7

Seasonal Climate Prediction for the Australian Sugar Industry Using Data Mining Techniques

Lachlan McKinna and Yvette Everingham
James Cook University,
Australia

1. Introduction

The ability to predict rainfall with adequate certainty and lead time is beneficial to both industry and public. Periods of high or low seasonal rainfall can have many follow on effects to agriculture, industry, public health and, water supply and management. In order to implement decisions, planning and management strategies to contend with these issues, the ability to predict seasonal rainfall quantities is of great importance (Klopper et al., 2006). Climate conditions are known to influence the cultivation of Sugarcane influencing planting, harvesting and milling (Muchow and Wood, 1996; Everingham et al., 2002; Jones and Everingham, 2005). Unforeseen climate events such as excessive rainfall, can adversely effect the agricultural practices related to Sugarcane cultivation. The Australian Sugarcane harvest period commences in May/June and aims to finish by November/December before the start of the rainy season (Everingham et al., 2002). The risk of excessive rainfall disrupting harvest operations is greatest towards the end of the sugarcane harvest period (Muchow and Wood, 1996; Everingham et al., 2002). Therefore, improved seasonal rainfall prediction during the October-December period is beneficial.

Statistical prediction of seasonal rainfall can be performed using a variety of techniques including: regression (Singhrattna et al., 2005), classification methods (Drosdowsky and Chambers, 2001), canonical correlation analysis (Landman and Mason, 1999) and neural networks (Mason, 1998). All statistical models require predictor variables which act as proxies for describing the behaviour of response variables (Hastie et al., 2001). When considering a seasonal forecast model, it is useful to draw predictor variables from a climate data set that is both historically and spatially complete (Washington and Downing, 1999). One of the most temporally and spatially resolute climate parameters is sea surface temperature (SST) data. Consequently, SST data are often used as an empirical measure of the ocean-atmosphere interaction in statistical climate models. However, a vast proportion of potential SST predictors may be redundant. Therefore employing data mining methods for the purpose of feature extraction and data reduction is advantageous.

Principal component analysis (PCA) is a commonly used feature extraction method that reduces data dimensionality whilst retaining the majority of the variability (Jolliffe, 1986). As sea surface temperature data sets are large, it is useful to perform PCA data reduction such that the bulk of the variability is contained in a small subset of variables (Wilks, 1995). PCA also referred to as empirical orthogonal function (EOF) analysis is commonly used throughout climate research (Wilks, 1995). PCA is popular because it is available in most

statistical software packages; is easy to self-program and can be applied to a variety of multivariate data from many disciplines. However, there are some disadvantages to PCA. In situations where there is ordering associated with the independent variables, then this ordering is ignored by PCA. This is pertinent when considering application of PCA to SST data where the variables are ordered in both a latitudinal and longitudinal direction. In order to perform PCA on a SST dataset, the spatial structure is reduced by "stringing-out" each 2D monthly piece of SST data into a one dimensional (1D) vector whose element order has no significance (Wilks, 1995). The 1-D vector then becomes a row in a large matrix to which the PCA is applied as illustrated in Fig. 1.

When analysing image type data, methods that extract information whilst maintaining the spatial structure are favourable. A method of image analysis known as the 2D discrete wavelet transform (DWT) has the ability to extract high and low frequency information from the image, and can perform dimension reduction whilst maintaining the spatial structure of data (Mallat, 1989a; Mallat, 1989b; Antonini et al., 1992). Within this chapter it is proposed that SST data be considered an image and analysed using a 2D DWT to maintain the spatial integrity of the dastaset. However, the literature involving the application of 2-D DWT methods to spatial climate data for the purpose of extracting useful features is scant.

Although feature extraction methods assist in mining useful features from data, they can still output high dimensional and collinear datasets. For the purposes of statistical modelling, a larger number of predictor variables than observations results in an ill-posed situation. A large number of variables can also have practical implications upon computational speed. Therefore, it is useful to employ data mining techniques to produce a smaller sub-set of predictor variables. Random forests (RF) analysis is a non-parametric approach which has emerged from classification and regression tree theory (Brieman, 2001). The RF method is robust to outliers, noise and is ideal for datasets of large dimension. The RF method is also useful for identifying variables of importance and hence can be used for data reduction. Firth et al. (2005) found RF to be a useful method for predicting the onset of the winter rain season for the wheat growing region of southwest Western Australia using climate indices including: SST data, mean sea level pressure (MSLP) and the southern oscillation index (SOI). Furthermore, Firth et al. (2005) found the RF method was able to locate regions of SST which were deemed to be important predictor variables.

Within this chapter we have developed a statistical model for the prediction of above median rainfall for Tully, in the northern part of the Australian sugarcane growing region. Data mining methods were explored for the purposes of feature extraction and variable reduction of SST data. The 2D DWT and PCA were both used for comparitive purposes of feature extraction upon SST data. We examined the RF algorithm for the purposes of variable reduction. Classification was performed using regularised discriminant analysis (RDA) and model performance was assessed based upon a 10-fold cross validated (CV) correct classification rate (CCR).

2. Principal component analysis

PCA performs a linear transform on an n-by-d data matrix $\mathbf{X}$ to produce a matrix $\mathbf{P}$ containing a set of d uncorrelated, independent variables of which the first few will contain the bulk of the variability exhibited in the original data (Jolliffe, 1986).

The complete n by d matrix of principal components $\mathbf{P}$ is given by

$$\mathbf{P} = \mathbf{X}^T\mathbf{A} \quad (1)$$

where $\mathbf{A}$ is a *d*-by-*d* matrix with columns being eigenvectors obtained by performing an eigenvalue decomposition (Jolliffe, 1986) on the covariance matrix $\mathbf{\Sigma}$ of $\mathbf{X}$, on the case of standardized variables, the eigen-decomposition will be on the correlation matrix. The columns of $\mathbf{A}$ are arranged such that the corresponding set of eigenvalues are in descending order (Jolliffe, 1986). As a consequence of arranging the columns of $\mathbf{A}$, the first principal component retains the largest amount of variability from the original data with the second principal component containing the next largest proportion of variability and so on (Jolliffe, 1986). Thus the bulk of the variance contained in the untransformed data comes to be contained in the first few PCs (Jolliffe, 1986).

In order to perform a PCA on a time series of 2D grided data, the dataset must be restructured into a single matrix (Wilks, 1995). For example, a time series of *j* 2D SST data observations must be rearranged into a single matrix $\mathbf{X}$ in order to perform a PCA. Typically the first 2D SST matrix of size *n-by-d* is reshaped into a single row vector which has the size *1-by-(nxd)*. The same method is followed for the j^{th} 2D SST data within the time series. The newly created row vectors are arranged to form the matrix $\mathbf{X}$ with the dimensions *j-by-(nxd)*. Thus, a single row reprents all the SST data from a point in time. Whereas, a column represents a time series of SST observations for a single geogrpahical location (see Fig. 1).

