Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

CUTOFF: A spatio-temporal imputation method

Lingbing Feng [a,*], Gen Nowak [b], T.J. O'Neill [b], A.H. Welsh [c]

[a] International Institute for Financial Studies, Jiangxi University of Finance and Economics, Nanchang, Jiangxi 330013, China
[b] Research School of Finance, Actuarial Studies and Applied Statistics, The Australian National University, Canberra, ACT 0200, Australia
[c] Centre for Mathematics and its Applications, The Australian National University, Canberra, ACT 0200, Australia

ARTICLE INFO

Article history:
Received 22 April 2014
Received in revised form 20 October 2014
Accepted 3 November 2014
Available online 8 November 2014
This manuscript was handled by Andras Bardossy, Editor-in-Chief, with the assistance of Niko Verhoest, Associate Editor

Keywords:
Spatio-temporal
Imputation
Cross-validation
Simulation

SUMMARY

Missing values occur frequently in many different statistical applications and need to be dealt with carefully, especially when the data are collected spatio-temporally. We propose a method called CUTOFF imputation that utilizes the spatio-temporal nature of the data to accurately and efficiently impute missing values. The main feature of this method is that the estimate of a missing value is produced by incorporating similar observed temporal information from the value's nearest spatial neighbors. Extensions to this method are also developed to expand the method's ability to accommodate other data generating processes. We develop a cross-validation procedure that optimally chooses parameters for CUTOFF, which can be used by other imputation methods as well. We analyze some rainfall data from 78 gauging stations in the Murray–Darling Basin in Australia using the CUTOFF imputation method and compare its performance to four well-studied competing imputation methods, namely, *k*-nearest neighbors, singular value decomposition, multiple imputation and random forest. Empirical results show that our method captures the temporal patterns well and is effective at imputing large gaps in the data. Compared to the competing methods, CUTOFF is more accurate and much faster. We analyze further examples to demonstrate CUTOFF's applications to two different data sets and provide extra evidence of its validity and usefulness. We implement a simulation study based on the Murray–Darling Basin data to evaluate the method; the results show that our method performs well in both accuracy and computational efficiency.

1. Introduction

Managing missing values is important when performing data analysis because, even though missing values are ubiquitous, most statistical methods assume a complete data matrix. Imputation has the goal of completing data by replacing missing values by valid estimates, preferably in an unbiased and computationally efficient way. A problem arises for imputation when the data are collected spatially and temporally, as in climatic and hydrological studies as well as in many other research fields, because the non-completeness of the data poses special challenges for statistical analysis and spatio-temporal modeling (Schneider, 2001).

Imputation methods attempt to fill in missing values using mathematical properties or logical relationships between data values. Two simple methods are the mean method of Linacre (1992), which substitutes missing values by the variable mean values and the linear interpolation method of Lowry (1972), which replaces the missing values by the mean values of the previous and next data in a chronological series (see also Ramos-Calzado et al., 2008). However, these methods may cause severely biased estimates when the data distribution is skewed or there are large consecutive missing gaps in the data, and we may be able to do better with methods that better reflect the correlation structure in spatio-temporal data.

Some recent work on the Expectation–Maximization (EM) algorithm and covariance estimation has led to the development of methods specifically to deal with missing values in spatio-temporal data. These methods treat the estimation of covariance matrices and imputation of missing values together, proceeding by estimating the covariance matrix and imputing missing values sequentially and iteratively (Schneider, 2006a). Schneider (2001) proposed a parametric method, which applies the EM algorithm and ridge regression (Hoerl and Kennard, 1970), to estimate the mean and covariance matrix of spatio-temporal data iteratively. Missing values are imputed in each iteration using regression of variables with missing values onto variables that are not missing. Kondrashov and Ghil (2006) proposed a novel, iterative form of the Multi-channel Singular Spectrum Analysis (M-SSA) approach which utilized temporal, as well as spatial correlations to fill in

[*] Corresponding author. Tel.: +86 0791 83802306.
E-mail addresses: lingbing.feng@anu.edu.au (L. Feng), gen.nowak@anu.edu.au (G. Nowak), terry.oneill@anu.edu.au (T.J. O'Neill), alan.welsh@anu.edu.au (A.H. Welsh).

http://dx.doi.org/10.1016/j.jhydrol.2014.11.012
0022-1694/© 2014 Elsevier B.V. All rights reserved.

missing points in geophysical data sets. The application of M-SSA, especially the choice of the parameter *M*, makes this method flexible in capturing varying temporal patterns (Schneider, 2006b). The main innovative point of the M-SSA method is the use of a lag-covariance matrix to capture the spectral information in the time series (Kondrashov and Ghil, 2007). These spatio-temporal imputation methods have rarely been used in real data analysis because of their complexity and the lack of readily available software.

Other methods that use matrix decomposition techniques rely on fewer prior assumptions about the covariance matrix. Beckers and Rixen (2003), for example, presented an iterative empirical orthogonal function (EOF) method to fill in an incomplete dataset, which does not need prior information about the covariance error structure and is parameter free. Fuentes et al. (2006,) proposed an intuitive algorithm which takes advantage of the singular value decomposition (SVD) and linear regression to impute missing values iteratively. This algorithm has been regarded as a stable and reasonable way of imputing spatio-temporal missing values, and is included as a standalone imputation function in the **SpatioTemporal** package (Lindström et al., 2013a) in the R statistical language (R Core Team, 2014). It is publicly available, widely used and computationally efficient. Further details about the application of this method can be found in Cohen et al. (2009); Lindström et al. (2013b) and Szpiro et al. (2009).

For spatio-temporal data, machine learning methods such as artificial neural networks (ANN) have been used to impute missing values because of their strength in capturing complex nonlinear relationships. Applications can be found in meteorological and hydrological science, where data are mostly spatio-temporal (Nourani et al., 2008). Rustum and Adeloye (2007) proposed using Kohonen self-ogranizing map (KSOM) to impute missing values and replace outliers in time series data, and demonstrated that it works well for an activated sludge data. They compared this method with multiple regression analysis and ordinary back-propagation ANN, and showed that KSOM performs better. More recently, Kim and Pachepsky (2010) proposed a two-step reconstruction method that combined regression trees and artificial neural networks (RT + ANN) to complete daily precipitation data. They compared the combination method to the stand-alone RT and ANN methods, and showed that the "RT + ANN" method is better in terms of predictive accuracy. More recently, Nourani et al. (2012) investigated the ability of ANN models to impute missing rainfall data and found the best imputation model to be the one hidden layer feed-forward network, trained by the Levenberg–Marquardt algorithm. In the literature on imputation by ANN methods, as far as we investigated, no authors have provided clear and readily implemented software (in R and other programming environment) for their proposed methods.

Stekhoven and Bühlmann (2012) proposed an iterative imputation method (*missForest*) based on random forest which can be used to impute mixed-type data. They found that *missForest* outperforms other methods like the k-nearest neighbors imputation (KNN, Troyanskaya et al., 2001), the Missingness Pattern Alternating Lasso (MissPALasso) algorithm (Städler and Bühlmann, 2010) and MICE (Buuren and Oudshoorn, 1999). This method has been successfully applied to deal with other missing value problems (see Andreis and Ferrari, 2012; Chen and Ishwaran, 2012; Reinhardt et al., 2012; Talbert and Sole, 2013 for more applications). Although the algorithm was originally used to impute gene expression data, it can also be used to impute spatio-temporal data, mainly due to the ability of random forest to model complex interactions and nonlinear relations. Additionally, the authors have developed the R package **missForest** that is freely available online (Stekhoven, 2012).

Another important set of methods for filling in missing data before modeling in spatio-temporal studies is based on the combination of correlation techniques and synchronicity measures. These methods include the simple arithmetic averaging (SAA) method (Xia et al., 1999; Xia et al., 2001; Aravena and Luckman, 2008), the inverse distance (ID) method (Xia et al., 1999; Eischeid et al., 2000; Teegavarapu and Chandramouli, 2005), the single best estimator (SBE) method (Xia et al., 1999; Eischeid et al., 2000) and the normal ratio (NR) method (Paulhus and Kohler, 1952; Abebe et al., 2000; Yozgatligil et al., 2012). The SAA method completes the data matrix using the arithmetic mean of the synchronous values for a fixed number of stations close to the target station. The ID method computes the missing value by a weighted mean of the values from neighboring stations with weights equal to the inverse of the distance to the target station raised to the power of one or two. The SBE method fills the gaps with the synchronous values from the surrounding station with maximum correlation. The idea of this method agrees with our yet to be described method in that it attempts to derive information from both spatial (by defining "surrounding station") and temporal dimensions (by averaging over the synchronous date points), but its "single best" choice might be too extreme for reliable imputation. The NR method was introduced to interpolate missing annual precipitation records and has since been treated as a classical imputation method (Paulhus and Kohler, 1952). They realized that a straight average of records at surrounding stations would not always yield satisfactory estimates in some regions, so a ratio method was proposed to weight across surrounding stations called "index stations". The weighting factor, called the normal ratio, is the ratio of the normal annual precipitation at the target station to that at the respective index stations. Young (1992) found that using objective weights based on the correlation coefficients is superior to using the normal ratios. These techniques, summarized by Ramos-Calzado et al. (2008) and Yozgatligil et al. (2012), motivated our work. The key weaknesses of these papers are that they: (1) have not provided an objective way to optimize correlation or distance measures for choosing surrounding stations; (2) did not evaluate imputation performance and compare different imputation methods in a consistent fashion; and (3) for the purpose of evaluating the uncertainty of the imputed values, they did not simulate missing datasets with similar missing patterns to the original data. In attempts to overcome these weaknesses, Ramos-Calzado et al. (2008) proposed a novel three-step framework for selecting similar neighboring stations, estimating missing values and assessing estimated uncertainties. However, in the procedure for selecting similar stations, the maximum distance parameter has to be chosen by trial and error which is unsatisfactory.

Our goal is to not only introduce a new imputation method which can handle spatio-temporal data but to also provide a reasonable framework for spatio-temporal imputation, which should consist of a complete procedure for parameter optimization, missing pattern simulation and method comparison. The literature on spatio-temporal cross-validation for imputation purposes and missing pattern simulation is rather scarce. To our knowledge, no literature addresses the topic of spatio-temporal cross-validation for imputation purposes and the literature on missing pattern simulation usually considers only data missing completely at random. In this paper, we propose a new framework for spatio-temporal imputation using an integrated imputation algorithm derived from many techniques (including SAA, KNN, SBE, NR, etc.) which takes into account the inherent spatial and temporal information within the data, a cross-validation technique for optimizing relevant parameters in the imputation algorithm, and a simulation procedure for assessing imputation uncertainty. The imputation algorithm is a procedure that can be applied before using statistical models, since the only parameter required to define spatial closeness between space elements is a cutoff for the correlation coefficient (so we name it CUTOFF).

