Data Min Knowl Disc (2016) 30:928–963
DOI 10.1007/s10618-015-0446-6

Mining significant association rules from uncertain data

Anshu Zhang[1] · Wenzhong Shi[1] ·
Geoffrey I. Webb[2]

Received: 21 January 2015 / Accepted: 30 November 2015 / Published online: 12 January 2016

Abstract In association rule mining, the trade-off between avoiding harmful spurious rules and preserving authentic ones is an ever critical barrier to obtaining reliable and useful results. The statistically sound technique for evaluating statistical significance of association rules is superior in preventing spurious rules, yet can also cause severe loss of true rules in presence of data error. This study presents a new and improved method for statistical test on association rules with uncertain erroneous data. An original mathematical model was established to describe data error propagation through computational procedures of the statistical test. Based on the error model, a scheme combining analytic and simulative processes was designed to correct the statistical test for distortions caused by data error. Experiments on both synthetic and real-world data show that the method significantly recovers the loss in true rules (reduces type-2 error) due to data error occurring in original statistically sound method. Meanwhile, the new method maintains effective control over the familywise error rate, which is the distinctive advantage of the original statistically sound technique. Furthermore, the method is robust against inaccurate data error probability information and situations not fulfilling the commonly accepted assumption on independent error probabilities of different data items. The method is particularly effective for rules which were most practically meaningful yet sensitive to data error. The method proves promising in enhancing values of association rule mining results and helping users make correct decisions.

Responsible editor: M.J. Zaki.

✉ Wenzhong Shi
lswzshi@polyu.edu.hk

[1] Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, People's Republic of China

[2] Faculty of Information Technology, Monash University, Melbourne, VIC 3800, Australia

Keywords Pattern discovery · Association rules · Statistical evaluation · Uncertain data

1 Introduction

Association rule mining has been ascendant since its introduction (Agrawal et al. 1993). It seeks *association rules*, or implicit data association patterns in an antecedent-consequent form, that meet certain interestingness measures. Compared with traditional exploratory data analysis, association rule mining is superior in investigating multiway interactions between numerous entities that are hard to represent by a single model. With ever complicated data interactions in modern databases, association rule mining has been in increasing demand, and has become a valuable tool for investigating interactions in complex data and supporting subsequent user decisions, for research and practical projects in many fields.

The key to the value of association rule mining is the reliability of its result. Such reliability is a balance between discovering authentic rules that can help decision making, and avoiding spurious rules that convey non-existent associations and mislead users into poor decisions. Due to the enormous number of potential rules that must be explored for high dimensional data, conventional association rule mining algorithms face a very high risk of falsely ‘discovering’ numerous spurious rules. As modern voluminous databases and rule mining algorithms produce ever greater numbers of rules, the problem of spurious rules becomes a critical barrier for reliability and value of association rule mining results. Another reliability issue arises from uncertainty in the data used, a major source of which is imprecision and error. The error propagates from source data and distorts the quantitative process at each stage of rule mining, eventually leading to loss of authentic rules and introduction of spurious rules.

Intensive association rule mining studies have been made on either avoiding spurious rules or handling data uncertainty. However, the two research branches rarely communicate, and it is hard to integrate their solutions for additive enhancement to association rule mining results.

Avoidance of spurious rules started from the original formulation of association rules (Agrawal et al. 1993), where two basic interestingness measures, support and confidence, were used to evaluate the rules. Numerous other measures have been proposed to augment support and confidence, such as leverage (Piatetsky-Shapiro 1991), R-interest (Srikant and Agrawal 1995), lift (International Business Machines 1996), interestingness (Gray and Orlowska 1998) and improvement (Bayardo et al. 2000). Rules with the measured values lower than specified minimums are considered uninteresting and removed from the result. There have been also quantitative measures for selecting rules of interest, such as non-redundant rules (Bastide et al. 2000; Zaki 2000), actionable rules (Liu et al. 2001) and productive rules (Webb 2007). These measures can be very useful, yet they frequently lack scientifically sound or empirically generic solutions to setting minimum thresholds for accepting interesting rules. The thresholds are usually up to subjective user specifications and bear high risk of being inappropriate and leading to questionable reliability of selected rules.

Studies have also utilized statistical hypothesis tests to evaluate the rules (Brin et al. 1997; Liu et al. 1999; Megiddo and Srikant 1998; Bay and Pazzani 2001; Zhang et al. 2004). The data, however large and comprehensive, is a representation, or can be regarded a sample, of associations between entities in the real world. The statistical tests assess whether a rule satisfies the interestingness measures in data purely by chance instead of due to real-world associations. If so, the rule is considered insignificant and pruned. However, it has been demonstrated that it is critical to adjust the significance level to allow for the multiple testing problem which occurs when many rules are tested (Webb 2007). Otherwise, the familywise error rate (FWER), or the risk that any spurious rules among all tested ones are accepted by the test, can be very large. Actually, spurious rules usually take a significant portion in the result. Statistically sound technique for adjusting the significance level has been proposed and achieved strict control over the FWER at a low user specified level, for example 5 % (Webb 2007). The technique will be elaborated in Sect. 2.2. While this technique is successful, it and other statistical evaluation methods have seldom, if ever, considered the impact of data error, though the error is often inevitable in data mining projects. As will be reasoned in Sect. 2.3 and confirmed in later experiments, data error can cause severe loss of true rules that may be discovered by the statistical tests.

Association rule mining with uncertain data has also attracted increasing research effort (Chui et al. 2007; Chui and Kao 2008; Aggarwal et al. 2009; Calders et al. 2010; Sun et al. 2010; Tong et al. 2012). 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. 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 the error as a major source of data uncertainty, they have yet addressed the random error behaviors 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 association rule mining algorithms employed the probabilistic data structure, and none was for random data error.

This study develops a new and robust method to conduct statistical test on association rules with uncertain data. The method aims at correcting computation of the statistical test for distortions made by data error, thereby obtaining resultant rules that are more reliable and valuable for decision making. The new method is based on an originally developed mathematical and statistical model for data error propagation through computation procedures of the test.

When experimented with computer synthesized data, the method recovered a considerable percentage of the true rules lost due to data error by existing statistically sound technique, whether the data error information was accurate or inaccurate, or the error probability was correlated to other factors in data. When applied to real-world data, the method improved the number of most practically valuable rules by 1–3 times. Also, it essentially preserved the low FWER of the existing statistically sound test, which is far stricter than most alternative methods that seldom get entire results free of false rules.

In the rest of this article, Sect. 2 introduces the basics of association rules and statistical tests on the rules. Section 3 presents the new statistical test method on association rules with uncertain data. Sections 4 and 5 respectively illustrate methods and results of, and discuss about the synthetic and real-world data experiments on the new test. Section 6 discusses the accuracy of data error information in reality and its implication to the practical value of the new method. Section 7 concludes the article and suggests further researches.

2 Statistical test for significant association rules

2.1 Association rules and interestingness measures

This study is described hereafter in terms of categorical data, one of the two most used data types in association rule mining. The categorical dataset D can be regarded as a set of records, where each record is a set of items, and each item is a value of an attribute. Transactional data, the other most used data types, may be handled as binary categorical data, where values 0 and 1 represent nonexistence and existence of an entry in the record. Numerical data is typically classified and transformed into categorical data before being explored for association rules.

An *association rule* is a pattern like $X \rightarrow Y$, where the *antecedent* X and *consequent* Y are non-empty sets of items in D. $X \cup Y$ contains at most one class of each attribute. The rule reflects an association between items in X and Y. This study limits Y to single-item consequent y.

Association rule mining aims at discovering all rules of interest that meet thresholds, mostly minimum, of specified interestingness measures, for example, *support*, *confidence* (Agrawal et al. 1993) and *improvement* (Bayardo et al. 2000):

$$\begin{aligned} support\,(X \rightarrow y) &= freq\,(X \cup y)/|D|;\\ confidence\,(X \rightarrow y) &= freq\,(y|X)/|D|;\\ improvement(X \rightarrow y) &= confidence(X \rightarrow y) - \max_{Z \subset X}(confidence(Z \rightarrow y)), \end{aligned}$$

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 $improvement(X \rightarrow y) > 0$, that is, every item in X improves the confidence of the rule. Unproductive rules include redundant items in X that are irrelevant to y, and are generally regarded as uninteresting and removed from final result.

