Estimation of missing precipitation records integrating surface interpolation techniques and spatio-temporal association rules

Ramesh S. V. Teegavarapu

ABSTRACT

Deterministic and stochastic weighting methods are the most frequently used methods for estimating missing rainfall values. These methods may not always provide accurate estimates due to their inability to completely characterize the spatial and temporal variability of rainfall. A new association rule mining (ARM) based spatial interpolation approach is proposed, developed and investigated in the current study to estimate missing precipitation values at a gauging station. As an integrated approach this methodology combines the power of data mining techniques and spatial interpolation approaches. Data mining concepts are used to extract and formulate rules based on spatial and temporal associations among observed precipitation data series. The rules are then used to improve the precipitation estimates obtained from spatial interpolation methods. A stochastic spatial interpolation technique and three deterministic weighting methods are used as interpolation methods in the current study. Historical daily precipitation data obtained from 15 rain gauging stations from a temperate climatic region (Kentucky, USA) are used to test this approach and derive conclusions about its efficacy for estimating missing precipitation data. Results suggest that the use of association rule mining in conjunction with a spatial interpolation technique can improve the precipitation estimates.

Key words | association rule mining, data mining, deterministic interpolation, missing precipitation data, ordinary kriging, spatial interpolation

Ramesh S. V. Teegavarapu
Department of Civil Engineering,
Florida Atlantic University,
777 Glades Road,
Boca Raton,
FL 33431-0991,
USA
Tel.: +1 561 297 3444
Fax: +1 561 297 0493
E-mail: *ramesh@civil.fau.edu*

INTRODUCTION

Deterministic weighting and stochastic interpolation methods (Wei & McGuinness 1973; Simanton & Osborn 1980; Tung 1983; Krajewski 1987; ASCE 1996; Vieux 2001) have been used in the past for spatial construction of rainfall fields or estimation of missing rainfall data at a point in space. Traditional weighting and data-driven methods are generally used for estimating missing precipitation. Weighting methods belong to a class of spatial interpolation techniques such as inverse-distance (Wei & McGuinness 1973), nonlinear deterministic and stochastic interpolation methods (e.g. kriging). Regression and time series analysis methods belong to data-driven approaches. The *Handbook of Hydrology* (ASCE 1996) recommends two methods for estimation of missing data. These methods are normal-ratio and inverse-distance weighting methods. Singh & Chowdhury (1986) compared thirteen rainfall estimation methods and found isohyetal method yielded higher estimates of mean daily, monthly aerial rainfall than other methods in the area of their study.

Tung (1983) compared five methods used for estimating point rainfall and indicated that arithmetic average and inverse-distance methods did not yield desirable results for mountainous regions. Variance-dependent surface interpolation methods that belong to the general family of kriging

doi: 10.2166/hydro.2009.009

have been applied for several geophysical interpolation problems in hydrology (Grayson & Bloschl 2001; Vieux 2001). These stochastic interpolation methods are based on the principle of minimizing estimation variances at points where measurements are not available. Kriging in various forms is applied for estimation of missing precipitation data and aerial precipitation from point measurements (Vieux 2001; Dingman 2002). Ashraf *et al.* (1997) compared interpolation methods (kriging, inverse distance and co-kriging) to estimate missing values of precipitation. They indicate that the kriging interpolation method provided the lowest root mean square error (RMSE). However, kriging methods are plagued by several limitations. Selection of semivariogram model, assignment of arbitrary values to sill and nugget parameters, and distance intervals, and the computational burden involved in interpolation of surfaces, are a few difficulties associated with this method.

Regression and time series models (Salas 1993) were used in the past for estimation of missing rainfall data. Global interpolation methods that use trend surface analysis and regression (Wang 2006) provide several advantages compared to deterministic weighting techniques. However, selection of the appropriate functional form to model the trend poses a major problem in trend surface analysis as there is an enormous range of candidate functions (Sullivan & Unwin 2003). Local interpolation methods such as thin-plate splines tend to generate steep gradients in data-poor areas and errors in the estimation process are compounded (Chang 2004).

Several limitations of spatial interpolation methods were reported by many researchers in many recent research studies. Vieux (2001) pointed out several limitations of the inverse-distance weighting method (IDWM), a major one being the "tent pole effect" that leads to higher estimates closer to the point of interest in space. Grayson & Bloschl (2001) list several limitations of Thiessen polygon and inverse-distance methods. They suggest that these methods are not recommended for spatial interpolation, considering the limitations. However, they recommend the thin-splines method and kriging for interpolation of hydrologic variables. The Thiessen polygon approach has a major limitation of not providing a continuous field of estimates if used for spatial interpolation (Sullivan & Unwin 2003). Brimicombie (2003) indicated that the main point of contention in application of IDWM for spatial interpolation is the selection of the number and relevant observation points for spatial interpolation at a point.

All these issues associated with spatial interpolation techniques may lead to under- or over-estimation of rainfall magnitudes at a gauge based on observations at all other gauges. Burrough & McDonnell (1998) reviewed several spatial interpolation techniques and concluded that geostatistical methods are superior to all other methods. Eischeid *et al.* (2000) report several interpolation methods for estimation of missing daily temperature and precipitation records and discuss their limitations. Underestimation of precipitation resulting from equating trace amount of precipitation to a zero value leads to significant errors in regional water balance assessments (Dingman 2002). Brown *et al.* (1968) reported that trace amounts of precipitation accounted for 10% of the summer precipitation on average for a region in Alaska that might be equal to one-third of the total precipitation observed over a few years. Underestimation of precipitation in minute amounts still leads to significant errors in water balance studies or in hydrologic modeling. Dingman *et al.* (1980) and Yang *et al.* (1998) suggest that corrections have to be made to account for underestimation associated with trace amounts of precipitation.

Overestimation of precipitation amounts by spatial interpolation techniques is not unusual. Traditional deterministic spatial interpolation techniques (e.g. distance-based weighting method) and stochastic techniques such as isotropic kriging do not consider the spatial variability of the precipitation patterns. In general, distance-based weighting methods suffer from one major conceptual limitation based on the fact that Eucledean distance is not always a definitive measure of the correlation among spatial point measurements. This also negates Tobler's first law of geography (Tobler 1970): "everything is related to everything else, but near things are more related than distant things", which forms the basis for many interpolation techniques. Also, the interpolation methods fail to estimate missing values correctly, if errors are introduced into the measurement process of rainfall at one or more rainfall stations. These are artifacts of interpolation techniques that cannot be avoided or eliminated all together in many situations.

