**Workflow: Annotated pdf, CrossRef and tracked changes**

# PROOF COVER SHEET

| | |
|---|---|
| Journal acronym: | TRES |
| Author(s): | Dugal Harris and Adriaan Van Niekerk |
| Article title: | Feature clustering and ranking for selecting stable features from high dimensional remotely sensed data |
| Article no: | 1500730 |
| Enclosures: | 1) Query sheet<br>2) Article proofs |

Dear Author,

**1. Please check these proofs carefully.** It is the responsibility of the corresponding author to check these and approve or amend them. A second proof is not normally provided. Taylor & Francis cannot be held responsible for uncorrected errors, even if introduced during the production process. Once your corrections have been added to the article, it will be considered ready for publication.

Please limit changes at this stage to the correction of errors. You should not make trivial changes, improve prose style, add new material, or delete existing material at this stage. You may be charged if your corrections are excessive (we would not expect corrections to exceed 30 changes).

For detailed guidance on how to check your proofs, please paste this address into a new browser window: http://journalauthors.tandf.co.uk/production/checkingproofs.asp

Your PDF proof file has been enabled so that you can comment on the proof directly using Adobe Acrobat. If you wish to do this, please save the file to your hard disk first. For further information on marking corrections using Acrobat, please paste this address into a new browser window: http://journalauthors.tandf.co.uk/production/acrobat.asp

**2. Please review the table of contributors below and confirm that the first and last names are structured correctly and that the authors are listed in the correct order of contribution.** This check is to ensure that your name will appear correctly online and when the article is indexed.

| Sequence | Prefix | Given name(s) | Surname | Suffix |
|---|---|---|---|---|
| 1 | | Dugal | Harris | |
| 2 | | Adriaan | Van Niekerk | |

Queries are marked in the margins of the proofs, and you can also click the hyperlinks below.

Content changes made during copy-editing are shown as tracked changes. Inserted text is in red font and revisions have a red indicator ⅄. Changes can also be viewed using the list comments function. To correct the proofs, you should insert or delete text following the instructions below, but **do not add comments to the existing tracked changes.**

# AUTHOR QUERIES

**General points:**

1. **Permissions:** You have warranted that you have secured the necessary written permission from the appropriate copyright owner for the reproduction of any text, illustration, or other material in your article. Please see http://journalauthors.tandf.co.uk/permissions/usingThirdPartyMaterial.asp.
2. **Third-party content:** If there is third-party content in your article, please check that the rightsholder details for re-use are shown correctly.
3. **Affiliation:** The corresponding author is responsible for ensuring that address and email details are correct for all the co-authors. Affiliations given in the article should be the affiliation at the time the research was conducted. Please see http://journalauthors.tandf.co.uk/preparation/writing.asp.
4. **Funding:** Was your research for this article funded by a funding agency? If so, please insert 'This work was supported by <insert the name of the funding agency in full>', followed by the grant number in square brackets '[grant number xxxx]'.
5. **Supplemental data and underlying research materials:** Do you wish to include the location of the underlying research materials (e.g. data, samples or models) for your article? If so, please insert this sentence before the reference section: 'The underlying research materials for this article can be accessed at <full link> / description of location [author to complete]'. If your article includes supplemental data, the link will also be provided in this paragraph. See <http://journalauthors.tandf.co.uk/preparation/multimedia.asp> for further explanation of supplemental data and underlying research materials.
6. The **CrossRef database** (www.**crossref**.org/) has been used to validate the references. Changes resulting from mismatches are tracked in red font.

**AQ1** Please note that the ORCID for Harris Dugal has been created from information provided through CATS. Please correct if this is inaccurate.

**AQ2** The disclosure statement has been inserted. Please correct if this is inaccurate.

**AQ3** Please note that the Funding sections have been created from information provided through CATS. Please correct if these are inaccurate.

**AQ4** The CrossRef database (www.crossref.org/) has been used to validate the references. Mismatches between the original manuscript and CrossRef are tracked in red font. Please provide a revision if the change is incorrect. Do not comment on correct changes.

**AQ5** Please provide missing editor name or institutional editor name for the “Cukur et al., 2015” references list entry.

**AQ6** The reference “Duin et al., 2005” is listed in the references list but is not cited in the text. Please either cite the reference or remove it from the references list.

**AQ7** Please provide missing access date.

**AQ8** Please provide missing editor name or institutional editor name for the “Kuncheva, 2007” references list entry.

**AQ9** Please provide missing access date.

**AQ10** Please provide missing editor name or institutional editor name for the “Sahu and Mishra, 2011” references list entry.

**AQ11** Please provide missing access date.

## How to make corrections to your proofs using Adobe Acrobat/Reader

Taylor & Francis offers you a choice of options to help you make corrections to your proofs. Your PDF proof file has been enabled so that you can mark up the proof directly using Adobe Acrobat/Reader. This is the simplest and best way for you to ensure that your corrections will be incorporated. If you wish to do this, please follow these instructions:

1. Save the file to your hard disk.

2. Check which version of Adobe Acrobat/Reader you have on your computer. You can do this by clicking on the "Help" tab, and then "About".

If Adobe Reader is not installed, you can get the latest version free from http://get.adobe.com/reader/.

3. If you have Adobe Acrobat/Reader 10 or a later version, click on the "Comment" link at the right-hand side to view the Comments pane.

4. You can then select any text and mark it up for deletion or replacement, or insert new text as needed. Please note that these will clearly be displayed in the Comments pane and secondary annotation is not needed to draw attention to your corrections. If you need to include new sections of text, it is also possible to add a comment to the proofs. To do this, use the Sticky Note tool in the task bar. Please also see our FAQs here: http://journalauthors.tandf.co.uk/production/index.asp.

5. Make sure that you save the file when you close the document before uploading it to CATS using the "Upload File" button on the online correction form. If you have more than one file, please zip them together and then upload the zip file.

If you prefer, you can make your corrections using the CATS online correction form.

### Troubleshooting

**Acrobat help:** http://helpx.adobe.com/acrobat.html
**Reader help:** http://helpx.adobe.com/reader.html

Please note that full user guides for earlier versions of these programs are available from the Adobe Help pages by clicking on the link "Previous versions" under the "Help and tutorials" heading from the relevant link above. Commenting functionality is available from Adobe Reader 8.0 onwards and from Adobe Acrobat 7.0 onwards.