D can be seen as a sample of the "real-world" situation. A rule satisfying specified interestingness measures in D may not satisfy the measures in reality. Instead, the measures may be fulfilled by chance due to random fluctuations of data. A rule in final resultant rule set is a *true discovery* if it indeed satisfies the specified interestingness measures in real-world situation; otherwise it is a *false discovery*.

2.2 Statistical tests for significant productive rules

A rule $X \rightarrow y$ is productive if and only if its confidence is higher than that of any of its generalizations formed by removing a subset from X. Following accepted practice

(Webb 2007) we perform a more computationally efficient test, which is similar to a full test for productivity, on those immediate generalizations formed by removing any one item in X. That is, for $X = \{x_1 \ldots x_n\}$, we test

$$\forall m = 1, \ldots, n, \ \Pr(y|X) > \Pr(y|X - \{x_m\}). \tag{1}$$

This ensures that all items in X must contribute to the strength of the association. The null hypothesis for this test is $\exists m = 1, \ldots, n, \ \Pr(y|X) \leq \Pr(y|X - \{x_m\})$, 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. The test result is the probability p that $X \rightarrow y$ has observed interestingness measure values in data if the null hypothesis is true. If p is below a specified significance level, such as 0.05, then $X \rightarrow y$ is *significant* and accepted as reflecting a real-world association.

Chi-square is commonly used for testing conditions like (1), yet it is widely criticized as inaccurate for small samples of sizes up to hundreds (McDonald 2014) which is a likely level of pattern supports in association rule mining. Also, the chi-square test is two-tailed, while (1) is a one-tailed condition. A more appropriate test for (1) is the Fisher exact test (Agresti 1992). This test is exact and thus 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:

$$\begin{aligned}
a &= freq\,(X \cup \{y\}) \\
b &= freq\,(X \cup \neg\{y\}) \\
c &= freq\,((X - \{x_m\}) \cup \neg\{x_m\} \cup \{y\}) \\
d &= freq\,((X - \{x_m\}) \cup \neg\{x_m\} \cup \neg\{y\}).
\end{aligned} \tag{2}$$

Where $\neg$ refers to that the record must not contain the item. The p value of this test is

$$p = \sum_{i=0}^{\min(b,c)} \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{(a+b+c+d)!(a+i)!(b-i)!(c-i)!(d+i)!}. \tag{3}$$

The multiple testing problem occurs when the Fisher exact test is 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 rules that are accepted 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. Detailed method of computing s can be referred in the above literature. With only a small 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 actually allows a substantial percentage of true rules to past the test and thus be discovered. This approach is highly effective and can achieve an FWER below 1 % with $\alpha = 0.05$.

2.3 Effect of data error on statistical test

Random error in data has no associations with any data items, therefore its major impact should be weakening actual associations in data. This can cause true rules to fail a significance test and thus be lost to the resultant rule set, thereby reducing the value of the result for decision support. This inference has been confirmed by experiments in Sects. 4 and 5. This problem cannot be alleviated by simply increasing the significance level for the test. As suggested by Webb (2007) and this study, only a small to moderate percentage of potential rules are actually true. If the significance level is increased to accept more rules indiscriminately, false discoveries can become substantial, or even the majority, among newly accepted rules. It may be acceptable to increase true discoveries at the price of slightly more false discoveries. Yet since false discoveries can be very harmful, increase in true discoveries needs to be much larger than, say dozens of times of, that in false discoveries. Such discriminate increment of true discoveries calls for a method to correct the test for impact of data errors according to their statistical behaviors, as to be presented in Sect. 3.

3 Statistical test for association rules with uncertain data

This section presents the new statistical test method for association rules which can get more accurate results in presence of uncertain, or erroneous data. The method achieves such improvement basically by correcting computational parameters in the test for the impact of data error, thus it will be referred as the *corrected test*, while the existing test without considering the data error will be referred as the *original test*. This study takes the statistically sound test described in Sect. 2.2 as the original test and the base of the corrected test. Yet the method of modelling and correcting the data error is applicable to other tests for significant rules.

Section 3.1 gives a mathematical model to describe the data error propagation to computational parameters in the statistical test. The model was originally developed, as past uncertain association rule studies did not provide a model exclusively for random behaviors of data error (see Sect. 1). Section 3.2 illustrates a method based on the error model to correct test parameters for the distortion made by the error, thereby recovering lost true discoveries. Section 3.3 discusses the technique to control the risk of false discoveries in the corrected test.

3.1 Modeling 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, …, k and any data record containing a. For $i, j = 1, \ldots, k$, 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, $p_{ij} = \Pr(\text{value in data} = i|\text{true value} = j)$. Then there is

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

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 $j = 1, \ldots, k$.

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, $s(c_j \to c_i)$, is the number of successes in $s_0(c_j)$ such independent Bernoulli experiments, and follows a binomial distribution: $s\left(c_j \to c_i\right) \sim B\left(s_0\left(c_j\right), p_{ij}\right)$. In association rule mining, normally $s_0(c_j) >> 30$, $s(c_j)p_{ij} >> 5$ and $s(c_j)\left(1 - p_{ij}\right) >> 5$, so the distribution of $s(c_j \to c_i)$ can be approximated by a normal distribution: $s\left(c_j \to c_i\right) \sim N\left(s_0(c_j)p_{ij}, s_0(c_j)p_{ij}\left(1 - p_{ij}\right)\right)$.

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

$$s(c_i) \sim N\left(\sum_{j=1}^{k} p_{ij}s_0(c_j), \sum_{j=1}^{k} p_{ij}(1 - p_{ij})s_0(c_j)\right). \tag{4}$$

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

$$E(s(c_i)) = \sum_{j=1}^{k} p_{ij} s_0(c_j), \tag{5}$$

$$\sigma^2(s(c_i)) = \sum_{j=1}^{k} p_{ij}(1 - p_{ij}) s_0(c_j). \tag{6}$$

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

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

$$\mathbf{E}(\mathbf{S}(a)) = \mathbf{P}\mathbf{S_0}(a), \tag{7}$$

$$\begin{aligned} \mathbf{\Sigma}(\mathbf{S}(a)) &= \begin{pmatrix} \sigma(s(c_1)) \\ \vdots \\ \sigma(s(c_k)) \end{pmatrix} \\ &= \begin{pmatrix} ((p_{11}(1-p_{11}))\,(s_0(c_1)) + \ldots + (p_{1k}(1-p_{1k}))\,(s_0(c_k)))^{1/2} \\ \vdots \\ ((p_{k1}(1-p_{k1}))\,(s_0(c_1)) + \ldots + (p_{kk}(1-p_{kk}))\,(s_0(c_k)))^{1/2} \end{pmatrix}. \end{aligned} \tag{8}$$

3.2 Recovering test parameters

Equation (7) 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 $\mathbf{S}_0(a)$**,** $\hat{\mathbf{S}}_\mathbf{0}(a)$**,** can then be solved:

$$\hat{\mathbf{S}}_\mathbf{0}(a) = \mathbf{P}^{-1}\hat{\mathbf{E}}(\mathbf{S}(a)). \tag{9}$$

When expanded, (9) is a matrix of k equations, each for one value in a. The ith row of the matrix form shows the *estimated true support* for value i:

$$\hat{s}_0(c_i) = \sum_{j=1}^{k} p_{ij}^{-1} \hat{E}\left(s(c_j)\right), \tag{10}$$

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

As $\hat{E}\left(s(c_j)\right)$ is the most probable value of the observed support $s(c_j)$, it is straightforward to take $\hat{E}\left(s(c_j)\right) = s(c_j)$. The probabilities that $s(c_j) > E(s(c_j))$ and

