Remote Sens. **2009**, *1*, 875-895; doi:10.3390/rs1040875

OPEN ACCESS

Remote Sensing

ISSN 2072-4292

www.mdpi.com/journal/Remote Sensing

Article

Supervised Classification of Agricultural Land Cover Using a Modified k-NN Technique (MNN) and Landsat Remote Sensing Imagery

Luis Samaniego [1] **and Karsten Schulz** [2,⋆]

[1] Department of Computational Hydrosystems, UFZ–Helmholtz-Centre for Environmental Research, Permoserstr. 15, 04318 Leipzig, Germany; E-Mail: luis.samaniego@ufz.de

[2] Department of Geography, Ludwig-Maximilians-Universität München, Luisenstr. 37, 80333 Munich, Germany

⋆ Author to whom correspondence should be addressed; E-Mail: k.schulz@lmu.de; Tel.: +49-89-2180-6681; Fax: +49-89-2180-6675.

Received: 9 September 2009; in revised form: 29 October 2009 / Accepted: 30 October 2009 / Published: 9 November 2009

Abstract: Nearest neighbor techniques are commonly used in remote sensing, pattern recognition and statistics to classify objects into a predefined number of categories based on a given set of predictors. These techniques are especially useful for highly nonlinear relationship between the variables. In most studies the distance measure is adopted a priori. In contrast we propose a general procedure to find an adaptive metric that combines a local variance reducing technique and a linear embedding of the observation space into an appropriate Euclidean space. To illustrate the application of this technique, two agricultural land cover classifications using mono-temporal and multi-temporal Landsat scenes are presented. The results of the study, compared with standard approaches used in remote sensing such as maximum likelihood (ML) or k-Nearest Neighbor (k-NN) indicate substantial improvement with regard to the overall accuracy and the cardinality of the calibration data set. Also, using MNN in a soft/fuzzy classification framework demonstrated to be a very useful tool in order to derive critical areas that need some further attention and investment concerning additional calibration data.

Keywords: land use classification; supervised classification; nearest neighbors; agricultural land cover; crops

---

1. Introduction

Remote sensing has become an important tool in wide areas of environmental research and planning. In particular the classification of spectral images has been a successful application that is used for deriving land cover maps [1, 2], assessing deforestation and burned forest areas [3], real time fire detection [4], estimating crop acreage and production [5] or the monitoring of environmental pollution [6]. Image classification is also commonly applied in other contexts such as optical pattern and object recognition in medical or other industrial processes [7]. As a result, a number of algorithms for supervised classification have been developed over the past to cope with both the increasing demand for these products and the specific characteristics of a variety of scientific problems. Examples include the maximum likelihood method [8], fuzzy-rule based techniques [9, 10], Bayesian and artificial neural networks [11, 12], support vector machines [13, 14] and the k-nearest neighbor (k-NN) technique [15, 16].

In general, a supervised classification algorithm can be subdivided into two phases: (i) the learning or "calibration" phase in which the algorithm identifies a classification scheme based on spectral signatures of different bands obtained from "training" sites having known class labels (e.g., land cover or crop types), and (ii) the prediction phase, in which the classification scheme is applied to other locations with unknown class membership. The main distinction between the algorithms mentioned above is the procedure to find relationships, e.g., "rules", "networks", or "likelihood" measures, between the *input* (here spectral reflectance at different bands, also called "predictor space") and the *output* (here the land use class label) so that either an appropriate discriminant function is maximized or a cost function accounting for misclassified observations is minimized [17–21]. In other words, they follow the traditional modeling paradigm that attempts to find an "optimal" parameter set which closes the distance between the observed attributes and the classification response [9, 22].