**Firefox users:** Firefox's inbuilt PDF Viewer is set to the default; please see the following for instructions on how to use this and download the PDF to your hard drive: http://support.mozilla.org/en-US/kb/view-pdf-files-firefox-without-downloading-them#w_using-a-pdf-reader-plugin

INTERNATIONAL JOURNAL OF REMOTE SENSING
https://doi.org/10.1080/01431161.2018.1500730

Taylor & Francis
Taylor & Francis Group

# Feature clustering and ranking for selecting stable features from high dimensional remotely sensed data

Dugal Harris [a] and Adriaan Van Niekerk [b]

AQ1

[a]Department of Geography and Environmental Studies, Stellenbosch University, Stellenbosch, South Africa; [b]Centre for Geographical Analysis, Stellenbosch University, Stellenbosch, South Africa

**ABSTRACT**

High dimensional remote sensing data sets typically contain redundancy amongst the features. Traditional approaches to feature selection are prone to instability and selection of sub-optimal features in these circumstances. They can also be computationally expensive, especially when dealing with very large remote sensing data sets. This article presents an efficient, deterministic feature ranking method that is robust to redundancy. Affinity propagation is used to group correlated features into clusters. A relevance criterion is evaluated for each feature. Clusters are then ranked based on the median of the relevance values of their constituent features. The most relevant individual features can then be selected automatically from the best clusters. Other criteria, such as computation time or measurement cost, can optionally be considered interactively when making this selection. The proposed feature selection method is compared to competing filter approach methods on a number of remote sensing data sets containing feature redundancy. Mutual information and naive Bayes relevance criteria were evaluated in conjunction with the feature selection methods. Using the proposed method it was shown that the stability of selected features improved under different data samplings, while similar or better classification accuracies were achieved compared to competing methods.

**ARTICLE HISTORY**
Received 22 March 2018
Accepted 2 July 2018

## 1. Introduction

In image classification, the amount of training data required to adequately represent class distributions in feature space increases exponentially as the number of features (variables) is increased – a phenomena known as the 'curse of dimensionality' (Bishop 2003). For finite training samples, increasing the features beyond a certain point results in over-training and a decrease in the classifier accuracy. This so-called 'peaking phenomenon' (Jain, Duin, and Mao 2000) requires the size of the feature set to be reduced to a salient minimum in order to achieve an accurate classification. Support vector machine (SVM) (Burges 1998) and random forest (Breiman 2001) classifiers have become popular in remote sensing, partly because of their lack of sensitivity to the peaking phenomenon (Guyon et al. 2002), but several studies have demonstrated the benefits of feature

**CONTACT** Dugal Harris ✉ dugalh@gmail.com PO Box 180, Newlands, Cape Town, 7725

reduction when these classifiers were applied to high-dimensional data (Guyon et al. 2002; Strobl et al. 2008; Tolosi and Lengauer 2011). Reducing the number of features is also beneficial from the perspective of measurement costs and feature computation time. This is particularly relevant in large-scale remote sensing studies involving Very High Resolution (VHR) imagery, as vast quantities of data require processing.

Two basic approaches to feature set reduction exist: feature selection and feature extraction. In feature extraction, the feature set is mapped into a new feature space of reduced dimensionality (Webb 2002). A major disadvantage of the feature extraction approach is that it requires measurements and computations to produce the full feature set, which can be prohibitively costly. Feature extraction also hinders interpretability as it alters the original representation of the features. A feature extraction approach was thus not followed in this study.

Feature selection involves the selection of a subset of features from the original set according to some criterion of subset performance. The number of possible subsets increases combinatorially with the size of the feature set and it is seldom practical to evaluate all possible subsets (Jain, Duin, and Mao 2000). A variety of search schemes exist for reducing the portion of feature space searched. The fastest and most straightforward search scheme is simply to rank features based on their individual performance and select the best *N*. However, feature ranking approaches are problematic for data sets containing feature redundancy. In these situations, correlated features are ranked similarly, resulting in sub-optimal and redundant feature sets.

More advanced search schemes use greedy sequential approaches, such as forward selection and backward elimination. Compared to the feature ranking approach, greedy search procedures are more likely to find the globally optimal feature set as they explore more of the search space and are less inclined to select multiple redundant features (Webb 2002). The forward selection (FS) approach starts with an empty feature set and proceeds in a number of steps where one feature is added to the selected set at each step. The feature whose selection most improves an accuracy criterion is the one that is selected for that step. The selection process proceeds for a set number of steps or until a stopping criterion is reached (Bishop 2003). The backward elimination (BE) method starts with the full set of features and proceeds in a number of steps where one feature is eliminated from the selected set at each step. The feature whose removal produces the best accuracy according to some criterion is the one eliminated for that step. Again, the BE selection process proceeds for a set number of steps or until a stopping criterion is reached (Bishop 2003). BE is computationally more costly than FS as it begins the evaluation on the full feature set. For the same reason, it also requires adequate data to represent the full feature set.

Feature selection methods can be grouped into filter, wrapper–, and embedded approaches. In the filter approach, generic measures of separability or importance are used to evaluate feature subsets, while in the wrapper approach, the accuracy of a specific classifier trained on the feature subset is used as the selection criterion (Duin and Tax 2005). An embedded approach is one where feature selection is incorporated into the classifier training procedure, such as with random forests (Breiman 2001). Filter approaches have the advantage over wrapper and embedded approaches of making feature selection independent of the classifier, thus allowing for greater flexibility in the choice of classifier (Brown et al. 2012). In general, filter approaches are also computationally more efficient than wrapper approaches. This is an important consideration for large and high dimensional data sets such as those often encountered in remote sensing. This study thus focuses on filter approaches.

High dimensional feature spaces typically contain feature redundancy (Cukur et al. 2015; Tolosi and Lengauer 2011; Yu and Liu 2004). Although feature correlation and redundancy are related, they are not strictly the same thing (Brown et al. 2012; Guyon and Elisseeff 2003). Features can help improve separability when the within-class correlation is stronger than the between-class correlation. We use the term 'redundancy' to refer to correlation of features between classes. The raw bands of multispectral imagery often have significant spectral overlap and consequently are correlated with one another. This spectral overlap will exacerbate the redundancy amongst features derived from these raw bands (Cukur et al. 2015). Hyperspectral imagery is also well-known for containing redundancy amongst the bands (Yuan, Zhu, and Wang 2015).

A number of authors have noted difficulties in selecting features from high dimensional data sets. Kononenko, Šimec, and Robnik-Šikonja (1997), Guyon et al. (2002), Yu and Liu (2004) and Yousef et al. (2007) noted that feature redundancy can have a negative impact on the optimality of feature selection. Feature redundancy not only leads to sub-optimal feature selection, but also makes selected features unstable and sensitive to small changes in the data used for selection (Guyon and Elisseeff 2003; Kalousis, Prados, and Hilario 2007; Li, James Harner, and A Adjeroh 2011; Tolosi and Lengauer 2011).