Teegavarapu & Chandramouli (2005) reported several limitations and advantages of deterministic and stochastic spatial interpolation techniques when missing precipitation data is estimated at a base station (i.e. station with missing data) using data at all other stations. They indicate that all interpolation techniques fail in estimation of missing precipitation data at a point in space in two situations: (1) when precipitation is measured at all or a few other stations and no precipitation occurred in reality at the base station; and (2) when precipitation is measured at the base station and no precipitation is measured or occurred at all the other stations. In case 1, all spatial interpolation techniques provide a positive value of estimate while in reality a zero value of precipitation is recorded at the base station. It is impossible to estimate missing precipitation data in the second case as the point observations are used to estimate the missing value at the base station by using spatial interpolation algorithms alone. All the interpolation techniques provide a zero value as an estimate for situations encountered in case 2. Data from other sources (e.g. radar-based precipitation estimates) can be used in the above situations to estimate the missing values. However, the reliability of radar-based precipitation measurements is a contentious issue (Young *et al.* 1999; Adler *et al.* 2001).

In two cases identified earlier a station closest to the base station is selected and the estimates provided by the spatial interpolation technique are revised using the observations recorded at the closest stations. A remedial strategy to address the first situation is proposed here. This strategy is based on the assumption that the station closest to the base station and the base station will experience similar precipitation magnitudes and patterns. The observations at a station that is closest to the base station by Euclidean distance or a station selected based on strongest correlation are used to modify the estimates provided by interpolation techniques or adopted as estimates. Teegavarapu (2007) verified this approach and proved that Euclidean distance is not always a surrogate measure of spatial correlation between observations recorded at any two stations. A similar approach was referred to as single best estimator (SBE) by Eischeid *et al.* (2000).

The main objective of this study is to evaluate the use of association rule mining (ARM), a data mining technique, in conjunction with a spatial interpolation technique to obtain the estimates of missing precipitation data and to overcome one of the major limitations of spatial interpolation techniques. Two of the four interpolation techniques used in the current study are improvised weighting methods recently reported by Teegavarapu & Chandramouli (2005). The contents of this paper are organized as follows. A brief introduction to the deterministic and stochastic interpolation techniques is provided first. Limitations of these methods and concepts related to data and association rule mining are discussed next. Development of association rule mining based interpolation is then presented along with a case study application. Finally results and analysis, general remarks and conclusions are presented.

INTERPOLATION TECHNIQUES

Deterministic spatial interpolation techniques such as the inverse-distance weighting method (IDWM) and its revised versions, such as the modified inverse-distance weighting method (MIDWM), the coefficient of correlation weighting method (CCWM), along with the ordinary kriging estimation method (KEM) are used in the current study to estimate missing precipitation data. The CCWM and MIDWM are two methods recently reported by Teegavarapu & Chandramouli (2005) for estimation of missing precipitation data. The following subsections provide brief details of these methods.

Inverse-distance weighting method (IDWM)

The inverse-distance (reciprocal-distance) weighting method (Simanton & Osborn 1980) is most commonly used for estimation of missing data. The weighting distance method for estimation of missing value of an observation, θ_m, using the observed values at other stations is given by

$$\theta_m = \frac{\sum_{i=1}^{n} \theta_i d_{m,i}^{-k}}{\sum_{i=1}^{n} d_{m,i}^{-k}} \quad (1)$$

where θ_m is the observation at the base station m; n is the number of stations; θ_i is the observation at station i; $d_{m,i}$ is the distance from the location of station i to station m; and k is referred to as the friction distance (Vieux 2001) that ranges

from 1.0–6.0. A variant of IDWM that incorporates local anisotropy was proposed and tested by Tomczak (1998).

Modifications are incorporated into the IDWM for estimation of missing data. These modifications are mainly related to distance and weight calculations. In two proposed variants of the IDWM, the distances are replaced by new parameters that can help in better estimation of missing data.

Modified inverse-distance weighting method (MIDWM)

The inverse-distance weighting method is modified by replacing the distance in Equation (1) by a distance defined by the property of the proximity (Thiessen) polygons. The distance is smaller than the actual distance measured from the base station. Since the distances are smaller compared to the original distances, the weights in MIDWM will now be higher for some of the stations than those used in the IDWM. Estimation of missing data value, θ_m, can be carried out using Equation (1) with the revised distance. Details of this method can be found elsewhere (Teegavarapu & Chandramouli 2005).

Coefficient of correlation weighting method (CCWM)

The success of the inverse-distance weighting method strongly depends on the existence of strong positive spatial autocorrelation. In the case of CCWM, the weighting factors are replaced by the correlation coefficients and the estimation method is given by

$$\theta_m = \frac{\sum_{i=1}^{n} \theta_i \rho_{mi}}{\sum_{i=1}^{n} \rho_{mi}} \tag{2}$$

where θ_{mi} is the coefficient of correlation, which is the ratio of covariance of two datasets to the product of standard deviations of datasets. The coefficient, θ_{mi}, is obtained by using the data at station m and any other station i. In applying this method available historical data are used for deriving the values of θ_{mi}. A similar approach of replacing Euclidean distance with statistical distance was reported by Ahrens (2006).

Kriging estimation method (KEM)

Kriging (Journel & Huijbregts 1978; Isaaks & Srivastava 1989; Webster & Oliver 2001) is widely recognized as the standard approach used for surface interpolation, based on scalar measurements at different points in space. This method is used to estimate missing data in the current study. Kriging is an optimal surface interpolation method based on spatially dependent variance (Vieux 2001). The degree of spatial dependence is generally expressed as a semi-variogram in kriging. The general expression that is used to estimate the semi-variogram is given by

$$\gamma(d) = \frac{1}{2n(d)} \sum_{dij=d} (\theta_i - \theta_j)^2 \tag{3}$$

where $\gamma(d)$ is the semi-variance which is defined over observations θ_i and θ_j lagged successively by distance d. Surface interpolation using kriging depends on the selected semi-variogram model and the semi-variogram must be fitted with a mathematical function or model. Depending on the shape of the semi-variogram several mathematical models are possible that include linear, spherical, circular, exponential and Guassian. A typical semi-variogram is shown in Figure 1. Three semi-variogram models, namely spherical, exponential and Gaussian given by Equations (4)–(6), are used in the current study:

$$\gamma(d)_1 = C_o + C_1 \left[\frac{1.5d}{a} - 0.5\left(\frac{d}{a}\right)^3 \right] \tag{4}$$

$$\gamma(d)_2 = C_o + C_1 \left[1 - \exp\left(-\frac{3d}{a}\right) \right] \tag{5}$$