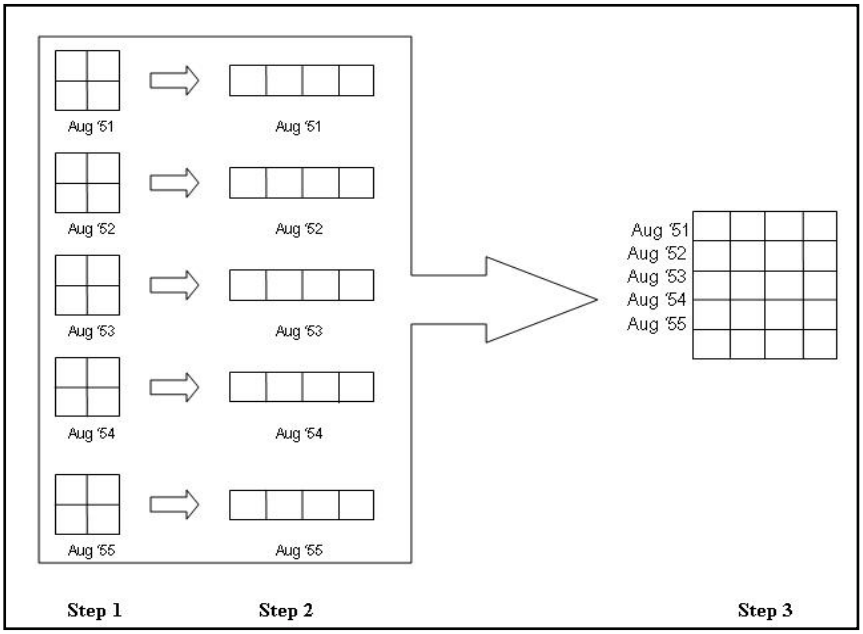


Fig. 1. A schematic diagram indicating the steps in preparing monthly 2-D spatial data for analysis in principal component analysis (PCA). Step 1 monthly spatial data matrices are arranged then in Step 2 are disassembled or "strung out" in the same manner to produce a 1-D vector. These vectors representing monthly data are then arranged as the rows of a large matrix as illustrated in Step 3.

3. 2-D discrete wavelet transform

An image can be considered a finite energy/intensity matrix of components *I(x,y)*, where *x* and *y* represent the horizontal and vertical directions respectively (Mallat, 1989a; Mallat, 1989b). The theory of the 2-D discrete wavelet transform closely follows the formulation of 1-D discrete wavelet transforms using multiresolution analysis (Mallat, 1989a; Mallat, 1989b).

A 1-D wavelet transform is similar to a Fourier transform, enabling the underlying frequencies within a signal to be identified. The Fourier transforms treats a signal as a whole or *globally* which can often cause small perturbations in the signal to be overlooked (Mallet et al., 2000). A 1-D wavelet transform allows a *localised* analysis of the signal using a *window function* which *translates* across the signal analysing discrete sections (Mallet et al., 2000). The continuous wavelet transform performed on a signal *f(t)* can be given as

$$S_{CWT}(a,b) = |a|^{-1/2} \int_{-\infty}^{\infty} f(t)\psi_{a,b}(t)\,dt \tag{2}$$

where $\psi_{a,b}(t)$ is the window analysing function and *a* and *b* are the *dilation* and *translation* parameters, respectively. The window function is referred to as the *mother* wavelet and has the form $\psi_{a,b}(t) = \psi\left(\frac{t-b}{a}\right)$ (Antonini et al., 1992). The wavelet transform is unique because the mother wavelet function has the ability to contract and dilate, allowing high and low frequencies in the signal to be well represented (Mallet et al., 2000). There are a number of families of wavelet functions including: Daubechies, coiflets and symlets each having its optimum use upon different signal types (Mallet et al., 2000). The symlet family however, has properties which lend themselves ideally to image analysis (Mallet et al., 2000).

An image *I(x,y)*, can be decomposed using a 2-D discrete wavelet transform (DWT) similar to how a 1-D signal *f(t)* can be analysed using a 1-D wavelet transform. The 2-D DWT analysis decomposes the image *I(x,y)* into four sub images, one *smooth* image and the three *detailed* images (Mallat, 1989a; Antonini et al., 1992; Mallet et al., 2000). The smooth image is representative of low frequency information and the three detailed images represent high frequency information from the original image (Mallat, 1989a; Antonini et al., 1992). The smooth image created is denoted as $S_j I$ and the three remaining detailed images $D_j^h I$, $D_j^v I$ and $D_j^d I$ capture high frequencies in the horizontal, vertical and diagonal directions respectively (Mallat, 1989a; Antonini et al., 1992). A 2-D wavelet transform also performs dimension reduction, with each sub-image being one quarter the size of the size of *I(x,y)*.

For further extraction of information, the smooth image $S_j I$ undergoes a successive transform, yielding another set of four sub images. The method of applying successive transforms is known as *multiscale pyramidal decomposition* (Mallat, 1989a; Antonini et al., 1992). We can consider the transform in a series of stages or levels. The original image is considered to be at the zeroth level of the transform denoted as $j = 0$. The first transform upon *I(x,y)* producing four sub-images is referred to as the first level transform (*j=1)* with sub images denoted as $S_1 I$, $D_1^h I$, $D_1^v I$ and $D_1^d I$ (Mallat, 1989a; Antonini et al., 1992). At each *level* of the multiscale pyramidal transform, further information about the horizontal, diagonal and vertical components is extracted, the dimensionality of the data is reduced and

the spatial structure is maintained (Antonini et al., 1992). Figure 2 illustrates schematically the process of a multiscale transform.

An example of an image decomposition using the discrete 2-D DWT is shown in Fig. 3. The original image *I(x,y)* is a *Mandrill face* from the Matlab® Image Processing Toolbox to which 2D DWT was applied using a symlet wavelet to the first level (*j=1*) of a multiscale pyramidal decomposition. The first sub-image S_1I is the *low-pass* image and represents the low frequencies or *smooth* details from the original image. The three remaining sub-images are the *detailed* images D_1^hI, D_1^vI and D_1^dI. The high frequency horizontal and vertical features of the Madrill face such as the whiskers and nose ridges are well represented in the images D_1^hI and D_1^vI respectively. Image D_1^hI represents the horizontal features of the Mandrill face whiskers and nose ridges.

From a climatology perspective, it is useful to locate regions of high frequencies in sea surface temperature anomaly data because temporal changes of frequencies in these regions may indicate the onset of a certain meteorological event. A useful tool for the analysis of sea surface temperature anomalies is then the 2-D discrete wavelet transform as it will detect high frequencies laterally, longitudinally and obliquely. An example of a 2D DWT decomposion of an SSTA image is given in Fig. 4.