Most imputation methods assume that missing values are missing completely at random (MCAR) or just missing at random (MAR). That is, the presence of missing values on a variable is unrelated to any other observed or unobserved variable (MCAR) or is related to other observed variables but not to its own unobserved values (MAR). In these situations, the missing data generating mechanism is ignorable and the analysis is simplified. We assume that the missing data are MAR given the spatial and temporal locations at which the observations were made or were intended to be made.

We describe and explain the CUTOFF spatio-temporal imputation procedure in Section 2.1. A description of four competing imputation methods based on KNN, SVD, MICE and random forest, chosen because they are representative, well-studied and freely available, is also included in Section 2. Section 3 includes an application of the CUTOFF method and the competing methods to a rainfall dataset from the Murray–Darling Basin in Australia and two other examples. Comparisons are performed within the proposed framework. A specific simulation study is given in Section 4 and we present some conclusions in Section 5.

2. CUTOFF Spatio-temporal imputation approach

2.1. Methodology

The new imputation method extends the idea of the SAA, SBE, and NR methods, extending the ratio equation idea of Aravena and Luckman (2008), to use both spatial and temporal information to impute missing values. To accomplish this, we impute missing values by first determining their spatial neighbors and then defining temporal neighborhoods within these spatial neighbors. To describe the procedure, assume that we have incomplete monthly spatio-temporal data in an $m \times n$ matrix $\mathbf{X}$ corresponding to m months (time points) and n stations (spatial locations). The method extends in a natural way to daily, weekly, quarterly or annual data, etc.

Let $x_{(i,j),k}$ be the observation in month i of year j at station k, for $i = 1, 2, \cdots, 12, j = 1, 2, \ldots, w$ and $k = 1, 2, \ldots, n$, and suppose that a specific observation $x_{(i^*,j^*),k^*}$ is missing. We call the month i^* a candidate month, the year j^* a candidate year and the station k^* a candidate station. We impute the missing observation by means of the following steps:

1. We create a reference list of stations that have a high correlation greater than some cutoff value r with the candidate station. For simplicity, we choose one cutoff value for imputing all missing values in a data matrix. Let L_{k^*} denote the set of reference stations for the candidate station k^*.
2. Let $J_{i,k}$ denote the set of reference years for which $x_{(i,j),k}$ is not missing for month i and station k, excluding year j^*. Let $\overline{R}$ denote the mean value of observations in month i^* across reference years excluding j^* for all reference stations. Let $\overline{C}$ denote the mean value of observations in month i^* across reference years excluding j^* for the candidate station k^*. Let R denote the mean value of observations in month i^* in year j^* for all reference stations. That is,

$$\begin{aligned} \overline{R} &= \frac{\sum_{k \in L_{k^*}} \sum_{j \in J_{i^*,k}} x_{(i^*,j),k}}{\sum_{k \in L_{k^*}} |J_{i^*,k}|} \\ \overline{C} &= \frac{\sum_{j \in J_{i^*,k^*}} x_{(i^*,j),k^*}}{|J_{i^*,k^*}|} \qquad (2.1) \\ R &= \frac{\sum_{k \in L_{k^*}} x_{(i^*,j^*),k}}{|L_{k^*}|} \end{aligned}$$

3. The imputed value $\hat{x}$ for $x_{(i^*,j^*),k^*}$ is calculated by assuming the ratio of the missing value to the mean of observations in candidate months for the candidate station is equal to the ratio of the mean of observations in the candidate month and candidate year for reference stations to the mean of observations in the candidate month for reference stations. That is,

$$\hat{x}/\overline{C} = R/\overline{R}, \qquad (2.2)$$

which yields

$$\hat{x} = R\left(\overline{C}/\overline{R}\right). \qquad (2.3)$$

Remarks:

- A candidate station k^* may have no stations in its reference file L_{k^*} when the cutoff value is too large. In this case, we use the most highly correlated station as the reference station.
- If the observations in the candidate month and year are missing for the reference station(s), then it may be impossible to compute R. This case is handled by searching down the reference list until a station is found that is not missing in the candidate month and year. Here, the reference list includes all stations ordered by their correlation with the candidate station.
- There are two further situations that would cause the imputation method to fail: (1) if an entire station is missing and; and (2) if a certain month is missing for every station. In these cases, we remove the entire station or month, respectively, from the data set.
- There may be situations where there are no neighboring stations with data sufficiently highly correlated with that of the candidate station or where the temporal correlation may be weak. In either case, CUTOFF might perform poorly.

2.2. Cross-validation algorithm

Before Eq. (2.3) can be applied to impute missing values in real data, the cutoff value r has to be determined. Although r can be chosen in a subjective way, we propose using cross-validation to choose the optimal cutoff value. In what follows, we describe the main steps of the cross-validation technique:

1. We create grids of possibly non-adjacent stations and possibly non-consecutive time points as follows: we randomly reorder both the rows and columns of the data matrix $\mathbf{X}$, keeping the labels which carry the location and time information with each observation. As the cutoff imputation method needs the spatio-temporal information to work, keeping the labels with each observation (including the missing observation) is essential.
2. Months (rows) are grouped into p folds and stations (columns) are grouped into q folds. After randomization and grouping, the data matrix consists of pq grids.
3. For a given value of r, and a given grid, in each grid, treat the observed values within the grid as missing and impute them by applying the CUTOFF method to the (observed) data from the remaining grids. The RMSE is then calculated by comparing the non-missing values to the imputed values and averaging over all grids. That is the within-grid RMSE,

$$\mathrm{RMSE} = \sqrt{\frac{\sum (\hat{x}_i - x_i)^2}{N_{\text{non-missing}}}}, \qquad (2.4)$$

where x_i denotes the observed values in this grid, $\hat{x}_i$ denotes the imputed value for x_i, and $N_{\text{non-missing}}$ is the total number of non-missing values in this grid. The RMSE is calculated in every grid and then averaged over pq grids to get the cross-validation error. The mean RMSE across all the grids (called CV-RMSE) is used to measure the imputation performance given r as the cutoff value.

4. For a range of cutoff values, we choose as the optimal cutoff value that gives the smallest CV-RMSE.

The above cross-validation procedure makes the CUTOFF imputation method less subjective. The simulation study in Section 5 provides evidence showing that this cross-validation technique is effective in choosing the optimal cutoff value and also for selecting tuning parameters optimally for some other imputation methods.

2.3. Some options for the CUTOFF method

We also explored some useful further options to extend the CUTOFF algorithm, which are useful in real data analysis. We considered an alternative to the correlation coefficient for finding the reference stations. Instead of using the classical Pearson correlation coefficient, Spearman's rho (ρ) rank correlation coefficient may be a better choice in some cases. For example, Presti et al. (2010) proposed using Spearman's rho to select similar stations for rainfall data. They also found that ρ larger than 0.75 between stations locates the similar stations within 10 km and then chose 0.75 as the threshold value for characterizing reference stations. In later sections, we employ the proposed cross-validation algorithm to find the optimal threshold value rather than setting it subjectively.

Apart from the simple averaging that we demonstrated in the original CUTOFF algorithm, we considered various weighting strategies for imputation. Firstly, rather than give all the points in similar temporal positions (in the monthly setting we mean the same month) equal weights, we can assign weights that decline smoothly with distance from the candidate points when calculating $\overline{R}$ and $\overline{C}$ in Eq. (2.1). We used a Nadaraya–Watson kernel weighting option for this purpose. When smoothing on $\overline{R}$, the kernel weighting process for imputing one missing value can be divided into two steps: apply Nadaraya–Watson kernel smoothing within every reference station separately and then average across all reference stations. Specifically, let

$$f(x_{i^*}) = \frac{\sum_{i \in N_{i^*}} K_\lambda(i^*, i) x_i}{K_\lambda(i^*, i)},$$

and then compute

$$\overline{R} = \frac{\sum_{k \in L_{k^*}} f(x_{i^*})}{N(L_{k^*})},$$

where i^* is the missing position in one candidate station, i represents all the other non-missing positions in the same month as the candidate station, x_i denotes observations in these non-missing positions, and $K_\lambda(\mu) = \frac{1}{\lambda} K(\frac{\mu}{\lambda})$ is the kernel function with bandwidth parameter λ.

We considered several common kernel functions to use with the CUTOFF method, namely, the *Uniform* kernel, the *Epanechnikov* kernel, the *Gaussian* kernel and the *Cosine* kernel. The *Epanechnikov* kernel has the form:

$$K(t) = \begin{cases} \frac{3}{4}(1 - t^2) & \text{if } |t| \le 1 \\ 0 & \text{otherwise} \end{cases}.$$

Apart from the four kernel functions provided, any user-defined kernel function can also be implemented in the R statistical language and then be passed to the CUTOFF algorithm we developed.

Rather than putting equal weights on the stations in the reference set, we can use a weighting strategy proposed by Young (1992) and Yozgatligil et al. (2012) to obtain better estimates. Assume there are h stations in the reference set L_{k^*} of station k^*, let r_{gk^*} denote the correlation between the reference station g and the candidate station k^*, and define the corresponding weight for the station g in L_{k^*} as

$$w_g = \frac{(n-2) \times r_{gk^*}^2}{1 - r_{gk^*}^2}, \quad g \in L_{k^*}.$$

This weight is only valid for those missing observations with more than 3 stations in the reference set; a simple average will still be used for missing observations with less than 3 reference stations.

Another important option that is available in our method is what we call the "adjacent method". In the original CUTOFF algorithm, only concurrent temporal information was used in calculating $\overline{C}$ and $\overline{R}$. In some cases, this may be too strict to get adequate temporal information. So we provide an option which allows the incorporation of adjacent months into the calculation. That is, if a value is missing in July, the observations in June and August (excluding the candidate year) will also be used in the calculation of $\overline{C}$ and $\overline{R}$, not just the candidate month July. We found in some real data analysis that the adjacent option performs almost as well as the original imputation and can be better at times without sacrificing much computational efficiency. We provide two adjacent scale sizes for the CUTOFF method, with scale 1 meaning two adjacent points are considered in imputation and scale 2 meaning that four adjacent points are considered. Generally, we noticed that in real data applications, scale sizes bigger than two did not improve the imputation performance for monthly data. We also found that a size of two is a good choice for monthly data.

When developing the CUTOFF algorithm, we found that some candidate stations have lower levels of correlation with other stations, and often have only one reference station when we use a fixed correlation coefficient cutoff for the whole dataset. We noted this issue in the methodology subsection, where we suggested setting the closest station to be the candidate station's reference station when the optimal cutoff value is too big to find any reference station for that candidate station. We explored another cutoff strategy which, rather than fixing the correlation coefficient, fixes the number of reference stations with respect to the correlation coefficient. The cross-validation procedure can still be used to find the optimal number of reference stations for this alternative method.

These options allow for much flexibility in the CUTOFF methods as most of them can be used individually or in combination with others. For example, the kernel and the adjacent options can be combined with Spearman's correlation strategy. We thoroughly explore these options and evaluate their performance on real data later in the paper.