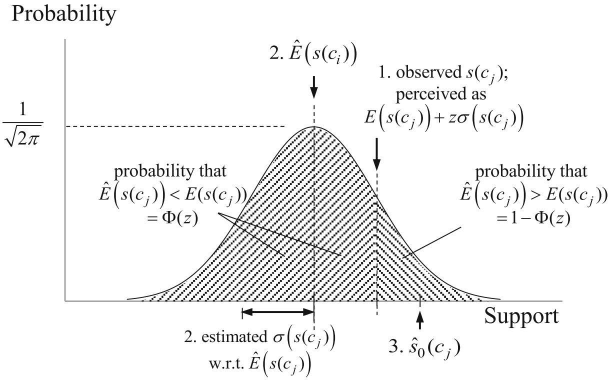


Fig. 1 Using $\sigma\left(s(c_j)\right)$ and z to control probability of overestimating $E(s(c_j))$ at arbitrary user specified value

$s(c_j) < E(s(c_j))$, 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 $\hat{E}\left(s(c_j)\right) > E\left(s(c_j)\right)$, or that $E\left(s(c_j)\right)$ 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 $E\left(s(c_j)\right) + z\sigma\left(s(c_j)\right)$, we take $\hat{E}\left(s(c_j)\right) = s(c_j) - z\sigma\left(s(c_j)\right)$. The probability that $s(c_j) > E(s(c_j)) + z\sigma\left(s(c_j)\right)$ is $1 - \Phi(z)$, where Φ is the cumulative distribution function of the standard normal distribution. The probability that $\hat{E}\left(s(c_j)\right) > E\left(s(c_j)\right)$, equivalent to $s(c_j) > E(s(c_j)) + z\sigma\left(s(c_j)\right)$ by this estimation, is also $1 - \Phi(z)$ (Fig. 1).

For (10), substitute $\hat{E}\left(s(c_j)\right)$ by $s(c_j) - z\sigma\left(s(c_j)\right)$, and $\sigma\left(s(c_j)\right)$ by its expression in (6):

$$\hat{s}_0(c_i) = \sum_{j=1}^{k}\left(p_{ij}^{-1}\left(s(c_j) - z\left(\sum_{l=1}^{k} p_{jl}(1-p_{jl})s_0(c_l)\right)^{1/2}\right)\right). \tag{11}$$

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

$$\hat{s}_0(c_i) = \sum_{j=1}^{k}\left(p_{ij}^{-1}\left(s(c_j) - z\left(\sum_{l=1}^{k} p_{jl}(1-p_{jl})\hat{s}_0(c_l)\right)^{1/2}\right)\right). \tag{12}$$

List all equations like (12) for $\hat{s}_0(c_l) = \hat{s}_0(c_1)\ldots\hat{s}_0(c_k)$ and combine them into a matrix:

$$\begin{pmatrix}\hat{s}_0(c_1)\\ \vdots \\ \hat{s}_0(c_k)\end{pmatrix} = \mathbf{P}^{-1}\left(\begin{pmatrix}s(c_1)\\ \vdots \\ s(c_k)\end{pmatrix} - z\begin{pmatrix}\left((p_{11}(1-p_{11}))\left(\hat{s}_0(c_1)\right)+\ldots+(p_{1k}(1-p_{1k}))\left(\hat{s}_0(c_k)\right)\right)^{1/2}\\ \vdots \\ \left((p_{k1}(1-p_{k1}))\left(\hat{s}_0(c_1)\right)+\ldots+(p_{kk}(1-p_{kk}))\left(\hat{s}_0(c_k)\right)\right)^{1/2}\end{pmatrix}\right). \tag{13}$$

(13) includes k equations and should have a unique solution for its k unknowns $\hat{s}_0(c_1)\ldots\hat{s}_0(c_k)$. However, an exact solution of (13) is complicated and computationally uneconomic. When only one $\hat{s}_0(c_i)$ is needed, all equations in (13) have to be solved, and all of $\hat{s}_0(c_1)\ldots\hat{s}_0(c_k)$ will be obtained. In the real operation, $\hat{s}_0(c_l)$ on the right side of (12) can be approximated by the observed support $s(c_l)$:

$$\hat{s}_0(c_i) = \sum_{j=1}^{k}\left(p_{ij}^{-1}\left(s(c_j) - z\left(\sum_{l=1}^{k} p_{jl}(1-p_{jl})s(c_l)\right)^{1/2}\right)\right). \tag{14}$$

An analytic evaluation of the discrepancy between the $\hat{s}_0(c_l)$ values solved from (14) and (13) 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 $s_0\left(I \cup \{c_i\}\right)$, and the observed support is $s\left(I \cup \{c_i\}\right)$. Under the assumption in Sect. 3.1 that items are independent in their chances of error occurrence, (14) still holds if c_i is substituted by $I \cup \{c_i\}$. Denote the estimated true value of $s\left(I \cup c_i\right)$ with respect to $\mathbf{P}$ and z by $\hat{E}(c_i, I, \mathbf{P}, z)$:

$$\begin{aligned}\hat{E}(c_i, I, \mathbf{P}, z) &= \hat{s}_0(I \cup \{c_i\})\\ &= \sum_{j=1}^{k}\left(p_{ij}^{-1}\left(s\left(I \cup \{c_j\}\right) - z\left(\sum_{l=1}^{k} p_{jl}(1-p_{jl})s\left(I \cup \{c_l\}\right)\right)^{1/2}\right)\right).\end{aligned} \tag{15}$$

Equations (9)–(15) 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 $i = 1, \ldots, k$, as $\sum_{j=1}^{k} p_{ji} = 1$ (see Sect. 3.1). According to Sect. 3.1, 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 50 % or more. 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 (9)–(15) 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 test for productivity of a rule $X \rightarrow y$ on one of its items $x_m \in X$. Parameters a, b, c and d for the test, as defined in (2), can be rewritten as

$$\begin{aligned} a &= s(X \cup \{y\}) \\ b &= s(X) - s(X \cup \{y\}) \\ c &= s((X - \{x_m\}) \cup \{y\}) - s(X \cup \{y\}) \\ d &= s(X - \{x_m\}) - s(X) - s((X - \{x_m\}) \cup \{y\}) + s\,(X \cup \{y\}), \end{aligned} \tag{16}$$

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 $a \sim d$. Applying (15) to $a \sim d$ 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.

$\hat{a}_0 \sim \hat{d}_0$ should be less distorted by data error than $a \sim d$. Therefore, replacing $a \sim d$ with $\hat{a}_0 \sim \hat{d}_0$ may lead to more accurate p value, and recover some true rules lost due to data error. According to (3), 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 false discoveries, 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 $\hat{a}_0 \sim \hat{d}_0$ values, as listed in Table 1. $\hat{a}_0 \sim \hat{d}_0$ values also need to be rounded to the closest integers in order to be used in the Fisher exact test.

3.3 Controlling false discoveries

For the method in Sect. 3.2, the z value is the key to the increase of true discoveries as well as the risk of false discoveries. A smaller z value leads to larger corrections to the Fisher exact test parameters, higher potential to recover true discoveries lost due to data error, yet also higher risk of overcorrecting these parameters and eventually false discoveries.

