Discovering Prediction Model for Environmental Distribution Maps

Ke Zhang[1,2], Huidong Jin[1,2], Nianjun Liu[1,2], Rob Lesslie[3], Lei Wang[1,2],
Zhouyu Fu[1,2] and Terry Caelli[1,2]

[1] Research School of Information Sciences and Engineering (RSISE) Australian National University
[2] National ICT Australia (NICTA), Canberra Lab, ACT, Australia
[3] Bureau of Rural Sciences (BRS), Canberra, Australia
{ke.zhang, huidong.jin, nianjun.liu, lei.wang, zhouyu.fu,
terry.caelli}@rsise.anu.edu.au
rob.lesslie@brs.gov.au

Abstract. Currently environmental distribution maps, such as for soil fertility, rainfall and foliage, are widely used in the natural resource management and policy making. One typical example is to predict the grazing capacity in particular geographical regions. This paper uses a discovering approach to choose a prediction model for real-world environmental data. The approach consists of two steps: (1) model selection which determines the type of prediction model, such as linear or non-linear; (2) model optimisation which aims at using less environmental data for prediction but without any loss on accuracy. The latter step is achieved by automatically selecting non-redundant features without using specific models. Various experimental results on real-world data illustrate that using specific linear model can work pretty well and fewer environment distribution maps can quickly make better/comparable prediction with the benefit of lower cost of data collection and computation.

Keywords: Environmental distribution map, prediction model, model selection, feature selection

1 Introduction

Technologies of analysing spatial data such as Environmental Distribution Maps (EDM) have raised great expectations for coping with the natural resource management and policy making. As social and ecological development are becoming more intensively linked through time, it would be very beneficial in the socio–environmental policy making if the human effects on ecosystem are evaluated/predicted with high accuracy. For example, in the planning of land use, we should carefully consider that which activities may generate negative effects on the regional and the local environmental issues. To make an appropriate land plan, the potential human influence to ecosystem need to be well predicted. As shown in Fig 1, the socio-environmental land use planning model [**?**] is dynamic, where the land planning decisions can be adjusted according to the prediction of

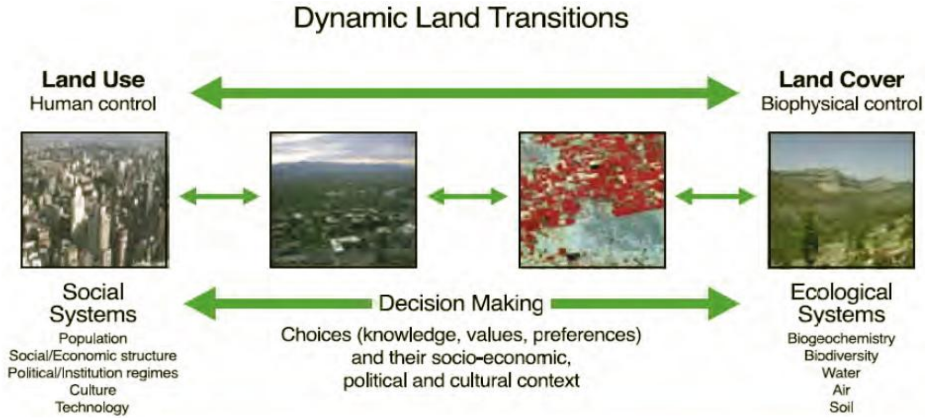


Fig. 1. The framework of socio-environmental land system [**?**].

environmental issues. In those prediction tasks, all decisive factors represented by large–scale spatial data, which may impact the target assessment and formulation, should be selected and considered carefully.