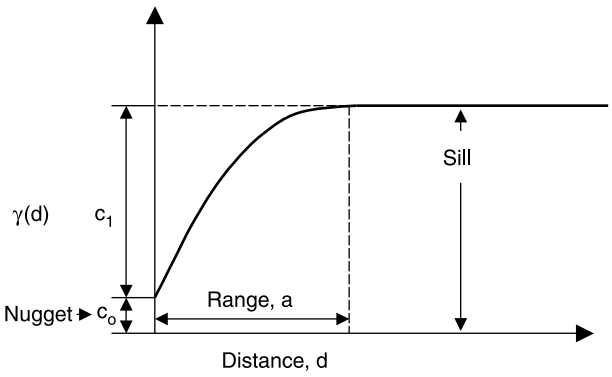

Figure 1 | Typical parameters (nugget and sill) of a semivariogram.

$$\gamma(d)_3 = C_o + C_1\left[1 - \exp\left(-\frac{(3d)^2}{a^2}\right)\right] \quad (6)$$

The parameters C_o, d and a are referred to as the nugget, distance and range and are shown in Figure 1. The summation of C_o and C_1 is referred to as the sill and the semi-variance at range, a, is equal to the sill value. The values of C_o and C_1 are obtained by trial-and-error procedures and the final values used in the study are 0.2 and 0.8, respectively. In ordinary kriging the weights are based not only on the distance between the measured points and the prediction location, but also on the overall spatial arrangement among the measured points and their values. The weights mainly depend on fitted model (i.e. semivariogram) to the measured points. The general equation for estimating missing value, θ_m, is given by

$$\theta_m = \sum_{i=1}^{n} \delta_i \theta_i \quad (7)$$

where θ_i is the weight obtained from the fitted semivariogram and θ_i is the value of the observation at location i. The observed data is used twice, once to estimate the semivariogram and then to interpolate the values.

DATA MINING

Data mining is the process of extracting interesting (nontrivial, implicit, previously unknown and potentially useful) information or patterns from large information repositories (Chen *et al.* 1996). It is also used to extract information and patterns derived by any knowledge discovery methods (Dunham 2002; Witten & Frank 2005). Knowledge discovery and data mining terms are often used interchangeably in data mining literature.

Association rule mining (ARM)

Association rule mining (ARM) is one of the popular data mining methods mainly aimed at extracting interesting correlations, frequent patterns, associations or causal structures among data available in databases (Agrawal & Srikant 1995; Zhang & Zhang 2002; Zhao & Bhowmick 2003). Association rule mining is regarded as an unsupervised knowledge discovery process. It has been successfully applied for deriving spatio-temporal relationships hidden in Earth science data (Zhang *et al.* 2005) and analysis of urban land cover change (Mennis & Liu 2005). Li *et al.* (2003) used data mining algorithms in conjunction with spatial interpolation to facilitate drought risk assessment using temperature and precipitation data.

ARM is carried out using an *a priori* algorithm developed by Agarwal *et al.* (1993). It is also referred to as a support–confidence framework for discovering association rules within a database. Association rules take the form "if antecedent then consequent". The format is generally expressed as $X \Rightarrow Y$, suggesting that event Y is expected to occur whenever event X is observed. The events X and Y are generally referred to as items or itemsets in traditional ARM literature. In the current context, an item refers to a specific event (e.g. occurrence or non-occurrence of rain at a station) and an itemset refers to a set of events (i.e. series of stations with occurrence or non-occurrence of rain). Before the details of the algorithm are explained, two important measures for association rules need to be discussed. These measures are support and confidence, and are discussed in relation to the current context.

Support

The support for an association rule $X \Rightarrow Y$ is the proportion of days (D) that contain both X and Y:

$$\alpha = p(X \cap Y) \quad (8)$$

Support is defined using itemsets and indicates the proportion of the total number of days which contain both X and Y. It is a measure of the significance or importance of an itemset. An itemset with a support greater than a minimum support threshold (θ_m) value is called a frequent or large itemset. One important property of support is the downward closure property that suggests that all subsets of a frequent set are also frequent. This property (i.e. no superset of an infrequent set can be frequent) is mainly used to reduce the search space in the *a priori* algorithm and prune the association rules.

Confidence

$$\beta = p\left(\frac{Y}{X}\right) = \frac{p(X \cap Y)}{p(X)} \tag{9}$$

Confidence is defined as the probability of seeing the rule's consequent under the condition that the transactions also contain the antecedent. It is important to note that confidence is directed and gives different values for the rules $X \Rightarrow Y$ and $Y \Rightarrow X$. Confidence is not downward closed and was developed together with support by Agrawal *et al.* (1993).

Support is initially used to find frequent (significant) itemsets exploiting its downward closure property to prune the search space. Then confidence is used in a second step to produce rules from the frequent itemsets that exceed a minimum confidence threshold. One main limitation of confidence parameter is that it is sensitive to the frequency of the consequent (Y) in the database. Consequents with higher support will automatically produce higher confidence values, even if there is no association between the items.

ASSOCIATION RULE MINING BASED SPATIAL INTERPOLATION

The association rule mining based spatial interpolation used in the current study is illustrated in Figure 2. A spatial interpolation technique is applied first to estimate missing data at a gauging station (i.e. base station) based on observations available at all other stations. The spatio-temporal database of historical precipitation observations are mined separately using the *a priori* algorithm and several rules are derived. The *a priori* algorithm as provided in the appendix is implemented in two major steps: (1) generation of frequent itemsets and (2) generation of association rules. In the current study the traditional distance weighting method (IWDM) and its variants (CCWM and MIDWM) and a stochastic interpolation approach, ordinary kriging (KEM), are investigated.

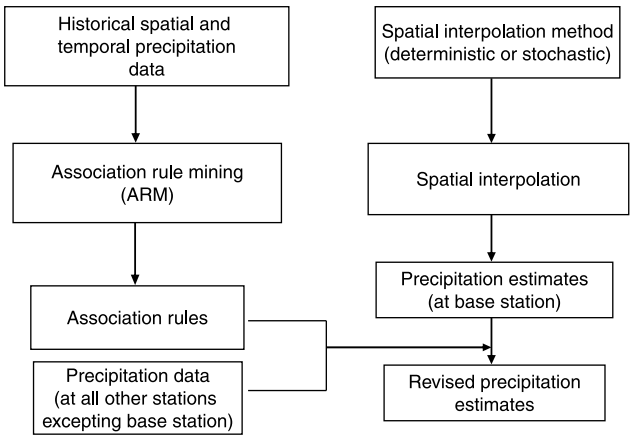

Figure 2 | Framework for integration of association rule mining (ARM) and spatial interpolation method for estimation of missing precipitation data.