Ideally, a quantitative relation shall be established between the z value and the risk of false discoveries, 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 Sect. 3.2, the z value only directly relates to the probability of overcorrecting each of the four Fisher exact test parameters. This probability then links to the probability of overcorrecting p value of the test, the risk of individual false discoveries, and finally the FWER. There are numerous uncertain factors in this

Table 1 Estimated true values of fisher exact test parameters with derivations

(**a**) Case 1:$c_i = x_m$	
Derivation	$a = s(X \cup \{y\})$ $= s((X - \{x_m\}) \cup \{x_m\} \cup \{y\})$ $= s((X - \{x_m\}) \cup \{y\} \cup \{c_i\})$
	$b = s(X) - s(X \cup \{y\})$ $= s((X - \{x_m\}) \cup \{x_m\}) - s((X - \{x_m\}) \cup \{x_m\}) \cup \{y\})$ $= s((X - \{x_m\}) \cup \{c_i\}) - s((X - \{x_m\}) \cup \{y\} \cup \{c_i\})$
	$a + c = s((X - \{x_m\}) \cup \{y\})$
	$b + d = (a + b + c + d) - (a + c)$ $= s(X - \{x_m\}) - s((X - \{x_m\}) \cup \{y\})$
Estimated true parameter values	$\hat{a}_0 = \hat{E}\,(c_i, (X - \{x_m\}) \cup \{y\}, \mathbf{P}, z)$
	$\hat{b}_0 = \hat{E}\,(c_i, X - \{x_m\}, \mathbf{P}, -z) - \hat{E}\,(c_i, (X - \{x_m\}) \cup \{y\}, \mathbf{P}, -z)$
	$\hat{c}_0 = a + c - \hat{a}_0$
	$\hat{d}_0 = b + d - \hat{b}_0$
(**b**) Case 2: $c_i = y$	
Derivation	$a = s(X \cup \{y\})$ $= s(X \cup \{c_i\})$
	$a + b = s(X)$
	$c = s((X - \{x_m\}) \cup \{y\}) - s(X \cup \{y\})$ $= s((X - \{x_m\}) \cup \{c_i\}) - s(X \cup \{c_i\})$
	$c + d = (a + b + c + d) - (a + b)$ $= s(X - \{x_m\}) - s(X)$
Estimated true parameter values	$\hat{a}_0 = \hat{E}\,(c_i, X, \mathbf{P}, z)$
	$\hat{b}_0 = a + b - \hat{a}_0$
	$\hat{c}_0 = \hat{E}\,(c_i, X - \{x_m\}, \mathbf{P}, -z) - \hat{E}\,(c_i, X, \mathbf{P}, -z)$
	$\hat{d}_0 = c + d - \hat{c}_0$
(**c**) Case 3: $c_i = x_e \in X - \{x_m\}$	
Derivation	$a = s(X \cup \{y\})$ $= s((X - \{x_e\}) \cup \{x_e\} \cup \{y\})$ $= s((X - \{x_e\}) \cup \{y\} \cup \{c_i\})$
	$b = s(X) - s(X \cup \{y\})$ $= s((X - \{x_e\}) \cup \{x_e\}) - s((X - \{x_e\}) \cup \{x_e\} \cup \{y\})$ $= s((X - \{x_e\}) \cup \{c_i\}) - s((X - \{x_e\}) \cup \{y\} \cup \{c_i\})$
	$c = s((X - \{x_m\}) \cup \{y\}) - s(X \cup \{y\})$ $= s((X - \{x_m\} - \{x_e\}) \cup \{x_e\} \cup \{y\}) - s((X - \{x_e\}) \cup \{x_e\} \cup \{y\})$ $= s((X - \{x_m\} - \{x_e\}) \cup \{y\} \cup \{c_i\}) - s((X - \{x_e\}) \cup \{y\} \cup \{c_i\})$
	$d = s(X - \{x_m\}) - s(X) - s((X - \{x_m\}) \cup \{y\}) + s(X \cup \{y\})$ $= s((X - \{x_m\} - \{x_e\}) \cup \{x_e\}) - s((X - \{x_e\}) \cup \{x_e\})$ $- s((X - \{x_m\} - \{x_e\}) \cup \{x_e\} \cup \{y\}) + s((X - \{x_e\}) \cup \{x_e\} \cup \{y\})$ $= s((X - \{x_m\} - \{x_e\}) \cup \{c_i\}) - s((X - \{x_e\}) \cup \{c_i\})$ $- s((X - \{x_m\} - \{x_e\}) \cup \{y\} \cup \{c_i\}) + s((X - \{x_e\}) \cup \{y\} \cup \{c_i\})$

Table 1 continued

Estimated true parameter values	$\hat{a}_0 = \hat{E}\left(c_i, (X - \{x_e\}) \cup \{y\}, \mathbf{P}, z\right)$
	$\hat{b}_0 = \hat{E}\left(c_i, X - \{x_e\}, \mathbf{P}, -z\right) - \hat{E}\left(c_i, (X - \{x_e\}) \cup \{y\}, \mathbf{P}, -z\right)$
	$\hat{c}_0 = \hat{E}\left(c_i, (X - \{x_m\} - \{x_e\}) \cup \{y\}, \mathbf{P}, -z\right) - \hat{E}\left(c_i, (X - \{x_e\}) \cup \{y\}, \mathbf{P}, -z\right)$
	$\hat{d}_0 = \hat{E}\left(c_i, X - \{x_m\} - \{x_e\}, \mathbf{P}, z\right) - \hat{E}\left(c_i, X - \{x_e\}, \mathbf{P}, z\right)$ $-\hat{E}\left(c_i, (X - \{x_m\} - \{x_e\}) \cup \{y\}, \mathbf{P}, z\right) + \hat{E}\left(c_i, (X - \{x_e\}) \cup \{y\}, \mathbf{P}, z\right)$

multi-step relation, for example, the value of each element in **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 from these uncertainties on the relation between z value and the FWER. If any factor is modeled 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 these 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 false discoveries. 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 occurs, the rules are mostly weaker and thus have smaller supports than rules in the actual data to be explored. As a is equal to the support of the rule, 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 (2). 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 false discoveries.

To address this issue, the simulation also recorded a *correctable range* for each test parameter, defined as the range between the maximum and minimum percentage changes in the parameter due to the correction in the simulation. The percentage changes can be positive or negative. 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 discoveries 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

$$\begin{aligned}\Pr(K \leq 1) &= \Pr(K = 0) + \Pr(K = 1)\\ &= C_n^0 r_{\max}^0 (1 - r_{\max})^{n-0} + C_n^1 r_{\max}^1 (1 - r_{\max})^{n-1}\\ &= (1 - r_{\max})^n + n r_{\max} (1 - r_{\max})^{n-1}. \end{aligned} \tag{17}$$

As reducing z value by a minimum enumeration step will lead to false discoveries, 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 the simulation is optimally used together with the statistically sound test which 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 false discoveries. Also, the two techniques should aim at achieving the same user specified maximum FWER (α or $r_{\max}$).

4 Experiment with synthetic data

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 early experiments of this study. The experiments 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 discoveries, hence confirming the need for the corrected test; and secondly evaluated the corrected test in terms of recovering true discoveries and controlling false discoveries. 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, or known true rules, thus the true and false discoveries 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 provides less confidence in evaluating the correctness of resultant rules, as true rules behind real data are rarely known.

4.1 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 classes with values from 0 to 4. The other seven were binary attributes.

In every record, the value of each attribute was assigned at random but 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 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

Table 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

factors establishes a new association with other factors. $x_0 \sim x_3$ were not in any predesigned data associations, and they simulated the numerous "noise" attributes irrelevant to interested associations in practical data mining.

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 $s = 10, 248$, and the statistically sound significance level was $\kappa = 4.88 \times 10^{-6}$. When *maxL* = 4, $s = 27, 608$ 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 statistical test was not expected to discover all predesigned rules; the larger the data size, the more likely the rules would pass the test. 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.

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

Table 3 Numbers of true discoveries from “ideal” data

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

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, each element in $\mathbf{P}(att_0)$ and $\mathbf{P}(att_3)$ was equal to the actual error probability between the 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 Sect. 6). 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 were assigned large inaccuracies in data 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 behaviors between data items (see Sect. 3.1). 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 $att_0 = 0 \sim 2$ and $att_0 = 3 \sim 4$, respectively.

In these four groups, att_3 had an aggregate error level of 10 %, and both attributes had perceived error levels of 10 %. The 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.