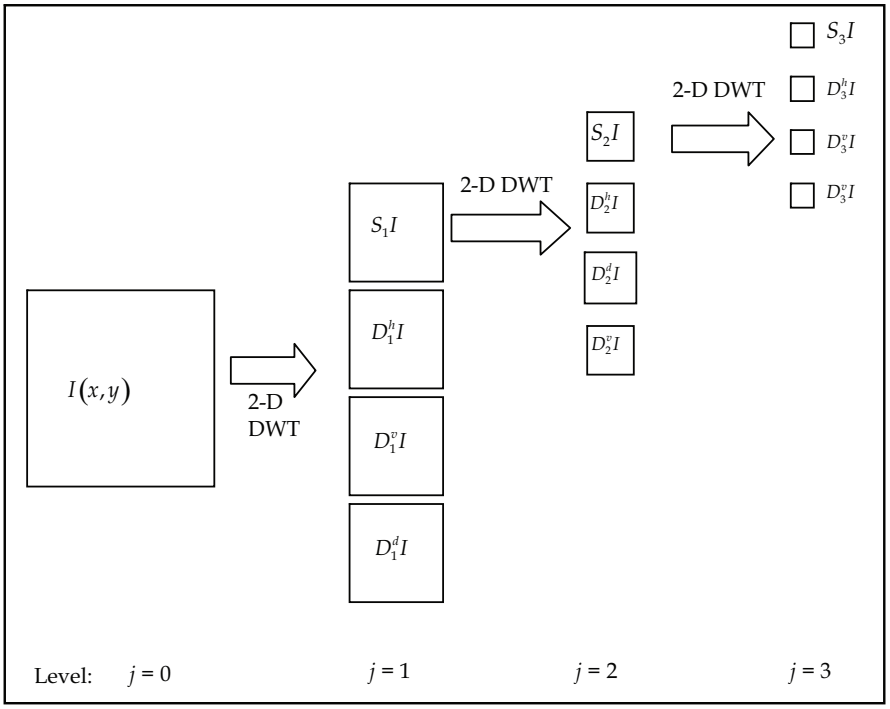
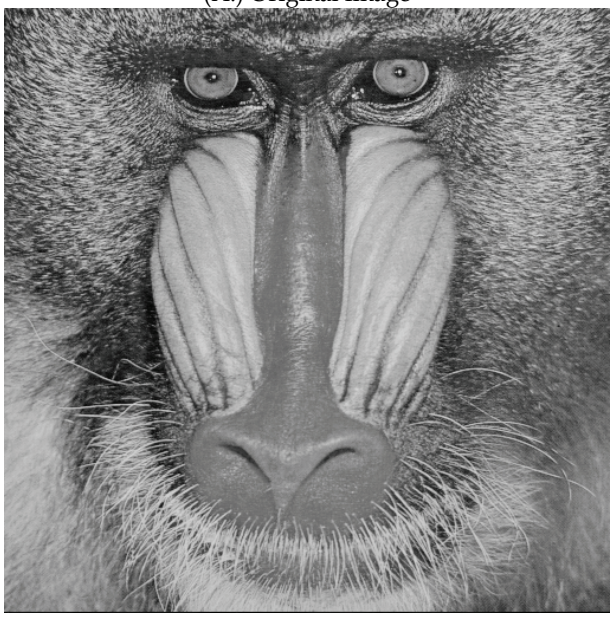


Fig. 2. A flow chart indicating a pyramidal, multiscale 2-D Wavelet transform. From the original image *I(x,y)*, four sub-images are produced one smooth image S_1I and three detailed images D_1^hI, D_1^vI and D_1^dI at level *j=1*. Successive transforms are performed upon the smooth image at each level *j* = 2, 3 in order to produce the multiscale pyramidal transform.

(A.) Original Image



(B.) Wavelet Transformed Sub-Images

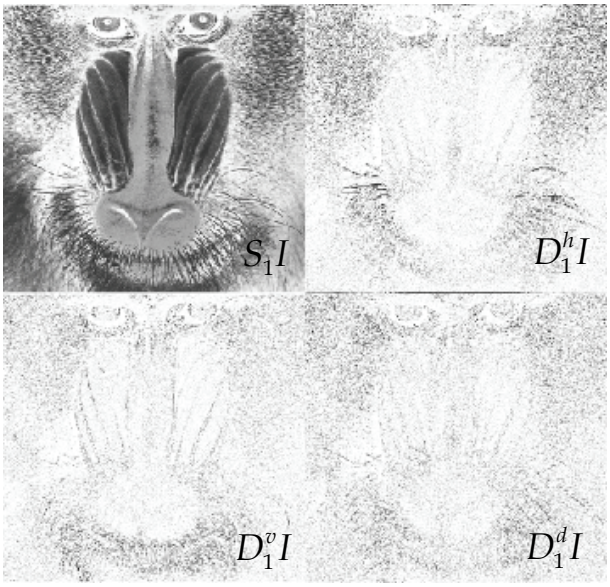
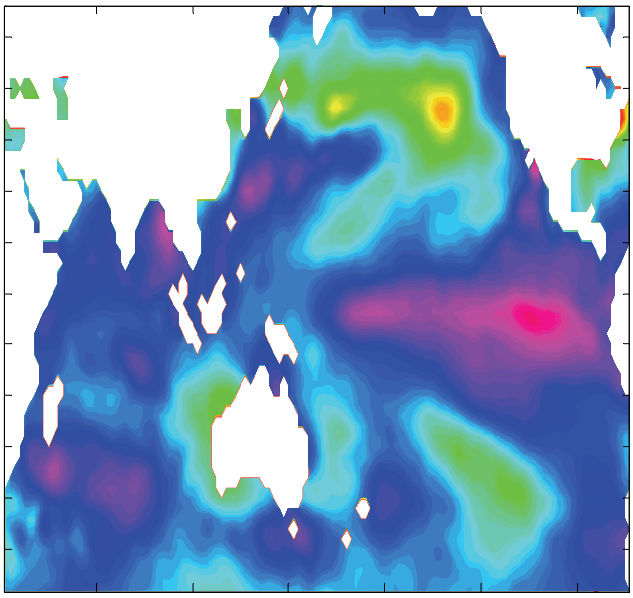


Fig. 3. (A.) Original image of a *Mandrill Face.* (B.) A first level, 2-D DWT representation of the *Mandrill* image using the Symlet wavelet with produced four sub-images S_1I, D_1^hI, D_1^vI and D_1^dI. Notice the horizontal, vertical and diagonal features extracted from the original image are emphasised in the sub-images D_1^hI, D_1^vI and D_1^dI respectively.