The increasing availability of high resolution imagery necessitates computationally efficient feature selection techniques robust to high dimensional redundant spaces. In this article, we propose a computationally efficient filter approach feature selection method for addressing the problems of sub-optimality and instability associated with high dimensional, redundant feature spaces of remotely sensed data. We adopt the filter approach due to its separation of feature selection and classification tasks. The method employs affinity propagation to identify clusters of redundant features, and redundancy is reduced by selecting a single representative feature from the most relevant clusters. The method requires no prior knowledge of the number of clusters. Correlation is used to measure feature similarity, which allows a broader encapsulation of feature redundancy than distance measures such as Euclidean distance (X. Chen et al. 2017). Assumptions of linear dependence between features and class labels made in structured sparsity regularization approaches (Gui et al. 2016) are also avoided by selecting features with a relevance heuristic, based on the use of naïve Bayes or mutual information criteria. The proposed method can be fully automated, or it can be used interactively to allow for consideration of computation time and measurement cost. The performance of the proposed method is compared to popular feature selection approaches, on a number of remote sensing data sets. The various feature selection methods are evaluated based on computation time, classification accuracy, and stability of selected features under different data samplings.

## 2. Methods

### 2.1. *Related work*

A number of feature selection approaches have been developed to address the issues of stability and sub-optimality encountered in high dimensional and redundant data. These methods consider the trade-off between feature relevance (i.e. how much information the feature contains about the class labels) and redundancy.

A means of selecting good features from redundant spaces was devised by Yousef et al. (2007), who used a *k*-means algorithm to produce a fixed number of clusters of correlated features. The accuracy of an SVM classifier is found for all the features of each cluster and the lowest performing clusters are eliminated, the remaining features combined, and the process is repeated until a desired number of clusters is reached. A related feature selection method that finds and removes redundancy by clustering features into similar groups was presented by Mitra, Murthy, and Pal (2002). They used a novel clustering algorithm to group correlated features based on a similarity measure they call 'maximal information compression index', which is the smallest eigenvalue of the feature covariance. Sahu and Mishra (2011) also used *k*-means clustering to group redundant features. The best feature, according to an importance measure, is then selected from each cluster. Cukur et al. (2015) proposed a similar method, where redundant features are clustered and top ranked features selected from each cluster using an importance measure called 'minimum redundancy maximum relevance' (mRMR).

A two-step procedure called the 'Fast Correlation-Based Filter' (FCBF) was developed by Yu and Liu (2004). It first creates a reduced set of relevant features, and then removes redundant features from this set using a search procedure based on Markov blanket filtering. A non-linear correlation measure, called symmetrical uncertainty, is used to measure both feature relevance and redundancy. Relevance is measured by how well features are correlated with class labels and redundancy is measured by how well features are correlated with each other.

Wu et al. (2013) compared a number of filter approach feature selection methods for reducing redundancy in three hyperspectral data sets. They used a number of performance measures for comparison, including classifier accuracy, feature stability, and their own criterion called the maximal minimal associated index quotient (MMAIQ). MMAIQ uses Cramer's *V*-test values to trade feature relevance against redundancy and is applied in a forward selection type routine. While the authors concluded that MMAIQ provided the best overall performance, it did not provide good stability for high dimensional data.

A number of feature importance measures (including the FCBF) were incorporated by Brown et al. (2012) into a common theoretical framework. These measures all consider both relevance and redundancy in some way. A comprehensive empirical study was used to compare the performance (in terms of stability and classifier accuracy) of these measures. The study tested the criteria in an FS search scheme, under varying conditions, including redundancy in high dimensional feature spaces. They concluded that joint mutual information (JMI) (Yang and Moody 1999) provides the best feature selection performance overall.

In recent years, a number of feature selection approaches based on structured sparsity regularization have been developed (Chen and Yanfeng 2015; Gui et al. 2016; Wang et al. 2010; X. Chen et al. 2017). Structured sparsity regularization modifies the traditional sparsity regularization approach by incorporating prior knowledge of the group structure of features to improve performance (Gui et al. 2016). The supervised multiview feature selection (SMFS) method of X. Chen et al. (2017) uses a structured sparsity approach that groups features by similarity and uses this similarity structure to address the trade-off between feature relevance and redundancy. In SMFS, features are clustered into homogenous groups or 'views' using affinity propagation (AP) with a squared Euclidean distance similarity measure. A sparse set of features is selected from these views by a joint ℓ1,2-norm minimization of an objective function comprised of loss function and regularization terms. The formation of the loss function assumes a linear dependence between features and class labels. Feature view structure is incorporated into the ℓ1,2-norms so as to encourage the sparsity of selected features within views, while retaining the information of multiple heterogeneous views. The objective function is minimized with quadratic programming, which is computationally expensive compared to greedy search type feature selection methods such as FS and JMI. Feature weights produced by the optimization procedure can be considered an importance measure that trades relevance against redundancy.

## 2.2. *Feature clustering and ranking*

Within the context of the related research overviewed in the previous section, our proposed feature selection method consists of the following three steps:

(1) Perform affinity propagation clustering (Frey and Dueck 2007) of the feature set using the absolute value of the correlation coefficient as the similarity metric.
(2) Rank each cluster's importance by finding the value of a relevance criterion for each individual feature and then finding the median of the feature relevance values in the cluster.
(3) Select a single feature from each of the *N* clusters with best importance scores.

Affinity propagation is a clustering technique that identifies cluster representatives ('exemplars'), and their corresponding clusters, by an iterative scheme of message passing between data points (Frey and Dueck 2007). A matrix of pair-wise similarities and a 'preference' parameter are required as inputs. The preference parameter affects the number of identified clusters and may be chosen automatically based on the values of the similarities. The proposed feature selection method sets the preference parameter to the median of the similarities, which results in a moderate number of clusters (Frey and Dueck 2007). Unlike clustering techniques such as *k*-means, affinity propagation does not require prior knowledge of the number of clusters.

Two kinds of messages, 'availability' and 'responsibility', are passed between data points at each iteration. The values of these messages express the current affinity one point has for choosing another as its exemplar. The responsibility $r(i,k)$ reflects the accumulated evidence that feature *k* is the exemplar for feature *i*, taking into consideration other possible exemplars for feature *i*. The availability $a(i,k)$ reflects the accumulated evidence for how appropriate it would be for feature *i* to choose feature *k* as its exemplar, taking into consideration support from other features for choosing *k* as their exemplar.

