} = (cr_R - cr_L) + 0.5[(cr_L - a) + (b - cr_R)]$ and $fuzziness = 0.5ft$ (Figure 5.3b). The proposed GA uses the simpler ft instead of $fuzziness$, as ft is easier for users to interpret and set a reasonable minimum threshold for.

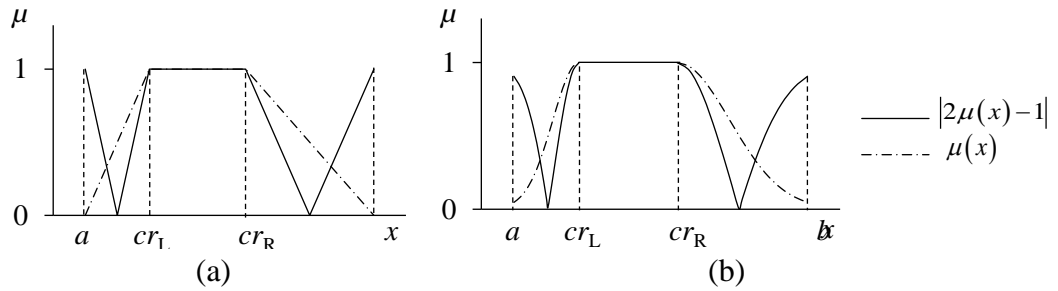


Figure 5.3 Relation between membership function and *fuzziness* for (a) linear and (b) Gaussian-curve-based discretization model

The GA does not directly reject crossover or mutation children that does not fulfilling ft_{\min} , as this will discard favourable data discretization scheme changes in these children towards larger RIMs, and thus considerably slows down the evolution. Instead, the GA first tries to manipulate the transition intervals of these children. For, say, the left transition, a and cr_L are respectively decreased and increased by equal magnitude to make $ft = ft_{\min}$. This does not change the concept of maximum membership degrees for any raw data value. A manipulation succeeds if it does not make core endpoint values conflict other concept cores or fall outside the attribute range. In pilot studies for experiments in Section 5.2, rejecting children with any

concepts not fulfilling ft_{\min} reduced the evolution speed by at least one half, that is, twice as many as generations were needed for discovering the same number of rules. Meanwhile, the concept manipulation succeeded in 95% of the cases, thereby making the evolution speed nearly unaffected by adding ft constraint.

(4) Computational considerations

The time complexity of the GA is $G \cdot N \cdot R$, where R is the number of rules in all chromosomes in each generation. As explained in Section 5.1.1, the currently used chromosome encoding based on main rules dramatically reduces R and thus saves the computation time. In the experiments later in this chapter, the proposed statistical tests took about 10% of the computation time and added to no computational burden compared with existing unadjusted test. Most computation time was spent in computing RIM values, as required by SARM in general.

5.2 Experiments: Hotel room price determinants and wildfire risk factors

5.2.1 Data collection and preprocessing

(1) Hotel experiment

The study area of Hotel experiment is metropolis Hong Kong, a special administrative region in southern China, and a world's leading financial centre and tourism destination. The city is centred around both sides of Victoria Harbour, where landmark scenic spots and luxury hotels are also the most concentrated.

Hotel room prices in Hong Kong Dollars were acquired from the online hotel agency Agoda, which included the largest number of Hong Kong hotels among popular online hotel platforms at the data collection time. With all-year large tourist flow in the city, the prices do not exhibit prominent seasonal change. However, some hotels sell rooms

at discounted rates 1–2 months ahead of, and significantly raise the prices towards check-in date, while other hotels do the contrary. To accurately measure room prices under these two pricing strategies, prices 3 and 7 weeks before check-in date were collected and averaged for the use of SARM. Midweek prices of the cheapest double rooms were collected following common practices in past studies. All prices were obtained within two hours on 1 April 2015, thereby minimizing possible price changes during data collection.

Accessibilities to various types of resources from hotels, represented by walkable road network distances in metres, are summarized in Table 5.1. A self-developed JavaScript program was used to search for the nearest resources to each hotel and measure corresponding distances on Google Maps data, with close human interventions to remove invalid resources and correct measured routes to walkable paths. Types of attractions, including shopping places, were defined according to and included most attraction types highlighted by the Hong Kong Tourism Board (HKTb).

Table 5.1 Accessibility attributes in Hotel experiment

	Name	Description
1–5	dist_top_spot1– dist_top_spot5	Distance to 1st–5th nearest ‘top 10 attractions’ receiving most visitors according to HKTb (2015), mostly landmarks or cultural spots; and major city parks and theme parks
6	dist_museum	Distance to nearest museums ^a
7	dist_worship	Distance to nearest temple/church/other worship places ^a
8	dist_beach	Distance to nearest beach ^a
9–13	dist_shop1– dist_shop5	Distance to 1st–5th nearest shopping centres/multi-storey specialty stores e.g. IT malls
14	dist_subway	Distance to nearest subway (Mass Transit Railway) stations
15–19	dist_bus1– dist_bus5	Distance to 1st–5th nearest bus stops; clustered stop boards for multiple bus routes were regarded as 1 stop

^a Only most significant 30 museums, 30 worship places and 10 beaches highlighted by HKTb

In densely populated Hong Kong, multiple 1 to 2-star cheap hotels are often located in the same multi-storey building. Such hotels are of identical resource accessibilities, similar room conditions and room prices highly dependent on each other, and some are actually operated as one by the same owner. However, statistical tests on variable associations, including those used in SARM, mostly assume mutual independence between tested subjects. Records for such hotels in the same building were thus merged into one new record, whose price was the average of these hotels weighted by numbers of rooms.

After removing hotels without available rooms and merging cheap hotels in the same buildings, the data contained 290 records covering around 68,000 rooms, among which 230 were 3-or-more-star hotels. This which respectively made up 83% and 94% the total number in Hong Kong by the end of 2014 (Census and Statistics Department, HKSAR 2015). The hotels and selected resources are mapped in Figure 5.4.

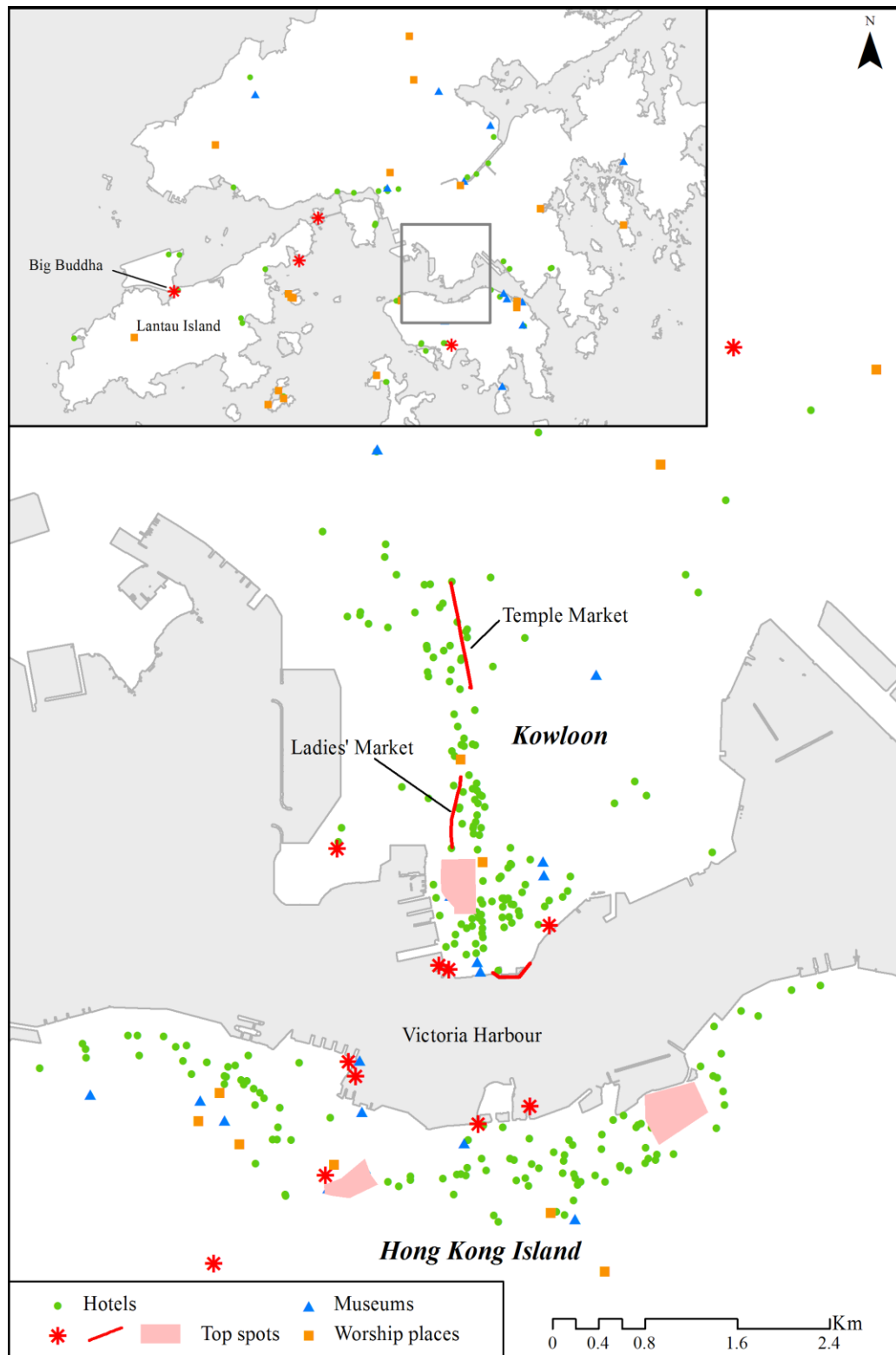


Figure 5.4 Hong Kong map with experimented hotel locations and selective resources

(2) Fire experiment

This experiment revisited Coverttype dataset used in Chapter 4 for investigating Colorado Front Range wildfire risks, and focused on all numerical attributes listed in Table 4.4 but aspect. Attributes above_water and below_water were handled as a single attribute verti_dist_to_water for representing vertical distances (positive or negative) from land cells to nearest water. The data then contained nine attributes in total.

The experiment used the data for Rawah, one of the four wilderness areas covered by the entire dataset, which contained 260,796 records. As will be shown in Section 5.2.4, when applied to Rawah data, the proposed GA already exhibited specific behaviours on large-sized data that were different from those on small-sized Hotel data. Thus, Rawah data alone was sufficient as a representative large-sized dataset for comprehensive evaluation of the proposed GA. Two randomly sampled datasets from the Rawah data, containing 2500 and 20,000 records, respectively, were also experimented for further investigation into behavioural changes of the GA against datasize variations.

5.2.2 Experiment specifications

The proposed GA was implemented and run on data for both experiments to find rules in specific forms. Targeted rule forms and other specifications for the experiments are listed in Table 5.2. A pilot was conducted to find the GA specifications. The el value was determined according to the requirement in Section 5.1.3; for Hotel data, the thereby determined el was equal to 40, and an extra setting of $el = 75$ was experimented in Section 5.2.2 for examining the robustness of spurious rule control by the statistical tests. The numbers of generations G were enough for the increase in numbers of significant rules nearly stall; in Hotel experiment, no more than 0.5

rules/run in the last 300 generations was discovered. In Fire experiment, the increase in the number of rules stopped before 500 generations. The k_{\max} value was set at 5, as more concepts in an attribute were found to produce very small rule supports relative to data size, and corresponding main rules hardly had sufficiently large total leverages to survive early generations of the GA.

Table 5.2 Specifications for Hotel and Fire experiments

	Hotel experiment	Fire experiment
<u>For targeted rules</u>		
Form of rules	resource accessibility(ies) → room price	fire risk factor(s) → horz_dist_to_fire
Max No. of items in antecedent	4	4
<u>For GA</u>		
Population size N , No. of elites el	120, 40 & 150, 75 - Sect. 5.2.2; 120, 40 - Sect. 5.2.3	300, 150
No. of generations G	3000	500
Crossover fraction p_c	0.8	0.8
Mutation scale SD_m	0.03	0.03
Mutation rate p_m	0.025	0.025
Max No. of concepts in an attribute k_{\max}	5	5
Population initialization	Based on equisize classification (see Sect. 5.1.3)	

Values of p_c , SD_m and p_m were generally uninfluential on the final result of the GA given enough generations, with their reasonable value ranges: 0.7–0.95 for p_c (in common practice its value is around 0.8) and 0.01–0.04 for SD_m and p_m . When SD_m , $p_m \geq 0.05$, the mutations were too large and often failed due to breaking the mathematical form of the data discretization model. The p_c , SD_m , p_m values did affect the evolution speed, that is, how many generations were needed for the evolution to stall. Also, the population size N should be at least $2el$, or the evolution would be slow, as more than 50% chromosomes are elites and survive without evolution. The selected

parameter settings in Table 5.2 were relatively good combinations for speeding up the evolution.