The description of the algorithm format shown in the appendix is a modified version of that presented by Zhao & Bhowmick (2003). In the first phase all the candidate one-item sets are identified. Using the parameters of minimum support and confidence large one-item sets are then selected. The process is continued to obtain two-item candidates and two-item large itemsets. A few steps associated with the implementation of the algorithm for the current study are shown in the appendix in Tables (a–d) respectively. Using the minimum confidence limit, association rules are derived from the large-1, large-2 and large-k itemsets. All the rules used in the current study are referred to as single consequent rules as the base station, Lexington, only features in the consequent part of all the rules.

Corrections are applied to the precipitation estimates provided by the spatial interpolation methods using ARM-derived rules. The ARM-based rules can be translated into mathematical forms as described by Equation (10) and conditions specified by inequalities (11)–(13):

$$\text{if}(\cap(\theta_i = 0)),\ \text{then}\ \theta_m^o = 0,\ \text{else}\ \theta_m^o = \theta_m\ \forall i,\ i \neq m \tag{10}$$

$$i \leq n - 1 \tag{11}$$

$$\alpha \geq \alpha_m \tag{12}$$

$$\beta \geq \beta_m \tag{13}$$

The variable, i, is the number of stations identified based on ARM of the database, θ_m is the estimated value using the spatial interpolation method at base station, m, θ_m^o is the revised estimate of the precipitation value, n is the total

number of stations, α, α_m, β and β_m are the support and minimum support, and confidence and minimum confidence levels, respectively.

CASE STUDY AREA

The case study area comprises of the eastern part of the state of Kentucky. The state-wide average annual precipitation based on data from 1971–2003 varied between 76.2 cm (30 in) and 193 cm (76 in), with values higher than 127 cm (50 in) in the southeastern region and lower than 107 cm (42 in) in the northeastern part. The statistics of rainfall data at different stations with the same historical record length are given in Table 1. The Cumberland (or Appalachian) Plateau dominates the eastern third of Kentucky and contains the highest point, Black Mountain, at 1,263 m above mean sea level. The Bluegrass Region (north-central) is a series of hills fronting the Ohio River. The far western corner includes the Mississippi River flood plain with the lowest elevation (78 m) in the state. The state with mean elevation of 229 m is dominated by the Ohio River forming its northern borders, and the Cumberland and Tennessee River systems, and their many spin-off lakes.

Table 1 | Statistics of observed daily and annual rainfall values at different stations

Station	$\bar{x}_1$	S_1	$\bar{x}_2$	S_2
Bardstown	0.986	1.290	125.456	23.602
Berea	0.904	1.158	119.019	21.460
Bowling Green	1.052	1.430	129.931	22.489
Buckhorn	0.861	1.062	118.562	19.403
Campbellsville	1.064	1.410	132.268	24.305
Covington	0.810	1.087	108.575	16.068
Cumberland	0.892	1.120	125.280	22.278
Grayson	0.864	1.074	107.513	14.709
Hardinsburg	1.024	1.283	122.428	19.093
Jackson	0.889	1.133	125.042	20.554
Lexington	0.881	1.209	116.271	20.947
London	0.785	1.097	118.087	20.340
Louisville	0.912	1.245	114.691	18.880
Somerset	1.016	1.290	129.080	19.667
Williamstown	0.765	1.074	113.373	17.493

$\bar{x}_1$=Mean rainfall (cm), daily; S_1 = Standard deviation (cm), daily; $\bar{x}_2$=Mean rainfall (cm), annual; S_2 = Standard Deviation (cm), annual.

Other major rivers include the Kentucky, Licking and Mississippi along its western border with the state of Missouri.

APPLICATION TO CASE STUDY

The association rule mining based spatial interpolation method is used to estimate missing rainfall data at a base station (i.e. Lexington, Kentucky). Data at the base station are assumed to be missing for the purpose of testing the improvised interpolation methods. Historical daily rainfall data from year 1971–2002 available at 15 rainfall gauging stations in the state of Kentucky are used for analysis. These gauging stations are shown in Figure 3 and are numbered for convenience. The data used in the current study are compiled and provided by the Kentucky Agricultural Weather Center, University of Kentucky. Approximately 67% of the historical data (7,800 d) is used for obtaining the association rules and 33% of the data (3,900 d) is used for testing the methods. This is consistent with procedures generally used for comparison of spatial prediction models (Kanevski & Maignan 2004). GSLIB (Duetsch & Journel 1992), a geo-statistical software library, is used to apply ordinary kriging with three semi-variogram models given by Equations (4)–(6). The existing code of GSLIB was modified and additional code was developed to use the GSLIB software library for temporal estimation of missing precipitation data.

The association rule mining is carried out using the WEKA (Waikato Environment for Knowledge Analysis)

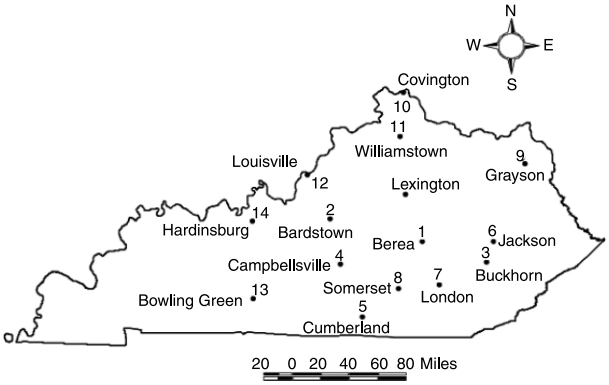

Figure 3 | Location of precipitation gauging stations in the eastern part of the state of Kentucky.

(WEKA 2001) modeling environment. Historical precipitation data at all stations are converted from continuous numerical values to categorical types of data. This process is referred to as discretization and it is needed for spatial association rule mining. The categorical data is specified as "no rain" and "rain" within the database of the modeling environment and they constitute only two class intervals. The ARM is used in a supervized manner in which the consequent (i.e. Lexington rain gauge station) is preselected. Once the association rules are extracted, they are applied to the estimates of the missing precipitation data to obtain revised estimates. The performances of ARM-based interpolation techniques are compared using two widely recognized and commonly used error measures, root mean square error (RMSE) and absolute error (AE), based on actual and estimated rainfall values at the base station. Several researchers (e.g. Chang 2004; Kanevski & Maignan 2004; Ahrens 2006) have recommended these two measures for comparison of spatial predictions of interpolation models for testing data. The error measures, RMSE and AE, are given by Equations (14) and (15), respectively:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{\phi}_i - \phi_i)^2} \tag{14}$$

$$\mathrm{AE} = \sum_{i=1}^{n}|\hat{\phi}_i - \phi_i| \tag{15}$$

where n is the total number of observations, $\hat{\phi}_i$ is the estimated value and ϕ_i is the actual value of the observation. The error measures, RMSE and AE, are provided in units of inches in this paper.