To initialise, the availabilities are set to zero, $a(i,k) = 0$. In our method, the similarity $s(i,k)$ between feature *i* and *k*, is set to the absolute value of the correlation coefficient, and the self-similarities, $s(k,k)$, are set to the preference value. At each iteration, the responsibilities are updated using the rule:

$$r(i,k) \leftarrow s(i,k) - \max_{k' \text{ s.t. } k' \neq k} \{a(i,k') + s(i,k')\} \tag{1}$$

The availabilities are correspondingly updated using the rule:

$$a(i,k) \leftarrow \begin{cases} \min\left\{0,\ r(k,k) + \sum\limits_{i' \text{ s.t. } i' \notin \{i,k\}} \max\{0, r(i',k)\}\right\}, & i \neq k \\ \sum\limits_{i' \text{ s.t. } i' \neq k} \max\{0, r(i',k)\}, & i = k \end{cases} \tag{2}$$

The exemplar for feature $i$ is identified by the value of $k$ that maximizes $a(i,k) + \ r(i,k)$. The iterations continue until the clusters (and their corresponding exemplars) remain stable for ten consecutive updates.

We investigated the performance of two different feature relevance measures: the accuracy of a naive Bayes classifier and the mutual information (MI) between the feature and the class labels. The naive Bayes classifier, using a histogram to model class densities, was chosen primarily because it makes no assumptions about the form of the class distributions and can thus provide a generic measure of separability. It is simple, fast, and recognized as being accurate for a variety of problems (Hand and Kerning 2001). The 'naive' assumption of feature independence is of no consequence when testing individual features.

MI is a measure of the dependence between two random variables (Brown et al. 2012). Given two random variables $X$ and $Y$, with probability distributions $p(x)$ and $p(y)$ and joint probability distribution $p(x,y)$, the MI between $X$ and $Y$ is defined as:

$$I(X;Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{3}$$

The MI between a feature and the class labels gives a useful indication of that feature's relevance or importance (Brown et al. 2012). The probability distributions in Equation (3) are not known and are estimated using histograms.

The cluster importance measure for the $k$th cluster is expressed as

$$C_k = \underset{X_j \in G_k}{\text{median}} \, R(X_j, Y) \tag{4}$$

where $G_k$ is the set of features in cluster $k$ ~~and~~ $R(X_j, Y)$ is the MI or naïve Bayes feature relevance measure for feature $X_j$ and class labels $Y$. Once the clusters of similarly relevant features have been ranked according to their importance measures, single features may be selected from the best clusters using an automatic procedure or by the user, taking measurement cost and computation time into account. The process of selecting a representative feature from each of the top ranked clusters reduces redundancy while retaining relevance. For automatic feature selection, the feature with the maximum relevance measure is selected from each of the $N$ best clusters. When criteria of measurement cost or computation time require consideration, the user should hand select features minimizing these values from the $N$ best clusters.

The number of clusters to select, $N$, can be specified by the user based on the size of the training set or by using a grid search with the final classifier accuracy as performance measure. To avoid biased accuracy estimates, all classifier accuracy evaluations, for cluster ranking or selection of $N$, are done on unseen test data using a five-fold cross-validation (Bishop 2003).

### 2.3. *Data sets*

Five remote sensing and one synthetic data set (Table 1) were used for comparing the proposed method against popular existing feature selection methods. The 'difficulty' in the last column of Table 1 is calculated as $N/(mc)$, as in Brown et al. (2012), where $N$ is the number of objects, $m$ the number of features and $c$ the number of classes. Smaller values indicate that the data is less representative of the underlying class distributions, which results in more challenging feature selection and classification tasks. The Spekboom set consists of 46 spectral and textural features derived from four band multispectral, 0.5 m spatial resolution aerial imagery. The classes represent three types of vegetation found in the Little Karoo, a semi-arid region in South Africa. It was created as part of a vegetation mapping project being conducted by the authors.

The two class synthetic data set was generated to have redundancy amongst the features. The first five features for class $j$ were generated from a normal distribution, $N(\mu_j, 1)$, with mean $\mu_j$ and standard deviation of one ($j = 1..2$). The mean, $\mu_j$, of each distribution, was generated from the standard normal distribution, $N(0, 1)$. The same number of objects were generated for each class. To introduce redundancy, an additional five features were generated by adding normally distributed noise, $N(0, 0.25)$, to the original five features. A further five redundant features were similarly generated, but by adding normally distributed noise, $N(0, 0.5)$, to the original features. Finally, two spurious features, sampled from a standard normal distribution, $N(0, 1)$, were added to the data set.

The Statlog Landsat and Urban Land Cover data sets were obtained from the UCI Machine Learning Repository (Lichman 2013). The Statlog Landsat features are generated from six band multispectral pixel values in three by three neighbourhoods. The data set consists of six land cover classes. The features of the Urban Land Cover data set are comprised of multi-scale spectral, size, shape, and textural measures, derived from high resolution aerial imagery (Johnson and Xie 2013).

Kennedy Space Centre (KSC) and Botswana are public hyperspectral data sets with vegetation and land cover classes (GIC 2014). The Botswana data were acquired by the Hyperion sensor on board the NASA EO-1 satellite and consist of 145 bands in the

Table 1. Data sets.

| Name | Abbreviation. | Number of Features | Number of Objects | Number of Classes | Difficulty |
|---|---|---|---|---|---|
| Spekboom | Spekboom | 46 | 57,877 | 3 | 419.40 |
| Synthetic | Synthetic | 17 | 10,000 | 2 | 294.12 |
| Statlog Landsat | Landsat | 36 | 3756 | 6 | 17.39 |
| Urban land cover | Urban | 147 | 261 | 9 | 0.20 |
| Botswana | Botswana | 145 | 1330 | 14 | 0.66 |
| Kennedy space centre | KSC | 176 | 1365 | 13 | 0.60 |

400–2500 nm portion of the spectrum, at a 30 m pixel resolution. The KSC data were acquired by the NASA AVIRIS (Airborne Visible/Infrared Imaging Spectrometer) sensor and consist of 176 bands in the 400–2500 nm range, acquired at a spatial resolution of 18 m.