5.2.3 Accessing control over spurious rules

The proposed technique was firstly experimented for its ability in controlling spurious rules. The conventional statistical test without adjustment for the multiple testing problem was also assessed to see whether it actually fails to control spurious rules, and thus the proposed adjusted tests are indispensable. To tackle the difficulty of accurately identifying true and spurious rules in real data, as has been stated in previous chapters, this chapter again referred to the approach of Webb (2007) and Chapter 3 of introducing random irrelevant data. Six out of the 19 accessibility attributes in Hotel experiment, and three out of the eight fire risk factor attributes in Fire experiment were randomly selected in each run. Data columns for the selected attributes were replaced by randomly and independently generated values. The replaced attributes were then referred to as ‘irrelevant’. Irrelevant attributes had no association with the rest of data, thus any rules involving them must be spurious.

For Hotel experiment, the proposed GA, in both generationwise and experimentwise adjustment approaches, was experimented with $ft_{\min} = 0.3, 0.5$ and 0.7 . Apart from the newly developed Gaussian-curve-based fuzzy data discretization model, models with triangular and trapezoidal fuzzy sets were also evaluated. Significance levels in the statistical tests were adjusted using Equations (5.4) and (5.6) which incorporated Holm procedure. Control experiment groups, or treatments, were set for the GA with unadjusted statistical tests ($\kappa = 0.05$) on rules using all three fuzzy data discretization models and three ft_{\min} values. Treatments with unadjusted tests took both crisp-fuzzy and conventional fuzzy approaches, the latter meant that p values of the tests were computed based on fuzzy pattern supports. Each treatment was repeated for 5 runs to produce average results.

Table 5.3 lists the results of treatments with statistically sound tests in generationwise approach, and those with unadjusted tests in conventional fuzzy approach. The ‘significant’ and ‘irrelevant’ columns respectively refer to numbers of all significant rules discovered and rules involving irrelevant attributes, the latter being the rules that must be spurious. Treatments with unadjusted tests in crisp-fuzzy approach resulted in at least dozens of irrelevant rules in every run, and thus obviously failed to control spurious rules. This was expected, since in Chapter 3 and 4, crisp unadjusted statistical tests already accepted dozens of spurious rules with even fixed, non-optimized data discretization schemes. Treatments with statistically sound tests in experimentwise adjustment approach produced zero irrelevant rules in all 45 runs, thus the FWER was likely below 5% as this approach was designed for.

Table 5.3 Result on control over spurious rules: Hotel experiment

Data discretization model	ft_{\min}	Generationwise adjusted κ , crisp-fuzzy SARM		Unadjusted κ , conventional fuzzy SARM	
		Significant	Irrelevant	Significant	Irrelevant
(a) $el = 40, N = 120$					
Tri. ^a	0.3	24.6	0.4	107.6	0.0
	0.5	23.4	0.2	108.6	2.2
	0.7	26.0	0.6	110.2	0.8
Trapez.	0.3	30.0	0.2	144.6	11.4
	0.5	23.6	0.0	121.2	3.6
	0.7	30.4	1.0	117.0	0.4
Gaus.	0.3	30.2	0.6	155.8	4.6
	0.5	26.8	0.6	136.8	13.8
	0.7	27.6	0.2	125.4	0.0
Average		27.0	0.4	125.2	4.1
% of irrelevant rules			1.6%	3.3%	
Minimum FWER ^b			24.4%	28.9%	
(b) $el = 75, N = 150$					
Tri.	0.3	21.8	0.2	168.8	4.6
	0.5	27.4	0.8	153.4	11.2
	0.7	23	0.4	155.2	5
Trapez.	0.3	23.8	0	254	12.4
	0.5	24.6	1.8	214.8	18.4
	0.7	19.8	0	173.8	14.6
Gaus.	0.3	22.4	0.2	238.6	17.2
	0.5	21.4	0	215.2	9.6
	0.7	19.8	0	186	6.8
Average		22.7	0.4	195.5	11.1
% of irrelevant rules			1.7%	5.7%	
Minimum FWER			22.2%	75.6%	

^a Tri. = triangular, trapez. = trapezoidal, Gaus. = Gaussian-curve-based; same in Table 5.4 and 5.5

^b = No. of runs containing irrelevant rules/total No. of runs (= 9 groups \times 5runs = 45)

The generationwise approach also appeared to well control the risk of spurious rules below 5% as designed, and such efficacy was robust against variant el values (Table 5.3). With the minimum FWER values below 25%, only 1–2 runs on average in the 5 runs of each treatment contained irrelevant rules. Thus, the variations in numbers of irrelevant rules in different treatments seemed not mainly attributed to data

discretization models or ft_{\min} values, but rather how many runs in each treatment happened to produce irrelevant rules. Regarding only the irrelevant rules as spurious ones, the generationwise adjusted test resulted in highly similar 1.6% and 1.7% spurious rules at $el = 40$ and 75, respectively.

Alternatively, users may want to estimate the percentage of spurious rules raising from original data without irrelevant attributes. As spurious rules rise purely by chance, the possibility of accepting spurious rules among all rules evaluated should be independent from whether the rules involve irrelevant attributes. Hence, the number of spurious rules that involved no irrelevant attributes could be estimated: computed using the method of Webb (2007), around 80% of all potential rules involved at least one irrelevant attributes when $k_{\max} = 5$. For example, when every attribute had five values, 19 attributes for the antecedent of up to four items and constituted 1.27×10^7 potential rules, out of which 1.03×10^7 rules contained items in the 6 irrelevant attributes. When every attribute had 2 values, there were totally 1.41×10^5 potential rules, 1.13×10^5 of which involved irrelevant attributes. As will be shown in 5.2.4, when $el = 40$, averagely 43.4 rules were discovered from original data using the generationwise adjusted test. Thus, the approximate percentage of spurious rules discovered from original data is $(0.4/80\%)/43.4 = 1.2\%$. Based on either evaluation method, the risk of spurious rules was far below 5%.

Meanwhile, unadjusted tests, even when applied on conventional fuzzy rules, were unable to control the spurious rules at 5% level. While the result contained 3.3% irrelevant rules at $el = 40$, the percentage quickly increased to 5.7% at a larger el value of 75. When $el = 40$, actually 31 out of the 45 runs contained irrelevant rules at the 50th generation, while most irrelevant rules were gradually phased out later, leaving only 12 runs with irrelevant rules at the end of GA. The el value of 40 was set to preserve main rules that contained any eligible rules using the proposed adjusted test. Unadjusted tests with a much higher significant level resulted in more than 40 main

rules with eligible rules, and those for irrelevant rules, typically of small summed leverages as they arose by chance from random data, tended to be phased out. With a higher el value, main rules for irrelevant rules would then have larger chance to survive, thereby increasing the risk of spurious rules. This was exactly the condition with $el = 75$ in this experiment.

According to Hotel experiment, the efficacy of the proposed GA in controlling spurious rules are not quite relevant to ft_{\min} values. Thus, Fire data were only experimented for three treatments (15 runs), namely those with all three data discretization models and $ft_{\min} = 0.5$. The result was listed in Table 5.4. Clearly, the generationwise adjusted test produced much smaller percentages of irrelevant rules than in Hotel experiment, averagely only 0.1–0.2% for all treatments with the three data discretization models in each datasize. The treatments with full data produced three irrelevant rules, which counted for $3/[15 \times (227.4 + 226.2 + 216.6)] = 0.09\%$ of all rules discovered. 90% of all potential rules involved the three irrelevant attributes, and the generationwise adjusted test accepted 788 significant rules on average from original data. The estimated percentage of spurious rules when exploring original data was then $(3/90\%)/(15 \times 788) = 0.03\%$. The lower risk of spurious rules in Fire experiment, compared with Hotel experiment, should be attributed to the richer data and more significant rules discovered in the former. With more rules discovered, spurious rules were likely a smaller portion of all rules newly discovered in each generation, while the generationwise adjusted test is designed to always cap the fraction of spurious rules at α even when all newly discovered rules in a generation are spurious if any of them are spurious. Also, p values of all irrelevant rules under the generationwise approach were much higher than the experimentwise adjusted significance levels for corresponding rules, in all generations since they were firstly discovered. Thus the experimentwise adjusted test should result in zero irrelevant rules and 0% FWER if applied to the data for evaluating the generationwise approach.

Table 5.4 Result on control over spurious rules: Fire experiment

Data size	Data discretization model	Generationwise adjusted κ , crisp-fuzzy SARM		Unadjusted κ , conventional fuzzy SARM	
		Significant	Irrelevant	Significant	Irrelevant
2500	Tri.	31.2	0.2 (0.6%)	108.2	16.2 (15.0%)
	Trapez.	33.2	0	144.6	50.8 (35.1%)
	Gaus.	40.8	0	150.2	56.4 (37.6%)
20,000	Tri.	103.2	0	178.2	18.4 (10.3%)
	Trapez.	108.8	0.2 (0.2%)	207.2	47.6 (23.0%)
	Gaus.	96.2	0	242.2	55.8 (23.0%)
Full data (260,796)	Tri.	227.4	0	- ^a	-
	Trapez.	226.2	0.6 (0.3%)	-	-
	Gaus.	216.6	0	-	-

^a Pilot runs produced >>5% false rules, as similar to other datasizes; experiment then discontinued, as the test obviously failed to control spurious rules

Unadjusted tests, in contrast, accepted far above 10% irrelevant rules. This result was even much worse than that for Hotel experiment, and obviously suggests the failure of unadjusted tests in controlling spurious rules. With up to four items for the eight fire risk factors in the antecedent and horz_dist_to_fire as the consequent, there were $C_8^1 + C_8^2 + C_8^3 + C_8^4 = 162$ possible main rules, and due to the rich data, most main rules contained eligible rules. Thus the GA was configured to keep 150 elites. As a result, irrelevant rules had little chance to be phased out during the evolution, and accumulated to a large amount by the end of the GA.

To sum up, experiments on both small (Hotel) and large (Fire) data show that the proposed statistically sound GA is capable and necessary for controlling spurious rules in SARM. Both approaches for adjusted statistical tests on the rules can control spurious rules below their respective aimed levels: $\alpha \times 100\%$ spurious rules in resultant rules for generationwise approach, and $\alpha \times 100\%$ FWER for experimentwise approach, while conventional unadjusted test cannot keep spurious rules under control. The proposed GA is more effective as the data enriches and more rules can be discovered, which is also the trend of modern SARM tasks.

5.2.4 Evaluating ability of discovering true rules

The second parts of the experiments evaluated the ability of the proposed GA in discovering true rules. GA was conducted on original data without introducing irrelevant attributes, and other specifications were the same as in 5.2.3. Hotel experiment also included corresponding treatments with conventional fuzzy SARM, for reconfirming the advantage of crisp-fuzzy SARM in discovering more rules, which was first revealed in Chapter 4.

Data for both experiments were also explored using predefined standard data discretization schemes for comparison with the proposed GA. Data were classified using both equisize and Jenks natural breaks schemes into 2–5 classes for all attributes. KORD algorithm and statistically sound test in direct adjustment approach (Webb 2007) were applied to the crisp discretized data to find significant rules. This was equivalent to the proposed GA in the experimentwise approach for crisp-fuzzy SARM, as the latter is also based on crisp pattern supports in the stage of finding true rules.

Figure 5.5 and Table 5.5 show the GA results of Hotel and Fire experiment, respectively. For Hotel experiment, only generationwise treatment results were plotted. The experimentwise approach, even with crisp-fuzzy SARM, produced only around 10 significant rule in each run, which indicates that this approach was too strict for the small Hotel data to accept sufficient rules for elaborated analysis. For Fire data, however, the over-conservativeness of the experimentwise approach quickly diminished with increasing sampled datasize. With the modest datasize of 20,000, the experimentwise approach already produced more than 3/4 the number of rules accepted with the generationwise approach. Thus the experimentwise approach is still useful when strict control over the FWER is desirable, unless with very small data.