RESULTS AND ANALYSIS

Historical precipitation data is used in the WEKA modeling environment to obtain association rules. Initially low values of support and confidence are used and the precipitation data are mined. The values are systematically increased in an iterative way and the association rules are obtained until no large itemsets are possible. Rules with different confidence and support values are used to improve the estimates provided by spatial interpolation methods. The rules obtained from the ARM process based on training data are provided in Table 2. Rules related to the maximum achievable confidence and support values based on the data are selected. Any rule generated by ARM has a general form with antecedent and consequent parts. For example, rule 2 (in Table 2) suggests that, in instances when no rainfall is recorded at rainfall gauging stations, Louisville and London, no rain was observed/recorded at the base station (i.e. Lexington). Also the generation of rules is stopped when the inequalities, Equations (12) and (13), are violated.

The ARM rules are transformed into mathematical expressions (Equations (10) and (11)). These expressions, along with the data at all other stations, are used to revise the estimates provided by the spatial interpolation methods. In the current study, three deterministic methods, namely IDWM, CCWM and MIDWM, and one stochastic interpolation approach, ordinary kriging, are used in conjunction with ARM to evaluate the proposed integrated methodology.

The results associated with these experiments are reported in Table 3. Two error measures, namely RMSE and AE, provided for three different methods for training and test data suggest that the use of ARM rules improves the overall estimation process. The performance is best for rule 1 and it decreased as the next-best rules are applied. However, the performance is better than the case when no rule was applied for revision of the precipitation estimates for all the rules. On average a decrease of 75 in for absolute error was evident from the application of the rules, with the lowest decrease for the CCWM method. Table 4 provides results for ordinary kriging with the application of rules.

Table 2 | Description of association mining rules along with support (α) and confidence (β) values for each rule

Rule	Description	α	β
1	Louisville = no rain ⇒ Lexington = no rain	0.55	0.89
2	Louisville = no rain, London = no rain ⇒ Lexington = no rain	0.50	0.95
3	Louisville = no rain, Berea = no rain ⇒ Lexington = no rain	0.50	0.95
4	Louisville = no rain, Somerset = no rain ⇒ Lexington = no rain	0.50	0.94

Table 3 | Performance of association rule mining based spatial interpolation methods using different rules for training and testing data

Error measure	Method ARM rule	Rule 1	Rule 2	Rule 3	Rule 4
	IDWM	IDWM^{+}			
RMSE	0.235	0.227	0.233	0.234	0.234
AE	777.376	636.125	714.773	731.219	726.390
RMSE*	0.243	0.238	0.242	0.243	0.242
AE*	365.054	310.450	348.253	351.649	349.822
	CCWM	CCWM^{+}			
RMSE	0.204	0.202	0.203	0.203	0.204
AE	639.095	570.285	605.578	609.320	610.248
RMSE*	0.226	0.225	0.226	0.226	0.226
AE*	320.907	294.135	311.394	310.921	311.984
	MIDWM	MIDWM^{+}			
RMSE	0.210	0.207	0.209	0.209	0.210
AE	672.598	581.511	630.505	634.716	636.675
RMSE*	0.229	0.228	0.229	0.230	0.229
AE*	331.841	296.767	320.507	319.853	321.367

IDWM^{+}, CCWM^{+}, MIDWM^{+}: improved estimates with ARM based spatial interpolation; RMSE*, AE*: error measures for test data in inches.

Average error (observed − estimated) can be used as one of the error measures to provide an assessment of bias. In general, all the methods resulted in over-estimation before the application of association rules and under-estimation after the application of rules. For example, application of the IDWM, MIDWM and CCWM without application of association rules to test data resulted in an over-estimation given by average errors as 0.001, 0.002, 0.001 in, respectively. With the application of rule, the methods (i.e. IDWM, MIDWM and CCWM) resulted in under-estimation with average errors of 0.023, 0.017 and 0.015 in, respectively.

Table 4 | Performance of association rule mining based spatial interpolation using kriging with different rules for test data

Error measure	Method ARM rule	Rule 1	Rule 2	Rule 3	Rule 4
	KEM*	KEM^{a+}			
RMSE	0.238	0.233	0.236	0.237	0.235
AE	424.912	318.582	352.503	360.250	354.27
	KEM†	KEM^{b+}			
RMSE	0.464	0.309	0.348	0.354	0.351
AE	1691.166	619.857	830.790	852.712	846.307
	KEM‡	KEM^{c+}			
RMSE	0.647	0.404	0.461	0.471	0.464
AE	2408.448	845.278	1136.891	1172.498	1157.256

*Spherical semi-variogram model.
†Exponential semi-variogram model.
‡Gaussian semi-variogram model.
KEM^{a+}, KEM^{b+}, KEM^{c+}: improved estimates with ARM based spatial interpolation.

To evaluate the effect of temporal scale on the number or the nature of rules, monthly historical data was used to develop ARM rules. The rules given in Table 5 suggest that, in all the months excepting in the month of April and June, the rules contained the rain gauging station, Louisville, in the antecedent part. Tables 6 and 7 provide results related to the months of April and June for ordinary kriging. It is interesting to note that rules 6 and 7 provided lower AE and RMSE values compared to those provided by rule 1. Results related to the three deterministic methods, IDWM, CCWM and MIDWM, are provided in Tables 8 and 9.

It is important to note that the error measures calculated are average values for the testing period of 3,900 d. Absolute errors calculated based on the daily estimated and observed values using the methods proposed in the current study ranged from 1,387 mm (55 in) to 680 mm (27 in) for 3,900 d. The highest absolute error resulted from the use of IDWM and the lowest absolute error from CCWM when these methods are used in conjunction with association rule mining. A 1% difference in RMSE value can suggest on average one interpolation method is either over-predicting or under-predicting rainfall values by 1%. Small variations in rainfall intensity can introduce significant changes in the runoff values generated from distributed rainfall–runoff models (Vieux 2001). Any improvement in the rainfall magnitude estimation, however minute it may be, can be considered significant, as rainfall is a crucial input that governs the response of hydrologic systems and the results of continuous simulation models. Similar arguments were made by Xu & Vandewiele (1994) to suggest that errors in precipitation values may lead to significant effects on the model performance and also on parameters.