### 2.4. *Experimental design*

The proposed Feature Clustering and Ranking (FCR) method was compared to a number of other established and competing feature selection methods. We adopted a similar, although reduced, evaluation approach to that of Brown et al. (2012) and Wu et al. (2013). The compared methods included the standard search approaches of ranking, FS and BE. These standard approaches and FCR were each evaluated with two different feature relevance criteria: MI and the naïve Bayes classification accuracy. The MI relevance criterion for FCR and ranking approaches finds the MI between individual features and the class labels. To integrate the MI relevance criterion into FS and BE, it is necessary to compute the MI of a set of multiple candidate features with the class labels. In this situation, the candidate features are first merged into a joint variable and then the MI of the class labels with this joint variable is computed (Brown et al. 2012). We used histograms with ten bins along each dimension to approximate probability densities for both the MI and naive Bayes criteria (Webb 2002), to avoid difficulties and inefficiencies associated with estimating probability densities for continuous variables (Brown et al. 2012).

In addition to the MI and naive Bayes criteria, two other criteria, namely JMI and maximum relevance minimum redundancy (mRMR), were included in our study to represent 'state of the art' performance. Brown et al. (2012) compared the performance of several feature selection criteria in redundant high dimensional spaces and found the JMI criterion gave the best overall performance in terms of classification accuracy and stability. The JMI measure for feature $X_k$ is

$$J_{\mathrm{JMI}}(X_k) = \sum_{X_j \in S} I(X_k X_j; Y) \tag{5}$$

where $Y$ are the class labels and $S$ is the set of previously selected features. JMI considers the MI between the class labels and the joint variables $X_k X_j$, which are the pairwise combinations of the candidate feature with each feature already selected. It measures how well the candidate feature complements selected features in describing the class labels.

The popular mRMR criterion, introduced by Peng, Long, and Ding (2005), expresses the trade-off between feature relevance and redundancy using mutual information measures. The mRMR measure for candidate feature $X_k$ is

$$J_{mRMR}(X_k) = I(X_k; Y) - \frac{1}{|S|} \sum_{X_j \in S} I(X_k; X_j) \tag{6}$$

The first term expresses relevance as the dependence between the candidate feature and class labels, while the second term approximates the redundancy as the mean pairwise dependencies between the candidate and previously selected features. The JMI and mRMR criteria are used in an FS search scheme. The evaluated methods are detailed in Table 2. In this evaluation, the term ‘method’ is used to refer to a combination of search scheme, such as FS, and criterion, such as JMI.

To quantify the stability of the selected features, we used the consistency index developed by Kuncheva (2007). If $A$ and $B$ and are subsets of the full feature set $X$, with $|A| = |B| = k$, $r = |A \cap B|$ and $0 < k < |X| = n$, the consistency index is

$$C(A, B) = \frac{rn - k^2}{k(n - k)} \tag{7}$$

Its value lies in the range $[-1, +1]$, where positive values indicate similar sets, zero indicates a random relation and negative values indicate an anti-correlation between the feature sets (Kuncheva 2007). To evaluate stability for a particular method, we select features from bootstrap samples of the data. The consistency index is found for each pairwise combination of selected features over ten bootstraps of the data and averaged to give a measure of overall stability.

A $k$-nearest-neighbour ($k$-NN) classifier (with $k = 3$) was used to evaluate the accuracy of the features selected by each method. $k$-NN is a generic classifier that makes no assumptions about the data and requires no tuning. While other classifiers may be more accurate in particular situations, $k$-NN allows a relative comparison of the feature selection methods, independent of the influence of classifier tuning for specific data. For each of the feature sets found from the bootstrap samples, the $k$-NN accuracy was found as the average per-class accuracy from a ten-fold cross-validation. For each method and data set combination, an overall accuracy was computed as the average of the bootstrap accuracies.

The number of features to select for each data set was fixed across methods. This parameter was selected by using the accuracy of a $k$-NN classifier ($k = 3$), trained on the first $N$ features selected by FS-NaiveBC, as the criterion in a grid search. A low value of $N$ that achieved good accuracy was selected for each data set. The number of features selected for each data set are detailed in Table 3.

The FCR methods (FCR-MI and FCR-NaiveBC) required some specific treatment to integrate them into the evaluation. After bootstrapping, clusters were assigned unique indices, ensuring identical clusters had the same index. The consistency index was then found using the selected cluster indices rather than feature indices. This was done to simulate hand-selection of preferred features from the best clusters for each bootstrap, while allowing the FCR algorithm to automatically choose the top ranked feature from

**Table 2.** Methods as combinations of search schemes and criteria.

| Search scheme | Criterion | Method Abbreviation |
|---|---|---|
| Forward selection | Joint m~~M~~utual ~~I~~information | FS-JMI |
| Forward selection | Maximum relevance minimum redundancy | FS-MRMR |
| Feature clustering and ranking | Naive Bayes | FCR-NaiveBC |
| Ranking | Naive Bayes | Rank-NaiveBC |
| Forward selection | Naive Bayes | FS-NaiveBC |
| Backward elimination | Naive Bayes | BE-NaiveBC |
| Feature clustering and ranking | Mutual information | FCR-MI |
| Ranking | Mutual information | Rank-MI |
| Forward selection | Mutual information | FS-MI |
| Backward elimination | Mutual information | BE-MI |

Table 3. Feature selection parameters.

| Data Set | Number of Features to Select |
|---|---|
| Spekboom | 6 |
| Synthetic | 5 |
| Statlog Landsat | 4 |
| Urban land cover | 4 |
| Botswana | 6 |
| Kennedy space centre | 7 |

each cluster (for the sake of simplicity and speed). In other words, the cluster index was used to represent the index of the preferred feature that could otherwise have been selected by hand from the cluster contents.

We followed a similar approach to that of Brown et al. (2012) for computing a single 'non-dominated' ranking of the methods that considers stability and accuracy performance simultaneously. The concept of 'Pareto optimality' was used to find a single optimal solution in terms of multiple criteria. In the context of our evaluation, the 'Pareto front' is the set of methods on which no other method can improve without degrading either the accuracy or stability. The methods in this set are called 'non-dominated' (Mishra and Harit 2010). Successive Pareto fronts can be formed iteratively by finding the current Pareto front of the set of methods that excludes members of the previous fronts. A method was thus given a non-dominated rank of *N* if it was a member of the $N^{th}$ Pareto front. The average of the non-dominated ranks for each method over the six data sets was used to produce an overall ranking.

The bulk of the software implementation was done in Matlab$^{TM}$, making use of the PRTools toolbox (TU Delft 2015). The MI, JMI, and mRMR criteria were computed using the FEAST (FEAture Selection Toolbox) C++ implementation (Brown et al. 2012).