2.4. Competing imputation methods

We compare CUTOFF with four well-studied imputation methods based on KNN, SVD, MI and random forest, respectively.

2.4.1. KNN Imputation

The k-nearest neighbor method is a non-parametric method for classifying values based on the closest neighboring values. It has been well studied and widely used as a basic machine learning algorithm, and it has also been applied as a useful imputation method. The basic idea of KNN imputation for spatio-temporal data is that a missing value in a target station is imputed by a weighted mean of data from the k nearest stations to the candidate station. The weights depend on the distance of the stations from the candidate station; Euclidean distance is mostly used and is used here.

The choice of the tuning parameter k usually has a large effect on the performance of KNN imputation. Stekhoven and Bühlmann (2012) proposed a cross-validation algorithm to obtain a suitable k by artificially introducing missing values to the validation sets; this is an alternative to the cross-validation we propose. We will use our cross-validation algorithm to find the optimal value of k. For implementation, we use the knnImputation function in the R package **DMwR** (Torgo, 2010).

2.4.2. SVD imputation method

When analyzing spatio-temporal data, it is common to model the observations as the sum of a spatially varying trend component and a spatio-temporal residual component. Empirical orthogonal functions (EOFs) can be used to model the trend component, with the goal of using as parsimonious a set of EOFs as possible to describe the leading modes of variability in the data. This is done by singular value decomposition (SVD). However, the SVD cannot be applied directly when the data matrix contains missing observations. With the intention of using SVD to get the EOFs, and facing the problem of missing values, Fuentes et al. (2006,) propose an algorithm to compute the SVD and impute the missing values iteratively. The method involves an iterative process of matrix decomposition and regression, and has been applied to deal with missing values in a complex spatio-temporal model for air quality data (Lindström et al., 2013b).

The tuning parameter is the number of singular vectors v, which can be chosen subjectively to be some small value or chosen objectively by our cross-validation procedure. The power of SVD, its nonparametric nature, computational efficiency and the pre-modeling functionality make this method popular among spatio-temporal analysts. Since the left singular vectors represent the spatial patterns, this method might work well when most of the monitoring sites share some analogous temporal patterns, giving a parsimonious set of basis functions to approximate the trend component. However, when observations vary substantially about the trend, the SVD approximation is a noisy representation of the real patterns and this method would potentially give poor results. As with the CUTOFF method, the SVD imputation method will fail if entire rows and/or columns are missing from the data matrix. For implementation, we use the SVDmiss function in the R package **SpatioTemporal** (Lindström et al., 2013a).

2.4.3. Multiple imputation

In multiple imputation, each missing value can be estimated by a set of plausible values and the final imputed values decided by “pooling” or “averaging” these values. Multiple imputation was originally developed for survey data (Rubin, 1987) and has been widely applied to handle complex missing value problems in diverse fields. Although it is not designed particularly for spatio-temporal imputation, it is a competing method because of its reportedly good performance and its availability. For example, S-PLUS, Stata, SPSS, SAS and R, all have implementations and it is often considered to be the standard missing value processing module in these software packages. In this work, we choose one of the MI methods called Multivariate Imputation by Chained Equations (MICE), which combines the two general approaches for imputation: joint modeling and a fully conditional specification. For a comprehensive introduction to software for multiple imputation and MICE, detailed procedures and some successful applications, see Buuren and Groothuis-Oudshoorn (2011) and Yozgatligil et al. (2012).

The function mice in the R package **mice** is used to implement this algorithm. An attractive feature of this function is that when imputing one column, instead of conducting a massive imputation, in which all other columns are used as predictors, a parsimonious MICE can be employed to improve efficiency. The way to implement this parsimonious version of MICE is to set a predictor matrix for the MICE algorithm which indicates which columns of predictors to use for specific missing variables. Not surprisingly, the choice of predictor matrix (called a *minor* parameter) applies a cut-off to the correlation coefficient. However, this choice does not have a substantial effect on the accuracy of the MICE algorithm. It is suggested that 0.5 is a reasonable choice and is the value used in our comparisons.

2.4.4. Imputation by random forest

Random forest (Breiman, 2001) is a popular machine learning algorithm that has been applied to regression and classification problems. The basic idea of random forest is to build a large collection of de-correlated regression trees to bootstrap versions of the training data, and average the result (Hastie et al., 2009). It has recently been extended to solve the missing data problem (Hapfelmeier, 2012; Stekhoven and Bühlmann, 2012). For a detailed implementation and a comprehensive algorithmic description of the function in R, refer to Stekhoven and Bühlmann (2012).

The missForest function in the R package **missForest** has two important parameters. The first parameter *ntree* is the number of trees to grow in each forest, the other parameter *mtry* is the number of variables randomly sampled at each tree split. The original intent of random forests is to choose these two parameters internally by checking the out-of-bag (OOB) error. We will use this recommended error checking to determine a good combination of *ntree* and *mtry*, and double check with our cross-validation choices.

3. Data analysis

3.1. Monthly rainfall data in Australia

The study area is the Murray–Darling Basin in southeastern Australia. The name of this basin is derived from its two major rivers, the Murray River and the Darling River. This basin is the most important agricultural area in Australia, producing over one-third of Australia’s food supply. Monthly rainfall data are available between the 1880s and 2010 with large sparse coverage during the early period. Because of the sparsity, we reduced the original data to a 100-year period, from 1911 to 2010. The original data sets were arranged in a $1200 \times n$ data matrix, where $n = 78$ is the number of gauging stations. All observed values were cube-root transformed because the data for nearly every station showed skewness. Using qq-plots, we explored several transformations (square-root, cube-root and log) and found that, as suggested in the literature, that the cube-root transformation worked best, both in terms of skewness and normality. There are many suggestions to perform cube-root transformation on rainfall data in the literature. For example, Garwood (1936) and Merrington (1941) found that the cube-root transformation is empirically useful for many practical problems and “better than Fisher’s square root transformation”. Howell (1965) found that the cube-root transformation was useful in dealing with rainfall data; see Kendall (1960) and Fu et al. (2010) for more applications.

Some detailed information about the missing data is provided in Table 1. A heat map of the data matrix for the Murray–Darling Basin is given in Fig. 1. Fig. 2 is a sunflower plot indicating each station’s geographical location and the number of missing observations. Some remarks regarding the missing portions of this data set are as follows:

- Fig. 1 shows that there is a significant number of missing values in our data. A large proportion of missing values are at the top section of the heat-map, indicating that most

Table 1
Missing proportion for monthly rainfall data in the Murray–Darling basin.

Proportion of missing	Number of stations
10% or more	2
[5%, 10%)	3
[2.5%, 5%)	8
[1%, 2.5%)	8
Less than 1%	33

missing values occur in the recent period. In the middle period, the data appear to be collected and recorded very well and there are approximately 50 years of complete data in this period. A few stations have missing values in the early period. In particular, station X051033 has 144 missing values all in the early period.

- Another missing pattern called "block missing" occurs when consecutive observations are missing, seen as yellow blocks in the heatmap. There has been much interest in whether imputation methods can successfully deal with these missing blocks or "gaps" in the data and it is another reason that more advanced imputation methods are desirable for filling spatio-temporal data (Kondrashov and Ghil, 2006; Lou and Obradovic, 2011).
- In Fig. 2, it appears that the missing value problem is more serious in the northern region of the basin than in the southern region. This implies some spatial correlation in the missing pattern.
- From Table 1, we see that of the 78 stations, 54 have missing values and 24 are complete. Most stations have less than 1% missing values (33 stations), three stations have between 5% and 10% missing and two stations have more than 10% missing.

3.2. Application of the CUTOFF Method

3.2.1. The default setting

The default CUTOFF method has a single parameter which is the cutoff value (r) for the correlation between stations. For the Murray–Darling Basin data, we consider a range of cutoff values from 0.55 to 0.95 in increments of 0.01. We performed cross-validation on both the rows and columns of the data matrix with the rows (months) grouped into p folds with three values for p, namely, 10, 20 and 100, corresponding to time periods of 10, 5, and 1 year, respectively, and columns (stations) grouped into $q = 10$ folds. Details are given in Table 2. Fig. 3 displays a plot of the CV-RMSE against the cutoff value for the three different grid sizes for the months. We can see from this plot that for all 3 grid sizes, the average CV-RMSE is minimized at a cutoff value of approximately $r = 0.75$. Further, the CV-RMSE seems to plateau with small fluctuations when the cutoff value is larger than 0.9. This is reasonable as the reference sets for each station do not change much when the cutoff value becomes large and remain unchanged when the cutoff value exceeds a threshold. So, for this dataset, the optimal cutoff value is chosen to be 0.75.

Fig. 4 shows the imputed values for the three stations with the largest number of missing observations. The top station (X051033) has all missing values in the beginning period; the middle station (X041079) has missing values mostly between 1988 and 1998, and also some at the end; the bottom station (X055063) has missing values mainly at the end. We can see from these reconstructed time series that the CUTOFF method seems to impute the missing values quite well, especially in the large gaps that exist at the beginning and end periods of the data. The imputed values, shown in red in Fig. 4, seem reasonable in the sense that, for each station, the variability in the imputed values is very similar to that of the non-missing values. Furthermore, some single point imputations,

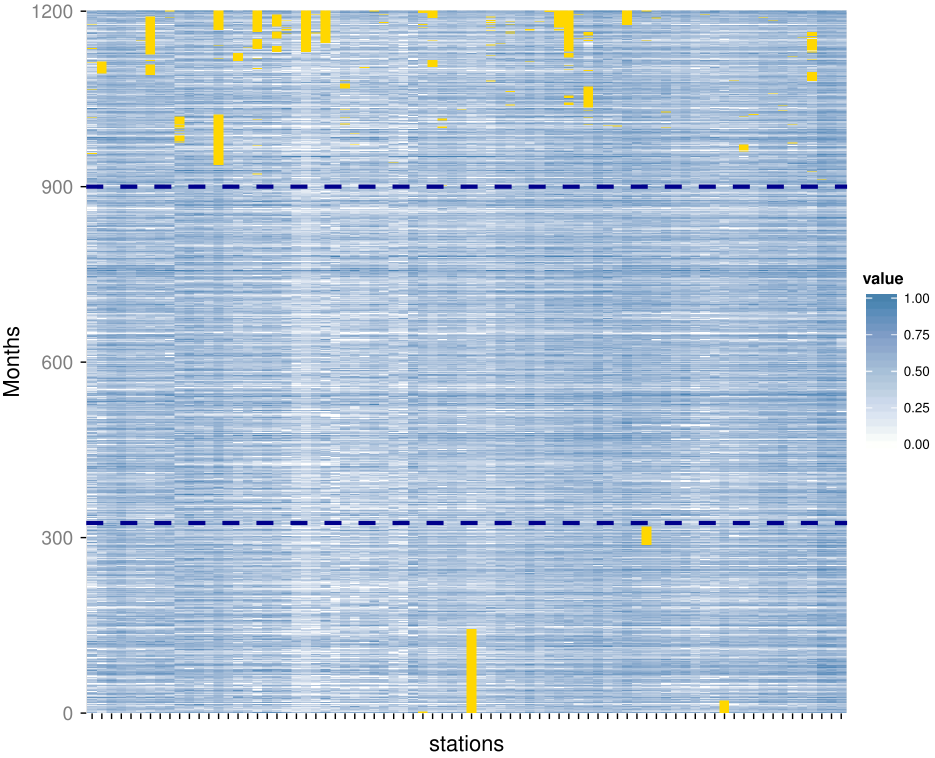

Fig. 1. A heatmap of the Murray–Darling Basin rainfall data matrix. The 78 stations are in the columns and the 1200 months are in the rows. Stations are ordered according to their station ID from the database. Rainfall values have been normalized to be between 0 and 1. Missing values are indicated in yellow and the middle section between the two horizontal dashed lines signifies a large contiguous portion of the data that contains no missing values. The lines separate the months into the early, middle and recent periods, corresponding to the bottom, middle and top part of the heatmap, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