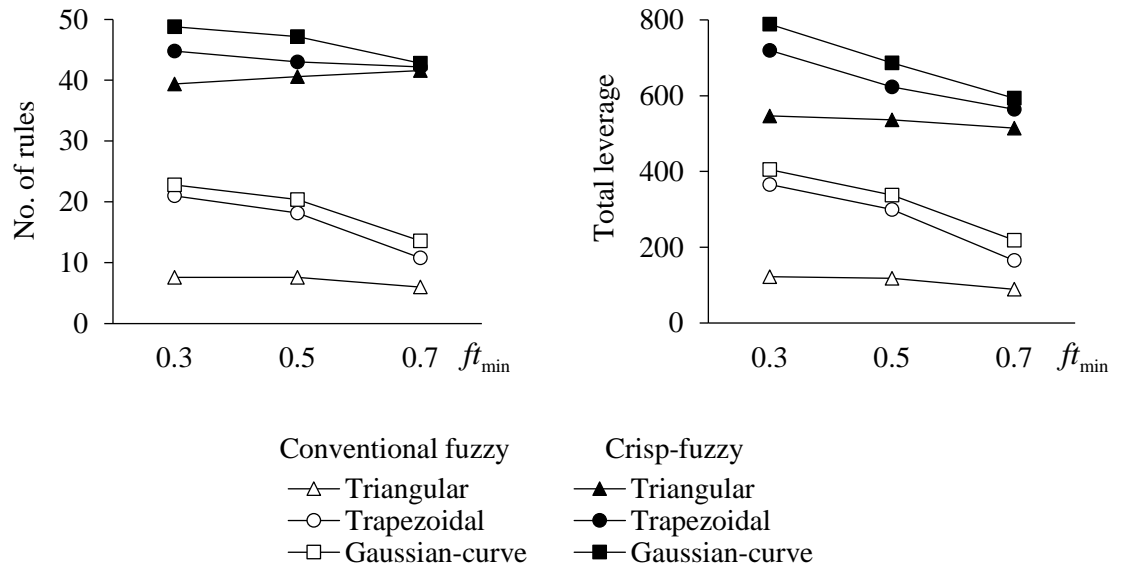


Figure 5.5 Result in discovering true rules: Hotel experiment, generationwise approach, $\alpha = 0.05$

Table 5.5 Result in discovering true rules: Fire experiment, $\alpha = 0.05$, $ft_{\min} = 0.5$

Data size	Data discretization model	Generationwise approach		Experimentwise approach	
		No. of rules	Total leverage	No. of rules	Total leverage
2000	Tri.	87.0	5.59	43.2	3.25
	Trapez.	81.3	5.86×10^3	34.0	2.91×10^3
	Gaus.	82.0	6.06	45.4	3.56
20,000	Tri.	349.2	1.43	271.6	1.20
	Trapez.	355.0	1.39×10^5	256.0	1.20×10^5
	Gaus.	347.0	1.49	261.0	1.20
Full data (260,796)	Tri.	793.8	2.58	730.2	2.50
	Trapez.	797.5	2.78×10^6	738.8	2.58×10^6
	Gaus.	771.4	2.94	740.8	2.64

With standard data discretization, the best result for Hotel data was obtained when all attributes had 3 classes in equisize schemes, and included 6 significant rules with summed leverage equal to 88.0. The best result for Fire data, obtained when all attributes had 3 classes in natural breaks schemes, included 306 rules with summed leverage equal to 9.09×10^5 . Compared with standard discretization results, GA with generationwise adjusted test discovered 2.5–8 times as many rules, and 3–9 times as high leverages. According to Section 5.2.3, only 1.2–1.6% of these discovered rules

were expected to be spurious. As Chapter 4 has shown that crisp rules generally exaggerate leverage, the standard discretization approach should obtain even lower summed leverages if it employed any fuzzy membership instead of the current crisp one. With the stricter experimentwise adjusted test, GA still discovered more rules than the standard discretization approach in Hotel experiment, and twice as many rules in Fire experiment. This clearly suggests the advantage of the GA in enriching resultant rules and optimizing RIM values.

Admittedly, the more abundant true rules discovered by the proposed GA than the predefined data discretization approach attributed to not only generational optimizations of concept intervals, but also higher flexibilities of the concept definitions that might be different for one attribute in different main rules. Also, the generationwise approach result took advantage of the looser statistical test that allowed for higher percentage of spurious rules than the experimentwise one. The direct adjustment approach for SARM with standard data discretization was equivalent to the experimentwise approach for GA, as they both control the FWER at α . However, this does not reduce merits of the proposed GA, as SARM with standard data discretization could achieve neither flexible concept definitions nor control over the percentage of spurious rules at α .

It was infeasible to find even optimal combination of numbers of concepts for individual attributes, let alone optimal intervals for these concepts: in order to find such an optimal combination, data with n attributes and four options of including 2–5 concepts for each attribute needs to be explored for rules 4^n times to try out all concept number combinations. Thus Hotel data needs to be explored using KORD for $4^{20}=1.10\times 10^{12}$ times, and Fire data needs to be explored $4^9 = 2.62\times 10^5$ times. Webb (2007) and Chapter 3 of this thesis have proven that unadjusted statistical tests for SARM with predefined data discretization cannot cap the percentage of spurious rules at the raw significance level α , and to the best knowledge of the author, this has not

been attained by other adjusted statistical test either, except for the statistically sound evaluation which directly controls the FWER at α .

In Hotel experiment, crisp-fuzzy treatments resulted in at least twice as many rules and as large summed leverages as corresponding conventional fuzzy treatments (Figure 5.5). This reconfirms the merit of crisp-fuzzy SARM over conventional fuzzy SARM in finding abundant true rules, and that the proposed GA should incorporate the crisp-fuzzy approach. Conventional fuzzy treatments were not applied in Fire experiment, but Chapter 4 has shown that crisp-fuzzy SARM doubled the number of resultant true rules compared with conventional fuzzy approach.

In both experiments, three forms of membership functions did not make large difference in number of rules discovered by the GA. Crisp-fuzzy SARM is expected to have such robustness against variations in membership function forms, as its statistical test stage is not based on the three fuzzy data discretization models, but instead based on the same crisp model matching the fuzzy ones. Although Gaussian-curve-based treatments resulted in more rules than trapezoidal ones at all ft_{\min} values (Figure 5.5), such difference might appear by chance, since chi-square tests for differences between mean numbers of resultant rules in the five runs of the two membership function forms resulted in $p > 0.2$ at all ft_{\min} values. Differences between Gaussian-curve-based and triangular treatments were larger, and chi-square tests suggested that the differences were significant at 0.1 significance level when ft_{\min} equalled to 0.3 or 0.5. The difference decreased to minimal at $ft_{\min} = 0.7$, when fuzzy sets in Gaussian-curve-based and trapezoidal data discretization models could include only narrow cores and became close to triangular ones. This indicates that for the proposed GA, cores in the discretization model is necessary for more flexible search for and more abundant resultant rules. Meanwhile, experiment results provided no evidence for comparing relative goodness of Gaussian-curve-based and trapezoidal models. The larger summed leverages in Gaussian-curve-based treatments (Figure 5.5,

Table 5.5) might simply reflect lower fuzziness degrees of Gaussian-curve-based fuzzy sets, which equal to $0.41ft$, while fuzziness degrees of trapezoidal sets are $0.5ft$. As suggested Section 4.2.3, given no expert knowledge indicating the membership function forms, the Gaussian-curve-based model is recommended in general, due to its widely recognized capability in representing linguistic concepts, especially geographical ones, and its higher tolerance to extraneous factors in imperfect data than trapezoidal models.

With increasing ft_{\min} values, Gaussian-curve-based and trapezoidal treatments in Hotel experiment resulted in fewer rules and smaller summed leverages. For crisp-fuzzy treatments, even rules were tested using crisp pattern supports that were irrelevant to ft_{\min} , larger ft_{\min} values could still slow down the evolutionary process in GA, as it caused more failures in fuzzy set manipulations for fulfilling ft_{\min} (see Section 5.1.3). Triangular treatments were largely unaffected, as triangular data discretization held $ft = 1$ for all concepts except for those with the smallest and largest value intervals in an attribute. Decreases in summed leverages should be due to dual effects of the slowed evolutions and fuzzier concepts under larger ft_{\min} values. The changing trend in number of rules did not appear in Fire experiment, where the GA focused on optimizing RIM values of resultant rules rather than numbers of rules due to data richness, as will be explained immediately below.

Figure 5.6 illustrates the generational changes in numbers of rules and total leverages of the GA in both experiments. In Hotel experiment, both numbers of rules and total leverages consistently increased during the evolution. In Fire experiment, however, numbers of rules stalled in early generations, and even decreased latter when $ft_{\min} = 0.3$ and $ft_{\min} = 0.5$; at $ft_{\min} = 0.5$, numbers of rules in 350th, 400–450th and 500th generation were 773, 782 and 771, respectively. At $ft_{\min} = 0.7$, main rules evolved slower than in treatments with smaller ft_{\min} values, thus the stage of decrease in numbers of rules might yet to be reached by the end of GA. With eight attributes in

the antecedent of up to four items and fixed attribute in the consequent, Fire data could constitute $C_8^1 + C_8^2 + C_8^3 + C_8^4 = 162$ main rules. As implied by the experiment setting of $el = 150$, around 150 main rules contained eligible rules by the end of pilot GA, and most of them were actually discovered in the first dozens of generations. The data was so rich that eligible rules were discovered under most possible attribute combinations, or main rules, and even subtle rules could be captured instead of rejected simply due to inadequate data and resultantly inadequate statistical significance. Thus, a large number of significant rules quickly saturated most possible main rules. Then the proposed GA shifted the focus of RIM optimization from discovering more rules to optimizing the fuzzy sets for concepts in existing main rules. The decrease in number of rules in later generations, with the proposed GA keeping effective in increasing total leverages of these rules (Figure 5.6d), was likely a streamlining process which resulted in more concise resultant rule set that tends to be easier for users to interpret and make decisions.

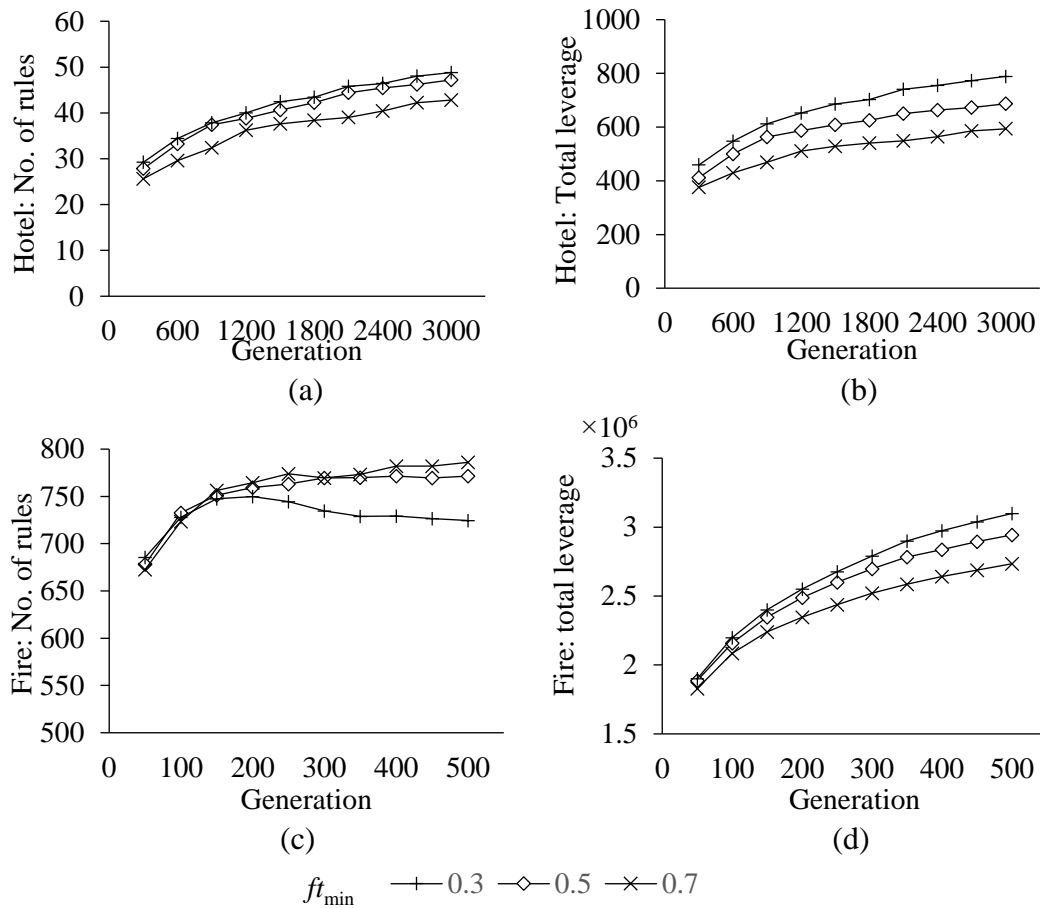


Figure 5.6 Evolutions in numbers of rules and total leverages in GA for (a) (b) Hotel and (c) (d) Fire experiments

5.2.5 Analysing practical implications for Hotel experiment

As the fire risk analysis with the Fire data has been conducted in Chapter 4, this section focused on practical implications of the Hotel experiment result. According to Section 5.2.4, statistically sound SARM with standard data discretization resulted in six rules at best. This is mainly due to limited data size in city-level hotel pricing studies, which also restricts the achievements of prior studies on this topic, as will be detailed in next paragraph. Thus, most analysis outcomes in this section are exclusively contributed by the proposed GA with much richer resultant rules, among which more than 98% are expected to be authentic, according to Section 5.2.3.

The current mainstream approach of hotel room pricing studies is hedonic price modelling. This approach evaluates impacts of various hotel attributes on room price, such as room conditions, services and accessibilities, by regressions with the price as dependent variable and hotel attributes as independent variables. Table 5.6 reviews some representative studies in this line that involved hotel accessibilities. Although room price usually shown positive correlation with accessibilities, sometimes insignificant or even negative correlations were also found, for accessibilities to both city centre/scenic spots and transport facilities. These inconsistent conclusions may be attributed to heterogonous impacts on room price of accessibilities of hotels in different levels (star ratings) or geographic locations, as indicated by results of Zhang H. *et al.* (2011) and Zhang Z. *et al.* (2011) (Table 5.6). This, however, is difficult to be confirmed by regressions. City-level hedonic hotel room price studies typically involve only dozens to hundreds hotels. Subdividing these hotels into groups by star ratings or locations will reduce statistical significance of regression models for each group, thus the models may lose the ability in identifying significant impacts of many hotel attributes. Besides, prior studies have seldom differentially investigated the impacts of accessibilities for detailed resource subtypes, such as landmarks and cultural relics under the type ‘attractions’.