As many geographic and data mining researchers have been working on the spatial data analysis, several prediction models/tools for crops production, plants in a certain ecosystem and other environmental specific fields have been proposed. In 1998, Priya et al. [**?**] proposed a multi-criteria prediction model for crop production, which linearly combines several decisive factors together. A GIS-based plant prediction model was proposed by Horssen et al. [**?**], which uses a geostatistical interpolation method to construct spatial patterns of relevant ecological factors. Besides those academic papers, prediction models based on spatial data analysis are also included in several GIS tools, such as IDRISI and MCAS-S [**?**], which are often customised for particular problems. However, there are two issues associated with these prediction models: (1) Do the domain experts need to specify a proper type of models for prediction tasks? (2) Are all inputting decisive factors they suggested necessary? The answers are normally no.

In this paper, based on a case study on the EDM data provided by Australian Bureau of Rural Sciences (BRS), we study the common problem existing in the current prediction models from two angles: model selection and model optimisation. Because of the similarity of prediction problems, the proposed methodology can also be applied on other prediction tasks.

The remainder of this paper is organized as follows. In Section 2, the BRS prediction problems and our abstraction framework for prediction problems are introduced. The method of model selection and the procedure of model optimisation are discussed in Section 3 and Section 4 respectively for a specific case of prediction problem. The experimental results are shown in Section 5 followed by concluding comments in Section 6.

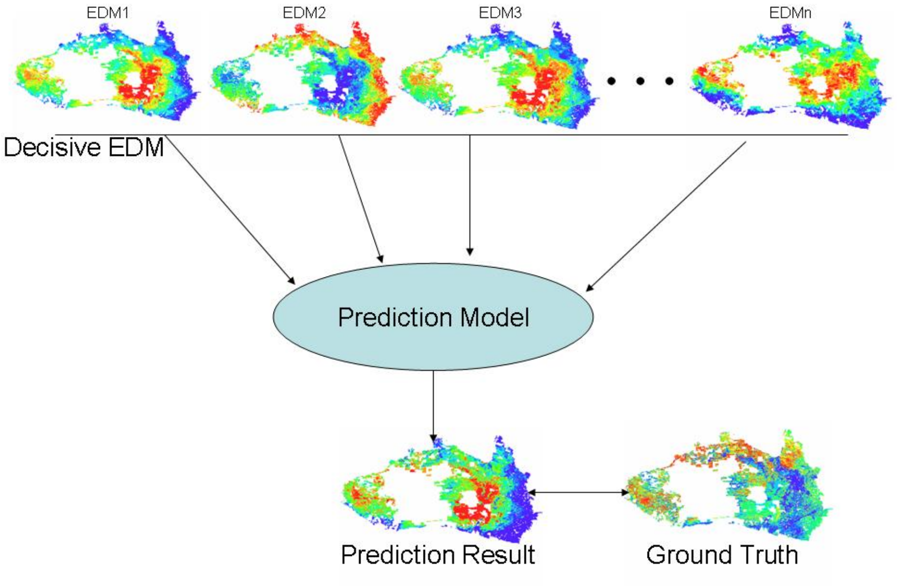


Fig. 2. The framework of our prediction model.

2 Problems and Framework

Australian Bureau of Rural Sciences (BRS), which developed the GIS-based decision making tool MCAS–S, provides the scientific advice for government environmental policy making by analysing the prediction of environmental issues. Generally, domain experts in BRS set up the prediction model manually. To illustrate the prediction model clearly, we take a specific case for the prediction of graze total stock in Australian BRS as an example. As shown in the top half of Fig. 2, the decisive EDMs are suggested by domain experts. They include 9 decisive factors: soil fertility/carbon/nitrogen/phosphor, annual/winter/spring rainfall amount, forage productivity, mean annual normalised difference vegetation index (NDVI mean) and mean annual net primary production (NPP mean). The decisive EDMs are assumed to be combined linearly, and their corresponding interactive weights were manually selected based on domain knowledge.