In general the RMSE and AE values decreased when rules based on ARM methodology are applied to revise the precipitation estimates from the spatial interpolation methods. It is interesting to note that the rainfall gauging station at Louisville appears in the antecedent part of

Table 5 | Rules based on monthly historical data along with support (α) and confidence (β) values

Month	Rule	Description	α	β
January	1	Louisville = no rain ⇒ Lexington = no rain	0.56	0.87
February	2	Louisville = no rain ⇒ Lexington = no rain	0.55	0.89
March	3	Louisville = no rain ⇒ Lexington = no rain	0.52	0.90
April	4	Louisville = no rain ⇒ Lexington = no rain	0.54	0.90
	5	Somerset = no rain ⇒ Lexington = no rain	0.53	0.82
May	6	Louisville = no rain ⇒ Lexington = no rain	0.53	0.87
June	7	London = no rain ⇒ Lexington = no rain	0.55	0.86
	8	Hardinsburg = no rain ⇒ Lexington = no rain	0.55	0.81
July	9	Louisville = no rain ⇒ Lexington = no rain	0.58	0.86
August	10	Louisville = no rain ⇒ Lexington = no rain	0.63	0.88
September	11	Louisville = no rain ⇒ Lexington = no rain	0.64	0.92
October	12	Louisville = no rain ⇒ Lexington = no rain	0.67	0.92
November	13	Louisville = no rain ⇒ Lexington = no rain	0.57	0.90
December	14	Louisville = no rain ⇒ Lexington = no rain	0.55	0.90

almost all of the ARM rules. One possible explanation is the highest correlation coefficient obtained based on observations at Louisville and the base station, Lexington. One might argue that there is no need for the whole exercise of developing and using association rules to obtain revised precipitation estimates in this situation. A plot of correlation coefficients between different stations and the base station is provided in Figure 4. The correlation coefficients based on observations at Lexington and between Bardstown, Bowling Green and Covington are 0.617, 0.593 and 0.589, respectively.

It is interesting to note that none of the stations were identified by the ARM procedure in antecedent parts of the rules. Louisville with the highest correlation coefficient of 0.691 appears in all the rules except for two rules derived for the month of June. This does not necessarily indicate consistency between the association rules and the correlation coefficients. Correlation is a measure of linear

Table 6 | Performance of association rule mining based spatial interpolation using kriging with different rules for the month of April

Error measure	Method ARM rule	Rule 4	Rule 5
	KEM*	KEM^{a+}	
RMSE	0.151	0.142	0.158
AE	32.085	23.213	27.365
	KEM†	KEM^{b+}	
RMSE	0.433	0.266	0.312
AE	136.660	50.012	66.654
	KEM‡	KEM^{c+}	
RMSE	0.615	0.371	0.416
AE	196.411	70.900	90.152

*Spherical semi-variogram model.
†Exponential semi-variogram model.
‡Gaussian semi-variogram model.
KEM^{a+}, KEM^{b+}, KEM^{c+}: improved estimates with ARM based spatial interpolation.

Table 7 | Performance of association rule mining (ARM) based spatial interpolation using ordinary kriging with different rules for the month of June

Error measure	Method ARM rule	Rule 7	Rule 8	Rule 1
	KEM*	KEM^{a+}		
RMSE	0.44	0.458	0.458	0.526
AE	61.056	53.59	53.251	77.301
	KEM†	KEM^{b+}		
RMSE	0.578	0.507	0.522	0.556
AE	156.895	83.378	86.904	102.034
	KEM‡	KEM^{c+}		
RMSE	0.727	0.575	0.599	0.642
AE	210.497	100.974	107.196	121.731

*Spherical semi-variogram model.
†Exponential semi-variogram model.
‡Gaussian semi-variogram model.
KEM^{a+}, KEM^{b+}, KEM^{c+}: improved estimates with ARM based spatial interpolation.

Table 8 | Performance of association rule mining (ARM) based spatial interpolation methods using different rules for month of April using test data

Error measure	Method ARM rule	Rule 4	Rule 5
	IDWM	IDWM$^+$	
RMSE	0.164	0.164	0.173
AE	29.101	28.079	29.202
	CCWM	CCWM$^+$	
RMSE	0.163	0.163	0.172
AE	27.850	27.113	27.959
	MIDWM	MIDWM$^+$	
RMSE	0.164	0.164	0.173
AE	28.748	27.813	28.888

IDWM$^+$, CCWM$^+$, MIDWM$^+$: improved estimates with ARM based spatial interpolation.

association and is based on observed data whereas association rules are developed based on categorical values. It should be noted that confidence measures the strength of the rule and support measures relevancy. A higher confidence value should not be mistaken for high correlation (Brijs *et al.* 2003). The relationships derived using the ARM approach does not represent any sort of causality or correlation (Dunham 2002).

The rules obtained from the ARM process can be ranked by a concept referred to as the usefulness of association rule. The usefulness is a measure given by the product of support and confidence. The rules provided in Table 2 have usefulness measures such as 0.489, 0.475, 0.475 and 0.47. Rule 1 is ranked highest in this case based on the usefulness metric and it also provided the lowest RMSE and AE values.

Table 9 | Performance of association rule mining (ARM) based spatial interpolation methods using different rules for the month of June for test data

Error measure	Method ARM rule	Rule 7	Rule 8
	IDWM	IDWM$^+$	
RMSE	0.451	0.468	0.469
AE	60.083	56.983	55.951
	CCWM	CCWM$^+$	
RMSE	0.432	0.450	0.451
AE	53.888	50.922	51.274
	MIDWM	MIDWM$^+$	
RMSE	0.436	0.454	0.456
AE	55.679	52.838	52.935

IDWM$^+$, CCWM$^+$, MIDWM$^+$: improved estimates with ARM based spatial interpolation.

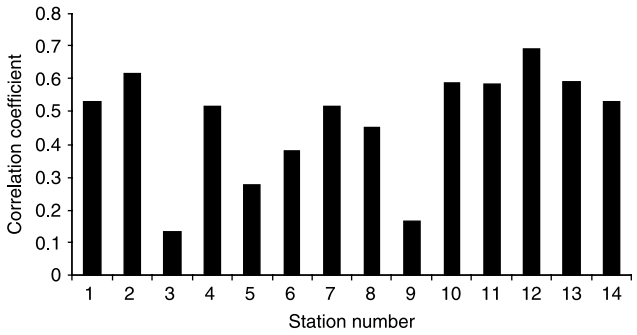

Figure 4 | Correlation coefficients based on observations at each station and the base station.