Table 5.6 Past hedonic hotel room price modelling studies that involved hotel accessibilities

Research	Study area	No. of hotels studied	Model ^a	Accessibility measure(s)	Correlation(s) with room prices ^b
Bull (1994)	Ballina, Australia	35 hotels/motels	Linear/quadratic/semilog/loglinear	Nearness to city centre	+
Thrane (2007)	Oslo, Norway	74	Semilog	Nearness to central railway station	+ (double room) I (single room)
Andersson (2010)	Singapore	69	Linear	1) Nearness to CBD 2) Beside Orchard Road (major shopping/entertainment district) or not	1) + 2) +
Zhang, H. <i>et al.</i> (2011)	Beijing, China	228, 3-star or above	Linear/semi-log/loglinear with geographically weighted regression (GWR)	Nearness to 1) closest scenic spots 2) transport hubs	Global average: 1) I 2) + Local GWR: – for some locations, + for others
Zhang, Z. <i>et al.</i> (2011)	New York, US	243, 1- to 5-star	Linear	Tourists' rating of location convenience	1- to 2.5-star hotels: I 3- to 5-star hotels: +
Lee and Jiang (2012)	Chicago, US	81, mid- to high-level	Linear, nonspatial/spatial lag models	Nearness to city centre	+
Park and Kim (2012)	Seoul, Koera	17, 5-star	Linear	Accessibility to road network	+
Balaguer and Pernías (2013)	Madrid, Spain	219, 2-star or above	Linear	Nearness to 1) city centre 2) airport	1) I (In most days of a week) 2) +
Napierala and Lesniewska (2014)	Lodz, Poland	155, all levels		Nearness to 1) city centre 2) nearest transport node	1) I 2) +

^a non-spatial models unless specified

^b +: significantly positive ($p < 0.05$), -: significantly negative, I: insignificant

The improved SARM result of the proposed GA helps address the two issues in prior hedonic analysis, and originally reveals the scale difference issue in analysing hotels of different price levels, as elaborated below.

Among all runs employing the proposed GA in generationwise approach, the result of the run producing the largest number of rules, with $f_{\min} = 0.5$ and 53 resultant rules, was used for analysis. Table 5.7 listed all resultant rules, with each item like

$$\text{attribute } a = \text{concept } c \mid \text{No. of concepts in } a \\ (\text{raw data interval of maximum membership} \cdot \\ \text{degree for } c \text{ among all concepts in } a)$$

The rules were grouped by and repeated under each resource subtype involved, with recurrences marked ‘*’. Concepts for accessibility attributes were assigned as below for attributes containing

- 2 concepts: near, far;
- 3 concepts: near, mid, far;
- 4 concepts: near, mid-near, mid-far, far;
- 5 concepts: near, mid-near, mid, mid-far, far;

Concepts in hotel room price were assigned similarly by replacing ‘near/far’ with ‘low/high’.

Table 5.7 Interpreted resultant rules of Hotel experiment

Antecedent (metre)		Consequent (HK\$) room_price =	Leve- rage	<i>p</i>
(a) Top attractions				
1	dist_top_spot1 = near 3 (<288)	low 3 (<566)	14.76	1.55×10^{-6}
2	dist_top_spot1 = near 3 \wedge dist_worship = near 3 (<546) (672)	low 3 (<942)	15.71	3.15×10^{-7}
3	dist_top_spot1 = near 3 \wedge dist_shop5 = near 2 (<434) (560)	low 3 (<566)	19.65	1.91×10^{-5}
4	dist_top_spot1 = near 2 \wedge dist_subway = near 2 (<527) (225)	low 3 (<701)	20.50	4.76×10^{-6}
5	dist_top_spot2 = near 2 (<1617)	high 3 (>1261)	13.79	3.95×10^{-6}
6	dist_top_spot2 = near 3 \wedge dist_worship = near 3 (<949) (672)	low 3 (<782)	12.88	1.33×10^{-5}
7	dist_top_spot2 = near 3 \wedge dist_worship = mid 3 (<949) (649–1533)	high 3 (>1193)	15.54	1.11×10^{-5}
8	dist_top_spot3 = near 2 (<1808)	high 3 (>1244)	15.87	5.54×10^{-8}
9	dist_top_spot3 = near 2 \wedge dist_top_spot2 = mid 3 (<1793) (763–1734)	high 2 (>940)	16.98	1.87×10^{-5}
10	dist_top_spot3 = near 2 \wedge dist_worship = mid 3 (<1801) (622–1781)	high 3 (>1128)	20.96	7.88×10^{-6}
11	dist_top_spot3 = mid-near 4 \wedge dist_bus4 = far 2 (896–1695) (>248)	high 5 (>1417)	11.98	5.02×10^{-6}
12	dist_top_spot4 = near 2 (<1951)	high 3 (>1222)	17.80	1.74×10^{-7}
13	dist_top_spot4 = near 2 \wedge dist_top_spot2 = mid 3 (<2082) (766–1866)	high 2 (>980)	16.94	1.41×10^{-6}
14	dist_top_spot4 = near 3 \wedge dist_museum = near 3 (<1861) (<915) \wedge dist_subway = mid 3 (208–591)	high 3 (>1077)	20.27	1.49×10^{-6}
15	dist_top_spot4 = near 2 \wedge dist_worship = mid 3 (<2410) (622–1781)	high 3 (>1160)	20.99	2.28×10^{-6}
16	dist_top_spot4 = near 2 \wedge dist_shop1 = near 3 (<2410) (<68)	high 3 (>1210)	13.43	1.54×10^{-5}
17	dist_top_spot4 = near 3 \wedge dist_subway = mid 3 (<1808) (223–598)	high 3 (>1022)	18.33	1.56×10^{-6}
18	dist_top_spot5 = near 2 (<3011)	high 3 (>1254)	15.02	1.44×10^{-6}
19	dist_top_spot5 = near 3 \wedge dist_worship = near 3 (<1810) (<628)	low 3 (<565)	11.39	1.36×10^{-5}
20	dist_top_spot5 = near 3 \wedge dist_worship = mid 3 (<1810) (628–1704)	high 3 (>1022)	14.40	1.23×10^{-5}

Table 5.7 (cont.)

	Antecedent (metre)	Consequent (HK\$) room_price =	Leve- rage	<i>p</i>
21	dist_top_spot3 = mid 3 ∧ dist_shop1 = near 3 (1598–2602) (<83)	low 3 (<511)	7.69	2.98×10 ⁻⁷
22	dist_top_spot3 = mid 3 ∧ dist_shop2 = near 3 (1563–2577) (<195)	low 3 (<511)	7.51	4.75×10 ⁻⁶
23	dist_top_spot1 = far 3 (>819)	mid 3 (566–1312)	14.32	2.52×10 ⁻⁵
24	dist_top_spot2 = far 2 (>1617)	mid 3 (538–1261)	18.58	1.11×10 ⁻⁷
25	dist_top_spot3 = far 2 (>1808)	mid 3 (424–1244)	18.38	7.03×10 ⁻⁹
26	dist_top_spot4 = far 2 (>1951)	mid 3 (516–1222)	20.27	9.11×10 ⁻⁸
27	dist_top_spot5 = far 2 (>3011)	mid 3 (498–1254)	16.68	1.20×10 ⁻⁶
(b) Museum and worship places				
28	dist_museum = near 3 (<1075)	high 2 (>1072)	15.40	5.89×10 ⁻⁵
29	dist_museum = near 4 ∧ dist_worship = mid-far 4 (<473) (693–1416)	high 3 (>1038)	13.18	7.74×10 ⁻⁶
30	dist_museum = near 3 ∧ dist_shop1 = near 3 (<922) (<67)	high 3 (>1222)	13.15	1.35×10 ⁻⁵
12*	dist_museum = near 3 ∧ dist_top_spot4 = near 3 (<915) (<1861) ∧ dist_subway = mid 3 (208–591)	high 3 (>1077)	20.27	1.49×10 ⁻⁶
31	dist_museum = mid 3 (1076–1927)	low 2 (<1072)	14.02	2.34×10 ⁻⁵
32	dist_museum = far 3 ∧ dist_shop1 = near 3 (>1307) (<67)	low 3 (<552)	5.94	1.99×10 ⁻⁶
33	dist_museum = far 3 ∧ dist_shop2 = near 3 (>1242) (<144)	low 3 (<552)	5.70	8.60×10 ⁻⁶
2*	dist_worship = near 3 ∧ dist_top_spot1 = near 3 (<672) (<546)	low 3 (<942)	15.71	3.15×10 ⁻⁷
6*	dist_worship = near 3 ∧ dist_top_spot2 = near 3 (<672) (<949)	low 3 (<782)	12.88	1.33×10 ⁻⁵
17*	dist_worship = near 3 ∧ dist_top_spot5 = near 3 (<628) (<1810)	low 3 (<565)	11.39	1.36×10 ⁻⁵
34	dist_worship = mid 3 (628–1371)	high 3 (>1213)	13.62	2.41×10 ⁻⁵
7*	dist_worship = mid 3 ∧ dist_top_spot2 = near 3 (649–1533) (<949)	high 3 (>1193)	15.54	1.11×10 ⁻⁵

Table 5.7 (cont.)

	Antecedent (metre)	Consequent (HK\$) room_price =	Leve- rage	<i>p</i>
9*	dist_worship = mid 3 \wedge dist_top_spot3 = near 2 (622–1781) (<1801)	high 3 (>1128)	20.96	7.88×10^{-6}
13*	dist_worship = mid 3 \wedge dist_top_spot4 = near 2 (622–1781) (<2410)	high 3 (>1160)	20.99	2.28×10^{-6}
18*	dist_worship = mid 3 \wedge dist_top_spot5 = near 3 (628–1704) (<1810)	high 3 (>1022)	14.40	1.23×10^{-5}
35	dist_worship = mid 3 \wedge dist_shop1 = near 5 (672–1403) (<40)	high 3 (>1316)	11.44	8.22×10^{-6}
36	dist_worship = far 3 (>1371)	mid 3 (>1213)	12.56	4.81×10^{-5}
(c) Shopping places				
37	dist_shop1 = near 3 (<35)	high 3 (>1312)	11.29	8.63×10^{-9}
38	dist_shop1 = mid 3 (35–178)	low 3 (<566)	12.87	9.38×10^{-6}
26*	dist_shop1 = near 3 \wedge dist_top_spot3 = mid 3 (<83) (1598–2602)	low 3 (<511)	7.69	2.98×10^{-7}
30*	dist_shop1 = near 3 \wedge dist_museum = near 3 (<67) (<922)	high 3 (>1222)	13.15	1.35×10^{-5}
32*	dist_shop1 = near 3 \wedge dist_museum = far 3 (<67) (>1307)	low 3 (<552)	5.94	1.99×10^{-6}
39	dist_shop1 = near 3 \wedge dist_bus4 = near 3 (<125) (<229)	low 3 (<566)	12.13	1.11×10^{-5}
40	dist_shop2 = near 3 (<190)	low 3 (<561)	12.10	4.51×10^{-5}
27*	dist_shop2 = near 3 \wedge dist_top_spot3 = mid 3 (<195) (1563–2577)	low 3 (<511)	7.51	4.75×10^{-6}
33*	dist_shop2 = near 3 \wedge dist_museum = far 3 (<144) (>1242)	low 3 (<552)	5.70	8.60×10^{-6}
41	dist_shop3 = near 2 (<354)	low 3 (<561)	12.81	9.02×10^{-6}
42	dist_shop4 = near 3 (<278)	low 3 (<565)	14.98	1.24×10^{-7}
43	dist_shop4 = near 3 \wedge dist_subway = near 3 (<277) (<201)	low 3 (<569)	16.39	7.87×10^{-6}
44	dist_shop5 = near 3 (<307)	low 3 (<561)	12.36	7.82×10^{-6}
3*	dist_shop5 = near 2 \wedge dist_top_spot1 = near 3 (<560) (<434)	low 3 (<566)	19.65	1.91×10^{-5}
45	dist_shop5 = near 2 \wedge dist_subway = near 4 (<552) (<219)	low 3 (<569)	21.67	6.61×10^{-6}

Table 5.7 (cont.)