However, we expected that the prediction model could input the decisive maps and output a predicted target map with out restriction on prediction model type. The purpose of the prediction modeling is to minimise the difference between the real–world target map (ground truth) and the output of the prediction model (predicted target map). The framework of our prediction model is illustrated in Fig. 2. In order to minimise the individual impacts on the prediction

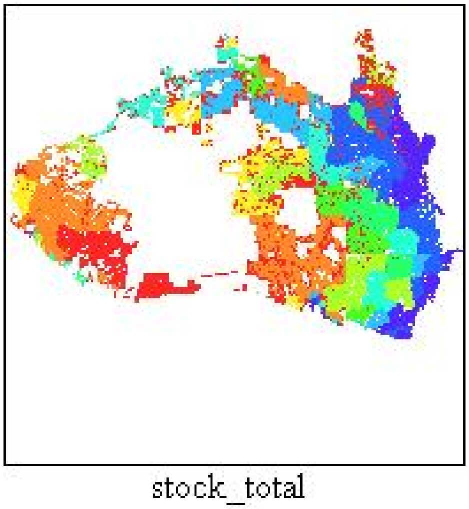


Fig. 3. The statistics of graze total stock as ground truth.

performance, we use a part of statistics data of graze total stock (ground truth) as training data to learn the parameters in our prediction model (see Fig. 3).

3 Prediction Model Selection

In the EDM prediction tasks, for the various kinds of decisive factors (inputting data), the accuracy of existing prediction methods may show significant differences. For the efficiency of the discovering procedure, the type of prediction models would be predetermined. Three widely used algorithms are employed for determining the model type.

3.1 Least Square Fitting (LSF)

The linear least squares fitting technique is the simplest and most commonly applied regression model and provides a solution to the problem of finding the best fitting straight line through a set of points [**?**]. In this method, we assume that those decisive EDMs are combined linearly:

$$AW = T \tag{1}$$

where A is the matrix of inputting data (all vectorised data of EDMs is ranged in column), T is the target data vector (vectorised from the target/training EDM), $W = [w_1, w_2, \cdots, w_N]^T$ is the weight vector of their corresponding decisive EDMs, and N is the number of decisive maps.

The LSF algorithm estimates W as follows:

$$W = (A^T A)^{-1} A^T T \tag{2}$$

where A^T is the transpose of A and A^{-1} is its inverse matrix.

Thus, given the inputting decisive data and the training target, the interactive weights of decisive EDMs can be estimated by Eq 2.

3.2 Support Vector Machine (SVM) Regression

The SVM algorithm is a generalisation of the *Generalised Portrait* algorithm. It was first developed at AT&T Bell Laboratories by Vapnik and his co-workers [**?**] [**?**]. Suppose we are given training data $\{(x_1, y_1), \cdots, (x_l, y_l)\} \subset X \times R$, where X denotes the space of the inputting patterns (e.g., $X = R^d$). The purpose of SVM is to find a function $f(x)$ that has at most ε deviation from the actually obtained targets y_i for all the training data, and at the same time is as flat as possible. Suppose function $f(x)$ is linear then it takes the form:

$$f(x) = \langle w, x \rangle + b \;\text{ with }\; w \in X, b \in R \tag{3}$$

where $\langle a, b \rangle$ denotes the dot product in X. The following optimisation problem is solved to obtain the weight vector w.

$$\begin{array}{ll} \text{Minimise} & \frac{1}{2}\langle w, w \rangle + C \times \sum_{i=1}^{l}(\xi_i + \xi_i^*), \qquad (4) \\ \text{Subject to:} & y_i - \langle w, x_i \rangle - b \leq \varepsilon + \xi_i \qquad (5) \\ & \langle w, x_i \rangle + b - y_i \leq \varepsilon + \xi_i^* \\ & \xi_i, \xi_i^* \leq 0. \end{array}$$

The non-linear SVM algorithm is similar with the linear one except that every inner product is replaced by a non-linear kernel function. In this paper we use Gaussian kernel function. This allows the algorithm to fit the maximum-margin hyper plane in the transformed feature space [**?**]. For the prediction task, the inputting data in Eq 3 is a matrix containing all data of decisive EDMs in the vector form. Let us say that $x = [x^1, x^2, \cdots, x^N]$, where x^n is the data vector of the n^{th} EDM. The expression of function $f(x)$ can be regarded as the prediction model that combines the inputting EDM data non-linearly to approximate the target.