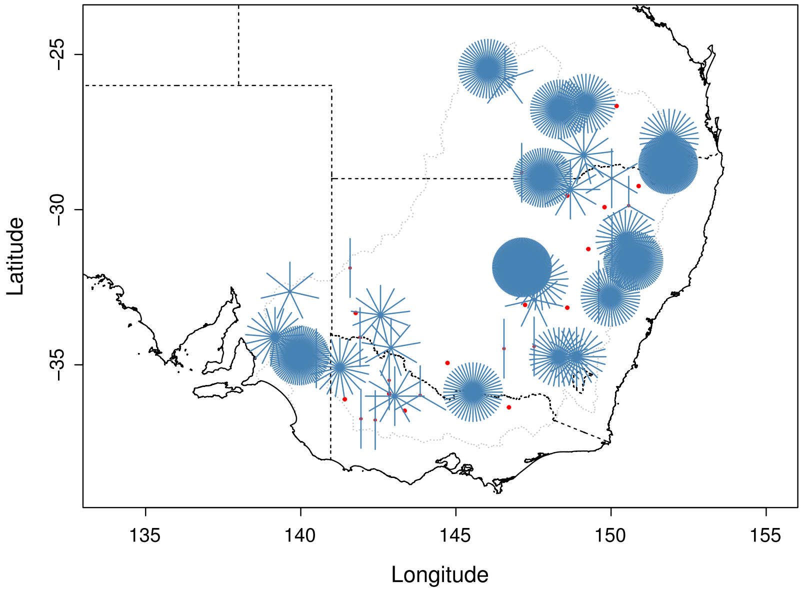

Fig. 2. Geographical map combined with a sunflower plot demonstrating the missing profile of the data. The number of missing observations at each station is indicated by the overlaid "sunflower". For a given station, the greater the number of "petals" on the sunflower, the greater the number of missing observations. For more information about sunflower plot, see Schilling and Watkins (1994).

Table 2
Specification of the partitions for cross-validation.

p (year size)	*q*	Number of grids
100 (1 year)	10	1000
20 (5 years)	10	200
10 (10 years)	10	10

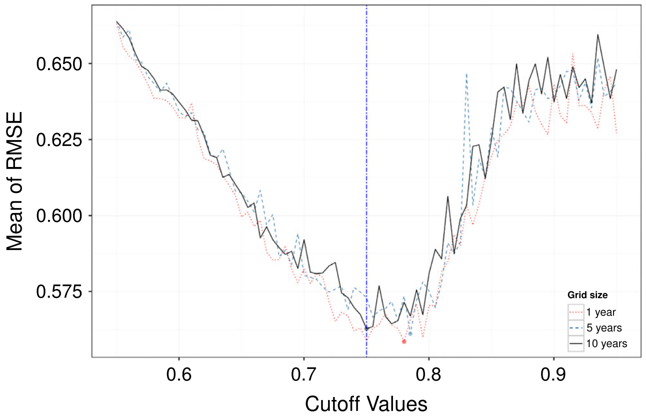

Fig. 3. The average CV-RMSE for varying cutoff values for three different grid sizes for the months.

such as those around 2005 in the middle station and around 2003 in the bottom station, are in line with the surrounding values (see Fig. 5).

3.2.2. Using options with the CUTOFF method

The original CUTOFF method described above is an intuitive and simple approach to impute missing values using both spatial and temporal information in a two-step fashion. The observed values derived from both dimensions are averaged to estimate the missing value. This method has performed well on imputing the Murray–Darling Basin rainfall data. However, by experimenting with the options that were described in Section 2.3, we find that there are many potential ways to improve the accuracy of the method. We describe here some possible options and option combinations, using our cross-validation procedure to compare the performance of these options on the same data matrix. We consider the default setting which uses CUTOFF by correlation as the benchmark method. Table 3 shows the mean RMSE (CV-RMSE) and the standard error of the RMSE (SE) for every option considered. We used a one-tailed Wilcoxon test to evaluate the improvement over the benchmark method.

We found that the benchmark method is difficult to significantly improve. The "adjacent" option attempts to use more temporal information by including more observed values from the missing month's adjacent months. However, this achieves only marginal improvement over the benchmark method. With the kernel method, we tried different kernel functions with optimal bandwidth values (λ) chosen by the same cross-validation strategy. Although the benchmark method outperforms the kernel options, kernel options are slightly more stable in terms of standard error.

The "CUTOFF by number" option, which, instead of cutting off by the correlation coefficient uses a fixed number of reference stations, gives 1.58% improvement over the benchmark method, and combined with the adjacent option gives improvements over the benchmark method of 2.15% and 2.22%. Wilcoxon tests indicate that the improvements are significant under the 5% significant level. For the benchmark method, we notice that the smallest number of reference stations for any candidate station is 1, the largest number of reference stations is 12 and the median number of reference stations is 6. Therefore, the underlying reason for the "CUTOFF by number" options outperforming the "CUTOFF by correlation" might be that more missing stations have been overfitted than under-fitted. Motivated by this insight, we then considered the "penalty" and the "relaxation" options. The penalty option fixes the number of reference station to *PN* for those candidate stations with more than 4 reference stations and the relaxation option relaxes the number of reference stations to *M* for those with less than 4 reference stations. However, from the results in Table 3, these two options perform no significantly better than the benchmark method, and actually significantly worse than the simpler "CUTOFF by number" option, which does not penalize or reward any stations.

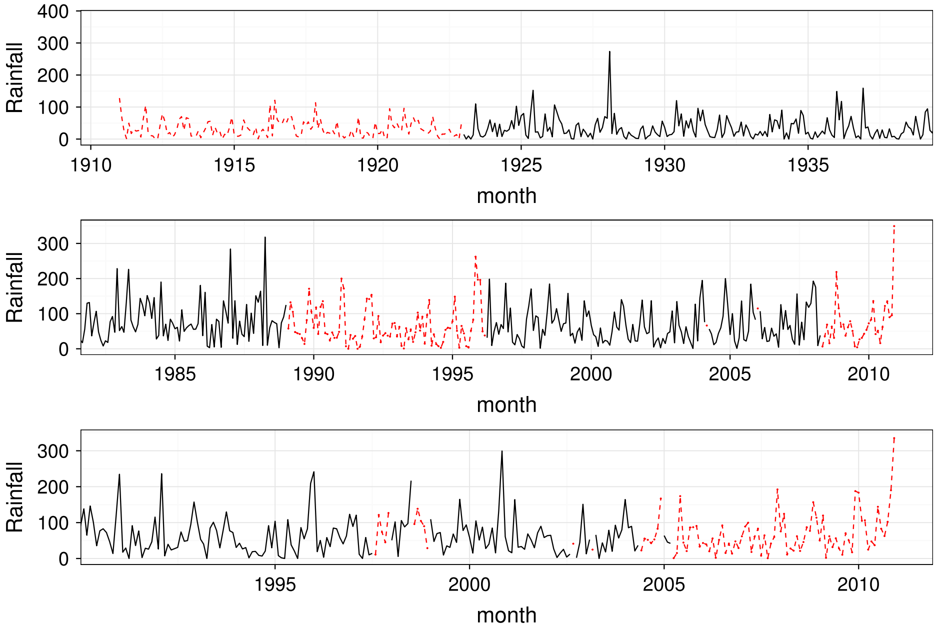

Fig. 4. Reconstruction of the time series using imputations from the CUTOFF method (red) for the three stations with the most missing observations, i.e., stations X051033, X041079 and X055063. Note that, for each time series plot, the time range has been chosen to highlight the missing period. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

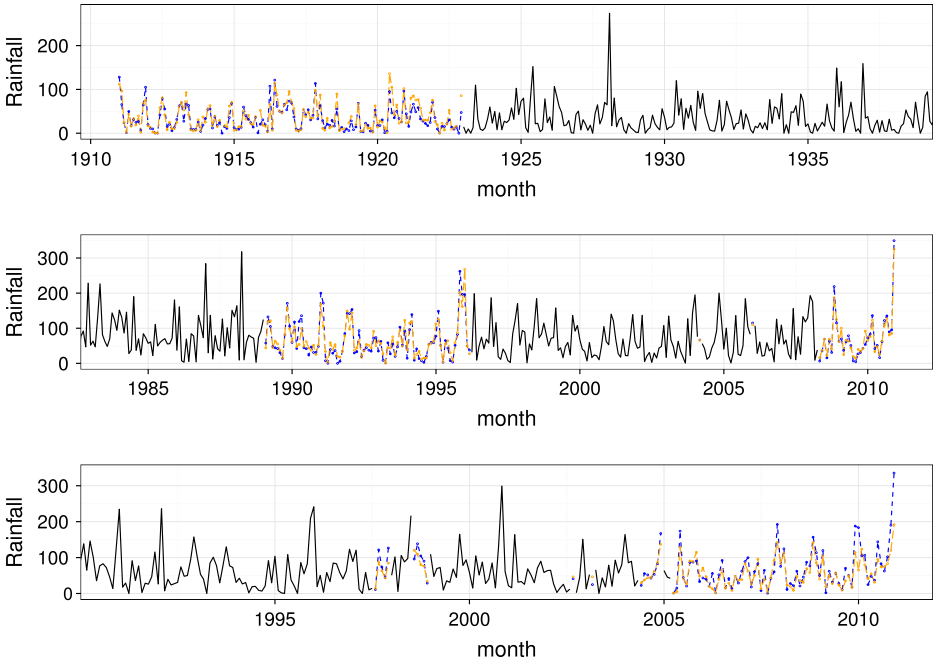

Fig. 5. Comparison of the imputations using the SVD method (yellow dashed line) and the CUTOFF method (blue dashed line) for the three stations with the most missing observations. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

We found that more complicated options like kernel smoothing, spatial weighting and penalizing work no better than the simple default method on the Murray–Darling Basin data. Moreover, these options may potentially improve imputation performance on different data.