	Antecedent (metre)	Consequent (HK\$) room_price =	Leve- rage	<i>p</i>
46	dist_shop1 = far 3 (>178)	mid 3 (566–1312)	17.41	2.11×10^{-5}
47	dist_shop2 = far 3 (>332)	mid 3 (561–1208)	21.15	2.79×10^{-8}
48	dist_shop3 = far 2 (>354)	mid 3 (561–1208)	19.02	3.21×10^{-7}
49	dist_shop4 = far 3 (>468)	mid 3 (565–1222)	17.60	1.28×10^{-6}
50	dist_shop5 = far 3 (>516)	mid 3 (561–1208)	17.49	9.79×10^{-6}
(d) Transports				
51	dist_subway = near 3 (<212)	low 3 (<562)	17.77	8.97×10^{-9}
4*	dist_subway = near 2 ^ dist_top_spot1 = near 2 (<225) (<527)	low 3 (<701)	20.50	4.76×10^{-6}
43*	dist_subway = near 3 ^ dist_shop4 = near 3 (<201) (<277)	low 3 (<569)	16.39	7.87×10^{-6}
45*	dist_subway = near 4 ^ dist_shop5 = near 2 (<219) (<552)	low 3 (<569)	21.67	6.61×10^{-6}
52	dist_subway = mid 3 (212–591)	high 3 (>890)	13.73	3.00×10^{-5}
15*	dist_subway = mid 3 ^ dist_top_spot4 = near 3 (223–598) (<1808)	high 3 (>1022)	18.33	1.56×10^{-6}
12*	dist_subway = mid 3 ^ dist_top_spot4 = near 3 (208–591) (<1861)	high 3 (>1077)	20.27	1.49×10^{-6}
	^ dist_museum = near 3 (<915)			
53	dist_subway = far 3 (>591)	mid 3 (562–890)	7.18	5.75×10^{-5}
39*	dist_bus4 = near 3 ^ dist_shop1 = near 3 (<229) (<125)	low 3 (<566)	12.13	1.11×10^{-5}
10*	dist_bus4 = far 2 ^ dist_top_spot3 = mid-near 4 (>248) (896–1695)	high 5 (>1417)	11.98	5.02×10^{-6}

For most rules in Table 5.7, the price was divided into three concepts, at around HK\$500–600 between ‘low’ and ‘mid’ and HK\$1200–1300 between ‘mid’ and ‘high’. Table 5.8 examines the three concepts against hotel room price distribution by star rating, and shows that the former is not equivalent to the latter, though star rating has been widely suggested the strongest determinant on room prices (Andersson 2010, Zhang, H. *et al.* 2011, Zhang, Z. *et al.* 2011, Balaguer and Pernías 2013). Particularly,

rules for ‘mid’ and ‘high’ prices may indicate accessibility characteristics of mid- to high-level hotels (3-stars or above) that are underpriced and sell good prices, respectively, and also link to profitability of these hotels.

Table 5.8 Hotel room price distribution by star rating

Hotel star rating	No. of hotels by price range (HK\$)					Total
	<500	500–600	600–1200	1200–1300	>1300	
5	0	0	5	2	31	38
4–4.5	0	6	69	8	30	114
3–3.5	4	14	52	1	4	75
2–2.5	3	1	1	0	0	5
1–1.5	36	12	13	0	0	58

(a) Top attractions

Hotels near top attractions, with distances from <1.6km for the second to <3km for the fifth nearest attractions, were generally associated with high room prices (rule 5, 8–18, Table 5.7). The associations are reconfirmed in rule 23–27, which indicate that farther hotels tend to sell medium instead of high prices. This supports the strategy of HKTb to focus on these attractions in local tourism promotion. In particular, rule 9 and 13 suggest that relative nearness to both the second and the third/fourth nearest attractions were associated with relatively high room prices, and since the rules have passed the productivity test, contributions of the second and the third/fourth nearest attractions to the price premium cannot replace mutually. Thus, for large tourism cities like Hong Kong, it is recommended to consider accessibilities of hotels to the nearest/second nearest scenic spots as well as existences of multiple spots in their wider surroundings, for more accurate evaluation of hotel locations in terms of price premium.

Meanwhile, hotels very close (<300m) to the nearest attractions were associated with low prices (rule 1). This, however, seems not to reflect any adverse effect of high

accessibility to attractions on room price, but rather because scales of distances and resultant nearness perceptions for high-level hotels are often larger than that for low-level ones. High-level hotels tend to locate in upscale commercial areas with large and relatively widely spaced buildings. Also, network distances in data were measured between entrances of hotels and spots, and might include up to hundreds of metres' walk around them for accessing their entrances. Thus, high-level hotels are unlikely within 300m to the nearest attractions, if not on immediate sides of them. Meanwhile, economic hotels in Hong Kong concentrate in crowded old districts with dense and relatively small buildings, where much larger numbers of buildings may locate within 300m to attractions. In West Kowloon area, for example, dozens of cheap hotel buildings are within 300m to Ladies' Market or Temple Street, two night market streets listed in 'top 10 attractions' (Figure 5.4). According to rule 23, hotels with far distances (>800m) from the nearest attractions typically sell medium prices. This implies that expensive hotels concentrate in 300–800m to the nearest spots. For most tourists, this distance range is of little difference from being within 300m.

Another evidence suggesting that rule 1 is due to scale effect is that when hotels had other features implying small scales of low-end hotel areas, as shown in other items of rule 3–4 and will be explained in (c)–(d), distance ranges to attractions indicating low prices were relaxed to be within around 500m. Rule 2, 6–7 and 19–20 are linked to worship places and will be elaborated in (b).

(b) Museums and worship places

Proximity to museums proved favourable for room price premium and were associated with high prices (rule 28–30, 12*), and areas of medium or far distances from museums were associated with low prices (rule 31–33). Proximity to worship places, by contrast, seems unfavourable. Hotels with near distances (<650m) to worship places were associated with low prices (rule 2*, 6*, 17*), while hotels with medium distances were connected to high prices (rule 34), especially those close to other favourable resources

(rule 7*, 9*, 13*, 18* and 35*). It should be noted that worship places in data were those featured by HKTb and of actual tourism value. As 81% of 3- to 5-star hotels in data are within 600m to their nearest attractions involved in (a) or (b), distance intervals for ‘near’ and ‘medium’ concepts for worship places are unlikely to suggest scale effect like in (a), but the real accessibility. A detailed view of the data revealed that many museums locate around downtown beside Victoria Harbour, while religious spots, except several churches in commercial areas, tend to locate in either near old districts with clustered cheap hotels as said in (a), or remote places with few nearby hotels, such as the famous Big Buddha on Lantau Island (Figure 5.4). Rule 36 associating farness to worship places and medium room prices seems to represent suburban mid- to high-star-rating hotels that are far from resources in general.

(c) Shopping places

Hotels from within 100m for the nearest shops to within 300m for the fifth nearest shops were associated with low room prices (rule 38–44, 26*, 30*, 32*, 27*, 33* and 3*). This seems to reflect a stronger effect of scale difference in nearness for high- and low-level hotels on evaluating their accessibilities for shopping places than for attractions. In high-class commercial districts, a hotel can achieve a network distance of about 300m to its fifth nearest shopping malls, if its adjacent buildings are mostly malls. Such condition is actually common in downtown of Hong Kong, which has very high shop density and attracts exceptional buying passion of visitors even among worlds’ popular tourism cities. Hotels with smaller distances, however, are less likely to be those in upscale districts. Another evidence that the above rules reflect the scale effect, similar to (a), is that additional conditions implying low-level areas, like being very close to spots (rule 3*) and subway stations (rule 45, to be explained in (d)), relaxed distance ranges to the fifth nearest shopping centres suggesting low prices to within around 550m. The exceptional rule 37 which associates hotels within 35m to shopping centres with high room prices should reflect luxury hotels built immediately

over a mall, which were assigned nominal distance of 20m from their lobbies to downstairs shops.

Beyond the scale of low-level districts, high accessibility to shopping places still contribute to room price premium, as implied by rule 46–50 associating hotels in even farther distances and medium prices.

(d) Transport facilities

Rules involving subways again exhibited the scale difference of high- and low-level hotels. The distance of within about 200m, associated with low room prices in rule 51, 4*, 43* and 45*, appears too short and more likely to suggest blocks of dense middle-sized buildings typically for economic hotels. Another proof is that other conditions implying small distance scales for crowded old districts, such as being very near to attractions or shops, further strengthen rule 4*, 43* and 45*. The ‘medium’ concept corresponding to about 200–600m, associated to high room prices in rule 52, 15* and 12*, still indicates convenient walking access and should reflect the degree of closeness in upscale areas. Hotels farther to subways appear less favourable and were associated with medium room prices (rule 53).

Meanwhile, accessibility to buses may not benefit room price premium. Rule 39* and 10* even suggest that hotels within 200–250m to the fourth nearest bus stops have relatively low room prices. Although the two rules might also partially be attributed to the previously stated scale effect, there are also no rules suggesting that hotels in any distance range to bus stops beyond 200–250m have price advantage over farther hotels.

The contrasting effects on room prices of accessibilities to subways and buses may be explained by that in Hong Kong, the former is a better indicator than the latter for convenience of transportation. Subways often lie along traffic arteries, with comparable fares but higher speed than buses, and thus are more preferable for visitors.

Buses have merits of denser networks and larger coverages, which, however, also indicates that proximities to bus stops are unnecessarily busy areas popular among visitors.

(e) Summary and recommendations

In Hong Kong, among various tourism resources investigated, high accessibilities to top attractions, museums, shopping centres and subways are significantly associated with hotel room price premium, while those to worship places and bus stops do not seem beneficial. In above analysis, rules containing accessibilities to multiple resource types prove to be helpful for reasoning the price determining effect of each resource type involved. Decision makers may refer to individual rules for distance ranges of particular resources for advantageous hotel locations, and estimate total room sale premium by selecting favourable or avoiding unfavourable hotel locations according to optimized leverages of corresponding rules.

The case study evokes two further recommendations for future studies on accessibilities to resources as hotel price determinates. First, in measuring accessibilities to a certain type of resources, the subtypes of resources to be included needs to be selective. It is suggested to firstly identify main subtypes of resources that actually contribute to price premium via SARM like that in this study or other techniques, and include only those subtypes in the accessibility measure. Otherwise, useless resource subtypes can hide real influences of other subtypes on room prices in subsequent hedonic price modelling. In the Hong Kong case, including worship places in the accessibility to attractions will not improve the price modelling. This might actually be one reason that some studies in Table 5.6 were unable to find significant correlations between nearness to attractions and hotel room price. Second, attentions need to be drawn to possible discrepant distance scales among hotels in different areas in the city. Regression models in prior hedonic analyses, linear or nonlinear, are mostly monotonic with respect to distances (Table 5.6). In this case, data for low-level

districts, which was found to have smaller distance scales in this study, can exhibit opposite accessibility-price relations to the rest data and worsen the regression results. It would be beneficial to identify possible different scales across the city districts, for which the proposed GA for SARM have been proven useful, and make differential analysis on hotels in districts of different scales.

5.3 Summary

This chapter develops a GA-based method for mining significant crisp-fuzzy SARs. With genetic optimizations, the proposed GA can significantly increase authentic resultant rules and fitness of RIM values. Meanwhile, the GA utilizes experimentwise and generationwise adjusted statistical tests on the rules, two tests newly developed for the GA, for strict control over the risk of spurious rules. The proposed GA also integrates and thus holds advantages of the Gaussian-curve-based data discretization and crisp-fuzzy SARM presented in Chapter 4.

Experiments show that the proposed GA can obtain 2.5–8 times as many rules, and 3–9 times as high RIM values as using statistically sound SARM with standard data discretization schemes. The GA can also effectively keep the FWER and percentage of spurious rules well below 5% user specified level, for both small and relatively large data. In the case study on hotel accessibilities to resources as room price determinants, the proposed GA revealed effects on room prices of resources in more detailed types than prior hotel pricing studies, helped resolve inconsistent outcomes about effects of resources in previous studies, and put new insight into the scale variation issue in analysing hotels in different districts and price levels.

Chapter 6 Conclusions and future work

SARM has become an important topic in GISc and a powerful tool for research, application, and user decision support relevant to spatial data. The usefulness of SARM results highly depends on their reliability, including abundance of authentic rules, risk of spurious rules, and goodness (accuracy and fitness for user needs) of RIM values. Meanwhile, such reliability can be greatly jeopardized by various types of uncertainties that can rise both in source data and each stage of SARM. Three pending uncertainty issues with particularly severe influences on reliability of SARM results are (random) data error, gradual/vague spatial concept, and uncertain concept modelling.

This thesis brings forward uncertainty-based SARM, which includes the development of new techniques, and improvement of existing techniques for handling uncertainties in SARM, and finally enhancing the reliability of SARM results on all three above aspects. Contributions and conclusions of the thesis are summarized in Section 6.1, and further studies are suggested in Section 6.2.

6.1 Research summary and significance

(1) Mining significant SARs from uncertain data

A method for SARM with uncertain erroneous data is innovated, for enrichment of authentic rules in the rule mining result, in the premise of strict control over spurious rules. The control of spurious rules employs statistically sound tests on rules, which has been proven effective for this purpose with determinate data. A statistical model is created to describe random data error propagation in computation procedures of the tests and measure resultant distortion to the test result. Based on this model, the corrected test for rules are designed. The corrected test combines both analytical and simulative techniques for correcting distortions of the test result and recovering loss

of true rules due to data error, while controlling the FWER at a low user specified level.