4.2 Results

Among resultant rules accepted by the statistical tests, a rule was a true discovery if it was also in the 61 predesigned productive rules, and was a false discovery if not. Figure 2 plotted the true discoveries, false discoveries and FWER for the E and R treatments. The corresponding numerical results are listed in Table 8, “Appendix 2”. 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 are listed later in Sect. 4.2. In the results, the number of rules counted only the 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 Sect. 4.1, the Original treatments produced a reference rule set for each data size in every dataset. The numbers of true discoveries containing att_0 and/or att_3 in each reference rule set, also reported in the “Original” rows in Table 8, were taken as 100 % for computing percentages of rules in Fig. 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 8, 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

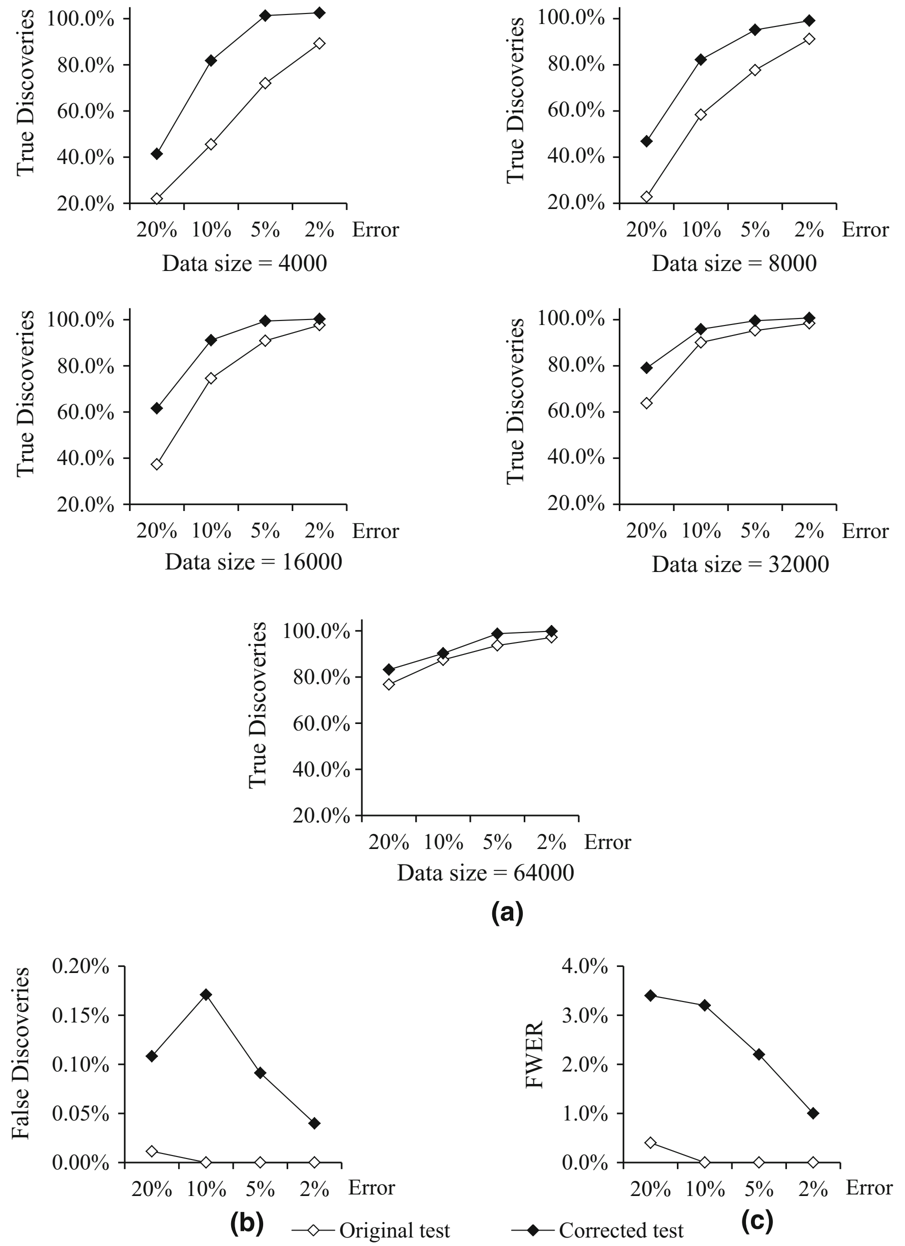


Fig. 2 Synthetic data experiment results. **a** True discoveries, **b** False discoveries, **c** FWER respect to total number of runs

Fig. 2, all results refer to averages for the two *maxL* values. In Fig. 2, percentages of true discoveries varied significantly with data sizes and thus are plotted for each data size; the false discoveries in different data sizes were relatively stable and thus aggregated. The essential numerical results are also summarized in Table 4.

Table 4 Summary of synthetic data experiment results

	Original data[a]	Original test (E treatments)	Corrected test (R treatments)
No. of true discoveries	1404.2	1100.6	1253.1
No. of false discoveries	0	0.04	1.44
% of false discoveries	0 %	0.01 %	0.10 %
FWER	0 %	0.10 %	2.50 %

[a] Values are (result in Original treatments) ×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 providing more runs in the experiment (Webb 2007). The numerous false discoveries in absence of, and the minimal false discovery rate 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 study.

When applied to erroneous data, the original test maintained strict control on false discoveries. With the maximum FWER set at 5 %, the actual FWER was only 0.1 %, and the percentage of false discoveries was 0.01 % (Table 4). That is, in terms of controlling false discoveries, the statistically sound test was robust to distortions posed by the data error to all patterns containing att_0 and att_3. Apparently, rule interestingness measure values computed from distorted pattern supports should also distort and cause the false discoveries to increase. However, under the assumption that the probability of error occurrence on each attribute was independent from values of other attributes (see Sect. 3.1), such error mostly distorts in proportion the supports of a rule and its sub-patterns containing the erroneous attributes. The proportional distortions largely cancel out each other when these support values are used together to compute rule interestingness. 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 when computing $confidence\,(X \rightarrow y) = support(X \cup y)/support(X)$. Similarly, error in y changes $confidence\,(X \rightarrow y)$ and $confidence\left(X - \underset{x \in X}{\{x\}} \rightarrow y\right)$ in rough proportion, while maintaining their differences. In both cases, the productivity of $X \rightarrow y$ with respect to its generalization $X - \underset{x \in X}{\{x\}} \rightarrow y$ would not be much affected, nor would relevant false discoveries be generated.

However, the data error did cause marked loss of true discoveries. The loss worsened with higher error levels and smaller data sizes. In E20 with data sizes 4000 and 8000, almost 80 % true discoveries lost (Fig. 2a) and only 3–6 rules were preserved (Table 8a). Such few true discoveries hardly made up a meaningful resultant rule set. While 90 %

Table 5 True discovery increases and recovery rates by error level

	R20	R10	R05	R02
z	0.78	0.27	0.06	0.01
True discovery 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 %			

data accuracy is satisfactory for many applications, the true discovery loss was still prominent in E10 and up to 50 % with small data sizes. Thus the original test may not obtain enough true discoveries 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 discoveries than the original test in E treatments for all error levels and data sizes (Fig. 2a). The true discovery increase was more significant when the true discovery loss in the original test was severer. For medium error levels and data sizes, true discovery rates raise from 60 to 70 % with the original test to 80–90 % with the corrected test.

The increase of true discoveries can be standardized by their loss in the original test into a *recovery rate*:

$$\text{recovery rate} = \frac{\text{No. of SR/DR true discoveries} - \text{No. of SE/DE true discoveries}}{\text{No. of reference rules } - \text{No. of SE/DE true discoveries}} \times 100\,\%. \tag{18}$$

The true discovery increases and recovery rates for various error levels are listed in Table 5. With smaller data error, the true discovery 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 discoveries, or the loss in value of resultant rules.