3.3. Comparison of methods

3.3.1. Choosing parameters via cross-validation

We chose four imputation methods to compare to the benchmark CUTOFF method. There are some key parameters in these methods that need to be specified beforehand. They are the nearest neighbor number k for the KNN method, the number of singular vectors v for the SVD method, the pooling size m for the MICE method and two parameters for random forest imputation: *mtry* for the number of variables randomly sampled at each tree split and *ntree* for controlling the number of trees to grow in each forest. We used our cross-validation algorithm to find the best k and v over a range of values from 1 to 20 for the KNN and the SVD method, respectively. Fig. 6 illustrates how the cross-validation error changes with varying parameter values for both KNN and SVD. For the KNN method, the CV-RMSE decreases from $k = 1$ to

Table 3
The performance of the CUTOFF method with different options for imputing the Murray–Darling Basin rainfall data. The benchmark method is the default method, CUTOFF by correlation, and is denoted by R. A is the number of adjacent points for the adjacent option (denoted by Adj), N is the number of fixed reference stations for the "CUTOFF by number" option (denoted by Num), PN is the penalty number on those which have more than four reference stations (penalty option is denoted by Pen), M is a relaxation operation on those which have less than four reference station (relaxation option is denoted by Relax), "space.weight" implements spatial weighting and "Kernel" indicates that kernel smoothing is used for temporal smoothing (kernel method is denoted by Kernel (a, λ) where a is the kernel used and λ is the bandwidth parameter for kernel method). The symbol "+" denotes a combination of options. *EpanK* refers to the *Epanechnikov* kernel. In the "Improvement" column, only significant improvements are shown. Non-significant ones are denoted by "–".

Options	CV-RMSE	SE	Improvement
R $(r = 0.75)$	0.5623	0.0435	(*Benchmark*)
R + Adj $(A = 1)$	0.5592	0.0434	–
R + Adj $(A = 2)$	0.5588	0.0431	–
R + Adj $(A = 3)$	0.5589	0.0431	–
Num $(N = 4)$	0.5534	0.0476	1.58%
Num $(N = 4)$ + Adj $(A = 1)$	0.5502	0.0474	2.15%
Num $(N = 4)$ + Adj $(A = 2)$	0.5498	0.0472	2.22%
R + Pen $(PN = 3)$	0.5549	0.0468	–
R + Pen $(PN = 4)$	0.5579	0.0499	–
R + Pen $(PN = 5)$	0.5574	0.0452	–
R + Pen $(PN = 4)$ + Relax $(M = 3)$	0.5554	0.0467	–
R + Pen $(PN = 4)$ + Relax $(M = 3)$	0.5534	0.0476	–
R + space.weight	0.5742	0.0698	–
R + Kernel (*EpanK*, $\lambda = 50$)	0.6547	0.0420	–
R(corr = "Spearman")	0.5699	0.0545	–

10, plateaus after 10 and does not increase even after 20. The CV-RMSE for the SVD method similarly decreases dramatically from $v = 1$ to 10, fluctuates between 10 and 17, but tends to increase after 17. The increase after $v = 17$ indicates that any v larger than 17 will overfit the data. The value 10 is approximately optimal for both the KNN method and the SVD method for imputing the Murray–Darling Basin rainfall data.

For the MICE method, we suggest using $m = 5$, not only because it is widely used and recommended in the literature, but also because the MICE method is very time-consuming compared to KNN and SVD. For the random forest method, we tried different pairs of values for *mtry* and *ntree* by checking the out-of-bag error (OOB error), which is provided as an internal test set error estimator for random forest. According to Breiman (2001), from each bootstrap training set, about one-third of the values are left out and not used in the construction of the trees (thus called out of bag), and these values are used for testing the error. He also provided a justification for using out-of-bag estimates to monitor the test error of random forest, emphasizing that the out-of-bag error is unbiased, unlike the cross-validation error estimates. We tried 25 combinations of *mtry* and *ntree* values and the OOB error and the time taken are recorded in Table 4.

Table 4 shows that the larger the two parameters, the smaller the OOB error. But it is also noticeable that the larger the two parameters, the longer the time taken for a single imputation. Stekhoven and Bühlmann (2012) used the same strategy to find the best values for the two parameters. They found that when *ntree* is 500 and *mtry* is bigger than the recommended value (one third of the number of columns, which in our case is 26), the OOB errors tend to decrease, which is an indication that the random forest imputation method overfits when the two parameters are unnecessarily large. However, in our case, there is no evidence of over-fitting. Even when *ntree* is 2000 and *mtry* is 50, the OOB error continuous to decrease. Generally, we consider $ntree = 100$, $mtry = 8$ is a reasonable choice. In view of the computational burden of random forest, we do not recommend using our own cross-validation algorithm to find the best parameter values.

With spatio-temporal data, it is arbitrary whether we assign time or space to the rows or columns of the data matrix. That is, we can apply random forest to analyze the data matrix or its transpose. When we applied random forest imputation to the transposed of the Murray–Darling Basin rainfall data matrix, some dramatic changes occurred, indicating that the OOB error may not be a good error estimator for this imputation problem. Table 5 lists the results. Compared with Table 4, we can see that transposing the original data matrix strongly influences both imputation error and computational time, in that both components are drastically reduced. Both the OOB error and computational time have been reduced significantly. To explain this, we applied our cross-validation procedure to both the original data matrix and the transposed data matrix using $ntree = 100$, $mtry = 8$, and found that the cross-validation error on the transposed data matrix is slightly larger than the cross-validation error on the original data matrix, the opposite of what we observed in Tables 4 and 5. Actually, the OOB error may be a good estimator when the original data are independent, but it is no longer valid for spatio-temporal data, and cross-validation might work better than OOB here. However, we still use $ntree = 100$, $mtry = 8$ for later analysis when using missForest.

3.3.2. Comparative results

The results from applying cross-validation to the CUTOFF method and the four other competing methods on the Murray–Darling Basin rainfall data matrix are shown in Table 6. The advanced imputation methods are better than naive imputation using mean or median values, at the cost of computational time.

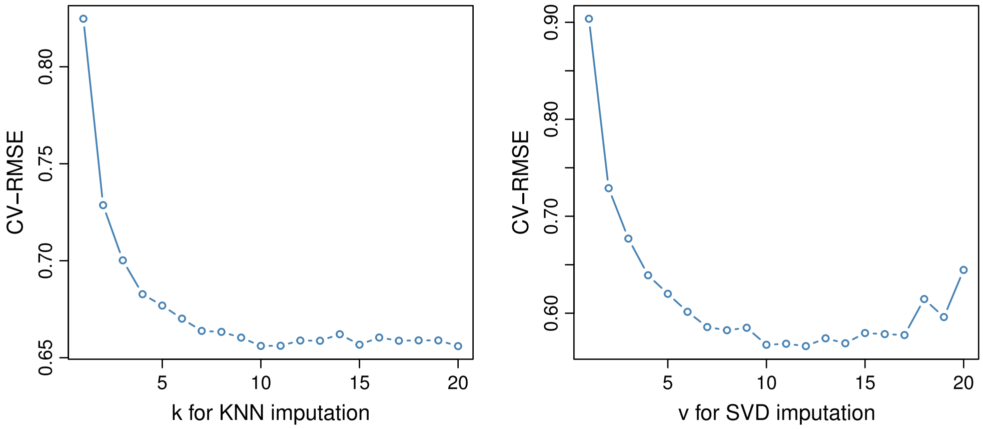

Fig. 6. Cross-validation results for choosing the optimal parameters for the KNN and the SVD imputation method.

Table 4
OOB error and runtime for different value of *ntree* and *mtry* applied to the data matrix of the Murray–Darling Basin rainfall data.

mtry	*ntree*				
	10	50	100	250	500
1	0.5408	0.4582	0.4484	0.4417	0.4399
	8.5 s	54.3 s	83.9 s	317.3 s	629.7 s
2	0.5003	0.4239	0.4157	0.4100	0.4083
	17 s	68.2 s	106.1 s	263.5 s	530 s
4	0.4712	0.3968	0.3886	0.3886	0.3818
	22.3 s	76.3 s	186.7 s	558.2 s	744.2 s
8	0.4416	0.3750	0.3673	0.3625	0.3613
	33.2 s	118.7 s	351.8 s	1168.1 s	1462.6 s
16	0.4212	0.3609	0.3535	0.3501	0.3485
	54.6 s	253.4 s	502.7 s	1253.7 s	2514.2 s

Table 5
OOB error and runtime for different value of *ntree* and *mtry* applied to the transposed data matrix of the Murray–Darling Basin rainfall data.

	ntree				
mtry	10	50	100	250	500
1	0.3812	0.3202	0.3139	0.3100	0.3086
	8.4 s	20.3 s	109.7 s	109.7 s	159.6 s
2	0.3650	0.3077	0.3001	0.2973	0.2953
	13.9 s	21.8 s	64.0 s	181 s	292.9 s
4	0.3563	0.2986	0.2925	0.2885	0.2872
	9.1 s	33.0 s	59.3 s	141.0 s	275.4 s
8	0.3495	0.2933	0.2870	0.2828	0.2820
	20.4 s	40.8 s	56.0 s	223.7 s	351.4 s
16	0.3431	0.2898	0.2834	0.2796	0.2785
	16.6 s	69.2 s	103.8 s	315.1 s	871.7 s

Table 6
Comparative results for cross-validation tests. CV-RMSE is the cross-validation RMSE for comparing the imputation accuracy and SE is the standard error of RMSE. Elapsed time is the runtime for each cross-validation procedure, performed on a Macbook Pro running Mac OS 10.8, with a 2.6 GHz Intel i7 quad-core CPU and 16 GB RAM.

Method	CV-RMSE	SE	Elapsed time (s)
CUTOFF($r = 0.75$)	0.5623	0.0435	53.9
KNN($k = 10$)	0.6562	0.03329	364.9
SVD($v = 10$)	0.5671	0.04218	321
MICE($m = 5$)	0.5704	0.04603	4166
missForest(*ntree* = 100, *mtry* = 8)	0.5405	0.04418	11780
Mean	1.1654	0.0599	4.7
Median	1.1683	0.0615	4.9

The KNN and the SVD methods have similar runtime, although the SVD method is slightly faster and more accurate than the KNN method. MICE and SVD perform similarly in terms of accuracy, but the MICE method takes much longer than the SVD method. Our CUTOFF method is slightly more accurate than the SVD method and is about 6 times faster. The missForest method has the smallest cross-validation error but it is very slow. For one cross-validation run, it takes about 3.5 h, which is unlikely to be practical in real data analysis.