Assessed with data in various sizes and error levels, the corrected test could consistently compensate rules lost due to data error and discover more true rules than the existing statistically sound test. Around 50% of the lost rules were recovered on average, given accurate error probability information. The efficacy of the corrected test is also largely robust against inaccurate error probability specifications and dependences among the error and attribute values, which makes this method practically useful with real-world imperfect data and error metadata. With the spurious rule rate below 0.2% and a FWER below 5%, the correct test basically reserved the distinctive advantage of existing statistically sound test on controlling spurious rules.

(2) Mining significant crisp-fuzzy SARs

Two techniques are developed in the context of fuzzy SARM. The first technique, a Gaussian-curve-based fuzzy data discretization model for SARM, improves previous models in terms of spatial semantics and relations between multiple spatial concepts. The thesis also conducts a systematic comparative study on RIM accuracy of ordinary SARM and fuzzy SARM with different data discretization models, which has rarely, if ever, been examined by empirical quantitative studies. The second technique, the crisp-fuzzy SARM, includes a statistically sound test stage based on crisp SARs, and an RIM evaluation stage based on fuzzy membership degrees. This method can overcome the difficulty of fuzzy rules to pass statistical tests, thus increasing authentic resultant rules, while integrating the more accurate RIM evaluation of fuzzy SARM for gradual or vague spatial concepts.

The two proposed techniques are experimented with data of various sizes, disturbances and spatial distributions of geographical objects. The techniques prove

to at least double the number of authentic rules, compared with conventional fuzzy SARM. The techniques also maintain the high accuracy of RIM values in fuzzy SARM, and avoid large overestimations of RIMs involving fuzzy concepts caused by ordinary SARM, typically by more than 50%. Rigorous control over spurious rules is realized by the statistically sound evaluation adopted into the techniques, with the FWER of below 1%.

(3) GA for mining significant crisp-fuzzy SARs

GA-based crisp-fuzzy SARM is established. This new algorithm has the merit of genetic optimizations for more authentic resultant rules and higher fitness of RIM values; low risk of spurious rules controlled by statistical testing; and further enrichment of true rules and more accurate RIM evaluation achieved by crisp-fuzzy SARM. Two approaches for statistical testing in the GA, experimentwise and generationwise adjustments, are designed based on statistically sound evaluation technique. The procedures for integrating the GA, crisp-fuzzy method and newly proposed Gaussian-curve-based data discretization are also presented.

The proposed GA can markedly improve the abundance of true rules and RIM goodness. Experimented with both small-size and large-size data, the proposed GA achieved 2.5–8 times as many rules, and 3–9 times as high RIM values as the SARM result without genetic optimization. The FWER and percentage of spurious rules are proven controlled below 5% user specified level, by the experimentwise and generationwise approach, respectively, as they are designed for.

(4) Practical implications of thesis work

New developments of this thesis are applied to and evaluated against a number of GISc case studies. The developments exhibit robustness to real-world imperfect

spatial data, and contribute improved SARM results as well as some insights to the case study topics.

In the spatio-temporal ARM with land use and socioeconomic change data, the corrected test discovered 2–4 times as many rules containing land use changes as by existing statistically sound test. Such improvement is more significant than the synthetic experiment result, suggesting that the corrected test has higher efficacy for more practically meaningful rules which are usually more also more sensitive to data error.

In the wildfire risk factor investigation, the Gaussian-curve-based fuzzy data discretization and crisp-fuzzy SARM doubled authentic rules as compared with conventional fuzzy SARM. Also, the resultant RIM values were more accurate than ordinary SARM, thereby making resultant rules more robust against variations in data discretization scheme, and discovering sensible fire risk factor interactions, such as unbalanced relief of risks above and below water areas.

In the study on hotel accessibilities to resources as room price determinates, non-GA SARM hardly found enough rules for a meaningful result, due to prevalent small sample sizes in hotel pricing studies. GA-based crisp-fuzzy SARM produced dozens of rules with low risk of being spurious. Based on this improved result, influences on room prices of resources in more detailed types than prior studies are analysed, some inconsistent findings on such influences in prior studies can be explained, and spatial variations in the scale of hotel accessibilities are revealed.

6.2 Future work

(1) Mathematical and computational improvement of corrected test

Existing statistically sound test for determinate data (Webb 2007) can strictly cap the FWER of resultant rules at arbitrary user specified upper limit. Among new developments of this thesis, the corrected test for SARM with uncertain data (Chapter 3) maintains statistical soundness at all stages except for the final simulation to determine the z value. The simulation seeks for a z value that has 50% probability to make at most one spurious rule accepted if the FWER is below the user specified level (see Section 3.1.4). Thus the z value is determined by average instead of maximum risk. The corrected test successfully controlled the FWER and percentage of spurious rules below the user specified maximum, in experiments with various treatments, thus its efficacy should be of genericity. However, they are still not theoretically guaranteed to always control spurious rules below arbitrary user specified levels, and might not achieve that given particularly unfavourable data. This calls for further improvement in statistical models of these techniques, even making them totally statistically sound, so as to confidently fulfil user requirement on spurious rule control, while keeping their current advantages in reliability on other aspects.

In the synthetic data experiment of SARM with uncertain data, when all data error probabilities were overestimated (in group R10+/+), the corrected test obtained even more true rules than when the error probabilities were accurate. The underlying reason is unclear, but may be some unrevealed mathematical characteristics in the corrected test that could be utilized to further improve its efficacy. This will be investigated for possible refinement of the mathematical model for the test.

The KORD algorithm used with the corrected test for searching for rules is efficient, and has linear time complexity with respect to datasize. By employing fast searching techniques, the actual time complexity of the experiments in Chapter 3 was less than

linear: as the datasize increased to 10 times as before, the time cost generally increased to 3–4 times. The simulation procedure, however, includes mining randomized data for 34 times (with $\alpha = 0.05$), and thus its time cost is over 30 times of mining the data once. The simulation based on sampled data is planned be investigated and should have significantly smaller computational overhead.

(2) Corrected test for GA

At the moment, the proposed GA does not employ the corrected test for further improving the abundance of true resultant rules, as these two methods are incompatible, for two reasons. First, the GA focuses on optimizing data discretization schemes for numerical data. There seems not an existing method to evaluate the error level of discretized data ready for rule mining based on the error level of raw numerical data. Second, the z value in the corrected test (Equation 3.12) is subject to the values of discretized data. In the GA, the data discretization scheme is variant in each individual and changes in each generation. If applied to the GA, the corrected test should run a simulation to recompute the z value for each individual and each generation, which is computationally unacceptable.

When mining numerical data, the data discretization schemes appear to be more influential than raw data noise on SARM results, given acceptable data quality. Also, as shown in Chapter 3, not adopting the corrected test will not increase the risk of spurious rules in the GA. Yet it is still necessary to develop a corrected test for the GA that overcomes the above two difficulties for the integration of the two techniques.

(3) Extension of new developments to other spatial pattern discovery problems

The developments in this thesis may be generalizable to the problems of mining other spatial patterns that require statistical tests for avoiding spurious patterns, especially

mining sequences and graphs. Like in SARM, the new methods may help these problems find more abundant and better patterns, while keeping low risk of spurious patterns. More reliable sequence and graph mining methods would be very useful for studying urban dynamics and human motilities.

An important challenge for generalizing these new methods lies in mining data streams, which is increasingly demanded in modern big data applications. The new methods must be upgraded to learning changing patterns from data streams of non-stationary distributions. A starting point to tackle this challenge may be the recently realized statistically sound test in data stream mining (Webb and Petitjean 2016).

(4) Extension to handling uncertainties exclusively for spatial data

As stated in Chapter 1, the thesis work focuses on spatial data that can be transformed into attribute-value data, including the nearness relation between spatial entities. The new methods are based on geographical theories and spatial data characteristics, and applied to GISc case studies, but they take data structures under which non-spatial attributes in spatial databases can be processed together. As uncertainty handling research is immature in SDM and even general data mining, SARM still face key uncertainty and reliability issues that hold for both spatial and non-spatial attributes. Uncertainty handling techniques exclusively for spatial (geometrical) data may not work well, or their efficacies may be hindered, unless such more general reliability issues are under control. Therefore, the current somehow ‘general’ solution in this thesis seems an inevitable first step in uncertainty-based SARM.

Exclusive spatial data uncertainty handling and reliability enhancement need to be the close next step. Forthcoming research would include, for example, fuzzy data discretization for directional relations, and statistical tests on rules involving uncertain topological relations.

A key challenge in the future research is that precise modelling of uncertain topology and concise rules are not yet concurrently achievable. As reviewed in Section 2.6, state-of-art uncertain topology studies basically divide spatial entities into certain and uncertain regions, and define more complex topology accordingly. For example, Clementini *et al.* (2000) extended the common nine topological relations between two areas into 56 uncertain ones. The increased numbers of topological relations and resultant rules can add to the difficulty for users to interpret and utilize the rules. Using upper-level topology in a taxonomy of uncertain topology can streamline the rules (Clementini *et al.* 2000), but also turns the rules back to a kind of certain form. It might be more desirable to develop a method that delivers the same number of rules as using certain topology but modifies rule interestingness measures to represent the degree of topological uncertainty.

Another challenge concerns how to recognize the places. Places often do not have precise boundaries, thus all spatial relations - distance, direction and topology between places are uncertain. The hotel room price case study in Chapter 5 involves only urban hotels, attractions and facilities, and distances between their entrances can be precisely measured for the use of the study. Yet the uncertainty issue will occur in studying prices of hotels serving countryside natural scenic spots, or the link between property price premiums and locations inside or near business areas, as natural spots and business areas are vaguely defined places. This uncertainty lies in all kinds of spatial analysis and spatial data mining problems and is linked to the recent proposal of “place-based GIS”.

Appendix Evaluating discrepancy between exact and approximate $\hat{s}_0(c_l)$ values

Let

$$f(x_1, \dots, x_k) = \sum_{j=1}^k p_{ij}^{-1} z \left(\sum_{l=1}^k p_{jl} (1 - p_{jl}) x_l \right)^{1/2}, \quad (\text{A1})$$

then the discrepancy between the exact solution to $\hat{s}_0(c_l)$ from Equation (3.10) and the approximate solution from Equation (3.11) is:

$$\begin{aligned} & \sum_{j=1}^k \left(p_{ij}^{-1} z \left(\left(\sum_{l=1}^k p_{jl} (1 - p_{jl}) s(c_l) \right)^{1/2} - \left(\sum_{l=1}^k p_{jl} (1 - p_{jl}) \hat{s}_0(c_l) \right)^{1/2} \right) \right), \quad (\text{A2}) \\ &= f(s(c_1), \dots, s(c_k)) - f(\hat{s}_0(c_1), \dots, \hat{s}_0(c_k)) \end{aligned}$$

and

$$\frac{\partial f(x_1, \dots, x_k)}{\partial x_l} = \frac{z}{2} \sum_{j=1}^k p_{ij}^{-1} \cdot [p_{jl} (1 - p_{jl})]^{1/2} \cdot x_l^{-1/2}. \quad (\text{A3})$$

Let $\Delta_l = s(c_l) - \hat{s}_0(c_l)$, then Equation (A2) may be estimated by the first degree Taylor polynomial of $f(s(c_1), \dots, s(c_k))$:

$$\begin{aligned} & \left(\Delta_1 \frac{\partial}{\partial \hat{s}_0(c_1)} + \dots + \Delta_k \frac{\partial}{\partial \hat{s}_0(c_k)} \right) f(\hat{s}_0(c_1), \dots, \hat{s}_0(c_k)) \\ &= \frac{z}{2} \sum_{j=1}^k \sum_{l=1}^k p_{ij}^{-1} \cdot [p_{jl} (1 - p_{jl})]^{1/2} \cdot \hat{s}_0(c_l)^{-1/2} \cdot \Delta_l \end{aligned} \quad (\text{A4})$$

The error of estimating Equation (A2) by Equation (A4) is the Lagrange form of the remainder of the first degree Taylor polynomial:

$$\begin{aligned} R_1(\hat{s}_0(c_1), \dots, \hat{s}_0(c_k)) &= \frac{1}{2!} \left(\Delta_1 \frac{\partial}{\partial \hat{s}_0(c_1)} + \dots + \Delta_k \frac{\partial}{\partial \hat{s}_0(c_k)} \right)^2 f(\hat{s}_0(c_1) + \theta \Delta_1, \dots, \hat{s}_0(c_k) + \theta \Delta_k) \\ &= -\frac{z}{8} \sum_{j=1}^k \sum_{l=1}^k p_{ij}^{-1} \cdot [p_{jl} (1 - p_{jl})]^{1/2} \cdot (\hat{s}_0(c_l) + \theta \Delta_l)^{-3/2} \cdot \Delta_l^2 \end{aligned} \quad (\text{A5})$$

where $0 \leq \theta \leq 1$. Each item in Equation (A5) with the same (j, l) value pair is equal to $-\left[(\hat{s}_0(c_l))^{1/2} \Delta_l \right] / \left[4(\hat{s}_0(c_l) + \theta \Delta_l)^{3/2} \right]$ times the corresponding item in Equation

(A4). Typically, $\hat{s}_0(c_l)$ is much larger than Δ_l , thus Equation (A4) is much larger than Equation (A5) and should be a reasonable estimator of Equation (A2).