3.3 Neural Networks

The neural network is a powerful data modeling tool that is able to capture and represent complex input/output relationships. In the EDM prediction case, we chose feed-forward neural network, which functions as follows: each neuron receives a signal from the neurons in the previous layer, and each of those signals is multiplied by a separate weight value. The weighted inputs are summed, and passed through a limiting function that scales the output to a fixed range of values. The output of the limiter is then broadcast to all of the neurons in the next layer [**?**]. When the decisive factors in a prediction task have a very complicated correlation, the predicting performance of Neural Networks may be better than the linear method and comparable to SVM. For our prediction tasks, each inputting data $x_1, x_2, \cdots, x_N$ can be regarded as a set of data extracted from decisive EDMs and they will be mapped non-linearly in the hidden nodes.

4 Model Optimisation

Since the data collection for the decisive factors takes the main proportion of the project cost, minimising the amount of inputting data will be significantly beneficial. Let F be a full set of decisive factors (can be regarded as features) and T is the target we want to predict. In general, the goal of feature selection can be formalised as selecting a minimum subset F^* such that $P(T|F^*)$ is equal or as close as possible to $P(T|F)$, where $P(T|F^*)$ is the posterior probability distribution of the target given the feature values in F [**?**]. We call such a minimum subset F^* an optimal subset. The feature selection method used in this paper is Redundancy Based Filter (RBF). The basic idea is using the concept of redundant cover to determine which features should be removed. The correlation between features and the target values are used to determine the features which form a redundant cover for others. There exist broadly two types of measures for their correlations: linear and non-linear [**?**]. Since linear correlation measures may not be able to capture correlations that are not linear in nature, in the approach we adopt a non-linear correlation measure based on the information-theoretical concept of entropy, a measure of the uncertainty of a random variable [**?**]. The entropy of a variable X is defined as

$$H(X) = -\sum_i P(x_i) \log_2 P(x_i) \tag{6}$$

and the entropy of X conditioned on variable Y is defined as

$$H(X|Y) = -\sum_j P(y_j) \sum_i P(x_i|y_j) \log_2 P(x_i|y_j) \tag{7}$$

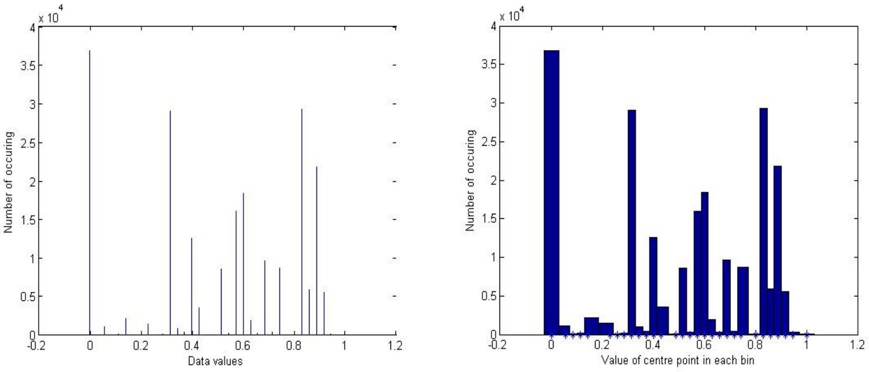
where $P(x_i)$ is the prior probability for all values of X, and $P(x_i|y_j)$ is the conditional probability of X given the values of Y. The amount by which the entropy decrease of X after conditioning reflects additional information about X provided by Y and is called information gain, given by

$$IG(X|Y) = H(X) - H(X|Y). \tag{8}$$

The information gain tends to favour variables with greater differences and can be normalised by their corresponding entropy. We use symmetrical uncertainty (SU) to measure information gains of features and it is defined as:

$$SU(X, Y) = 2\left[\frac{IG(X|Y)}{H(X) + H(Y)}\right]. \tag{9}$$

The value of SU is ranged within [0, 1]. A value of 1 indicates that knowing the values of either feature completely predicts the values of the other; a value of 0 indicates that variables X and Y are independent.