A somewhat different approach has recently been proposed by [23] as the modified nearest-neighbor (MNN) technique within the context of land cover classification in a meso-scale hydrological modeling project. Their MNN algorithm is a hybrid algorithm that combines algorithmic features of a dimensionality reduction algorithm and the advantages of the standard k-NN being: (i) an extremely flexible and parsimonious—perhaps the simplest [24]—supervised classification technique which requires only one parameter (k the number of nearest neighbor), (ii) extremely attractive for operational purposes because it does neither require preprocessing of the data nor assumptions with respect to the distribution of the training data [25], and (iii) a quite robust technique because the single 1-NN rule guarantees an asymptotic error rate at the most twice that of the Bayes probability of error [24].

However, k-NN also has several shortcomings that have already been addressed in the literature. For instance, (i) its performance highly depends on the selection of k; (ii) pooling nearest neighbors from training data that contain overlapping classes is considered unsuitable; (iii) the so-called *curse of dimensionality* can severely hurt its performance in finite samples [25, 26]; and finally (iv) the selection of the distance metric is crucial to determine the outcome of the nearest neighbor classification [26].

In order to overcome at least parts of these shortcomings, the MNN technique abandons the traditional modeling approach in finding an "optimal" parameter set that minimizes the distance between the observed attributes and the classification response. The basic criteria within MNN is to find an

embedding space, via (non-)linear transformation of the original feature space, so that the cumulative variance of a given class label for a given portion of the closest pairs of observations should be minimum. In simple words, a feature space transformation is chosen in a way that observations sharing a given class label (crop type) are very likely to be grouped together while the rest tend to be farther away.

In [23] the MNN method has been introduced within a meso-scale hydrologically motivated application where a Landsat scene has been used to derive a 3-class (forrest, pervious, impervious) land cover map of the Upper Neckar–Catchment, Germany. MNN has also been extensively compared to other classification methods and showed superior performance. In the following sections we extend the analysis of the MNN method and demonstrate its performance to a smaller scale agricultural land use classification problem of a German test-site using mono-temporal and multi–temporal Landsat scenes. First, Section 2 briefly reviews the main features of the MNN method. In Section 3, the application including test site and data acquisition are described, followed by an extensive illustration of the MNN performance in Section 4. Results and future research directions are finally discussed in Section 5.

2. The MNN Method

2.1. Motivation

A very important characteristic of the proposed MNN technique, compared to conventional NN approaches, is the way distances between observations are calculated. NN methods require the *a priori* definition of a distance in the feature- or $\mathbf{x}$-space. Because the selection of the distance affects the performance of the classification algorithm [25, 27], a variety of reasonable distances is usually tested, for instance: the L_1 Norm (or Manhattan distance), the L_2 Norm (or Euclidian distance), L_∞ norm (or Chebychev distance), the Mahalanobis distance [28], and similarity measures such as the Pearson-, and the Spearman-Rank correlation.

In MNN, on the contrary, we search for a metric in a lower dimensional space or *embedding* space which is suitable for separating a given class. In essence, MNN finds a nonlinear transformation (or *mapping*) that minimizes the variability of a given *class membership* of all those pairs of points whose cumulative distance is less than a predefined value (D). It is worth noting that this condition does not imply finding clusters in the predictor space that minimize the total intra-cluster variance, which is the usual premise in cluster analysis (e.g. in k-means algorithm) [8]. A lower dimensional *embedding* space is preferred in MNN, if possible, because of the empirical evidence supporting the fact that the intrinsic manifold dimension of the multivariate data is substantially smaller than that of the inputs [25]. A straightforward way to verify this hypothesis is calculating the dimensionality of the covariance matrix of the standardized predictors and to count the number of dominant eigenvectors [29].

An important issue to be considered during the derivation of an appropriate transformation or embedding space is the effect of the intrinsic nonlinearities present in multivariate data sets (e.g., multi–temporal and/or multi–/hyperspectral imagery). It is reported in the literature that various sources of nonlinearities affect severely the land cover classification products [30]. This in turn would imply that dissimilar class labels might depend differently upon the input vectors. Consequently, it is better to find *class specific* embeddings and their respective metrics rather than a global embedding.