For each specific attribute in data, the elements of \mathbf{P} and \mathbf{P}^{-1} and $\hat{s}_0(c_1) \cdots \hat{s}_0(c_k)$ values can be substituted into Equation (A4) for evaluating the discrepancy. An exemplary evaluation was made with the item “ $att_3 = 1$ ” in the synthetic experiment (see Section 3.2.1), one of the most affected items by the discrepancy. The computation used the “ideal” data defined in Section 3.2.1 as the original data, and the average z value actually used in the experiment at each error level in Table 3.6. At the highest error level with 20% records contained erroneous att_3 values, the relative discrepancy with respect to the correction to $s(att_3 = 1)$ was only -0.19% and -0.06% for the data size of 4000 and 64,000, respectively. At lower data error levels, the relative discrepancy was even smaller as the z value decreased.

References

- Aggarwal, C.C., Li, Y., Wang, J., and Wang, J., 2009. Frequent pattern mining with uncertain data. In: *17th international conference on knowledge discovery and data mining (KDD 2009)*, 29–38.
- Agrawal, R., Imielinski, T., and Swami, A., 1993. Mining associations between sets of items in massive databases. In: *1993 ACM-SIGMOD International Conference on Management of Data*, Washington, DC, 207–216.
- Agrawal, R. and Srikant, R., 1994. Fast algorithms for mining association rules in large databases. In: *The 20th International Conference on Very Large Databases (VLDB '94)*, 12–15 December 1994, Santiago, Chile. San Francisco: Morgan Kaufmann Publishers Inc., 487–499.
- Agresti, A., 1992. A survey of exact inference for contingency tables. *Statistical Science*, 7(1), 131–153.
- Alcalá-Fdez, J. Alcalá, R., Gacto, M.J., and Francisco Herrera, F., 2009. Learning the membership function contexts for mining fuzzy association rules by using genetic algorithms. *Fuzzy Sets and Systems*, 160, 905–921.
- Alhajj R. and Kaya, M., 2008. Multi-objective genetic algorithms based automated clustering for fuzzy association rules mining. *Journal of Intelligent Information Systems*, 31, 243–264.
- Anchor Point Group, 2010. *Anchor Point national wildfire hazard/risk rating model* [online]. Available from: [http://www.anchorpointgroup.com/images/APG%20National%20Fire %20Model%20-%20Public.pdf](http://www.anchorpointgroup.com/images/APG%20National%20Fire%20Model%20-%20Public.pdf) [Accessed 3 May 2015].
- Andersson, D.E., 2010. Hotel attributes and hedonic prices: an analysis of internet-based transactions in Singapore's market for hotel rooms. *The Annals of Regional Science*, 44, 229–240.
- Appice, A., Ceci, M., Lanza, A. Lisi, F.A., and Malerba, D., 2003. Discovery of spatial association rules in geo-referenced census data: A relational mining approach. *Intelligent Data Analysis*, 7, 541–566.
- Bai, H., Ge, Y., Wang, J., Li, D., Liao, Y., and Zheng, X., 2014. A method for extracting rules from spatial data based on rough fuzzy sets. *Knowledge-Based Systems*, 57, 28–40.
- Balaguer, J. and Pernías, J.C, 2013. Relationship between spatial agglomeration and hotel prices. Evidence from business and tourism consumers. *Tourism Management*, 36, 391–400.
- Baralis, E., Cagliero, L., Cerquitelli, T., and Garza, P., 2012. Generalized association rule mining with constraints. *Information Sciences*, 194, 68–84.
- Barb, A. and Kilicay-Ergin, N., 2013. Genetic optimization for associative semantic ranking models of satellite images by land cover. *ISPRS International Journal of Geo-Information*, 2(2) 531–552.

- Bastide, Y., Pasquier, N., Taouil, R., Stumme, G., and Lakhal, L., 2000. Mining minimal non-redundant association rules using frequent closed itemsets. In: *First international conference on computational logic*, 972–986.
- Bay, S.D. and Pazzani, M.J., 2001. Detecting group differences: Mining contrast sets. *Data Mining and Knowledge Discovery*, 5(3), 213–246.
- Bayardo, R. J., Jr., Agrawal, R., and Gunopulos, D., 2000. Constraint-based rule mining in large, dense databases. *Data Mining and Knowledge Discovery*, 4(2/3), 217–240.
- Beaubouef, T., Ladner, R., and Petry, F., 2004. Rough set spatial data mining for data mining. *International journal of intelligent systems*, 19(7), 567–584.
- Ben-Israel, A. and Greville, T.N.E., 2003. *Generalized inverses: Theory and applications*. Springer-Verlag, New York.
- Bishop, G., 2009. *Assessing the likely quality of the statistical longitudinal census dataset*. Research paper, Australian Bureau of Statistics.
- Blackard, J.A., 1998. *Comparison of neural networks and discriminant analysis in predicting forest cover types*. Thesis (PhD). Colorado State University.
- Bogorny, V., Kuijpers, B., and Alvares, L.O., 2008. Reducing uninteresting spatial association rules in geographic databases using background knowledge: A summary of results. *International Journal of Geographical Information Science*, 22(4), 361–386.
- Bordogna, G., Carrara, P., and Pasi, G, 1991. Query term weights as constraints in fuzzy information retrieval. *Information Processing & Management*, 27(1), 15–26.
- Bordogna, G. and Pasi, G, 1993. A fuzzy linguistic approach generalizing Boolean information retrieval: A model and its evaluation. *Journal of the American Society for Information Science*, 44(2), 70–82.
- Bosc, P., Dubois, D., HaddjAli, A., Pivert, O., and Prade, H., 2007. Adjusting the core and/or the support of a fuzzy set - A new approach to fuzzy modifiers. In: *IEEE International Fuzzy Systems Conference 2007*, 23-26 July 2007 London, 1–6.
- Bradshaw, L.S., Deeming, J.E., Burgan, R.E., and Cohen, J.D., compilers, 1984. *The 1978 National Fire-Danger Rating System: Technical Documentation*. Ogden, UT: US Forest Service.
- Brin, S., Motwani, R. and Silverstein, C., 1997. Beyond market baskets: Generalizing association rules to correlations. In: *SIGMOD 1997, ACM SIGMOD international conference on management of data*, 265–276.
- Bull, A.O., 1994. Pricing a motel's location. *International Journal of Contemporary Hospitality Management*, 6(6), 10–15.

- Burda, M., Pavliska, V., and Valasek, R., 2014. Parallel mining of fuzzy association rules on dense data sets. In: *2014 IEEE International Conference on Fuzzy Systems*, 6–11 July 2014 Beijing.
- Calders, T., Garboni, C., and Goethals, B., 2010. Approximation of frequentness probability of itemsets in uncertain data. In: *The 10th IEEE international conference on data mining (ICDM 2010)*, 749–754.
- Calkin, D.E., Cohen, J.D., Finney, M.A., and Thompson, M.P., 2014. How risk management can prevent future wildfire disasters in the wildland-urban interface. *Proceedings of the National Academy of Sciences of the United States of America*, 111(2), 746–751.
- Carmona, C.J., Gonzalez, P., del Jesus, M.J., and Herrera, F., 2010. NMEEF-SD: Non-dominated multiobjective evolutionary algorithm for extracting fuzzy rules in subgroup discovery. *IEEE Transactions on Fuzzy Systems*, 18(5), 958–970.
- Casillas, J. and Martinez-Lopez, F.J., 2009. Mining uncertain data with multiobjective genetic fuzzy systems to be applied in consumer behaviour modelling. *Expert Systems with Applications*, 36(2), 645–1659.
- Carvalho, J.V. and Ruiz, D.D., 2013. Discovering frequent itemsets on uncertain data: A systematic review. In: *The 9th international conference on machine learning and data mining*, 390–404.
- Census and Statistics Department, HKSAR, 2015. *Hong Kong Annual Digest of Statistics 2015* [online]. Available from: <http://www.statistics.gov.hk/pub/B10100032015AN15B0100.pdf> [Accessed 12 Dec 2016].
- Chambers, M.E., Fornwalt, P.J., Malone, S.L., and Battaglia, M.A., 2016. Patterns of conifer regeneration following high severity wildfire in ponderosa pine dominated forests of the Colorado Front Range. *Forest Ecology and Management*, 378, 57–67.
- Chen, C., Hong, T., Tseng, S., and Chen, L., 2008. A Multi-objective genetic-fuzzy mining algorithm. In: *2008 IEEE International Conference on Granular Computing*, 26-28 August 2008 Hangzhou, 115–120.
- Chui, C. and Kao, B., 2008. A decremental approach for mining frequent itemsets from uncertain data. In: *The 12th Pacific-Asia conference on knowledge discovery and data mining (PAKDD 2008)*, 64–75.
- Chui, C., Kao, B., and Hung, E., 2007. Mining frequent itemsets from uncertain data. In: *The 11th Pacific-Asia conference on knowledge discovery and data mining (PAKDD 2007)*, 47–58.
- Clementini, E., Di Felice, P., and Koperski, K., 2000. Mining multiple-level spatial association rules for objects with a broad boundary. *Data and Knowledge Engineering*, 34, 251–270.
- Cohn, A. and Gotts, N., 1996. The “egg-yolk” representation of regions with indeterminate boundaries. In: Burroughs, P. and Frank, A., editors.

Proceedings of GISDATA Specialist Meeting on Geographical Entities with Undetermined Boundaries. Taylor & Francis, 1996, 171–187.

Colorado Climate Center, 2016. *Climate of Colorado* [online]. Available from: <http://climate.colostate.edu/climateofcolorado.php> [Accessed 10 October 2016].

Fan, B., 2014. Hybrid spatial data mining methods for site selection of emergency response centers, *Natural Hazards*, 70(1), 643–656.

Farzanyar, Z. and Kangavari, M., 2012. Efficient mining of fuzzy association rules from the pre-processed dataset. *Computing and Informatics*, 31, 331–347.

Fazzolari, M., Alcalá, R., Nojima, Y., shibuchi, H., and Herrera, F., 2013. A review of the application of multiobjective evolutionary fuzzy systems: Current status and further directions. *IEEE Transactions on Fuzzy Systems*, 21(1), 45–64.

Feng, T., Zeng, Z., Wu, X., Liu, R., and Gao, L., 2010. Discovery of multi-level spatial association rules based on DE-9IM. In: *2010 International Conference on Management and Service Science*.

Foody, G.M., 2002. Status of land cover classification accuracy assessment. *Remote Sensing of Environment*, 80, 185–201.

Fosu, G.B., 2001. *Evaluation of population census data through demographic analysis*. In: *Symposium on global review of 2000 round of population and housing censuses: Mid-decade assessment and future prospects*. Available from: http://unstats.un.org/unsd/demographic/meetings/egm/symposium2001/docs/symposium_11.htm#_Toc7406238 [Accessed 22 July 2015].

Gerdzheva, A.A., 2014. A comparative analysis of different wildfire risk assessment models (a case study for Smolyan district, Bulgaria). *European Journal of Geography*, 5 (3): 22–36.

Ghobadi, G.J., Gholizadeh, B., and Dashliburun, O.M., 2012. Forest fire risk zone mapping from geographic information system in northern forests of Iran (case study, Golestan province). *International Journal of Agriculture and Crop Sciences*, 4 (12), 818–824.

Gonzales, E. and Zettsu, K., 2012. Association rule mining from large and heterogeneous databases with uncertain data using genetic network programming. In: *The Fourth International Conference on Advances in Databases, Knowledge, and Data Applications (DBKDA 2012)*, 74–80.

Gray, B. and Orlowska, M., 1998. CCAIIA: Clustering categorical attributes into interesting association rules. In: *The 2nd Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'98)*, 132–143.

Herrera, F. and Martinez, L., 2000. A 2-tuple fuzzy linguistic representation model for computing with words. *IEEE Transactions on Fuzzy Systems*, 8 (6), 746–752.