While the data error mostly led to loss of true discoveries, productive rules that were not discovered in Original treatments were in fact occasionally discovered in E and R treatments. We call such true discoveries "gained" from erroneous data *serendipitous discoveries*. These rules are favourable but contrary to the expectation that random data error should cause loss of true discoveries. Still, serendipitous discoveries do result from the random nature of data error. As explained in Sect. 3.1, 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 $\Phi\left(\left(E(s(S)) - s_0(S)\right)/\sigma(s(S))\right)$ 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 past 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 Sect. 4.1 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 the number of reference rules (Table 5). However, the increases were actually sharp and around 2–10 times of the number in E treatments (Table 8b). 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 5, 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 then the original test. At low data error levels, as the true discovery increase dropped due to reduced room of improvement, the serendipitous discoveries became more important, and took as many as 50 % of the true discovery increase in R02. This is also the reason of 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 discoveries.

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 false discoveries. Then z value had to be larger to control the FWER, as can be seen in Table 5. The large z value also limited the true discoveries, 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 Sect. 3.3, when $z = 0$ the potential of the corrected test in increasing true discoveries 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 false discoveries below 0.2 % (Fig. 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 4), indicating that on 97.5 % occasions there were not any false discoveries. Computed from Table 4, the ratio between true and false discovery increases was about 109:1. Users would obtain 109 more true discoveries

Table 6 Synthetic data experiment results with inaccurate error specifications or dependent data error

	R20	R20–/–	R20+/–	R10	R10+/+	R10+/–	R10_ ErrDep	R10_ ValDep
(**a**) Results of corrected test								
Recovery rate	34.1 %	26.5 %	20.3 %	55.8 %	70.4 %	42.8 %	65.3 %	38.3 %
% of false discoveries	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
	E20	E10	E10_ErrDep	E10_ValDep				
(**b**) True discoveries in original test								
True discovery rate	51.4 %	77.0 %	79.3 %	70.5 %				

at the risk of one more false discovery. Statistically unsound tests on association rules usually had nearly 100 % FWER, as stated in Sect. 1, and percentages of false discoveries were often high as well. Compared with unsound tests, the corrected test can be regarded to have essentially equal advantage in false discovery control to the original 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 6a. 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 discoveries, 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 discoveries in E20, the 20.3 % increase of true discoveries was still significant.

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 6a). This again enlarged the corrections to the parameters, as explained in Sect. 3.3. R20+/– lost more recovery rate than R20–/– as its z value had no obvious reduce compared to R20 (Table 6a). 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 Sect. 4.1, 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 discoveries than E10 (Table 6b). 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 discovery rate in the original test (E10_ValDep) than E10 (Table 6b). According to Table 2, 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 $att_0 = 3 \sim 4$, 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 false discoveries in R10_ErrDep and R10_ValDep. The robustness of the corrected test in controlling false discoveries is also expected, as the FWER was controlled by the simulation which worked regardless of the efficacy in increasing true discoveries.

5 Experiment on real-world data: Mining spatiotemporal association rules

This experiment investigates how the corrected statistical test improves the value of real-world association rule mining results. The case study targeted at spatiotemporal association rules between land use and socioeconomic changes in Massachusetts, the U.S., in 1985 to 1999, using data in a geographical information system (GIS). Previous data analyses, including some association rule mining 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 association rule mining 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.

5.1 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 GIS 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 7. Figure 3 shows an overview of the study area and land uses in 1985 and 1999 of a small locality within the area. Clearly, the locality experienced land use transformations towards urban area. Such a geographic region can be suitable for investigating associations between urbanization and socioeconomic changes.

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.

Table 7 Land use classes of study area

Urban land uses (13 classes)	Non-urban land uses (9 classes)
Participation recreation	Forest
Spectator recreation	Cropland
Water based recreation	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
Urban open area	
Transportation	
Waste disposal	

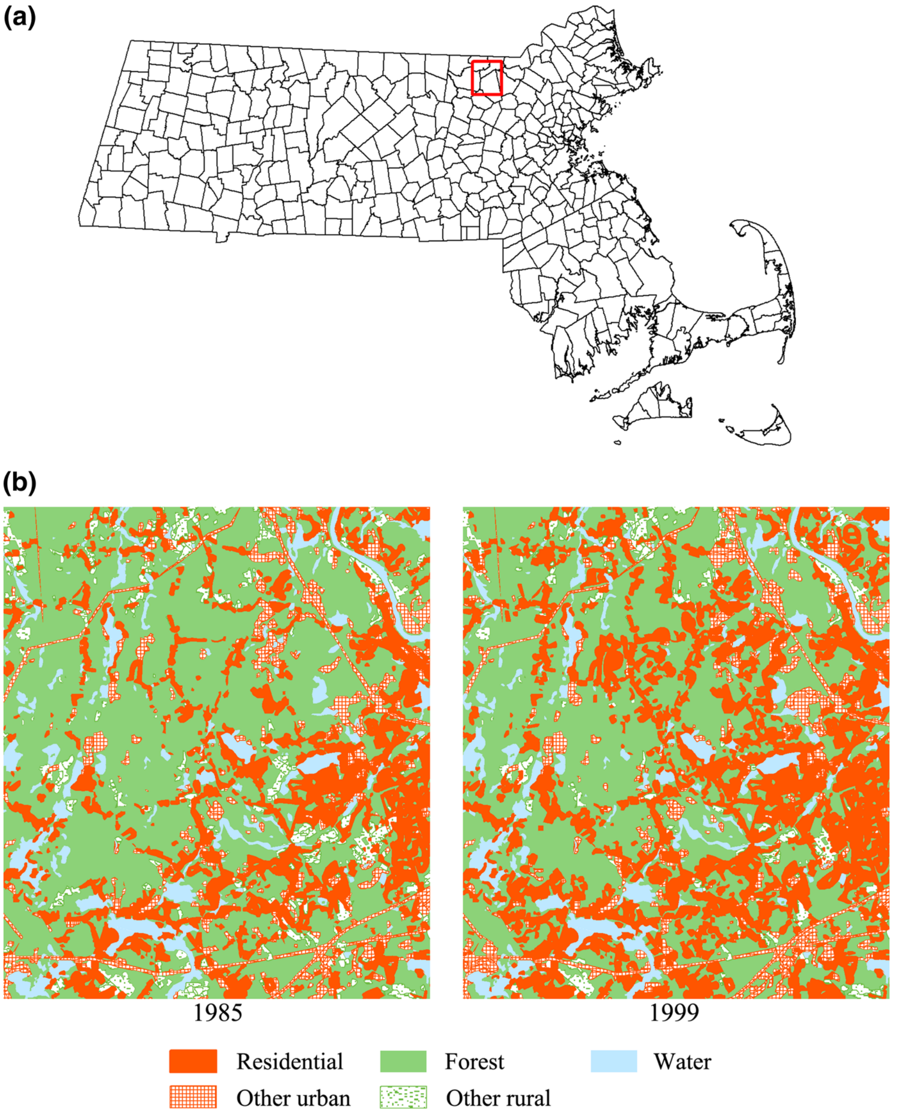


Fig. 3 **a** Overview of Massachusetts with town boundaries. **b** Land uses of a small locality. The locality is marked by rectangular box in (**a**)

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 to 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 polygon was further split into several small ones with areas as close as possible to 10,000 m^2. 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 2 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 false discoveries, 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 10,000 productive rules with at most four items in the antecedents in the Original treatment was determined. The value was 7.06×10^6 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 false discoveries. 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.

5.2 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 false discoveries, 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 false discoveries. 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.

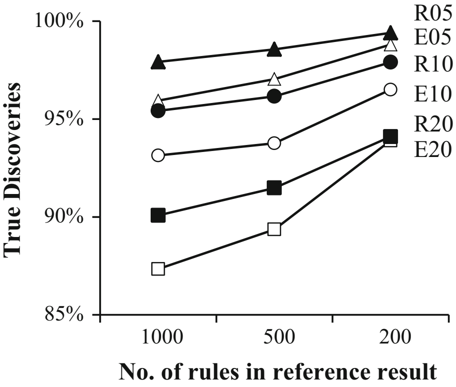
As commonly in practical data mining, besides the statistical test, more basic rule interestingness measures were needed to trim resultant rules to the amount that human users could consider. Here the leverage (Piatetsky-Shapiro 1991) measure was used:

$$leverage(X \rightarrow y) = support(X \rightarrow y) - support(X)support(y). \tag{19}$$

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

Fig. 4 Recovery of true discoveries by corrected test in real-world data experiment



regarded as true discoveries if they were also among the reference results. According to Sect. 4.2, 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 Sect. 4.2 also suggests that serendipitous discoveries take small fractions of true discoveries, and their absence should only cause slight underestimation on the merit of the corrected test, rather than overestimation.