There are two possible ways to speed up the computation time for missForest. As described in Stekhoven and Bühlmann (2012), the first is to reduce the number of trees and the second is to reduce the number of variables resampled. However, using both either separately or in combination would diminish its slight accuracy advantage over the CUTOFF method, and it is still much slower than the CUTOFF method. It is also possible to speed up the KNN and the SVD methods by reducing k and v, but we can see roughly from Fig. 6 that there is a loss of accuracy.

Overall, the CUTOFF method with the default setting performs better than almost all the competing methods in terms of accuracy and computation time. missForest can do slightly better than the CUTOFF method in accuracy, but with greater demand for computational resources. The SVD method is a good alternative to our method when the data matrix is not too large.

3.4. Further examples

We also assessed the performance of the CUTOFF method and the four competing methods on two more monthly rainfall data sets from the southwest coast division (SWC) and the northeast coast division (NEC) of Australia. To be consistent, the study period has been set to be the same as the previous Murray–Darling Basin data, that is, the 100 years from January 1911 to December 2010. The cube-root transformation was applied to both data sets. There are 45 gauging stations in the southwest coast division and 775 observations are missing. In the northeast coast division, there are 22 gauging stations and 526 observations are missing. To find the optimal parameter for the CUTOFF, the KNN and the SVD methods, we used cross-validation. Additionally, we chose the "CUTOFF by number" option which we described and discussed in Section 2.3, to extend the CUTOFF method. Cross-validation was also applied to find the best N for this option. Figs. 7 and 8 illustrate the cross-validation results for the southwest coast division and the northeast coast division, respectively. For the south west coast division data, the optimal cutoff value for the correlation coefficient in the default CUTOFF method is 0.85, the optimal N for the "CUTOFF by number" method is 4, the optimal value for the KNN method is $k = 9$ and the optimal value for the SVD method is $v = 5$. For the northeast coast division data, the optimal values for these four methods are 0.7, 5, 6 and 3, respectively. Because there are only 22 stations in the northeast coast division data set, the cross-validation results tend to be unstable. It is difficult to choose an optimal value for the KNN method by looking at the cross-validation plot, although it is still straightforward to pick up minimum points from the other three cross-validation plots.

These optimal choices are then used when we compute the cross-validation error. The results are given in Table 7. missForest is the best imputation method for both the southwest coast division data and the northeast coast division data. The default CUTOFF method performs relatively well on the southwest coast division data but not as well as the other competing methods for the northeast coast division data. The "CUTOFF by number" method is the second best method for imputing the southwest coast division data but is the worst for the northeast coast division data. The default CUTOFF method is the fastest method for both data sets, but it seems that the CUTOFF method is not as powerful for imputing small data sets. We consider the SVD method as the most competitive method to our method in view of its stable performance in both accuracy and computation time.

4. Simulation analysis

We conducted some simulations to compare the performance of the CUTOFF method with the four competing methods. The principle of simulation is to generate multiple data sets that resemble the real data. This is quite difficult when we are simulating data with nonrandom missing patterns. We propose a simulation procedure which, to some degree, is able to approximate the original missing pattern. Although the procedure is designed particularly for the Murray–Darling Basin rainfall data, the idea is flexible and can be extended to simulate data with other missing patterns.

The simulated data sets are derived from a reduced portion of the real data that contains no missing values. As described in Fig. 1, the middle portion of data between the two dashed lines has no missing values and can therefore be used for simulation

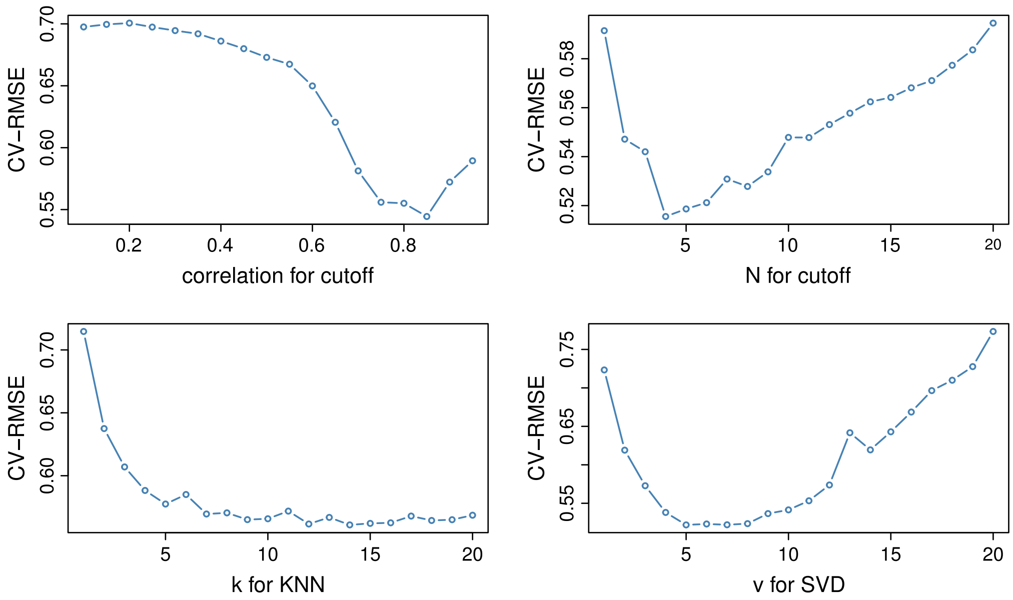

Fig. 7. Cross-validation results for finding the best parameter choices (southwest coast division).

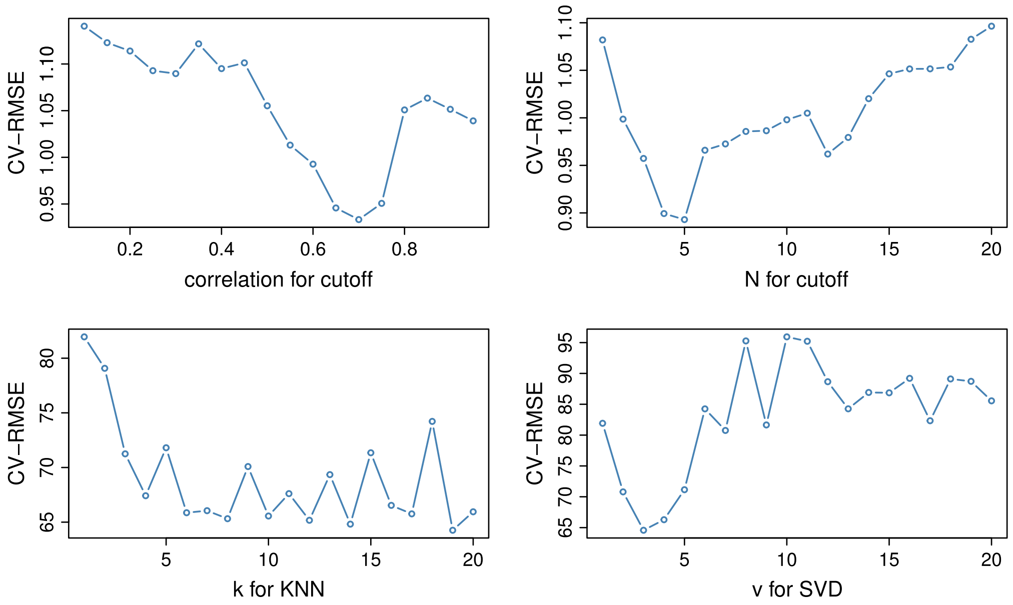

Fig. 8. Cross-validation results for finding the best parameter choices (northeast coast division).

Table 7
Cross validation error (RMSE) and runtime results for applying the imputation methods on two areas in Australia. SWC denotes the southwest coast division and NEC the northeast coast division.

Methods (with optimal parameters)	CV-RMSE		SE		Runtime	
	SWC	NEC	SWC	NEC	SWC	NEC (s)
CUTOFF by Corr	0.5427	0.9389	0.0962	0.1475	31.3 s	14.7
CUTOFF by N	0.5132	0.8958	0.0812	0.1493	65.7 s	33.6
KNN	0.5714	0.9390	0.0802	0.1344	165.9 s	65.78
SVD	0.5367	0.916	0.0682	0.0939	120.7 s	23.7
MICE	0.5228	0.9172	0.0576	0.135	2587.3 s	626.0
missForest	0.4991	0.8616	0.068	0.131	4.5 h	5266.6

purposes. The goal is then to induce missing observations in this reduced portion of data in a manner that closely mimics the missing pattern in the original data. Fig. 9 shows a heatmap of the original data matrix with the stations reordered according to the number of missing values. The data matrix has also been divided into four sections with distinct missing patterns. Specifically, the left top section (S1) has the most missing values, the section below (S2) has some blocks of missing values. There are some scattered

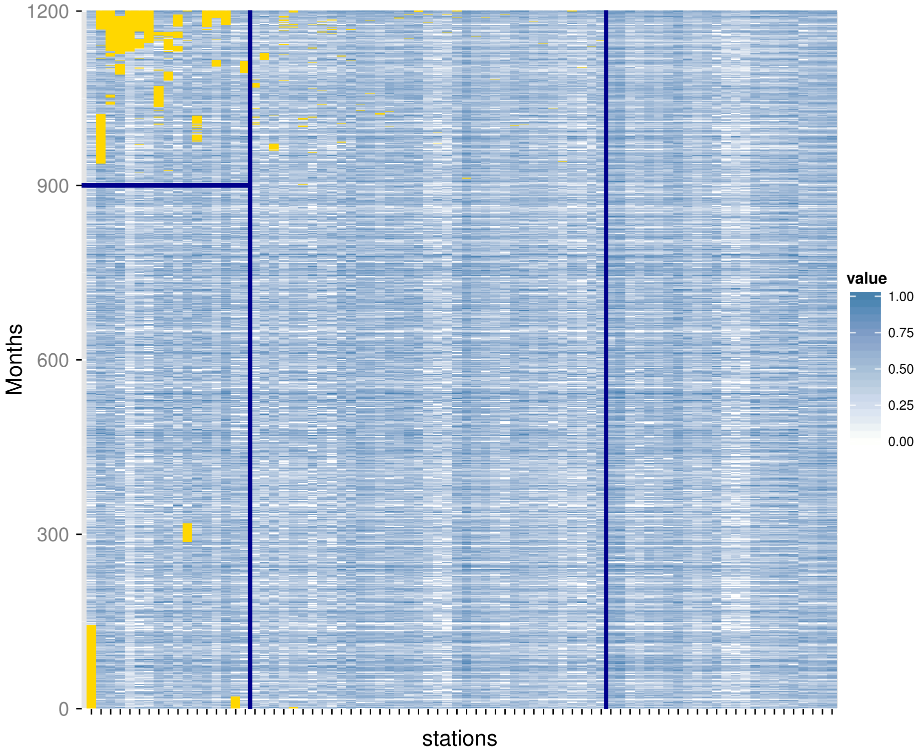

Fig. 9. A heatmap of the data matrix with stations (columns) reordered by the number of missing observations. The data matrix has been partitioned into four sections. Each section represents one type of missing pattern that is simulated separately.

small missing blocks in the middle section (S3), and the rightmost section has no missing values (S4). The simulation will be done under what we called "block missing", with the only differences in terms of missing patterns between these four sections being the block size and the probability of an observation being missing. For example, compared to the other three sections, observations in S1 are more likely to be missing and the block size in S1 tends to be larger. Also, S2 and S3 are similar both in block size and missing probability, and observations in these two sections are more prone to be missing than observations in S4. Therefore, to generate the missing pattern in the reduced portion of the data with no missing values, this portion will similarly be divided into four sections. For each section, the missing pattern will be generated separately and then combined to produce the final missing pattern.