2.2. Problem Formulation

Supervised classification algorithms aim at predicting the class label $h \in \{1,\ldots,l\}$—from among l predetermined classes —that correspond to a query $\mathbf{x}_0$ composed of m characteristics, i.e., $\mathbf{x}_0 = \{x_1,\ldots,x_m\}_0 \in \mathbb{R}^m$. To perform this task, a classification scheme c "ascertained" from the training set $\mathcal{T} = \{(\mathbf{x}_i, \mathbf{e}_i) : \quad i = 1,\ldots n\}$ is required. Here, to avoid any influence from the ordering, the membership of the observation i to the class h is denoted by a unit vector $\mathbf{e}_i \in \{\mathbf{e}^h : \quad h = 1,\ldots,l\}$ [e.g., class 3 is coded in a three-class case as $\mathbf{e}^3 = \{0,0,1\}$]. Moreover, each class should have at least one observation, i.e. $n_h > 0$, and $\sum_{h=1}^{l} n_h = n$, where n is the sample size of the training set.

In general, the proposed MNN algorithm estimates the classification response $\mathbf{y}_0 = c(\mathbf{x}_0) = \{y_1,\ldots,y_l\}_0$ as a combination of selected membership vectors $\mathbf{e}_i = \{e_{i1},\ldots,e_{il}\}$ whose respective "transformed" inputs $B_h[\mathbf{x}_i]$ are the k-nearest neighbors of the transformed value of the query $B_h[\mathbf{x}_0]$.

Since there are l possible embedding spaces in this case, then l possible neighborhoods of this query have to be taken into account in the estimation. Here B_h denotes a class specific transformation function. A stepwise meta–algorithm for the estimation of the classification response can be formulated as follows:

1. An attribute (crop/landuse class) h is selected.
2. An optimal distance d_h for the binary classification problem is identified.
3. The binary classification is performed using k-NN with the distance d_h leading for each vector $\mathbf{x}_0$ a partial membership vector $\{y_{h'}\}_{0h}$, $h' = 1,\ldots,l$.
4. Steps 1) to 3) are repeated for each attribute h.
5. The partial memberships are aggregated into the final classification response $\mathbf{y}_0$.
6. The class exhibiting the highest class membership value is assigned to $\mathbf{x}_0$.

In total, h optimal metrics have to be identified observing some conditions. A distance measure d_h is considered suitable for the binary classification of a given attribute h if the following discriminating conditions are met:

(a) The distances in the transformed space between pairs of points in which both correspond to the selected attribute are small, and

(b) The distances in the transformed space between pairs of points in which only one of them correspond to the selected attribute are large.

These conditions ensure that the k-NN classifier works well at least for a given attribute in its corresponding embedded space. The distance between two points in MNN is defined as the Euclidean distance between their transformed coordinates

$$d_{B_h}(\mathbf{u}_i, \mathbf{u}_j) = \|B_h[\mathbf{x}_i] - B_h[\mathbf{x}_j]\| \tag{1}$$

where B_h is a transformation (possibly nonlinear) of the m-dimensional space $\mathbf{x}$ into another κ - dimensional space $\mathbf{u}_h$ usually with ($\kappa \leq m$):

$$\mathbf{u}_h = B_h[\mathbf{x}] \tag{2}$$

As a result, MNN requires finding a set of transformations B_h so that the corresponding distance d_h performs well according to the conditions (a) and (b) mentioned above. In the present study, this was attained by a *local variance* function $G_{B_h}(p)$ [31] that accounts for the increase in the variability of the class membership e_h with respect to the increase in the distance of the nearest neighbors in a nonparametric form.