A look at the observed data in the test period at the base station indicates that there are 2,495 d of no precipitation out of a total of 3,900 d. The CCWM provided a zero value of precipitation estimates for 1,106 d and a positive value for 1,389 d. The over-estimation of precipitation for 1,389 d is 51.71 in. Use of ARM rules can help eliminate this over-estimation. In this context, the use of ARM rules eliminated 1,459 d of positive precipitation. Under-estimation of precipitation is possible due to the application of ARM-based spatial interpolation. This happens when no rain occurred at one of the stations identified in the antecedent part of the ARM rule and precipitation occurred at the base station. In spite of under-estimation, the overall absolute error obtained is lower than the one obtained by using any interpolation technique. Two error measures (root mean square error and absolute error) were used to assess and compare the performance of ARM-based interpolation techniques in this study. However, these global error measures may not provide a complete assessment of methods as they are average measures calculated for a specific period of time.

GENERAL REMARKS

The use of association rule mining to improve estimates of missing precipitation data by interpolation methods is demonstrated in this study. Three main factors will affect the nature of the rules generated from the ARM process. These include: (1) the length of the spatio-temporal data, (2) the minimum threshold support (α) and confidence (β)

limits and finally (3) the discretization of the data into categorical data classes. In general pseudo-association is expressed by either distance in IDWM (or MIDWM) or by correlation coefficient in CCWM. By integrating ARM with spatial interpolation techniques in the current study, associations between observations are expressed in the form of rules. Historical data are required to estimate missing precipitation values using CCWM and KEM and also to obtain association rules using ARM.

The threshold values of support and confidence factors influence the number of best rules found by the ARM process. Selection of final rules at the end of the ARM completion process can be a contentious yet crucial issue that might affect the revised estimates obtained from spatial interpolation techniques. However, the rules obtained from the historical data can be used to obtain revised estimates from IDWM, CCWM, MIDWM and KEM, and performance of these methods can be evaluated before they can be applied to the test data. In the current study strong association is detected based on the data mining approach between observations at any two stations. The rules thus generated based on the ARM concept are limited in number. However, the case may be different if an entirely separate rain gauging network is used. It is possible that many interesting rules are pruned or not reported, as the support and confidence values are restricted to a few pre-specified limits. The number of association rules grows exponentially based on the number of stations and the categorical attribute values. In the current case, 15 stations with binary attributes (i.e. yes or no) associated with rainfall occurrence and non-occurrence, a total of $15 \times 2^{15-1}$ association rules are possible. Also the discretization scheme will affect the nature of association rules. Exhaustive studies using ARM concepts need to be conducted before any recommendations can be made about the transferability of the approach discussed in this study to other climatic regions under different meteorological conditions.

CONCLUSIONS

An association rule mining (ARM) based spatial interpolation approach is presented in this paper. This innovative approach is used to improve the precipitation estimates provided by traditional and improved deterministic and stochastic spatial interpolation techniques. The ARM methodology, besides offering insights into the spatio-temporal precipitation data patterns and the associations among observations, also helps in addressing one major ubiquitous limitation of all spatial interpolation techniques in accurately estimating missing precipitation records. The use of ARM is not equivalent to the use of correlation analysis to revise estimated precipitation values obtained from deterministic and stochastic interpolation techniques. Considerable improvements in the estimates were achieved when ARM is used in conjunction with interpolation techniques. However, the uncertainty in the revised estimates can only be assessed by using support and confidence indices available within the ARM framework and ultimately through the use of calibrated and validated hydrological simulation models.

ACKNOWLEDGEMENTS

The author thanks the Kentucky Agricultural Weather Center, University of Kentucky, for providing the data required for the research study reported in this paper.

REFERENCES

Adler, R. F., Kidd, C., Petty, G., Morissey, M. & Goodman, H. M. 2001 Intercomparison of global precipitation products: the third precipitation inter-comparison project (PIP-3). *Bull. Am. Meteor. Soc.* **82**, 1377–1396.

Agrawal, R., Imielinski, T. & Swami, A. 1993 Mining association rules between sets of items in large databases. In *Proceedings of the ACM SIGMOD International Conference on Management of Data* ACM, Washington, DC. pp. 207–216.

Agrawal, R. & Srikant, R. 1995 Fast algorithms for mining association rules. In *Proceedings of 20th Conference on Very Large Databases* (ed. J. B. Bocca, M. Jarke & C. Zaniolo), pp. 487–499. Morgan Kaufmann Publishers Inc., San Francisco, CA.

Ahrens, B. 2006 Distance in spatial interpolation of daily rain gauge data. *Hydrol. Earth Syst. Sci.* **10**, 197–208.

ASCE 1996 *Hydrology Handbook*, 2nd edition. American Society of Civil Engineers ASCE, New York.

Ashraf, M., Loftis, J. C. & Hubbard, K. G. 1997 Application of geostatisticals to evaluate partial weather station network. *Agric. Forest Meteor.* **84**, 255–271.

Brijs, T., Vanhoof, K. & Wets, G. 2003 Defining interestingness for association rules. *Int. J. Inf. Theor.* **10** (4), 370–376.

Brimicombie, A. 2003 *GIS, Environmental Modeling and Engineering*. Taylor and Francis, London.

Brown, J., Dingman, S. L. & Lewellen, R. J. 1968 *Hydrology of a drainage basin on the Alaskan coastal plains*. U.S. Army Cold Regions Research and Engineering Lab., Hanover, NH, Research Report 240.

Burrough, P. A. & McDonnell, R. A. 1998 *Principles of Geographical Information Systems*. Oxford University Press, Oxford.

Chang, K.-T. 2004 *Introduction to Geographic Information Systems*. McGraw-Hill, New York.

Chen, M.-S., Han, J. & Yu, P. S. 1996 Data mining: an overview from a database perspective. *IEEE Trans. Knowledge Data Eng.* **8**, 866–883.

Deutsch, C. V. & Journel, A. G. 1992 *Geostatistical Software Library and User's Guide*. Oxford University Press, New York, USA.

Dingman, S. L. 2002 *Physical Hydrology*. Prentice Hall, Englewood Cliffs, NJ.

Dingman, S. L., Barry, R. G., Weller, G., Benson, C., LeDrew, E. F. & Goodwin, C. W. 1980 Climate, snow cover, microclimate, and hydrology. In *An Arctic ecosystem: The Coastal Tundra at Barrow, Alaska* (ed. J. Brown, P. C. Miller, L. L. Tieszen & F. Bunnell), Dowden, Hutchinson and Ross, Inc., PA. pp. 30–65.

Dunham, M. 2002 *Data Mining: Introductory and Advanced Topics*. Prentice Hall, Englewood Cliffs, NJ.

Eischeid, J. K., Pasteris, P. A., Diaz, H. F., Plantico, M. S. & Lott, N. J. 2000 Creating a serially complete, national daily time series of temperature and precipitation for the western United States. *J. Appl. Meteor.* **39**, 1580–1591.

Grayson, R. & Bloschl, G. 2001 *Spatial Patterns in Catchment Hydrology: Observations and Modeling*. Cambridge University Press, Cambridge.