## 3. Results and discussion

The results of the stability and accuracy evaluations for each method and data set combination are shown in Figures 1 and 2 respectively. The methods appear along the x-axis in order of their mean stability in Figure 1, and mean accuracy in Figure 2, over the six data sets. The wide range of stabilities confirms the sensitivity of some methods to variations in the data. The method accuracies span a smaller range than the method stabilities. Nonetheless, there were substantial differences in accuracy between the best and worst methods. Compared to the other data sets, the stability of the Spekboom, Synthetic, and Landsat data was noticeably superior. As reflected in the 'difficulty' values in Table 1, these data sets are more representative of the underlying distributions and are thus less sensitive to disturbances.

FCR-NaiveBC and FCR-MI occupy the top two positions for both performance measures. The ranking methods, Rank-MI and Rank-NaiveBC both had poor accuracy performance. This was expected as these methods do not consider feature complementarity and only measure relevance of features in isolation. The relatively poor accuracy and stability of FS-JMI was surprising in the context of the results of Brown et al. (2012), where it produced the best overall performance. The FS-JMI results nevertheless provide a benchmark that helps confirm the usefulness of the FCR method for the type of data investigated in our study.

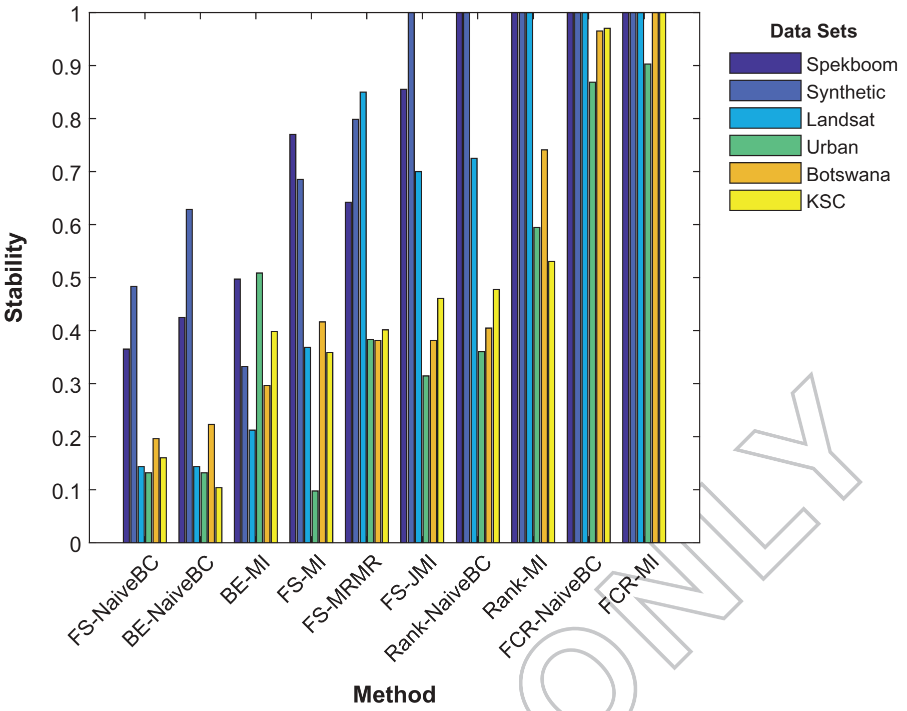

**Figure 1.** Method stability per data set (methods along the x-axis are ordered by their mean stability over the data sets).

As with classifier design, there is a 'curse of dimensionality' problem with computing the MI of joint variables. As the number of features increases, the number of objects needed to adequately represent the feature distribution increases exponentially (Brown et al. 2012). For this reason, the MI criterion is not well suited for evaluating the BE search scheme, which requires computation of the relevance criterion for the full feature set. This likely explains the poor performance of BE-MI in terms of both accuracy and stability. Part of the motivation for the JMI and mRMR formulations is to circumvent this kind of representivity issue by using a low dimensional approximation to MI.

The method execution times, summed over the six data sets, are provided in Table 4. The execution time of FCR competed well with the other methods, although mRMR was the fastest overall. The naïve Bayes criterion is slower to compute than the MI criterion as it uses a five-fold cross-validation to evaluate the classification accuracy, while MI is computed once-off. Methods using the naïve Bayes criterion are consequently slower than their MI counterparts. FS-JMI is faster than the related FS-MI method, as the criterion only requires MI computations between pairwise combinations of features and the class labels, while the MI criterion is evaluated on the combination of all selected features. BE is known to be less efficient than FS (Guyon and Elisseeff 2003) and was the slowest of the tested search schemes.

Table 5 presents the non-dominant ranking of the methods, in terms of both accuracy and stability. The best ranked method overall was FCR-MI, followed by FCR-NaiveBC. The Rank-NaiveBC, Rank-MI, BE-NaiveBC , and BE-MI methods were ranked lowest due to the

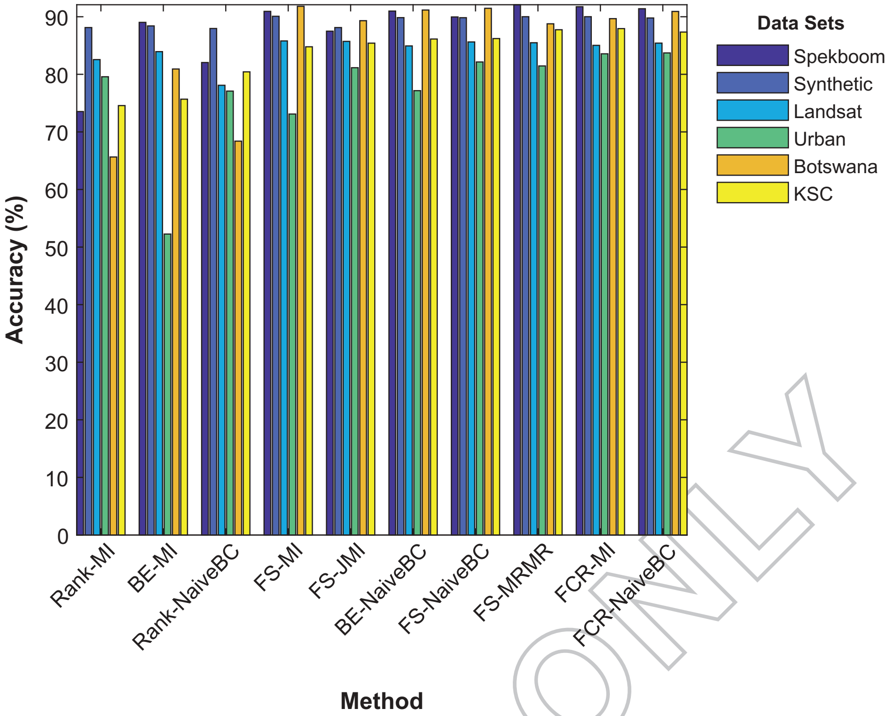

**Figure 2.** Method accuracy per data set (methods along the x-axis are ordered by their mean accuracy over the data sets).