Figure 4 shows true discoveries 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 discoveries, and recovered 20–50 % of true discovery loss in the original test. Compared with the synthetic data experiment, the real data experiment had less significant true discovery 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 Sect. 4.2, more sufficient data leads to more true discoveries in the test. Recovery rate of the corrected test might also be underestimated due to the exclusion of serendipitous discoveries.

It may be argued that over 85 % true rules discovered by the original test (Fig. 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 discoveries involving land use changes were plotted in Fig. 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

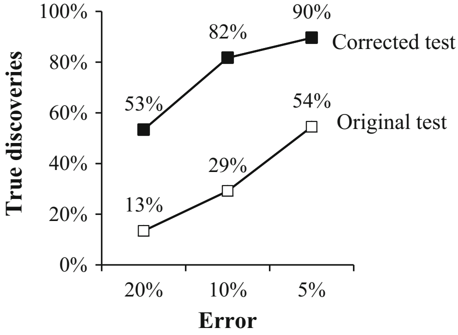


Fig. 5 Recovery of true discoveries involving land use changes in real data experiment

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 discoveries 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 discoveries as the original test.

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

- Land use changed from Forest to Residential, > 1/2 acre lots ^ 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 practise 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 association rule mining 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 discoveries at 10 % error level, far above the 54 % true discoveries 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.

6 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 proposes 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 Sect. 4.2, 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, maching-ing 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 discoveries with underestimated error probabilities (see Sect. 4.2). 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).

7 Conclusions and further research

This article presents a novel method for testing statistical significance of association rules with uncertain data. The method aims at improving the reliability of association 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. An original mathematical model was established to describe the propagation of data error through computational processes in the statistical test on association rules, and finally to the test result. Based on this model, techniques were developed to recover true rules via correcting the test computation for impact of data error, as well as to control the risk of spurious rules.

When assessed with synthetic data, the new method recovered around half of the true rules lost due to data error with accurate error probability information. It also largely maintained its ability for recovering true rules with inaccurate error probability specifications and dependences among the error and attribute values. Meanwhile, the new method maintained superior control on spurious rules by existing statistically sound technique, and achieved a familywise error rate (FWER), or the risk that the result included any spurious rules, of below 5 %. When experimented with real-world data, the new method discovered several times as many as by the existing technique those most practically valuable rules containing temporal changes in data.

Despite of its efficacy, the new method needs to be matured by further developments. Existing statistically sound test for determinate data (Webb 2007) can strictly cap the FWER of resultant rules at arbitaray user specified upper limit. The new method for uncertain data maintains statistical soundness at all stages except for the final simulation to determine the z value which controls the conservativeness of correction to the test parameters (see Sect. 3.3). 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. As the z value is determined by average instead of maximum risk, even the FWER was successfully controlled below the user specified maximum of 5 % in the synthetic data experiment, it is not theoretically guaranteed below arbitrary user specified levels. On rare occasions, the simulation produced rather small z values, which led to over-corrections to the test and multiple false discoveries. Improvement is needed to stabilize the determined z value, even to make the simulation statistically sound, so as to confidently fulfill user requirement on the maximum FWER, while keeping strong ability to recover lost true rules.

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

Acknowledgments We wish to thank the action editor Mohammed J. Zaki and the anonymous reviewers for their insightful comments which have helped very much on improving the manuscript. This study is partially supported by the Special Fund for Technological Leading Talents, National Administration of Surveying, Mapping and Geoinformation, P.R. China.

Appendix 1: Evaluating discrepancy between exact and approximate $\hat{s}_0(c_l)$ values

Let $f\left(x_1,\ldots,x_k\right) = \sum_{j=1}^{k} p_{ij}^{-1} z \left(\sum_{l=1}^{k} p_{jl}(1-p_{jl})x_l\right)^{1/2}$, then the discrepancy between the exact solution to $\hat{s}_0(c_i)$ from (13) and the approximate solution from (14) is:

$$\sum_{j=1}^{k}\left(p_{ij}^{-1}z\left(\left(\sum_{l=1}^{k}p_{jl}(1-p_{jl})s(c_l)\right)^{1/2}-\left(\sum_{l=1}^{k}p_{jl}(1-p_{jl})\hat{s}_0(c_l)\right)^{1/2}\right)\right)$$
$$= f\left(s(c_1),\ldots,s(c_1)\right) - f\left(\hat{s}_0(c_1),\ldots,\hat{s}_0(c_k)\right), \tag{20}$$

and

$$\frac{\partial f\left(x_1,\ldots,x_k\right)}{\partial x_l} = \frac{z}{2}\sum_{j=1}^{k} p_{ij}^{-1}\cdot\left[p_{jl}\left(1-p_{jl}\right)\right]^{1/2}\cdot x_l^{-1/2}. \tag{21}$$

Let $\Delta_l = s\left(c_l\right) - \hat{s}_0\left(c_l\right)$, then (a1) may be estimated by the first degree Taylor polynomial of $f\left(s(c_1),\ldots,s(c_1)\right)$:

$$\left(\Delta_1\frac{\partial}{\partial\hat{s}_0(c_1)}+\cdots+\Delta_k\frac{\partial}{\partial\hat{s}_0(c_k)}\right)f\left(\hat{s}_0(c_1),\cdots,\hat{s}_0(c_k)\right)$$
$$= \frac{z}{2}\sum_{j=1}^{k}\sum_{l=1}^{k}p_{ij}^{-1}\cdot\left[p_{jl}\left(1-p_{jl}\right)\right]^{1/2}\cdot\hat{s}_0(c_l)^{-1/2}\cdot\Delta_l. \tag{22}$$

The error of estimating (20) by (22) is the Lagrange form of the remainder of the first degree Taylor polynomial:

$$R_1\left(\hat{s}_0(c_1),\cdots,\hat{s}_0(c_k)\right) = \frac{1}{2!}\left(\Delta_1\frac{\partial}{\partial\hat{s}_0(c_1)}+\cdots+\Delta_k\frac{\partial}{\partial\hat{s}_0(c_k)}\right)^2 f$$
$$\times\left(\hat{s}_0(c_1)+\theta\Delta_1,\cdots,\hat{s}_0(c_k)+\theta\Delta_k\right)$$
$$= -\frac{z}{8}\sum_{j=1}^{k}\sum_{l=1}^{k}p_{ij}^{-1}\cdot\left[p_{jl}\left(1-p_{jl}\right)\right]^{1/2}$$
$$\cdot\left(\hat{s}_0(c_l)+\theta\Delta_l\right)^{-3/2}\cdot\Delta_l^2, \tag{23}$$

where $0 \le \theta \le 1$. Each item in (23) with the same (j,l) value pair is equal to $-\left[\left(\hat{s}_0(c_l)\right)^{1/2}\Delta_l\right]/\left[4\left(\hat{s}_0(c_l)+\theta\Delta_l\right)^{3/2}\right]$ times the corresponding item in (22). Typically $\hat{s}_0\left(c_l\right)$ is much larger than Δ_l, thus (22) is much larger than (23) and should be a reasonable estimator of (20).

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 (22) for evaluating the discrepancy. An exemplary evaluation was made with the item "$att_3 = 1$" in the synthetic experiment (see Sect. 4.1), one of the most affected items by the discrepancy. The computation used the “ideal” data defined in Sect. 4.1 as the original data, and the average z value actually used in the experiment at each error level in Table 5. 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.