Formally, $G_{B_h}(p)$ can be defined for $0 < p \leq 1$ as [22]

$$G_{B_h}(p) = \frac{1}{\mathcal{N}(p)} \sum_{\substack{d_{B_h}(i,j) < D_{B_h}(p) \\ (i,j) \in \mathcal{C}_h}} (e_{ih} - e_{jh})^2 \tag{3}$$

where p is the proportion defined as the ratio of $\mathcal{N}(p)$ to $\mathcal{N}_h$, $d_{B_h}(i,j)$ is the distance between points i and j in the transformed space, $D_{B_h}(p)$ is the p percentile of the d_{B_h} distribution, and $\mathcal{C}_h$ is the set of all possible pairs having at least one element in class h, formally

$$\mathcal{C}_h \equiv \{(i,j) : \ (\mathbf{e}_i = \mathbf{e}^h) \vee (\mathbf{e}_j = \mathbf{e}^h) \ \wedge \ j > i\} \ \forall h \in \{1, \ldots, l\} \tag{4}$$

here $\mathcal{N}_h$ denotes the cardinality of the set $\mathcal{C}_h$ given by

$$\mathcal{N}_h = |\mathcal{C}_h| = \frac{n_h}{2}(2n - n_h + 1) \tag{5}$$

and $\mathcal{N}(p)$ refers to the number of pairs $(i,j) \in \mathcal{C}_h$ that have a $d_{B_h}(i,j)$ smaller than the limiting distance $D_B(p)$, formally defined as

$$\mathcal{N}(p) = |\{(i,j) \ ; \ d_{B_h}(i,j) < D_B(p) \ \wedge (i,j) \in \mathcal{C}_h \ \forall h\}| \tag{6}$$

where $|\{.\}|$ denotes the cardinality of a given set.

The transformation B_h can be identified as the one which minimizes the integral of the curve G_{B_h} vs. p up to a given threshold proportion p^*, subject to the condition that each G_{B_h} is a monotonic increasing curve (see later Figure 4 in Section 4.) for an illustration of this minimization problem). In simple words this procedure aims at close neighbors having the same class membership, whereas those further away should belong to different classes.

Formally, this can be written as follows: Find $B_h, \quad h = 1, \ldots, l \quad$ so that

$$\min \int_0^{p^*} G_{B_h}(p) dp \tag{7}$$

subject to

$$G_{B_h}(p) \leq G_{B_h}(p + \delta) \quad \forall \ p, \ \delta > 0 \tag{8}$$

where δ is a given interval along the axis of p. It is worth noting that the objective function denoted by (7) will be barely affected by outliers (i.e. those points whose close neighborhood is relatively quite

far), which appear often in unevenly distributed data. For computational reasons, equ. (7) can be further simplified to

$$\sum_{q=1}^{Q} G_{B_h}(p_q) \rightarrow \min \tag{9}$$

where $\{p_q : \ q = 1, \ldots, Q\}$ is a set of Q portions such as $p_{q-1} < p_q < 1, \ \forall q$, and $p_Q = p^*$. The total number of pairs within the class h can be defined as $\mathcal{N}_h^* = |\{(i,j) \ (\mathbf{e}_i = \mathbf{e}^h) \wedge (\mathbf{e}_j = \mathbf{e}^h) \ \wedge \ j > i\}| \ \forall h$, thus this threshold for the class h can be estimated by

$$p^h = \frac{\mathcal{N}_h^*}{\mathcal{N}_h} = \frac{n_h - 1}{2n - n_h + 1} \tag{10}$$

In summary, this method seeks different metrics, one for each class h, on their respective embedding spaces defined by the transformations B_h so that the cumulative variance of class attributes for a given portion of the closest pairs of observations is minimized. Consequently these metrics are *not global*. The approach to find class specific metrics in different embedding spaces is what clearly differentiates MNN from other algorithms such as the standard k-NN or AQK. In contrast to MNN, k-NN employs a global metric defined in the input space, whereas AQK uses a kernel function to find ellipsoidal neighborhoods [25]. The kernel, in turn, is applied to find implicitly the distance between points in a feature space whose dimension is larger than the input space $\mathbf{x}$ [26].