To simulate the missing values within each section, we generate a vector V of 1s and 0s for each station in the section (with 1 indicating missing). As most missing observations tend to occur in blocks, each 1 or 0 represents three consecutive months. That is, missing observations are simulated in groups of 3 months. Further, the 1s and 0s will be simulated such that if a group of 3 months is missing, then the probability that the next group of 3 months is also missing increases. There are four parameters, listed below, that control the simulation of each missing vector V, and the details of the simulation are given in Table 8.

1. n: length of the missing vector V.
2. $maxlen$: maximum length of a missing block (i.e., maximum length of a string of consecutive 1s in V).
3. $prob$: probability that a given element of V is equal to 1 when the previous element is equal to 0.
4. $cnst$: a constant used to control the increase in the probability that a given element of V is equal to 1 when the previous elements are a string of 1 s. $prob^*$ is the increased probability, defined as: $prob^* = \frac{count+cnst}{maxlen+cnst}$, in which $count$ is an intermediate quantity in the algorithm that we will explain below.

Algorithm 1. Simulating V

```
Initialize $V$ to be an $n \times 1$ vector of 0 s.
Set $V[1] = 1$ with probability $prob$, where $V[i]$ denotes the
  $i$th element of $V$, and set $count = V[1]$.
for $i = 2$ to $n$ do
    If $count > maxlen$ or $count = 0$, set $V[i] = 1$ with
      probability $prob$.
    If $0 < count < maxlen$, set $V[i] = 1$ with probability
      $prob^* = \frac{count+cnst}{maxlen+cnst}$.
    If $V[i] = 0$ set $count = V[i]$, else set $count = count + V[i]$
end for
```

Fig. 10 shows an example of the simulated missing data matrix, as produced by Algorithm 1, in the reduced portion of the data with no missing values. The parameter values used to simulate the missing values in each section are listed in Table 8. The simulated dataset shown in Fig. 10 has an obvious block missing pattern in the top-left section and some random missing values scattered throughout the rest of the data matrix. Compared with the original data matrix in Fig. 9, the algorithm has done reasonably well in simulating the original missing pattern.

In order to assess the relative performance of the various imputation methods, we will conduct a simulation study based on Algorithm 1. The imputation methods are then used to complete the missing data. For each simulated matrix, the prediction error is obtained by calculating the difference between the imputed values and the true values for the simulated missing observations. Here we still use root mean squared error (RMSE) as a measure of prediction accuracy. RMSE is used widely in statistics to evaluate the performance of predictive models. In the literature of imputation, including hydrological imputation, it is also considered one of the standard measures of imputation accuracy. For example, some

Table 8
The parameters used to simulate the missing data in each of the four sections of the reduced portion of the data with no missing values.

Section	*n*	*maxlen*	*prob*	*cnst*	Years	Stations
S1	48	12	0.05	12	12	17
S2	432	3	0.005	3	36	17
S3	576	3	0.005	3	48	37
S4	576	3	0.0005	3	48	24
Total	–	–	–	–	48	78

Table 9
Comparison between the CUTOFF method and the four competing methods in terms of mean RMSE and the standard error of the RMSE.

	Measure		
Method	RMSE	SE	Runtime (s)
CUTOFF	0.5658	0.0198	58.3
KNN	0.7141	0.0204	171.7
SVD	0.5794	0.0170	115.1
MICE	0.5918	0.02	11556.3
missForest	0.5654	0.0190	31723

recent work in hydrological imputation by Teegavarapu (2014) and Ferrari and Ozaki (2014) choose RMSE as the primary measure of imputation performance. The adoption of RMSE can also be found in the cited work of these two papers. In each imputed data matrix, let $\hat{x}_i$ denote an imputed value, x_i denote the corresponding real value and N_{missing} be the number of missing values in that matrix. The simulation RMSE for that imputed data matrix is then defined as:

$$\text{RMSE} = \sqrt{\frac{\sum(\hat{x}_i - x_i)^2}{N_{\text{missing}}}}.$$

We constructed 500 simulated matrices and calculated the average RMSE for each imputation; the results are summarized in Table 9. The missForest method is slightly better than our CUTOFF method but is extremely slow. The CUTOFF method performs similarly to the SVD method and is slightly better than the MICE method. In terms of computation time, the CUTOFF method is the fastest; it is twice as fast as the SVD method and three times faster than the KNN method. The SVD method has the smallest standard errors.

When analyzing the rainfall data from the Murray–Darling basin in Section 4, cross-validation was used to choose the optimal cutoff value for the correlation between stations. With the simulated data, we can explore in more detail how the performance of the CUTOFF method varies for different cutoff values. We simulated 100 data matrices with missing values, using Algorithm 1 as described above, and applied the CUTOFF method to complete these matrices for varying cutoff values ranging from 0.55 to 0.95. For each cutoff value, the mean RMSE and the standard error were calculated and the results displayed in Fig. 11. We chose the SVD method with $v = 4$ as the reference method as this small choice of v makes the speed of the SVD method comparable to CUTOFF. It can be seen that, taking account of imputation accuracy and computation time, the SVD method is close to the CUTOFF method. For 100 simulated data matrices, the mean RMSE for the SVD method is 0.6433, while the smallest mean RMSE for the CUTOFF method is 0.5633. Taking the standard errors into consideration, the CUTOFF method outperforms the SVD method for cutoff values between approximately 0.625 and 0.865. In addition, the cutoff value that produces the smallest RMSE in the simulated data is 0.75, which is the same value that was chosen in the real data analysis by cross-validation. This provides further support for using the cross-validation approach for choosing the optimal cutoff value when analyzing real data.

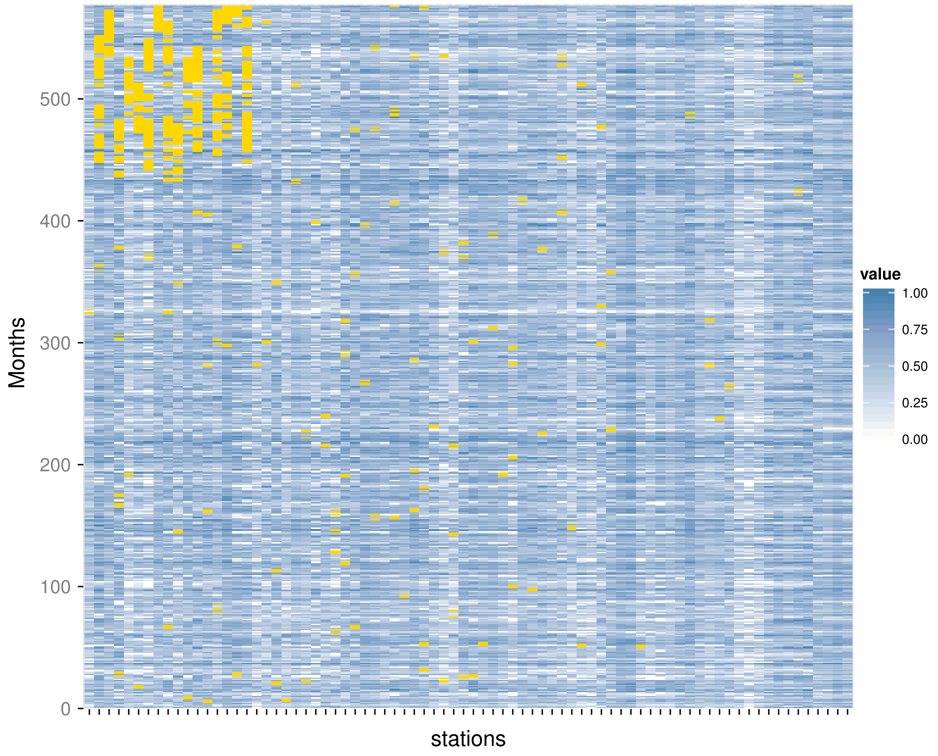

Fig. 10. Heatmap of the simulated missing pattern induced in the reduced portion of the data with no missing values. Missing values are denoted in yellow and observed values (which have been normalized) in blue. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

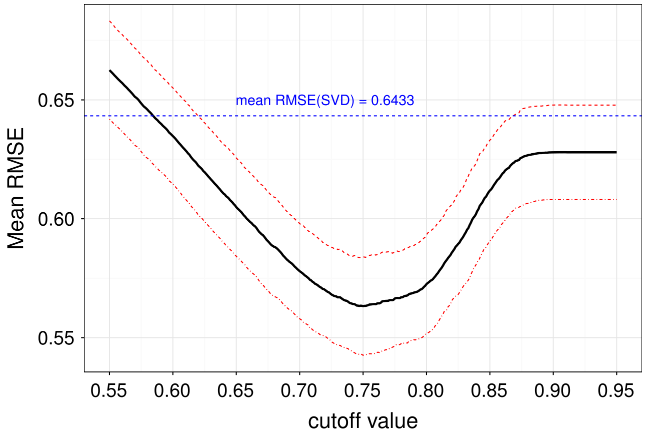

Fig. 11. Simulation using a sequence of cutoff values from 0.55 to 0.95, with steps of 0.001. The dashed blue line represents the mean RMSE level for the SVD method from 100 simulations. The two red dashed lines represent 1.96 standard error intervals. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

5. Conclusion

In this paper, we proposed a new cutoff method for imputing spatio-temporal data and conducted a comprehensive study on some real data. The results show that the cutoff spatio-temporal imputation procedure successfully completes the data matrix using both spatial and temporal information. The procedure seems fairly robust to the length of the large gaps of missing data, and the different temporal variations in each single time series.

We believe that the proposed imputation framework will find wide applicability in the analysis of spatio-temporal data. In this framework, the proposed imputation method is both easy to understand and fast to implement. We developed a cross-validation procedure to optimally choose parameters and also proposed a simulation algorithm, which can be used to further assess the performance of the imputation method. Using the cross-validation approach to find the optimal cutoff value is a crucial part of the procedure. Using this optimal cutoff value, the cutoff imputation can do better than most advanced imputation methods. Although we found that some machine learning imputation methods such as missForest are moderately better in terms of accuracy, they are much slower. Another advantage of the CUTOFF method is that it is an automatic procedure and can be applied before modeling. The key parameter that needs specifying is the cutoff value, which is chosen by a grid-wise cross-validation technique that seems to be robust to the grid-size.