**Table 4.** Method cumulative execution time over all data.

| Method | Time (s) |
|---|---|
| FS-MRMR | 1.40 |
| Rank-MI | 2.11 |
| FS-JMI | 2.46 |
| FCR-MI | 3.01 |
| FS-MI | 12.08 |
| FCR-NaiveBC | 72.61 |
| Rank-NaiveBC | 73.05 |
| BE-MI | 242.39 |
| FS-NaiveBC | 429.76 |
| BE-NaiveBC | 3340.84 |

known limitations of these methods. The FS-MRMR method was competitive on all measures, and was the third ranked method overall.

If the clustering step were omitted, FCR-MI and FCR-NaiveBC would simplify to Rank-MI and Rank-NaiveBC, respectively. FCR-MI and FCR-NaiveBC showed a substantial improvement in performance compared to Rank-MI and Rank-NaiveBC, which lends support to the effectiveness of the clustering step. Considering the combination of the MI and naive Bayes criteria with each search scheme in isolation, there was a general trend for MI to produce better stability and naive Bayes to produce better accuracy. While FCR worked well with either criterion, the results favoured the use of MI as it is

**Table 5.** Non-dominated ranking of methods by accuracy and stability.

| Method | Rank |
|---|---|
| FCR-MI | 1 |
| FCR-NaiveBC | 1.67 |
| FS-MRMR | 2 |
| FS-JMI | 2.50 |
| FS-MI | 2.83 |
| FS-NaiveBC | 2.83 |
| Rank-MI | 3 |
| BE-NaiveBC | 3.33 |
| Rank-NaiveBC | 3.33 |
| BE-MI | 4 |

faster and produced a better non-dominant ranking than naive Bayes. On the whole, the evaluations demonstrate that the proposed FCR method is effective at selecting accurate and stable features from high dimensional remote sensing data containing redundancy.

## 4. Conclusions

Small changes in data sets containing redundancy can result in substantial changes in selected features. Feature redundancy is also known to cause selection of sub-optimal features. This study presented and evaluated a new method for selecting stable and informative features from redundant data by ranking correlated clusters of features. The method uses affinity propagation to identify a moderate number of clusters of correlated and similarly relevant features. It then ranks the clusters using an importance measure, calculated as the median of a relevance criterion evaluated on each individual feature in the cluster. By selecting individual features from the best clusters, a set of informative features is found while simultaneously removing redundancy from the data. These features may be selected automatically based on their relevance, or interactively, taking criteria such as computation time and measurement cost into account. The option to include factors other than relevance and redundancy in determining selected features distinguishes FCR from related feature selection methods; although this option is currently limited to a manual procedure. Future work will investigate ways of combining relevance with other criteria to allow feature selection without user input.

The effectiveness of the proposed FCR method was evaluated by comparing its accuracy, stability, and execution time to a set of popular feature selection methods. Two criteria were tested for measuring feature relevance: the MI between the candidate feature(s) and the class labels, and the accuracy of a naive Bayes classifier trained on the candidate feature(s). Unlike structured sparsity regularization approaches, these relevance criteria do not assume a linear dependence between features and class labels. FCR performed well overall, with both naive Bayes and MI criteria and was the highest ranked method when considering the accuracy and stability measures in combination. Another benefit of FCR is its relative speed compared to greedy search FS and BE type methods. Ever increasing quantities of high spatial and spectral resolution remote sensing data are being produced and require interpretation (Chi et al. 2016). In this context, instability and sub-optimality associated with feature selection from high dimensional redundant data will become increasingly important. Computationally efficient techniques, such as FCR, are required to address these challenges.

## Acknowledgments

This work was supported by funding from the Department of Environmental Affairs (DEA) via the Working for Natural Resources Programme. DEA was otherwise not involved in this research. The financial assistance of the National Research Foundation (NRF) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the authors and are not necessarily to be attributed to the NRF. We would like to thank www.linguafix.net for language editing.

## Disclosure statement

AQ2 No potential conflict of interest was reported by the authors.

## Funding

AQ3 This work was supported by the Department of Environmental Affairs;National Research Foundation.

## ORCID

Dugal Harris http://orcid.org/0000-0003-3480-5707
Adriaan Van Niekerk http://orcid.org/0000-0002-5631-0206

## References

Bishop, C. M. 2003. *Neural Networks for Pattern Recognition*. New York: Oxford University Press. AQ4 ~~doi:10.1002/0470854774.~~

Breiman, L. 2001. “Random Forests.” *Machine Learning* 45 (1): 5–32. doi:10.1023/A:1010933404324.

Brown, G., A. Pocock, M.-J. Zhao, and M. Lujan. 2012. “Conditional Likelihood Maximisation: A Unifying Framework for Mutual Information Feature Selection.” *Journal of Machine Learning Research* 13: 27–66. doi:10.1016/j.patcog.2015.11.007.

Burges, C. J. C. 1998. “A Tutorial on Support Vector Machines for Pattern Recognition.” *Data Mining and Knowledge Discovery* 2 (2): 121–167. doi:10.1023/A:1009715923555.

Chen, X., and G. Yanfeng. 2015. “Class-Specific Feature Selection with Local Geometric Structure and Discriminative Information Based on Sparse Similar Samples.” *IEEE Geoscience and Remote Sensing Letters* 12 (7): 1392–1396. doi:10.1109/LGRS.2015.2402205.

Chen, X., G. Zhou, Y. Chen, G. Shao, and G. Yanfeng. 2017. “Supervised Multiview Feature Selection Exploring Homogeneity and Heterogeneity with L1,2 -Norm and Automatic View Generation.” *IEEE Transactions on Geoscience and Remote Sensing* 55 (4): 2074–2088. doi:10.1109/TGRS.2016.2636329.

Chi, M., A. Plaza, J. A. Benediktsson, Z. Sun, J. Shen, and Y. Zhu. 2016. “Big Data for Remote Sensing: Challenges and Opportunities.” *Proceedings of the IEEE* 104 (11): 2207–2219. doi:10.1109/JPROC.2016.2598228.

Cukur, H., H. Binol, F. S. Uslu, Y. Kalayci, and A. Bal. 2015. “Cross Correlation Based Clustering for Feature Selection in Hyperspectral Imagery.” In *2015 9th International Conference on Electrical and Electronics Engineering (ELECO)*, 232–236. Bursa: IEEE. doi:10.1109/ELECO.2015.7394552.