(A.) Original SSTA Image



(B.) Wavelet Transformed Sub-images

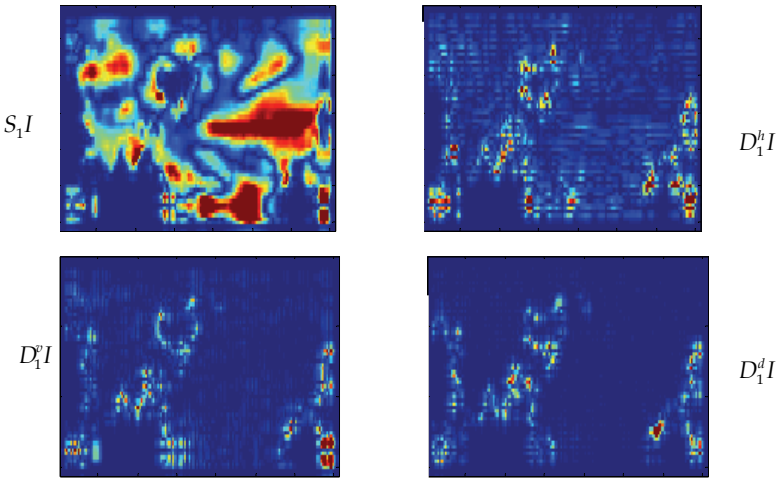


Fig. 4. (A.) An example image of a single SSTA map from the Indian and Pacific Oceans surrounding the Australian continent. (B.) A first level 2-D DWT representation of the SSTA map using a Symlet wavelet. Four sub-images S_1I, D_1^hI, D_1^vI and D_1^dI were produced. Low frequency or *smooth* features are emphasised in the sub-image S_1I whereas high frequency features in the horizontal, vertical and diagonal directions are emphasised in the sub-images D_1^hI, D_1^vI and D_1^dI respectively.

4. Random forests and classification and regression trees

Feature extraction methods do not necessarily reduce the dimensionality of the data. In high dimensional and low observational settings, model performance can be adversely affected. Therefore, following feature extraction, a feature selection method which reduces the dimensionality of the data can be applied (Svetnik et al., 2004). Svetnik et al. (2004) investigated a feature selection method based upon the random forest (Brieman, 2001) technique. Random forests (Brieman, 2001) are often used as a method for classifying data into groups for the situation where there exists many predictor variables. A favourable attribute of the random forest technique is its ability to identify a subset of variables that best classify objects into groups (Brieman, 2001). The variable selection algorithm performs a random forest analysis which is indicative of the feature variables most important for classifying an observation (Svetnik et al. 2004). A fraction of the least important variables are then removed and the random forest is re-implemented. This routine is continued until an assessment criteria called the *out-of-bag error rate* (Brieman, 2001) is minimized, at which point the variables of most importance for classification are determined. This process of variable selection using random forests is contained in a package called varSelRF which performs variable selection procedure using R statistical software (Diaz-Uriarte and Alvarez de Andres, 2006). This is a very useful tool for dimension reduction in the situation where there exists many predictor variables (Svetnik et al., 2004). We will now briefly overview classification and regression trees, and random forests.

4.1 Decision trees

Classification and regression trees (Hastie et al., 2001) are collectively known as decision trees and can be used both for classification and prediction. The benefit of decision trees is that they are a non-linear method and have the ability to handle different types of data. An added benefit of classification and regression trees is their ability to handle missing data within predictor variables (Hastie et al., 2001).

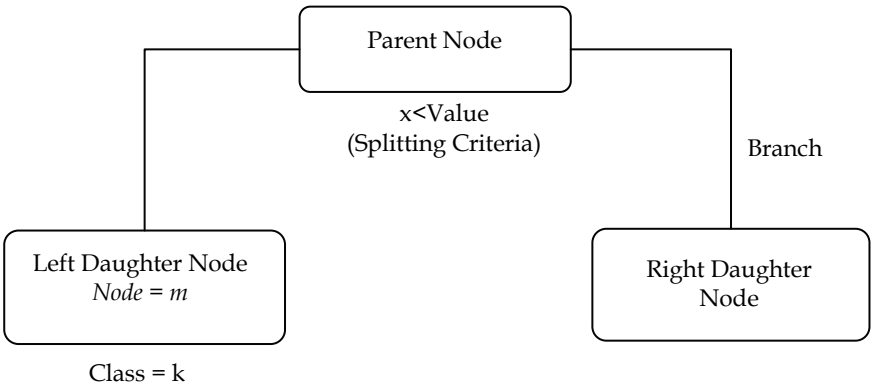


Fig. 5. Decision tree terminology: A *parent node* in a decision tree is split due to some splitting criteria into *left and right daughter nodes* which are connected to the mother node via *branches*. Eventually splitting will continue until the *terminal nodes* are reached, at which point the data should be split into distinct classes.

Decision tree methodology can be summarised into three steps; (i) splitting criteria, (ii) pruning and (iii) tree Selecting (Hastie et al., 2001).

i. The splitting criterion dictates how data is to be partitioned into new groups at each node. Splitting is performed in a *greedy* fashion at a *parent node* from which data is split into two *daughter* groups (Hastie et al., 2001). Splitting in this manner continues until *terminal* nodes are reached where only a small number of observations of the same distinct class reside (Hastie et al., 2001).
ii. *Pruning* is carried out to reduce the number of nodes in the large tree that has been created (Hastie et al., 2001). Pruning ensures the tree is not overfitted, whilst ensuring the tree is large enough to avoid biases occurring when used to make predictions (Hastie et al., 2001).
iii. Tree selection finds the optimum tree model which is often determined by examining the cross-validated error rate (Hastie et al., 2001). The tree that presents the lowest cross-validated error rate is often chosen as optimal (Hastie et al., 2001).

4.2 Tree splitting criteria

The difference between *classification* and *regression* trees is the splitting criteria used for each. For *classification trees* there are several splitting criteria, of which the most commonly used is the known as the *Gini* split criteria and is defined as

$$i(k) = \sum_{k \neq k'} \hat{p}_{mk} \hat{p}_{mk'} = \sum_{k=1}^{K} \hat{p}_{mk} \left(1 - \hat{p}_{mk}\right) \tag{3}$$

Where $\hat{p}_{mk}$ is the probability that an item in node *m* is of class *k*. The impurity measure *i(k)* is also known as the misclassification error (Hastie et al., 2001). The optimum split of the data from the parent (P) node to the left (L) and right (R) child nodes is based upon the impurity measures at each node. The change in the impurity Δ, is calculated as

$$\Delta = i(k)_P - \left[p_L i(k)_L + p_R i(k)_R \right] \tag{4}$$

where, $i(k)_P$, $i(k)_R$ and $i(k)_L$ are the impurities at the parent node and the right and left child nodes respectively (Hastie et al., 2001). The proportions of data in the left and right child nodes are denoted as p_L and p_R (Hastie et al., 2001). The split that produces the greatest change in impurity is ultimately chosen ensuring that the impurity at the child nodes is much less than that of the parent node (Hastie et al., 2001).