(a) (b)

Fig. 4. (a) is the histogram of the Soil Fertility Environmental Distribution (SFED) expression data. (b) is the histogram of SFED after equal frequency discretisation into 30 intervals.

Since the huge differences of the number of discrete values among EDM data, varying from 70 to 2,000, the EDM data is required to discretise into a same scale before calculating their entropy. For the convenience of computation, we discretise each feature into 30 bins with equal occurring frequency. A result of discretisation for one EDM is illustrated in Fig. 4.

In order to select the non-redundant features explicitly, we differentiate two types of correlation between the features and the target [**?**]:

Individual T-correlation Symmetrical Uncertainty (ISU): The correlation (represented by SU) between any feature F_i and the target T is denoted by ISU_i.

Combined T-correlation Symmetrical Uncertainty (CSU): The correlation (represented by SU) between any pair of features F_i and F_j $(i \neq j)$ and the target T is denoted by CSU_{ij}. In the computation of CSU, we treat the pair of features F_i and F_j as one single feature $F_{i,j}$.

We assume that a feature with a larger individual T-correlation value contains by itself more information about the target than a feature with a smaller individual T-correlation value. For two features F_i and F_j $(i \neq j)$ with $ISU_i > ISU_j$, we choose to evaluate whether feature F_i can form an approximate redundant cover for feature F_j in order to maintain more information about the target. In addition, if combining F_j with F_i does not provide more predictive power in determining the target than F_i alone, we heuristically decide that F_i forms an approximate redundant cover for F_j. Therefore, an approximate redundant cover can be defined as: For two features F_i and F_j, if $ISU_i \geq ISU_j$ and $ISU_i \geq CSU_{i,j}$, F_i forms an approximate redundant cover for F_j.

The RBF feature selection algorithm can be expressed as follows [**?**]:

1. Order features based on decreasing ISU values.

2. Initialise F_i with the first feature in the list.
3. Find and remove all features for which F_i forms an approximate redundant cover.
4. Set F_i as the next remaining feature in the list and repeat step 3 until the end of the list.

The algorithm described above can determine the redundant features automatically and it can select non-redundant features independent of the prediction model. For the independency of prediction models, the RBF shows an obvious advantage for EDM feature selection. Since the uncertainty of combination models in the environmental prediction problems, we may not guarantee the performance of selected features by only using a specified prediction model. Therefore, based on the RBF algorithm, the comparably reliable features could be selected among the batch of decisive EDM factors, as substantiated in Table 2.

5 Experimental Results

5.1 Data Description

As mentioned in Section 2, 9 decisive EDMs were recommended to predict the potential production of graze, and we had an EDM of statistics of graze total stock as the training/testing data. All of those EDM data were provided by Australian BRS, with the size of 700×880, float format. Those EDM data had been normalised into the range of [0, 1] as a pre-processing step. The algorithms mentioned in this paper were implemented by MATLAB, and performed on a computer with 3.2GHz CPU.

5.2 Prediction Model Selection

In our experiments, we randomly selected 80% of the original data (graze stock statistics) as training data and take the rest as testing data, and independently repeated this procedure for 30 times with random partitions of training and testing data for each run. The performances of evaluation method were measured by mean square error and their standard deviation (Std.) as well as running time. Table 1 shows the results of training and testing errors for 3 different models.

Table 1. The prediction accuracy for each method based on 30 independent runs. The best one within a row is indicated in bold.