2.3. Deriving an Optimal Transformed/Embedding Space

There are infinitely many possibilities of selecting a transformation B_h. Among them, the most parsimonious ones are the linear transformations [22]. Here, for the sake of simplicity, l matrices of size $[\kappa \times m]$ are used:

$$\mathbf{u}_h = \mathbf{B}_h \mathbf{x} \quad h = 1, \ldots, l \tag{11}$$

It is worth mentioning that the standard k-NN method is identical to the proposed MNN if $\mathbf{B}_h = \mathrm{diag}(1, \ldots, 1)$. Appropriate transformations $\mathbf{B}_h$ have to be found using an optimization method. Du to the high dimensionality of the resulting optimization problem (see (9)), the following global optimization algorithm based on simulated annealing [32] is proposed:

1. Select a set of threshold proportions $p_1 < p_2 < \ldots < p_Q < 1$. The value of p_Q can be chosen to be equal $\frac{k}{n}$ with k being the number of the nearest-neighbors to be used for the estimation. For the other proportions, $p_q = \frac{q}{Q} p_Q$ is a reasonable choice.
2. Select $h = 1, \ldots, l$.
3. Select the dimension of the $\mathbf{u}_h$ space κ.
4. Fix an initial annealing temperature t.
5. Randomly fill the matrix $\mathbf{B}_h$.

6. Check the monotonic condition using: $W_h = \prod_{q=2}^{Q} \max(1, \frac{G_{B_h}(p_{q-1})}{G_{B_h}(p_q)})$.
7. Calculate the objective function $O = W_h \sum_{q=1}^{Q} G_{B_h}(p_q)$.
8. Randomly select a pair of indices (i_1, i_2) with $1 \leq i_1 \leq \kappa$ and $1 \leq i_2 \leq m$.
9. Randomly modify the element b_{i_1,i_2} of the matrix $\mathbf{B}_h$ and formulate a new matrix $\mathbf{B}_h^*$.
10. Calculate the penalty for non-monotonic behavior: $W_h^* = \prod_{q=2}^{Q} \max\left(1, \frac{G_{B_h^*}(p_{q-1})}{G_{B_h^*}(p_q)}\right)$
11. Calculate $O_h^* = W_h^* \sum_{q=1}^{J} G_{B_h^*}(p_q)$.
12. If $O^* \leq O$ then replace $\mathbf{B}_h$ by $\mathbf{B}_h^*$. Else calculate $R = \exp(\frac{O-O^*}{t})$. With the probability R, replace $\mathbf{B}_h$ by $\mathbf{B}_h^*$.
13. Repeat steps (8)-(12) M times (with M being the length of the Markov chain [32]).
14. Reduce the annealing temperature t and repeat steps (8)-(13) until the objective function O achieves a minimum.

It should be stated here, that in case a nonlinear function B_h would be required, the proposed algorithm is not affected, in general, and the generation of new solutions can be done as proposed in steps 8) to 9) after some modification. Also, any other global optimization method, such as Genetic Algorithms (GA) [33] or the Dynamically Dimensioned Search (DDS) algorithm [34] (among many others) could have been used here in principal, but good experience with the simulated annealing method in earlier applications has led to that choice.

2.4. Selection of an Estimator

The prediction of the classification vector $\mathbf{y}_0$ based on a given vector $\mathbf{x}_0$ is done with an ensemble of predictions (one with each transformation $\mathbf{B}_h$) to ensure interdependency among the l classes. This characteristic of the system was already assumed during the selection of the matrices $\mathbf{B}_h$ by means of their simultaneous optimization achieved in Equation (7). Consequently, the local estimator should also consider this fact.