The splitting criteria in *regression trees* depends upon the residual sum of squares (Hastie et al., 2001). The split considers all the possible variables as predictors for the split and chooses the one which minimises the residual sum of squares error at the child nodes. The impurity measure *i(t)* for a variable *y* in a regression tree is given by:

$$i(t) = \sum \left\{ y_j - \bar{y}(t) \right\}^2 \tag{5}$$

Where, $\bar{y}(t)$ is the mean of an observation in node t and y_j represents j^{th} observation of variable *y* in node *t* (Hastie et al., 2001). The best split at a parent node for a regression tree is determined by examining the change in impurity Δ in terms of residual sum of squares error as below

$$\Delta = SSE_P - [P_L \cdot SSE_L + P_R \cdot SSE_R] \quad (6)$$

where SSE_P is the within groups sum of squares of the parent node and SSE_L, SSE_R are the residual sum of squares error of the left and right child nodes respectively (Hastie et al., 2001). The best split occurs when the change in impurity is maximised, which means that we desire the residual sum of squares error in the child nodes to be minimised for an optimal split (Hastie et al., 2001).

4.3 Random forests

One way to improve the decision tree method is by creating an ensemble of *n* decision trees. An ensemble classification can then be determined by a majority vote amongst the *n* trees created. This is the basis for random forests, a technique that can greatly improve data classification, does not overfit and is relatively robust to noise and outliers (Brieman, 2001). The n^{th} tree within the random forest is unpruned and grown from the n^{th} bootstrap (Hastie et al., 2001) sample of the data. At each node of the n^{th} tree, a sub-set of all variables *mtry* is selected randomly to determine the splitting criteria. The parameters *n*, *mtry* and number of nodes within each tree *nodesize* are user inputs.

Random forest performance is assessed using a measure known as the out-of-bag error rate (OOB). The OOB is a form of cross validation. OOB of the n^{th} tree is determined when those data left out of the n^{th} bootstrap are passed down the tree and classification is performed. The proportion of times that observations are not allocated to their true groups forms the OOB.

4.4 Variable selection using random forests

Svetnik et al. (2004) developed a method for feature selection based upon the RF technique. The method performs random forest upon the data set ($\mathbf{x}_1,\ldots,\mathbf{x}_p$; y) and indicates which of variables $\mathbf{x}_1,\ldots,\mathbf{x}_p$ are of most importance for classifying an observation $\mathbf{y}$ (Svetnik et al., 2004). A fraction of the least important variables are then removed and the random forest is re-implemented. This routine is continued until the OOB is minimized. The result is then a reduced subset of predictor variables (Svetnik et al., 2004; Diaz-Uriarte, 2005). A R statstical package known as varSelRF has been developed which determines variables of most importance using RF (Diaz-Uriarte, 2005). Within the package varSelRF, the user must define the fraction of least important variables dropped at each iteration.

5. Discriminant analysis

Discriminant analysis is a statistical technique used to classify observed data into one of two or more discrete, uniquely defined groups using an allocation rule (Duda and Hart, 1973; Johnson and Wichern, 2002). Allocation or discriminant rules are developed from randomly sampled "learning" or "training" data drawn from *k* known populations, $\pi_1,\ldots,\pi_K$ and based upon the allocation rules, future observations are placed into groups $\omega_1,\ldots,\omega_K$ (Johnson and Wichern, 2002; Rencher, 2002; Afifi and Clark?et al., 2004).

The Regularised Discriminant Analysis (RDA) algorithm formulates a classification score $cf(x_i)$, for allocation of a test object $\mathbf{x}_i = (x_{i1},\ldots,x_{ip})^T$ to class ω_k based upon the training data set (Wu et al., 1996; Johnson and Wichern, 2002; Afiti et al., 2004). The observed object x_i is assigned to the class ω_k which produces the lowest classification score (Wu et al., 1996). The classification score for RDA is given by

$$cf(\mathbf{x}_i) = (\mathbf{x}_i - \boldsymbol{\mu}_k)^T \hat{\boldsymbol{\Sigma}}_k^{-1}(\lambda,\gamma)(\mathbf{x}_i - \boldsymbol{\mu}_k) + \ln\left|\hat{\boldsymbol{\Sigma}}_k(\lambda,\gamma)\right| - 2\ln P(\omega_k) \quad (7)$$

where $\boldsymbol{\mu}_k$ is the class mean vector and $P(\omega_k)$ is the prior probability of an object belonging to class *k*, and $\hat{\Sigma}_k(\lambda,\gamma)$ is the regularised class covariance matrix which is a function of two regularisation parameters are introduced λ and γ (Friedman, 1989).

As the prior values of λ and γ are unknown, a training set is required such that optimum values of the regularisation parameters can be obtained (Friedman, 1989). A grid spanning $0 \le \lambda \le 1$ and $0 \le \gamma \le 1$ is formulated creating a two-parameter optimisation problem whereby a search for the best values of λ and γ is performed (Friedman, 1989). The best regularisation parameters are obtained by minimizing the misclassification risk associated with cross-validation (Hastie et al., 2001) of the training data (Friedman, 1989). Regularisation parameters of $\lambda = 0$ and $\gamma = 0$ represent quadratic discriminant analysis (QDA), $\lambda = 1$ and $\gamma = 0$ represent linear discriminant analysis (LDA), and $\lambda = 1$ and $\gamma = 1$ represents a nearest mean classifier which assigns an observation to a class with the nearest (Euclidean distance) mean (Duda and Hart, 1973; Friedman, 1989).

6. Data

6.1 Rainfall data

Sugarcane cultivation is prevalent along the east coast of Australia between the latitudes of 16° S and 25° S. We have selected Tully (17.56° S, 146.56° E) as a case study location (Fig. 6.). Tully is a very wet sugarcane growing region with an annual median rainfall total of 4000 mm. Tully was selected as a case study location because the authors have engaged participatively with industry consultative groups within this region. Monthly rainfall data was obtained from the Australian Bureau of Meteorology (BOM) for the Tully Sugar Mill, BOM station number 32042. Total October-November-December (OND) rainfall between 1950 and 1999 inclusively was calculated and converted into categories of either (i) below median rainfall or (ii) above median, rainfall after the rainfall data was median filtered to remove any long term trends.

6.2 Sea surface temperature data

The sea surface temperature (SST) data used in this investigation was the Extended Reconstructed SST dataset (ERSST version 2.0) (Smith and Reynolds, 2004) for the years 1950 – 1999. Given that the objective was to predict rainfall for the October-December period, sea surface temperatures prior to October are needed if the model is to be temporally predictive. We decided to use August sea surface temperatures so that industry would have approximately a one month lead-time to react to the prediction. Following Drosdowsky and Chambers (2000), a subset of ocean covering 60°N – 55°S and 30°E – 70°W was selected which encompassed the Indian and Pacific Oceans adjacent to the Australian Continent. The temporal and spatial resolution of the ERSST dataset is monthly, with 2° by 2° grid spacing. A median filter was passed over the data to remove any long term trends. August Sea surface temperature anomalies (SSTA) were calculated for a given SST grid point by subtracting the long term August SST average of that grid point. To ensure SSTA at higher latitudes were not overemphasised, SSTA data were scaled by the cosine of latitude.