	LSF	SVM	NN
Training error(mean)	**3.38e-6**	6.67e-5	9.38e-6
Training error(Std.)	2.96e-6	1.89e-5	7.81e-6
Testing error(mean)	**9.53e-6**	3.07e-5	2.63e-5
Testing error(Std.)	1.02e-5	1.13e-4	8.11e-6

As shown in Table 1, the difference of testing errors between LSF and SVM is quite significant: $SVM_{TestingError} - LSF_{TestingError}$=2.11e-5, which is much larger than the standard deviation of LSF's testing error. Similarly, we can say that the accuracy of LSF is significantly higher than that of Neural Networks. Thus, LSF shows obvious advantages compared with the other two methods.

The curve of computing time for each prediction method is illustrated in Fig. 5.

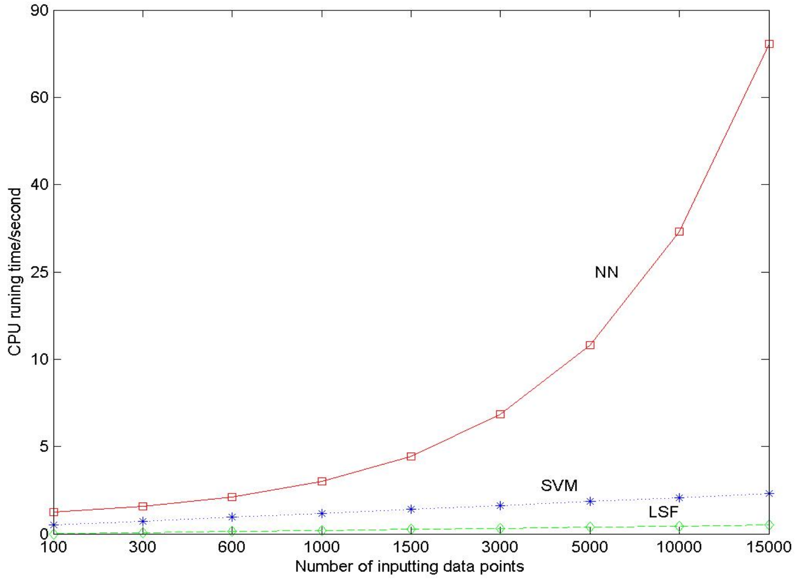


Fig. 5. Curves of computing time to training prediction methods with different training sample sizes.

As shown in Table 1 and Fig. 5, it is clear that the linear method (LSF) has higher efficiency in evaluating the prediction model in this project, which has the highest prediction accuracy with least computing cost among those experimented methods. Thus, we can suggest that it would be beneficial if the LSF prediction method can be used to set up the prediction model rather than using a more complicated non-linear prediction methods. This is also confirmed by domain experts.

However, we cannot guarantee that the linear method is suitable to each GIS-based prediction case because of different kinds of decisive factors and large variations in their correlations. From this viewpoint, we may suggest that, for every GIS-based prediction case, model selection should be pre-performed in order to find the prediction method which suits the current case best.

5.3 Model Optimisation

As mentioned in the previous section, a model-independent feature selection method was used in the model optimisation. In order to make the entropy of features comparable, the histograms of all EDMs and the target were discretised into 30 bins with an approximately equal occurring frequency. And then, the conditional probability tables of each feature/combined features given target could be calculated by counting their corresponding discretised data. In this prediction case, we calculated 9 conditional probability tables $P(F_i|T)$, and 81 combined conditional probability tables $P(F_{i,j}|T)$. In the experiments, considering acceptable computational load, we randomly cropped 100,000 data points from the 9 decisive EDMs and the target map. Based on the algorithm described in Section 4, the feature selector removed the decisive EDM of "rainfall amount in winter/spring" as a redundant feature. To check the reliability of the selection result, we had performed a performance comparison of all methods (LSF, SVM and NN) for using all the features and only the 8 selected features. The comparison results are listed in Table 2.

Table 2. Performance comparison by using the 9 and the 8 features based on 30 independent runs. The best one within a row indicated in bold.