Isaaks, H. E. & Srivastava, R. M. 1989 *An Introduction to Applied Geostatistics*. Oxford University Press, Oxford.

Journel, A. G. & Huijbregts, C. J. 1978 *Mining Geostatistics*. Academic, New York.

Kanevski, M. & Maignan, M. 2004 *Analysis and Modelling of Spatial Environmental Data*. EPFL Press, Lausanne, Switzerland.

Krajewski, W. F. 1987 Co-kriging of radar and rain gage data. *J. Geophys. Res.* **92** (D8), 9571–9580.

Li, D., Harms, S., Goddard, S., Waltman, W. & Deogun, J. 2003 Time-series data mining in a geospatial decision support system. In *Proceedings of National Conference on Digital Government Research*, Digital Government Society of North America, Boston. pp. 1–4.

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

Salas, J. D.-J. 1993 Analysis and modeling of hydrological time series. In *Handbook of Hydrology* (ed. D. R. Maidment), McGraw-Hill, New York. pp. 19.1–19.72.

Simanton, J. R. & Osborn, H. B. 1980 Reciprocal-distance estimate of point rainfall. *J. Hydraul. Eng. Div.* **106** (HY7), 1242–1246.

Singh, V. P. & Chowdhury, K. 1986 Comparing some methods of estimating mean areal rainfall. *Water Res. Bull.* **22**(2), 275–282.

Sullivan, D. O. & Unwin, D. J. 2003 *Geographical Information Analysis*. John Wiley & Sons, New York.

Teegavarapu, R. S. V. 2007 Use of universal function approximation in variance dependent surface interpolation technique: an application in hydrology. *J. Hydrol.* **332**, 16–29.

Teegavarapu, R. S. V. & Chandramouli, V. 2005 Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records. *J. Hydrol.* **312**, 191–206.

Tobler, W. R. 1970 A computer movie simulating urban growth in the Detroit region. *Econ. Geogr.* **46**, 234–240.

Tomczak, M. 1998 Spatial Interpolation and its uncertainty using automated anisotropic inverse distance weighting-cross-validation/jackknife approach. *J. Geogr. Inf. Decision Anal.* **2**, 18–30.

Tung, Y. K. 1983 Point rainfall estimation for a mountainous region. *J. Hydraul. Eng. ASCE* **109** (10), 1386–1393.

Vieux, B. E. 2001 *Distributed Hydrologic Modeling Using GIS. Water Science and Technology Library*. Kluwer. Amsterdam.

Wang, F. 2006 *Quantitative Methods and Applications in GIS*. Taylor and Francis, London, UK.

Webster, R. A. & Oliver, M. A. 2001 *Geostatistics for Environmental Scientists*. John Wiley & Sons, Chichester.

Wei, T. C. & McGuinness, J. L. 1973 *Reciprocal Distance Squared Method, A Computer Technique for Estimating Area Precipitation*. Technical Report ARS-Nc-8. US Agricultural Research Service, North Central Region, Ohio.

WEKA 2001 *Waikato Environment for Knowledge Analysis*. The University of Waikato, New Zealand.

Witten, I. H. & Frank, E. 2005 *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco.

Xu, C.-Y. & Vandewiele, G. L. 1994 Sensitivity of monthly rainfall-runoff models to input errors and data length. *Hydrol. Sci. J.* **39** (2), 157–176.

Yang, D., Goodiso, B. E., Ishida, S. & Benson, C. S. 1998 Adjustment of daily precipitation data at 10 climate stations in Alaska: application of World Meteorological Organization intercomparison results. *Water Resour. Res.* **34**, 241–256.

Young, C., Nelson, B., Bradley, A., Smith, J., Peters-Lidard, C., Kruger, A. & Baeck, M. 1999 An evaluation of NEXRAD precipitation estimates in complex terrain. *J. Geophys. Res.* **104**, 19691–19703.

Zhang, C. & Zhang, S. 2002 *Association Rule Mining: Models and Algorithms. Lecture Notes in Artificial Intelligence*. Springer, New York.

Zhang, P., Steinbach, M., Kumar, V., Shekar, S., Tan, P.-N., Klooster, S. & Potter, C. 2005 Discovery of patterns in earth science data using data mining. In *Next Generation of Data-Mining Applications* (ed. M. M. Kantardzic & J. Zurada), Wiley-IEEE Press, New Jersey. pp. 167–187.

Zhao, Q. & Bhowmick, S. S. 2003 *Association Rule Mining: A Survey*. Technical report, CAIS, Nanyang Technological University, Singapore, report no. 2003116.

First received 22 January 2008; accepted in revised form 6 September 2008.

APPENDIX

Structure of *a priori* algorithm used in ARM implementation

Input:

Spatio-temporal database of historical precipitation records at all stations (D)

Support (minimum value) = α_m

Confidence (minimum value) = β_m

Method:

L_1 = large 1-itemsets;

C_1 = candidate itemsets;

for ($k = 2$; $L_{k-1} \neq \emptyset$; $k{+}{+}$) do begin

$C_k = a$ *priori*-generation(L_{k-1}); {Generation of new candidates from L_{k-1}}

for all data $T \in D$ do begin

C_t = subset (C_k, T);

for all candidates $C \in C_t$ do

Count = Count (C) +1; {Increment support count of C by 1}

end

$L_k = \{C \in C_t | \text{Count } (C) \geq \alpha_m \times |D|\}$

end

$L_f = \cup_k L_k$

R_t = Generate-rules (L_f, β_m)

Output:

Association rules (R_t) derived from the spatio-temporal precipitation database

α_m: Minimum Support

β_m: Minimum Confidence

T: Observed precipitation data represented in categorical form for one day.

D: Number of days of spatial precipitation data

Progress of association rule mining (ARM) process based on *a priori* algorithm

(a) C_1

Items	Count
Louisville (no rain)	4,000
Lexington (no rain)	3,500
London (no rain)	3,000
Berea (no rain)	2,000
Somerset (no rain)	1,000
Willimastown (no rain)	900

(b) L_1

Large 1 items
Louisville (no rain)
Lexington (no rain)
London (no rain)
Berea (no rain)

(c) C_2

Items	Count
Louisville (no rain), Lexington (no rain)	3,000
Berea (no rain), Lexington (no rain)	2,500
London (no rain), Lexingotn (no rain)	1,900
Louisville (no rain), Berea (no rain)	1,200

(d) L_2

Large 2 items	Count
Louisville (no rain), Lexington (no rain)	3,000
Berea (no rain), Lexington (no rain)	2,500
London (no rain), Lexingotn (no rain)	1,900