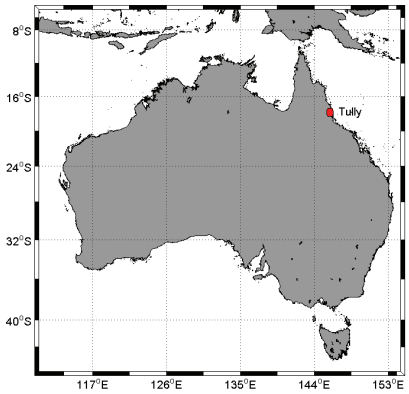


Fig. 6. Study location Tully is located on the north east coast of Australia.

7. Data mining and modelling approach

A model for the prediction of October-November-December (OND) seasonal rainfall for Tully was developed. The variables used to predict above median rainfall were data mined from August SSTA data. Two models were formulated to assess the performance of the PCA relative to the 2D DWT as a feature extraction method. Model development followed five stages as outlined in Fig. 7.: (1) rainfall and SSTA data input, (2) feature extraction, (3) feature selection, (4) discriminant analysis and (5) model validatation.

The data mining approach comprised feature extraction and feature selection steps. Feature selection was performed using the random forest variable selection technique outlined by Svetnik et al (2004). Prior to feature selection, both PCA and 2D DWT were separately performed on the August SSTA data from 1950 – 1999. Whilst there exists many types of wavelet analysing functions, a s*ymlet* with four vanishing moments was chosen as it had symmetrical properties which are considered suitable for image analysis (Mallet et al., 2000). The multiscale 2D DWT was then computed to the 4th level yielding the sub-images (matrices): S_4I, D_4^hI, D_4^vI and D_4^dI. The feature selection step was performed using the RF variable selection algorithm: varSelRF. This process identified the optimum variables for the PCA and 2D DWT feature extracted SSTA data sets respectively. The varSelRF model parameters used are outlined in Table 1. The feature extracted subset of SSTA variables that best predicted above median rainfall were chosen to train the classification rules for RDA.

varSelRF Parameter	Value
n	5000
ntree.iterate	2000
mtry	$\sqrt{\text{number of variables}}$
vars.frac.dropped	0.02

Table 1. Parameters set for varSelRF variable selection algorithm where, *n* is the number of trees in the original random forest, *ntree.iterate*, is the number of trees to use for all additional forests, *mtry* is the number of variables to randomly select at each node split and, *vars.frac.dropped* is the fraction of least important variables dropped at each iteration.

The final step of model development was model validation. This was performed using a 10-fold cross validation approach. After the feature extraction step of model development, the PCA and 2D DWT data were randomised and split into ten equal sized groups for the purpose of cross validation. A single group representing 10% of all data was isolated and kept aside as test data. The remaining 90% of data became the training data set and was used in the feature selection and discriminant analysis steps of model development. After model training was complete, the test data set was input to assess predictive skill. Model predictive skill was quantified using the percentage of observations that were correctly classified, referred to as the correct classification rate (CCR). The process of cross validation was repeated 10 times (10-fold cross validation) and predictive skill was assessed based upon the overall average CCR. Whilst there exist many measures for comparing forecasting performance we elected to use an accuracy measure based on the CCR as it provided a direct and intuitive way to compare the data mining approaches.

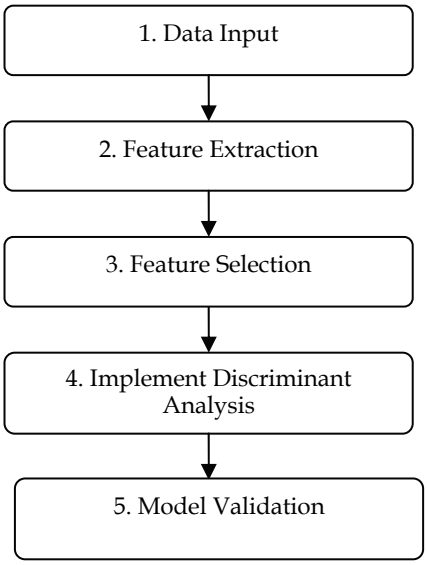


Fig. 7. A schematic diagram showing the steps for the construction of a seasonal rainfall forecast model. Data is first collected and processed with useful features/information extracted in step 2. A feature selection method is then implemented in step 3 to reduce the dimensionality of the data set. Step 5 implements a statistical learning model establish rules from previous data allowing for future prediction. Finally the model is validated to assess its performance in step 5 and feedback loop indicates the modelling process is repeated and augmented if model is deemed unskilful.

8. Results and discussion

Results displayed in Table 2 detail the 10-fold CCR for predicting above median rainfall during OND for Tully. Two different methods of feature extraction: PCA and 2-D DWT were used, whilst the best predictor variables were selected using a RF algorithm: varSelRF.

The model developed using 2D DWT feature extraction produced a 10-fold CCR of 82 % whereas, the the model developed using PCA feature extraction yielded a 10-fold CCR of 67 %. Therefore, the combined data mining approach of using 2D DWT and RF was found to give a predictive skill 20% higher than the the combined PCA and RF data mining approach. Table 2 also details the average number of variables selected using the varSelRF algorithm from each cross validation training set. The number of PCs selected was significantly less than the subset of wavelet coefficients selected. On average, only 3 PCs were selected as best predictor variables. Conversely, the the average number of wavelet coefficients selected as best predictors was 46. This may explain, in part, the improved predictive skill of the 2D DWT – RF method relative to the PC – RF method.

Feature Extraction Method	PCA	2D-DWT
10-Fold Cross-Validated CCR	67 %	82 %
AverageNo.of RF-Slected Variables	3	46

Table 2. Correct classification rates (CCR) indicating predictive skill of two of the model contrasting the feature extraction methods of PCA and 2D-DWT. The average number of variables selected from each cross validation training set using varSelRF method is also detailed.

For data reduction purposes, the number of PCs selected is usually determined by examining a scree plot. The first few PCs that explain the largest proportion of the total variance are typically selected. However, within this investigation we allowed the RF algorithm varSelRF to select the PCs of most importance for prediction from each cross validation set. It was found that during the cross validation process, PCs 4 and 8 were consistently among the set of best predictor variables. PCs 4 and 8 explained 5.7 and 3.9 % of the total variance respectively. Noteworthy was that the PCs that exlained more of the total variance (ie. PCs 1 – 3) were never selected using varSelRF. The number of PCs selected from each cross validation training set using varSelRF is shown in a bar chart (Fig. 8). The bar chart also indicates that the cummulative amount of the total variance explained by the selected subset of PCs. In a previous model for the prediction of Australian seasonal rainfall (Drosdowsky and Chambers, 2000) used PCs of SSTA data computed over the same geographical domain as we have used in this chapter. Drosdowsky and Chambers (2000), used the first two variamax rotated SSTA PCs as predictor variables which explained 11.5 and 4.3 % of the total vriability respectively. Spatial loadings plots of the first two PCs indicated they were related to the El Nino – Souther Oscillation (ENSO) and Indian Ocean SST patterns respectively (Drosdowsky and Chambers, 2000). To give some climatological understanding to the variables slected using varSelRF, we have examined the spatial loadings of PCs 4 and 8.