Duin, R. P. W., and ~~M. J. T. David~~. 2005. “Statistical Pattern Recognition.” In *Handbook of Pattern Recognition and Computer Vision*, C. H. Chen and P. S. P. Wang edited by, *3rd Ed.*, 1–21. Singapore:World Scientific. doi:10.1142/9789812775320_0001

Frey, B. J., and D. Dueck. 2007. “Clustering by Passing Messages between Data Points.” *Science* 315 (5814): 972–976. doi:10.1126/science.1136800.

GIC. 2014. “Hyperspectral Remote Sensing Scenes.” http://www.ehu.eus/ccwintco/index.php?title=Hyperspectral_Remote_Sensing_Scenes.

Gui, J., Z. Sun, J. Shuiwang, D. Tao, and T. Tan. 2016. "Feature Selection Based on Structured Sparsity: A Comprehensive Study." *IEEE Transactions on Neural Networks and Learning Systems* 28 (7): 1–18. doi:10.1109/TNNLS.2016.2551724.

Guyon, I., and A. Elisseeff. 2003. "An Introduction to Variable and Feature Selection." *Journal of Machine Learning Research* 3: 1157–1182. doi:10.1016/j.aca.2011.07.027.

Guyon, I., J. Weston, S. Barnhill, and V. Vapnik. 2002. "Gene Selection for Cancer Classification Using Support Vector Machines." *Machine Learning* 46 (1–3): 389–422. doi:10.1023/A:1012487302797.

Hand, D. J., and Y. Kerning. 2001. "Idiot's Bayes - Not So Stupid After All?" *International Statisitical Review* 69 (3): 385–398.

Jain, A. K., R. P. W. Duin, and J. Mao. 2000. "Statistical Pattern Recognition: A Review." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22 (1): 4–37. doi:10.1109/34.824819.

Johnson, B., and Z. Xie. 2013. "Classifying a High Resolution Image of an Urban Area Using Super-Object Information." *ISPRS Journal of Photogrammetry and Remote Sensing* 83 (September): 40–49. doi:10.1016/j.isprsjprs.2013.05.008.

Kalousis, A., J. Prados, and M. Hilario. 2007. "Stability of Feature Selection Algorithms: A Study on High-Dimensional Spaces." *Knowledge and Information Systems* 12 (1): 95–116. doi:10.1007/s10115-006-0040-8.

Kononenko, I., E. Šimec, and M. Robnik-Šikonja. 1997. "Overcoming the Myopia of Inductive Learning Algorithms with RELIEFF." *Applied Intelligence* 7 (1): 39–55. doi:10.1023/A:1008280620621.

Kuncheva, L. I. 2007. "A Stability Index for Feature Selection." In *International Multi-Conference: Artificial Intelligence and Applications*, 390–395. Innsbruck, Austria: IASTED.

Li, S., E. James Harner, and D. A Adjeroh. 2011. "Random KNN Feature Selection - a Fast and Stable Alternative to Random Forests." *BMC Bioinformatics* 12 (1): 450. doi:10.1186/1471-2105-12-450. BioMed Central Ltd.

Lichman, M. 2013. "UCI Machine Learning Repository." http://archive.ics.uci.edu/ml.

Mishra, K. K., and S. Harit. 2010. "A Fast Algorithm for Finding the Non Dominated Set in Multi Objective Optimization." *Multi-Objective Optimization Using Evolutionary Algorithms* 1 (25): 35–39.

Mitra, P., C. A. Murthy, and S. K. Pal. 2002. "Unsupervised Feature Selection Using Feature Similarity." *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI* 24 (3): 301–312. doi:10.1109/34.990133.

Peng, H., F. Long, and C. Ding. 2005. "Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy." *IEEE Trans. On Pattern Analysis and Machine Intelligence* 27 (8): 1226–1238. doi:10.1109/TPAMI.2005.159.

Sahu, B., and D. Mishra. 2011. "A Novel Approach for Selecting Informative Genes from Gene Expression Data Using Signal-to-Noise Ratio and t-Statistics." In *2011 2nd International Conference on Computer and Communication Technology (ICCCT-2011)*, 5–10. Allahabad, India: IEEE. doi:10.1109/ICCCT.2011.6075207.

Strobl, C., A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis. 2008. "Conditional Variable Importance for Random Forests." *BMC Bioinformatics* 9 (January): 307. doi:10.1186/1471-2105-9-307.

Tolosi, L., and T. Lengauer. 2011. "Classification with Correlated Features: Unreliability of Feature Ranking and Solutions." *Bioinformatics (Oxford, England)* 27 (14): 1986–1994. doi:10.1093/bioinformatics/btr300.

TU Delft. 2015. "PRTools." http://prtools.org/prtools/.

Wang, H., F. Nie, H. Huang, S. Risacher, A. J. Saykin, and L. Shen; ADNI. 2010. "Efficient and Robust Feature Selection via Joint ℓ2, 1-Norms Minimization." *Advances in Neural Information Processing Systems* 23 (March): 1813–1821. doi:10.1016/j.neuroimage.2010.10.081.

Webb, A. R. 2002. *Statistical Pattern Recognition*. Chichester, UK: John Wiley & Sons. doi:10.1002/0470854774.

Wu, B., C. Chen, T. M. Kechadi, and L. Sun. 2013. “A Comparative Evaluation of Filter-Based Feature Selection Methods for Hyper-Spectral Band Selection.” *International Journal of Remote Sensing* 34 (22): 7974–7990. doi:10.1080/01431161.2013.827815.

Yang, H. H., and J. Moody. 1999. “Data Visualization and Feature Selection: New Algorithms for Nongaussian Data.” *Advances in Neural Information Processing Systems* 12 (Mi): 687–693.

Yousef, M., S. Jung, L. C. Showe, and M. K. Showe. 2007. “Recursive Cluster Elimination (RCE) for Classification and Feature Selection from Gene Expression Data.” *BMC Bioinformatics* 8: 144. doi:10.1186/1471-2105-8-144.

Yu, L., and H. Liu. 2004. “Efficient Feature Selection via Analysis of Relevance and Redundancy.” *Journal of Machine Learning Research* 5 (2004): 1205–1224.

Yuan, Y., G. Zhu, and Q. Wang. 2015. “Hyperspectral Band Selection by Multitask Sparsity Pursuit.” *IEEE Transactions on Geoscience and Remote Sensing* 53 (2): 631–644. doi:10.1109/TGRS.2014.2326655.