- Hollister, J.W., Gonzalez, M.L., Paul, J.F., August, P.V., and Copeland, J.L., 2004. Assessing the accuracy of National Land Cover Dataset area estimates at multiple spatial extents. *Photogrammetric Engineering and Remote Sensing*, 70, 405–414.
- Holm, S., 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6, 65–70.
- Hüllermeier, E., 2009. Fuzzy methods in data mining. In: John Wang (ed.) *Encyclopedia of Data Warehousing and Mining, 2nd Edition*. IGI Global: Hershey, USA, 907–912.
- International Business Machines, 1996. *IBM intelligent miner user's guide, version 1, release 1*.
- Jiang, B., 2013. Head/tail breaks: A new classification scheme for data with a heavy-tailed distribution. *The Professional Geographer*, 65(3), 2013.
- Jones, N. and Lewis, D. (eds, with Aitken, A., Hörngren, J., and Zilhão, M.J.), 2003. *Handbook on improving quality by analysis of process variables*. Final report, Eurostat.
- Kaufmann, M.R., Veblen, T.T., Romme, W.H., 2006. *Historical fire regimes in ponderosa pine forests of the Colorado Front Range, and recommendations for ecological restoration and fuels management* [online]. Available from: http://coloradoforestrestoration.org/CFRIdpdfs/2006_HistoricalFireRegimesFrontRange.pdf [Accessed 12 May 2015].
- Kaya, M., 2006. Multi-objective genetic algorithm based approaches for mining optimized fuzzy association rules. *Soft Computing*, 10, 578–586.
- Koperski, K. and Han, J., 1995. Discovery of spatial association rules in geographic information databases. In: *The 4th International Symposium on Large Spatial Databases (SSD '95)*, 6–9 August 1995, Portland, Maine, US. *Lecture Notes in Computer Science*, vol. 951. Springer, 47–66.
- Krasnow, K., Schoennagel, T., and Veblen, T.T., 2009. Forest fuel mapping and evaluation of LANDFIRE fuel maps in Boulder County, Colorado, USA. *Forest Ecology and Management*, 257, 1603–1612.
- Ladner, R., Petry, F.E., and Cobb, M.A., 2003. Fuzzy set approaches to spatial data mining of association rules. *Transactions in GIS*, 7(1), 123–138.
- Laube, P., Berg, M., and Kreveld, M., 2008. Spatial support and spatial confidence for spatial association rules. In: *The 13th international symposium on spatial data handling: Headway in spatial data mining*. Springer, 575–593.
- Lee, A.J.T., Hong, R., Ko, W., Tsao, W., and Lin, H., 2007. Mining spatial association rules in image databases. *Information Sciences*, 177, 1593–1608.

- Lee, S. and Jang, S., 2012. Premium or discount in hotel room rates? The dual effects of a central downtown location. *Cornell Hospitality Quarterly*, 53(2), 165–173.
- Lein, J.K. and Stump, N.I., 2009. Assessing wildfire potential within the wildland-urban interface: A southeastern Ohio example. *Applied Geography*, 29, 21–34.
- Lisi, F.A. and Malerba, D., 2004. Inducing multi-level association rules from multiple relations. *Machine Learning*, 55, 175–210.
- Liu, B., Hsu, W., and Ma, Y., 1999. Pruning and summarizing the discovered associations. In: *The fifth ACM SIGKDD international conference on knowledge discovery and data mining*. New York: AAAI, 125–134.
- Liu, B., Hsu, W., and Ma, Y., 2001. Identifying non-actionable association rules. In: *The 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'01)*, 329–334.
- McDonald, J.H., 2014. *Handbook of biological statistics*. 3rd ed. Baltimore: Sparky House Publishing.
- Megiddo, N. and Srikant, R., 1998. Discovering predictive association rules. In: *The fourth international conference on knowledge discovery and data mining*. Menlo Park: AAAI, 27–78.
- Mennis, J. and Liu, J., 2005. Mining association rules in spatio-temporal data: An analysis of urban socioeconomic and land cover change. *Transactions in GIS*, 9(1), 5–17.
- Mitchell, A., 2005. *The ESRI guide to GIS analysis, Volume 2: Spatial measurements and statistics*. Redlands: ESRI Press.
- Mitchell, M., 1996. *An Introduction to Genetic Algorithms*. Cambridge, MA: MIT Press.
- Napierala, T. and Lesniewska, K., 2014. Location as a determinant of accommodation prices: Managerial approach. In: *The 7th World Conference for Graduate Research in Tourism, Hospitality and Leisure*, 3–7 June 2014, Istanbul, Turkey, 687–692.
- Noble, I.R., 1980. McArthur's fire-danger meters expressed as equations. *Australian Journal of Ecology*, 5, 201–203.
- Office for National Statistics, UK, 2014. *2011 Census quality survey* [online]. Available from: <http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-user-guide/quality-and-methods/quality/quality-measures/assessing-accuracy-of-answers/2011-census-quality-survey-report.pdf> [Accessed 22 July 2015].
- Olson, C.E., 2008. Is 80% accuracy good enough? In: *The 17th William T. Pecora Memorial Remote Sensing Symposium*. Available from: <http://www.asprs.org/a/publications/proceedings/pecora17/0026.pdf> [Accessed 27 February 2014].

- Park, E. and Kim, Y., 2012. An analysis of urban hotel location focusing on market segment and local & foreign guest preference. In: *The Eighth International Space Syntax Symposium*, 3–6 January 2012 Santiago, Chile.
- Penrose., R, 1955. A generalized inverse for matrices. *Mathematical Proceedings of the Cambridge Philosophical Society*, 51, 406–413.
- Piatetsky-Shapiro, G., 1991. Discovery, analysis, and presentation of strong rules. In: Piatetsky-Shapiro, G. and Frawley, J. (Eds.), *Knowledge Discovery in Databases*, 229–248. Menlo Park: AAAI/MIT Press.
- Randell, D., Cui, Z., and Cohn., A, 1992. A spatial logic based on regions and connection. In: *Third International Conference on Knowledge Representation and Reasoning*, San Mateo, CA, 165–176.
- Rao, C.R. and Mitra, S.K., 1972. Generalized inverse of a matrix and its applications. In: *The Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Theory of statistics*, 601–620.
- Rodman, L.C., Jackson, J., Huizar, R., and Meentemeyer, R.K., 2006. An association rule discovery system for geographic data. In: *IEEE International Conference on Geoscience and Remote Sensing Symposium*, 3461–3464.
- Robinson, V.B., 2000. Individual and multipersonal fuzzy spatial relations acquired using human-machine interaction. *Fuzzy Sets and Systems*, 113, 133–145.
- Saaroni, H. and Ziv, B., 2003. The impact of a small lake on heat stress in a Mediterranean urban park: The case of Tel Aviv, Israel. *International Journal of Biometeorology*, 47, 156–165.
- Sadat, Y., Nikaein, T., and Karimipour, F., 2015. Fuzzy spatial association rule mining to analyze the effect of environmental variables on the risk of allergic asthma prevalence. *Geodesy and Cartography*, 41(2), 101–112.
- Salleb-Aouissi, A., Vrain, C., and Nortet, C., 2007. Quantminer: A genetic algorithm for mining quantitative association rules. In: *International Joint Conference on Artificial Intelligence 2007*, 1035–1040.
- Shaffer, J.P., 1995. Multiple hypothesis testing. *Annual Review of Psychology*, 46, 561–584.
- Sherriff, R.L., *et al.*, 2014. Historical, observed, and modeled wildfire severity in montane forests of the Colorado Front Range. *PLoS ONE*, 9(9), e106971.
- Shi, W., 2010. *Principles of modeling uncertainties in spatial data and spatial analyses*. Boca Raton: CRC Press.
- Shi, W., Wang, S., Li, D., and Wang, X., 2003. Uncertainty-based spatial data mining. In *Proceedings of Asia GIS Association 2003 Conference*, Wuhan, China, October 16-18, 2003. Wuhan: Wuhan University, 124–135.

- Shu, H., Zhu, X., and Dai, S., 2008. Mining association rules in geographical spatio-temporal data. In: *The 21st International Society of Photogrammetry and Remote Sensing Congress*, Part B2, 225–228.
- Smith, J.H., Stehman, S.V., Wickham, J.D., and Yang, L., 2003. Effects of landscape characteristics on land-cover class accuracy. *Remote Sensing of Environment*, 84, 342–349.
- Stehman, S.V., Wickham, J.D., Wade, T.G., and Smith, J.H., 2008. Designing a multi-objective, multi-support accuracy assessment of the 2001 National Land Cover Data (NLCD 2001) of the conterminous United States. *Photogrammetric Engineering and Remote Sensing*, 74, 1561–1571.
- Sun, L., Cheng, R., Cheung, D., and Cheng, J. 2010. Mining uncertain data with probabilistic guarantees. In: *The 17th international conference on knowledge discovery and data mining (KDD 2010)*, 273–282.
- Sun, Q., Fang, T., and Guo, D., 2005. Study on scale transformation in spatial data mining. In: *International Symposium in Spatio-temporal Modeling, Spatial Reasoning, Analysis, Data Mining and Data Fusion*.
- Taussky, O., 1949. A recurring theorem on determinants. *The American Mathematical Monthly*, 56(10), 672–676.
- Tew, C., Giraud-Carrier, C., Tanner, K., and Burton, S., 2014. Behavior-based clustering and analysis of interestingness measures for association rule mining. *Data Mining and Knowledge Discovery*, 28(4), 1004–1045.
- The Executive Office for Administration and Finance, Commonwealth of Massachusetts, 2012. *MassGIS datalayers*. Available from: <http://www.mass.gov/anf/research-and-tech/it-serv-and-support/application-serv/office-of-geo-graphic-information-massgis/datalayers/layerlist.html> [Accessed 26 September 2013].
- The Virginia Department of Forestry, 2003. *GIS FAQs: Statewide wildfire risk assessment* [online]. Available from: <http://www.dof.virginia.gov/gis/download/Statewide-faq.htm> [Accessed 11 May 2015].
- Thompson, W.A., Vertinsky, I., Schreier, H., and Blackwell, B.A., 2000. Using forest fire hazard modelling in multiple use forest management planning. *Forest Ecology and Management*, 134(1–3), 63–176.
- Thrane, C., 2006. Examining the determinants of room rates for hotels in capital cities: The Oslo experience. *Journal of Revenue and Pricing Management*, 5(4), 315–323.
- Ting, K.M., 2011. Confusion matrix. In: Sammut, C., and Webb, G.I., (eds.) *Encyclopedia of Machine Learning*. 1st ed. Springer, New York, 209.
- US Forest Service, 2010. *Northeast wildfire risk assessment* [online]. Available from: http://www.na.fs.fed.us/fire/pubs/northeast_wildfire_risk_assess10_hr.pdf [Accessed 12 May 2015].

- Verhein, F. and Chawla, S., 2008. Mining spatio-temporal patterns in object mobility. *Data Mining and Knowledge Discovery*, 16, 5–38.
- Versichele, M., De Groote, L., Bouuaert, M.C., Neutens, T., Moerman, I., and Van de Weghe, N., 2014. Pattern mining in tourist attraction visits through association rule learning on Bluetooth tracking data: A case study of Ghent, Belgium. *Tourism Management*, 44, 67–81.
- Wang, H., Fu, B., 1989. The effects of water body on temperature (in Chinese). *Scientia Meteorologica Sinica*, 11(3), 233–243.
- Webb, G.I., 2007. Discovering significant patterns. *Machine Learning*, 68, 1–33.
- Webb, G.I. and Petitjean, F., 2016. A multiple test correction for streams and cascades of statistical hypothesis tests. In: *The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16)*.
- Webb, G.I. and Zhang, S., 2005. *K*-optimal rule discovery. *Data Mining and Knowledge Discovery*, 10(1), 39–79.
- White, P.J. and Mulligan, G.F., 2002. Hedonic estimates of lodging rates in the four corners region. *The Professional Geographer*, 54(4), 533–543.
- Worboys, M.F., 2001. Nearness relations in environmental space. *International Journal of Geographical Information Science*, 15(7), 633–651.
- Wu, B., Li, R., and Huang, B., 2014. A geographically and temporally weighted autoregressive model with application to housing prices. *International Journal of Geographical Information Science*, 28(5), 1186–1204.
- Yager, R.R., 1979. On the measure of fuzziness and negation, Part I: Membership in the Unit Interval. *International Journal of General Systems*, 5, 221–229.
- Yang, G., Shu, L., Di, X., and Heemun, C., 2013. Korean forest fire danger rating index overview (in Chinese). *World Forestry Research*, 3(26), 64–68.
- Yang, L., Stehman S.V., Smith J.H., and Wickham J.D., 2001. Thematic accuracy of MRLC land cover for eastern United States. *Remote Sensing of Environment*, 76, 418–422.
- Zaki, M.J., 2000. Generating non-redundant association rules. In: *The 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2000)*, 34–43.
- Zhang, A., Shi, W., and Webb, G.I., 2016. Mining significant association rules from uncertain data. *Data Mining and Knowledge Discovery*, 30(4), 928–963.
- Zhang, H., Padmanabhan, B., and Tuzhilin, A., 2004. On the discovery of significant statistical quantitative rules. In: *The tenth international conference on knowledge discovery and data mining*, New York: ACM, 374–383.

- Zhang, Z., Ye, Q., and Law, R., 2011. Determinants of hotel room price: An exploration of travelers' hierarchy of accommodation needs. *International Journal of Contemporary Hospitality Management*, 23(7), 972–981.
- Zhang, H., Zhang, J., Lu, S., Cheng, S., and Zhang J., 2011. Modelling hotel room price with geographically weighted regression. *International Journal of Hospitality Management*, 30, 1036–1043.
- Zhu, Q., Pan, D., and Yang, G., 2010. A sampling based algorithm for finding association rules from uncertain data. In: *2010 International Conference on Artificial Intelligence and Computational Intelligence (AICI'10)*, 124–131.
- Zhu X., Wu X., 2006. Error awareness data mining. In: *2006 IEEE international conference on granular computing*, 269–274.
- Zhu, X., Wu X., and Yang, Y., 2004. Error detection and impact-sensitive instance ranking in noisy datasets. In: *The 19th national conference on artificial intelligence*, 378–383.