	LSF		SVM		NN	
Number of features	9	8	9	8	9	8
Training error(mean)	**3.37e-6**	1.12e-5	6.67e-5	3.03e-5	9.38e-5	1.33e-5
Training error(Std.)	2.95e-6	3.65e-6	1.89e-5	1.03e-4	7.81e-6	1.33e-6
Testing error(mean)	9.53e-6	**2.69e-6**	3.07e-5	1.23e-5	2.63e-5	7.28e-6
Testing error(Std.)	1.01e-5	1.09e-5	1.13e-4	3.32e-4	8.11e-6	9.47e-6
CPU running time/s	0.0381	**0.0349**	0.1536	0.1390	32.6086	27.0955

As shown in Table 2, the feature removed in the RBF algorithm is redundant, and the rest of features can make a better/comparable prediction with lower computation load.

In order to further confirm the effect of the RBF method, we employed "wrapper" feature selector to validate the result of RBF. Our experiments showed that reducing the decisive EDM of "rainfall amount in winter/spring" was the only positive action comparing with the other 8 features. To verify this conclusion, we performed additional experiments that randomly reduced a pair of features and calculated the influence of the rest of 7 features. And also the experiments showed that none of set of 7 features has a comparable prediction accuracy with that of using 9 features or 8 features.

6 Conclusions

We have presented a discovering method to choose a prediction model for EDM data for Australian Bureau of Rural Sciences. The discovering procedure consists of two procedures: model selection by comparing the performance of 3 prediction models which can be learned from ground truth data; and model optimisation which aims to use less environmental data for prediction but without any loss on accuracy. Various experimental results have shown that using a specific linear model can work pretty well and fewer EDMs can quickly make better/comparable prediction with lower cost of data collection and computation. It means that this model may help Australian BRS save a lot of resources in the real-world application. In the future work, we will incorporate spatial information into the prediction model to enhance the accuracy.

Acknowledgement Financial support for this project from the Australian Research Council is gratefully acknowledged. NICTA is funded by the Australian Government's Department of Communications, Information Technology, and the Arts and the Australian Research Council through Backing Australia's Ability and the ICT Research Centre of Excellence programs.

References

1. Hill, M.: The global land project: An international context for australian analysis of human transformation of ecosystems and landscapes. In: Australian Bureau of Rural Sciences Seminar Series Presents. (2005) Accessed on 30 Jan 2007. `http://www.affashop.gov.au/PdfFiles/brs_seminar_25nov05.pdf`.
2. Priya, S., Shibasaki, R., Ochi, S.: Soil erosion and crop production: A modeling approach. In: Proceeding of Global Environmental Symposium organized by Japanese Society of Civil Engineers. (1998) 175–180
3. Horssen, P., Schot, P., Barendregt, A.: A gis-based plant prediction model for wetland ecosystems. Landscape Ecology **14** (1999) 253–265
4. Hill, M., Lesslie, R., Barry, A., Barry, S.: A simple, portable, spatial multi–criteria analysis shell–MCAS–S. In: International Symposium on Modelling and Simulation, University of Melbourne. (2005)
5. Chatterjee, S., Hadi, A., Price, B.: Simple linear regression. Regression Analysis by Example, 3^{rd} ed, New York:Wiley (2000) 21–50
6. Vapnik, V., Lerner, A.: Pattern recognition using generalized portrait method. Automation and Remote Control. **24** (1963) 774–780
7. Vapnik, V., Golowich, S., Smola, A.: Suppor vector method for function approximation, regression estimation and signal processing. Advances in Neural Information Processing Systems. **9** (1997) 281–287
8. Cristinaini, N., Shawe-Taylor, J.: An introduction to support vector machines and other kernel–based learning methods. Cambridge University Press. ISBN 0–521–78019–5 (2000)
9. McCollum, P.: An introduction to back-propagation neural networks. encoder. The Newsletter of Seattle Robotics Society. (1997)
10. Yu, L., Liu, H.: Redundancy based feature selection for microarray data. In: KDD. (2004) 737–742
11. He, H., Jin, H., Chen, J.: Automatic feature selection for classification of health data. In: Australian Conference on Artificial Intelligence. (2005) 910–913