Spatial loadings plots of PCs 4 and 8 and are presented in Fig. 9A and 9B respectively. The loading plots indicated that PC 4 explains variability in the central-equatorial and northern Pacific Ocean, the equatorial and southwestern Indian Ocean, and the west coast of Central America. The loadings plot of PC 8 indicated it explains variability in the Southern Ocean to the east and west of the Australian contient, and also the western Pacific Ocean. From this we can assume that these regions are likely to be of importance to OND seasonal rainfall in Tully. In contrast, a spatial loadings plot of PC 1 has been included (Fig. 9C). Although PC 1 was never selected by the varSelRF algorithm, we see it strongly related to variability in the ENSO region which agrees with results of Drosdowsky and Chambers (2000). These results

thus, suggest that SST variability wihin the ENSO region of the eastern tropical Pacific Ocean may not be strongly related to OND season rainfall in Tully.

It was also of interest to investigate the spatial significance of each 2D DWT coefficient selected using varSelRF. In order to do this, an inverse wavelet transform was performed. Binary matrices (0,1) of equal size to the fourth level, 2D DWT coefficient matrices S_4I, D_4^hI, D_4^vI and D_4^dI were constructed. Wavelet coefficients identified as best predictors were given a value unity, an all other coefficients were set to zero. The inverse 2D wavelet transform was then performed upon the binary wavelet coefficient matrices to derive an image *I(x,y)* with the same dimensions as the original SSTA data. The inverse wavelet derived image (Fig. 10.), revealed regions of importance lay in: the central Indian Ocean, Southern Ocean, the Coral Sea adjacent to Papua New Guinea, the Northern Pacific Ocean, and the west coast of the Central America. A region of most of importance was also identified in the equatorial eastern Pacific Ocean. Strikingly, the regions identfied as best predictors from 2D DWT coefficients were very similar to the spatial loading plots of PC 4 and PC 8.

These results suggest that the combined 2D DWT and RF approach was a useful tool for data mining teleconnections between seasonal rainfall and SST data. The results also suggest that the PCs that explain most variance in the data may not necessarily form the best set of predictor variables. As, such a variable selection method such as the RF or similar may be of benefit when choosing a sub-se tof PCs.

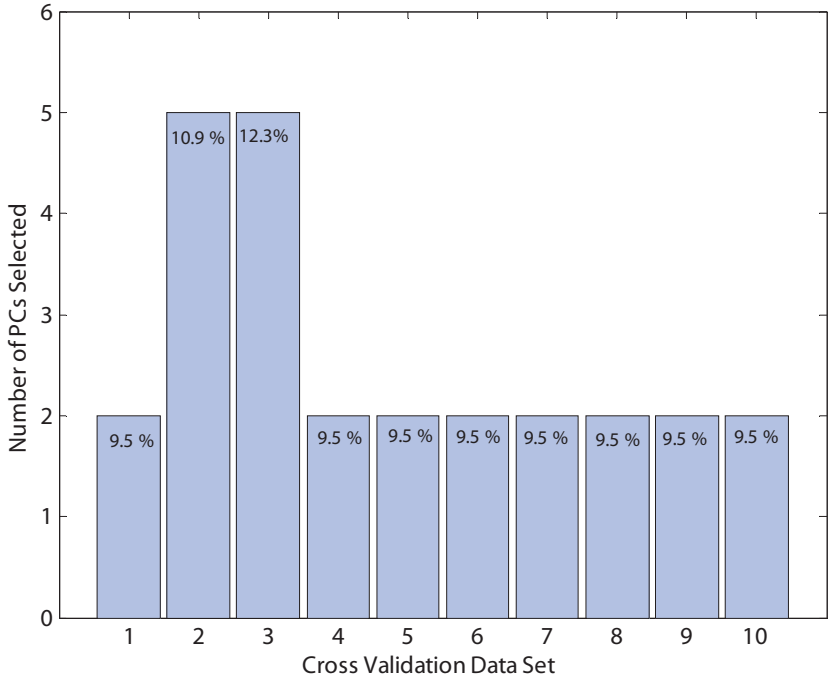


Fig. 8. Number of PCs selected using varSelRF algorithm for each cross validation set of August SSTA PCs. The cumulative percentage of total variance explained by each set of PCs is also given.

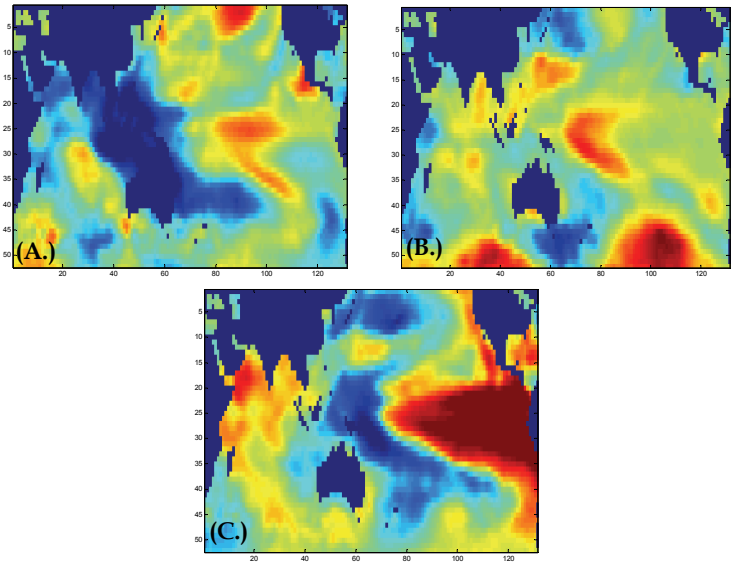


Fig. 9. Loadings plots of (A.) PC 4 and (B.) PC 8 which explained 5.7 and 3.9 % of the total variance respectively. (C.) The loadings plot of PC 1, which explained 18.2 % of the total variance respectively. Within the spatial loadings plots, warm colours indicate regions of high positive loading. Cool colours indicate regions of negative loadings.

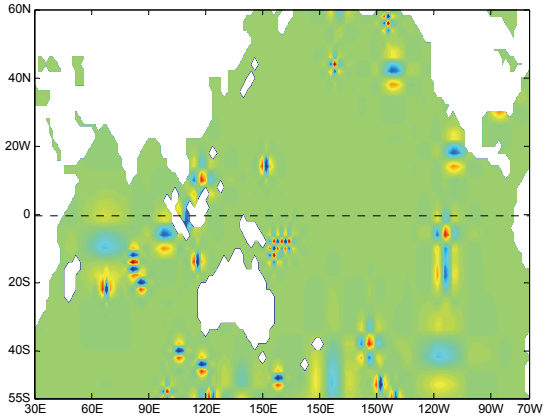


Fig. 10. Regions of most importance for prediction of Tully OND seasonal rainfall. Regions of importance were identified from an inverse wavelet transform of coefficients identified as of importance using the varSelRF algorithm. Points of colour against the green background indicate predictor locations.