We have focused on prediction accuracy as a criterion for comparing imputation procedures. This is useful and important for obtaining good estimates but it does not address the question of how to make valid inferences with the imputed data. Generally, treating imputed values as observed values underestimates the level of uncertainty and produces over-optimistic inferences. Tackling this problem is beyond the scope of this paper, but it is worth noting that repeated stochastic imputation provides one way of making valid inferences through multiple imputation (MI) (Rubin, 1978).

The imputation framework is open to future generalization, possibly at the cost of its simplicity. CUTOFF can be used for rainfall data, and it is equally applicable to other spatio-temporal data sets in hydrology, such as ground water measurements, and soil moisture. CUTOFF is also expected to be useful for spatio-temporal data sets in general, such as temperature, air pollution and so on. We have provided some useful options which can be used to extend the original CUTOFF method. Although we did not see any large improvements on the data we analyzed by trying these options, we believe they could be useful in some other situations with different data generating processes. An R package, **cutoffR**, which includes the implementations of the CUTOFF method, the cross-validation algorithm and the simulation algorithm, is available on CRAN (http://cran.r-project.org/).

Acknowledgments

This work was supported by an ARC Discovery Project grant (No. DP1092801). We acknowledge the helpful suggestions of the Editor, Associate Editor and Reviewers which improved this paper.

References

Abebe, A., Solomatine, D., Venneker, R., 2000. Application of adaptive fuzzy rule-based models for reconstruction of missing precipitation events. Hydrol. Sci. J. 45 (3), 425–436.

Andreis, F., Ferrari, P.A., 2012. Missing data and parameters estimates in multidimensional item response models. Electron. J. Appl. Stat. Anal. 5 (3), 431–437.

Aravena, J., Luckman, B., 2008. Spatio-temporal rainfall patterns in Southern South America. Int. J. Climatol. 29 (14), 2106–2120.

Beckers, J., Rixen, M., 2003. EOF calculations and data filling from incomplete oceanographic datasets. J. Atmos. Ocean. Technol. 20 (12), 1839–1856.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.

Buuren, S., Groothuis-Oudshoorn, K., 2011. Mice: multivariate imputation by chained equations in R. J. Stat. Softw. 45 (3).

Chen, X., Ishwaran, H., 2012. Random forests for genomic data analysis. Genomics 99 (6), 323–329.

Cohen, M., Adar, S., Allen, R., Avol, E., Curl, C., Gould, T., Hardie, D., Ho, A., Kinney, P., Larson, T., et al., 2009. Approach to estimating participant pollutant exposures in the multi-ethnic study of atherosclerosis and air pollution (MESA air). Environ. Sci. Technol. 43 (13), 4687–4693.

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. Meteorol. 39 (9), 1580–1591.

Ferrari, G.A.T., Ozaki, V., 2014. Missing data imputation of climate datasets: implications to modeling extreme drought events. Rev. Bras. Meteorol. 29, 21–28.

Fuentes, M., Guttorp, P., Sampson, P.D., 2006. Using transforms to analyze space-time processes in: Finkenstadt, B., Held, L., V.I. (Ed.), Statistical Methods for Spatio-Temporal Systems, CRC/Chapman and Hall, 2006, pp. 77–150.

Fu, G., Viney, N., Charles, S., 2010. Evaluation of various root transformations of daily precipitation amounts fitted with a normal distribution for Australia. Theoret. Appl. Climatol. 99 (1), 229–238.

Garwood, F., 1936. Fiducial limits for the poisson distribution. Biometrika 28 (3/4), 437–442.

Hapfelmeier, A., 2012. Analysis of Missing Data with Random Forests. Ph.D. Thesis, Ludwig Maximilian University of Munich.

Hastie, T., Tibshirani, R., Friedman, J.J.H., 2009. The Elements of Statistical Learning: Data Mining, Inference and Prediction, second ed. Springer.

Hoerl, A., Kennard, R., 1970. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12 (1), 55–67.

Howell, W., 1965. Cloud seeding against the 1964 drought in the Northeast. J. Appl. Meteorol. 4, 553–559.

Kendall, G., 1960. The cube-root-normal distribution applied to Canadian monthly rainfall totals. Int. Assoc. Sci. Hydrol. 53, 250–260.

Kim, J., Pachepsky, Y.A., 2010. Reconstructing missing daily precipitation data using regression trees and artificial neural networks for SWAT streamflow simulation. J. Hydrol. 394 (3), 305–314.

Kondrashov, D., Ghil, M., et al., 2006. Spatio-temporal filling of missing points in geophysical data sets. Nonlinear Processes Geophys. 13 (2), 151–159.

Kondrashov, D., Ghil, M., et al., 2007. Reply to T. Schneider's comment on 'Spatio-temporal filling of missing points in geophysical data sets'. Nonlinear Processes Geophys. 14 (1), 3–4.

Linacre, E., 1992. Climate Data and Resources: A Reference and Guide. Routledge.

Lindström, J., Szpiro, A., Sampson, P.D., Bergen, S., Oron, A.P., 2013a. SpatioTemporal: Spatio-Temporal Model Estimation. R Package Version 1.1.7. <http://CRAN.R-project.org/package=SpatioTemporal>.

Lindström, J., Szpiro, A.A., Sampson, P.D., Oron, A.P., Richards, M., Larson, T.V., Sheppard, L., 2013b. A flexible spatio-temporal model for air pollution with spatial and spatio-temporal covariates. Environ. Ecol. Stat., 1–23.

Lou, Q., Obradovic, Z., 2011. Modeling multivariate spatio-temporal remote sensing data with large gaps. Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence, vol. 2. AAAI Press, pp. 1711–1716.

Lowry, W.P., 1972. Compendium of Lecture Notes in Climatology for Class III Meteorological Personnel. No. 335. Secretariat of the World Meteorological Organization, Geneva.

Merrington, M., 1941. Numerical approximations to the percentage points of the χ^2 distribution. Biometrika 32 (2), 200–202.

Nourani, V., Mogaddam, A.A., Nadiri, A.O., 2008. An ANN-based model for spatiotemporal groundwater level forecasting. Hydrol. Process. 22 (26), 5054–5066.

Nourani, V., Baghanam, A., Gebremichael, M., 2012. Investigating the ability of Artificial Neural Network models to estimate missing rain-gauge data. J. Environ. Inf. 19 (1).

Paulhus, J.L., Kohler, M.A., 1952. Interpolation of missing precipitation records. Mon. Weather Rev. 80, 129–133.

Presti, R.L., Barca, E., Passarella, G., 2010. A methodology for treating missing data applied to daily rainfall data in the Candelaro River Basin (Italy). Environ. Monit. Assess. 160 (1-4), 1–22.

Ramos-Calzado, P., Gómez-Camacho, J., Pérez-Bernal, F., Pita-López, M., 2008. A novel approach to precipitation series completion in climatological datasets: application to Andalusia. Int. J. Climatol. 28 (11), 1525–1534.

R Core Team, 2014. R: A Language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org/>.

Reinhardt, J.D., von Elm, E., Fekete, C., Siegrist, J., 2012. Social inequalities of functioning and perceived health in Switzerland–A representative cross-sectional analysis. PloS one 7 (6), e38782.

Rubin, D.B., 1978. Multiple imputations in sample surveys – a phenomenological bayesian approach to nonresponse. In: Proceedings of the Section on Survey Research Methods. American Statistical Association, p. 20.

Rubin, D.B., 1987. Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons, New York.

Rustum, R., Adeloye, A.J., 2007. Replacing outliers and missing values from activated sludge data using Kohonen self-organizing map. J. Environ. Eng. 133 (9), 909–916.

Schilling, M., Watkins, A., 1994. A suggestion for sunflower plots. Am. Stat., 303–305.

Schneider, T., 2001. Analysis of incomplete climate data: estimation of mean values and covariance matrices and imputation of missing values. J. Clim. 14 (5), 853–871.

Schneider, T., 2006a. Analysis of incomplete data: readings from the statistics literature. Bull. Am. Meteorol. Soc. 87, 1410–1411.

Schneider, T., 2006b. Comment on Spatio-temporal filling of missing points in geophysical data sets by D. Kondrashov and M. Ghil. Nonlinear Processes Geophys. 13, 151–159.

Städler, N., Bühlmann, P., 2010. Pattern Alternating Maximization Algorithm for High-Dimensional Missing Data. arXiv preprint arXiv:1005.0366.

Stekhoven, D.J., 2012. missForest: Nonparametric Missing Value Imputation Using Random Forest. R Package Version 1.3. <http://CRAN.R-project.org/package=missForest>.

Stekhoven, D.J., Bühlmann, P., 2012. MissForest – nonparametric missing value imputation for mixed-type data. Bioinformatics 28 (1), 112–118.

Szpiro, A., Sampson, P., Sheppard, L., Lumley, T., Adar, S., Kaufman, J., 2009. Predicting intra-urban variation in air pollution concentrations with complex spatio-temporal dependencies. Environmetrics 21 (6), 606–631.

Talbert, S., Sole, M.L., 2013. Too much information: research issues associated with large databases. Clinical Nurse Specialist 27 (2), 73–80.

Teegavarapu, R.S.V., 2014. Missing precipitation data estimation using optimal proximity metric-based imputation, nearest-neighbour classification and cluster-based interpolation methods. Hydrol. Sci. J. 59 (11), 2009–2026, <http://dx.doi.org/10.1080/02626667.2013.862334>.

Teegavarapu, R.S., Chandramouli, V., 2005. Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records. J. Hydrol. 312 (1), 191–206.

Torgo, L., 2010. Data Mining with R: Learning with Case Studies. Chapman and Hall/CRC, <http://www.dcc.fc.up.pt/ltorgo/DataMiningWithR>.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., Altman, R.B., 2001. Missing value estimation methods for DNA microarrays. Bioinformatics 17 (6), 520–525.

Van Buuren, S., Oudshoorn, K., 1999. Flexible Multivariate Imputation by MICE. TNO Prevention Center, Leiden, The Netherlands.

Xia, Y., Fabian, P., Stohl, A., Winterhalter, M., 1999. Forest climatology: estimation of missing values for Bavaria, Germany. Agric. For. Meteorol. 96 (1), 131–144.

Xia, Y., Fabian, P., Winterhalter, M., Zhao, M., 2001. Forest climatology: estimation and use of daily climatological data for Bavaria, Germany. Agric. For. Meteorol. 106 (2), 87–103.

Young, K.C., 1992. A three-way model for interpolating for monthly precipitation values. Mon. Weather Rev. 120 (11), 2561–2569.

Yozgatligil, C., Aslan, S., Iyigun, C., Batmaz, I., 2012. Comparison of missing value imputation methods in time series: the case of Turkish meteorological data. Theoret. Appl. Climatol., 1–25.