Appendix 2: Numerical synthetic data experiment results

Table 8 Numerical synthetic data experiment results for E and R treatments

Data size	4000		8000		16,000		32,000		64,000	
maxL	3	4	3	4	3	4	3	4	3	4
(**a**) True discoveries										
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										
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**) False discoveries										
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

References

Aggarwal CC, Li Y, Wang J, Wang J (2009) Frequent pattern mining with uncertain data. In: Proceedings of 17th international conference on knowledge discovery and data mining (KDD 2009), pp 29–38

Agrawal R, Imielinski T, Swami A (1993) Mining associations between sets of items in massive databases. In: Proceedings of 1993 ACM-SIGMOD international conference on management of data, pp 207–216

Agresti A (1992) A survey of exact inference for contingency tables. Stat Sci 7(1):131–153

Bastide Y, Pasquier N, Taouil R, Stumme G, Lakhal L (2000) Mining minimal non-redundant association rules using frequent closed itemsets. In: Proceedings of first international conference on computational logic, pp 972–986

Bay SD, Pazzani MJ (2001) Detecting group differences: mining contrast sets. Data Min Knowl Disc 5(3):213–246

Bayardo RJ Jr, Agrawal R, Gunopulos D (2000) Constraint-based rule mining in large, dense databases. Data Min Knowl Disc 4(2/3):217–240

Ben-Israel A, Greville TNE (2003) Generalized inverses: theory and applications. Springer, New York

Bishop G (2009) Assessing the likely quality of the statistical longitudinal census dataset. Research paper, Australian Bureau of Statistics

Brin S, Motwani R, Silverstein C (1997) Beyond market baskets: generalizing association rules to correlations. In: SIGMOD 1997, proceedings ACM SIGMOD international conference on management of data, pp 265–276

Calders T, Garboni C, Goethals B (2010) Approximation of frequentness probability of itemsets in uncertain data. In: Proceedings of IEEE international conference on data mining (ICDM 2010), pp 749–754

Carvalho JV, Ruiz DD (2013) Discovering frequent itemsets on uncertain data: a systematic review. In: Proceedings of 9th international conference on machine learning and data mining, pp 390–404

Chui CK, Kao B (2008) A decremental approach for mining frequent itemsets from uncertain data. In: Proceedings of 12th Pacific-Asia conference on knowledge discovery and data mining (PAKDD 2008), pp 64–75

Chui CK, Kao B, Hung E (2007) Mining frequent itemsets from uncertain data. In: Proceedings of 11th Pacific-Asia conference on knowledge discovery and data mining (PAKDD 2007), pp 47–58

Foody GM (2002) Status of land cover classification accuracy assessment. Remote Sens Environ 80:185–201

Fosu GB (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. http://unstats.un.org/unsd/demographic/meetings/egm/symposium2001/docs/symposium_11.htm#_Toc7406238. Accessed 22 July 2015

Gray B, Orlowska M (1998) CCAIIA: clustering categorical attributes into interesting association rules. In: Proceedings of 2nd Pacific-Asia conference on knowledge discovery and data mining (PAKDD'98), pp 132–143

Hollister JW, Gonzalez ML, Paul JF, August PV, Copeland JL (2004) Assessing the accuracy of National Land Cover Dataset area estimates at multiple spatial extents. Photogramm Eng Remote Sensing 70:405–414

International Business Machines (1996) IBM intelligent miner user's guide, version 1, release 1

Jones N, Lewis D (eds, with Aitken A, Hörngren J, Zilhão MJ) (2003) Handbook on improving quality by analysis of process variables. Final report, Eurostat

Mennis J, Liu JW (2005) Mining association rules in spatio-temporal data: an analysis of urban socioeconomic and land cover change. Trans GIS 9(1):5–17

McDonald JH (2014) Handbook of biological statistics, 3rd edn. Sparky House Publishing, Baltimore

Shaffer JP (1995) Multiple hypothesis testing. Annu Rev Psychol 46:561–584

Liu B, Hsu W, Ma Y (1999) Pruning and summarizing the discovered associations. In: Proceedings of 5th ACM SIGKDD international conference on knowledge discovery and data mining (KDD '99), pp 125–134

Liu B, Hsu W, Ma Y (2001) Identifying non-actionable association rules. In: Proceedings of the 7th ACM SIGKDD international conference on knowledge discovery and data mining (KDD'01), pp 329–334

Megiddo N, Srikant R (1998) Discovering predictive association rules. In: Proceedings of 4th international conference on knowledge discovery and data mining (KDD '98), pp 27–78

Office for National Statistics, The United Kingdom (2014) 2011 Census quality survey. 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 CE (2008) Is 80% accuracy good enough? In: Proceedings of 17th William T. pecora memorial remote sensing symposium. http://www.asprs.org/a/publications/proceedings/pecora17/0026.pdf. Accessed 27 Feb 2014

Piatetsky-Shapiro G (1991) Discovery, analysis, and presentation of strong rules. In: Piatetsky-Shapiro G, Frawley J (eds) Knowledge discovery in databases. AAAI/MIT Press, Menlo Park, pp 229–248

Penrose R (1955) A generalized inverse for matrices. Math Proc Cambridge Philos 51:406–413

Rao CR, Mitra SK (1972) Generalized inverse of a matrix and its applications. In: Proceedings of the sixth Berkeley symposium on mathematical statistics and probability, volume 1: theory of statistics, pp 601–620

Smith JH, Stehman SV, Wickham JD, Yang L (2003) Effects of landscape characteristics on land-cover class accuracy. Remote Sens Environ 84:342–349

Srikant R, Agrawal R (1995) Mining generalized association rules. In: Proceedings of 21st international conference on very large data bases, pp 407–419

Stehman SV, Wickham JD, Wade TG, Smith JH (2008) Designing a multi-objective, multi-support accuracy assessment of the 2001 National Land Cover Data (NLCD 2001) of the conterminous United States. Photogramm Eng Remote Sensing 74:1561–1571

Sun L, Cheng R, Cheung DW, Cheng J (2010) Mining uncertain data with probabilistic guarantees. In: Proceedings of 17th international conference on knowledge discovery and data mining (KDD 2010), pp 273–282

Taussky O (1949) A recurring theorem on determinants. Am Math Mon 56(10):672–676

The Executive Office for Administration and Finance, Commonwealth of Massachusetts (2012) MassGIS datalayers. http://www.mass.gov/anf/research-and-tech/it-serv-and-support/application-serv/office-of-geographic-information-massgis/datalayers/layerlist.html. Accessed 26 Sept 2013

Ting KM (2011) Confusion matrix. In: Sammut C, Webb GI (eds) Encyclopedia of machine learning, 1st edn. Springer, New York

Tong Y, Chen L, Ding B (2012) Discovering threshold-based frequent closed itemsets over probabilistic data. In: Proceedings of 28th international conference on data engineering, pp 270–281

Webb GI (2007) Discovering significant patterns. Mach Learn 68:1–33

Webb GI, Zhang S (2005) *K*-optimal rule discovery. Data Min Knowl Disc 10(1):39–79

Yang L, Stehman SV, Smith JH, Wickham JD (2001) Thematic accuracy of MRLC land cover for eastern United States. Remote Sens Environ 76:418–422

Zaki MJ (2000) Generating non-redundant association rules. In: Proceedings of 6th ACM SIGKDD international conference on knowledge discovery and data mining (KDD-2000), pp 34–43

Zhang H, Padmanabhan B, Tuzhilin A (2004) On the discovery of significant statistical quantitative rules. In: Proceedings of 10th international conference on knowledge discovery and data mining (KDD 2004), pp 374–383

Zhu XQ, Wu XD (2006) Error awareness data mining. In: 2006 IEEE international conference on granular computing, pp 269–274

Zhu XQ, Wu XD, Yang Y (2004) Error detection and impact-sensitive instance ranking in noisy datasets. In: Proceedings of 19th national conference on artificial intelligence, pp 378–383