9. Conclusion

The purpose of this chapter was to investigate methods of data mining sutiable for developing a seasoanl rainfall predictive model for Tully, Australia. Two data mining

approaches were used to determine a subset of predictor variables from SSTA data: (i) PCA feature extraction with RF variable selection and, (ii) 2D DWT feature extraction with RF variables selection. Two separate models were then developed to predict above median OND rainfall for Tully, Australia using RDA. A 10-fold cross validation was performed upon each model to assess performance. The CCR scores were 67 % and 82 % for the PCA - RF, and 2D DWT – RF data models resepectively. The results indicated that 2D DWT –RF data mining approach typically produced a larger subset of predictor variables than the PCA – RF method. This extra degree of information may explain the enhanced predictive skill of the 2D DWT – RF predictor data set.

The RF algorithm consistently chose PC 4 and PC 8 as predictor variables, which together explained 9.5 % of total variance. Typically, variable selection is performed by selecting the first few PCs which explain the largest proportion of total variance. However, within this study PCs 1 – 3 were never selected using RF variable selection. This suggested that the spatial loadings of the PCs may have been of greater importance than the proportion of variance explained by the PC. Inverse 2D DWT allowed the wavelet variables of most importance to be spatially mapped. Interestingly, the spatial loadings of PC 4 and PC 8 were very similar to the spatial locations identified from the inverse 2D DWT. This provided further evidence to suggest that the spatial location of predictors was of greater importance than the amount of variance explained.

This research concerned constructing forecast models for the prediction of above median rainfall for OND seasonal rainfall for a single case study location: Tully, with a lead time of one month. It would be useful to extend the modelling and data mining methods of this work to other sugar growing regions across Australia and assess predictive skill. Moreover, the technique outlined in this paper need not be limited to sugarcane growing regions, but may be applicable to other locations and agricultural industries where knowledge about the future climate is paramount for enhancing forward planning activities.

10. References

Afifi, A., V. A. Clark, et al. (2004). *Computer Aided Multivariate Analysis* Chapman and Hall/CRC. USA

Antonini, M., M. Barlaud, et al. (1992). Image Coding Using Wavelet Transform. *IEEE Transactions on Image Processing* 1(2): 205-220.

Brieman, L. (2001). Random Forests. *Machine Learning* 45(1): 5-32.

Dash, M. and H. Liu (1997). Feature Selection for Intelligent Data Analysis. 1: 131-156.

Diaz-Uriarte, R. and S. Alvarez de Andres (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics* 7(1): 3.

Drosdowsky, W. and L. E. Chambers (2001). Near-Global Sea Surface Temperature Anomalies as Predictors of Australian Seasonal Rainfall. *Journal of Climate* 14(7): 1677-1687.

Duda, R. O. and P. E. Hart (1973). *Pattern Classification and Scene Analysis*, John Wiley and Sons. Canada

Everingham, Y. L., R. C. Muchow, et al. (2002). Enhanced risk management and decision-making capability across the sugarcane industry value chain based on seasonal climate forecasts. *Agricultural Systems* 74: 459-477.

Firth, L. and M.L. Hazelton et al. (2005). Predicting the Onset of Australian Winter Rainfall by Nonlinear Classification. *Journal of Climate*, 18(6), 772 - 781

Friedman, J. H. (1989). Regularized Discriminant Analysis. *Journal of the American Statistical Society* 84(405): 165-175.

George, E. I. (2000). The Variable Selection Problem. *Journal of the American Statistical Society* 95(452): 165 - 175.

Hastie, T., R. Tribshirani, et al. (2001). *The Elements of Statistical Learning: Data Mining, Interference and Prediction*, Springer-Verlag. New York, USA

Johnson, R. A. and D. W. Wichern (2002). *Applied Multivariate Statistical Analysis* Prentice Hall. Upper Saddle River

Jolliffe, I. T. (1986). *Principal Component Analysis*, Springer-Verlag. New York, USA

Jones, K. and Y. L. Everingham (2005). Can ENSO combined with Low-Frequency SST signals enhance or suppress rainfall in Australian sugar growing areas? MODSIM 2005 International Congress on Modelling and Simulation, Melbounre, Modelling and Simulation Society of Australia.

Klopper, E., C. Vogel, et al. (2006). Seasonal Climate Forecasts – Potential Agricultural-Risk Management Tools? *Climatic Change* 76(1): 73-90.

Kumar, P. and E. Foufoula-Georgiou (1997). Wavelet Analysis for Geophysical Applications. *Reviews of Geophysics* 35(4): 385–412.

Landman, W. A. and S. J. Mason (1999). Operational long-lead prediction of South African rainfall using canonical correlation analysis. *International Journal of Climatology* 19(10): 1073-1090.

Mallat, S. G. (1989a). A Theory for Multiresolution Signal Decomposition: The Wavelet Representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 11(7): 674-693.

Mallat, S. G. (1989b). Multiresolution Approximations and Wavelet Orthonormal Bases of $L^2(R)$. *Transactions of the American Mathematical Society* 315(1): 69-87.

Mallet, Y., O. de Vel, et al. (2000). Fundamentals of Wavelet Transforms. In. *Wavelets in Chemistry*, B. Walczak (Eds), Elselvier: Netherlands: 57-79.

Mason, S. J. (1998). Seasonal forecasting of South African rainfall using a non-linear discriminant analysis model. *International Journal of Climatology* 18(2): 147-164.

Muchow, R. C. and A. W. Wood (1996). Rainfall Risk and Scheduling the Harvest of Sugarcane. In. *Sugarcane: Research Towards Efficient and Sustainable Production*, J. R. Wilson, D. M. Hagarth, J. A. Campbell and A. L. Garside (Eds), CSIRO: Brisbane, Australia.

Rencher, A. C. (2002). *Methods of Multivariate Analysis*, John Wiley and Sons. Canada

Singhrattna, N., B. Rajagopalan, et al. (2005). Seasonal forecasting of Thailand summer monsoon rainfall. *International Journal of Climatology* 25(5): 649-664.

Smith, T. M. and R. W. Reynolds (2004). Improved Extended Reconstruction of SST (1854–1997). *Journal of Climate* 17(12): 2466–2477.

Svetnik, V., A. Liaw, et al. (2004). Application of Breiman's Random Forest to Modeling Structure-Activity Relationships of Pharmaceutical Molecules. In. *Multiple Classifier Systems*(Eds): 334-343.

Washington, R. and T. E. Downing (1999). Seasonal Forecasting of African Rainfall: Prediction, Responses and Household Food Security. *The Geographical Journal* 165(3): 255-274.

Wilks, D. S. (1995). *Statistical Methods in the Atmospheric Sciences*, Academic Press Inc. San Diego, USA

Wu, W., Y. Mallet, et al. (1996). Comparison of regularized discriminant analysis linear discriminant analysis and quadratic discriminant analysis applied to NIR data. *Analytica Chimica Acta* 329(3): 257-265.