There are many types of local estimators that can be employed, for instance: the nearest neighbor, mean of close neighbors, local linear regression, local kriging, among others, as presented in [22]. For the classification problem stated in section 1., however, the mean of the k closest neighbors is the most appropriate one because it considers several observations. Thus, the expected value of the observation 0 can be estimated as:

$$\begin{aligned} y_{0h'} &= \frac{1}{l\,k} \sum_{h=1}^{l} \sum_{j} e_{jh'} \\ j &\in \{j' : \quad d_{B_h}(0, j') < D_h(k) \quad j' = 1, \ldots, n\} \\ h' &= 1, \ldots, l \end{aligned} \tag{12}$$

where $D_h(k)$ is the distance of the k-nearest neighbor of observation $\mathbf{x}_0$ in the transformed space $\mathbf{u}_h$.

The advantage of this approach is that one can always find an estimate for a given observation. Its main disadvantage, however, occurs for those points that do not have close enough neighbors (i.e., false neighbors), which, in turn, might lead to an erroneous estimation.

As a result of Equation (12), the classification vector $\mathbf{y}_0$ has values in the interval $0 \leq y_{h0} \leq 1$ $h = 1, \ldots, l$ hence $\sum_{h=1}^{l} y_{h0} = 1$, i.e., it is a soft classification. Since these values have been derived from a sample, they could be understood as the empirical probability that the observation 0 belong to the class h. If a hard classification is required, then the class having the highest y value can be assigned to the given observation $\mathbf{x}_0$. This procedure to obtain a binary classification also allows the estimation of the measure of ambiguity $b(\mathbf{y}_0)$ as

$$b(\mathbf{y}_0) = 1 - \max_h y_{h0} \tag{13}$$

which indicates whether or not the classifier c has yielded a clear response [9].

2.5. Validation

As the transformation $\mathbf{B}_h$ is derived from observations it is necessary to validate it. Possible methods for validation are cross-validation and split sample testing. If the latter method is used, then $\mathbf{x}_0 \in \mathcal{V}$ and $\mathcal{T} \cap \mathcal{V} \equiv \emptyset$. $\mathcal{V}$ is commonly known as the validation set and should be similar in composition to the calibration set. Its cardinality, however, may vary. Here, the sample size of $\mathcal{V}$ is denoted by $v = |\mathcal{V}|$.

3. Applications

3.1. Study Area and Data Availability

The proposed method was applied to two different land cover classification examples using i) single, and ii) multi-temporal Landsat TM/ETM+ images for the year 1999. Both examples use data from agricultural test sites located in the Parthe-catchment, southeast of Leipzig, Germany (see Figure 1). The study area is part of the German ”loess band” consisting of highly productive soils with land use dominated by agricultural activities. The elevation above sea level ranges from 150 to 190 m with only very mild slopes. The yearly mean air temperature is 9.8 °C, the mean annual precipitation is 610 mm.

The test sites comprise 23 fields with a total area of 426 ha. Main soil types are Cambisol, Luvisol, and Stagnic Gleysol; and dominating crops are summer and winter cereals with a total fraction of over 70%. Table 1 summarizes the considered crops and land use classes as well as their spatial fractions. These data, as well as detailed tillage and management information for the year 1999, were collected and provided by the farm cooperative *Landwirtschafts GmbH KÖG Kleinbardau*.

In addition, three Landsat images of the study region have been made available for the growing season of 1999: two Landsat 5-TM scenes from 30. April and 3. July (path/row 193/025) and a Landsat 7-ETM+ scene from 13. September 1999 (path/row 193/024). All three images were geo-referenced within the ERDAS imagine software using a digital topographic map at the scale 1:25,000 (*Landesvermessungsamt Sachsen*) and 40 significant reference points resulting in a maximum RMSE-error of 7.8 m. For both examples we only used the 6 bands in the visible and near infrared region, thus ignoring the thermal

bands of TM and ETM+ as well as the panchromatic channel of ETM+. Figure 2 summarizes the vegetation periods of considered crops, the time schedule of soil and crop management activities and the dates of available remote sensing images. An atmospheric correction has not been considered here.

Figure 1. Distribution and location of the test sites.

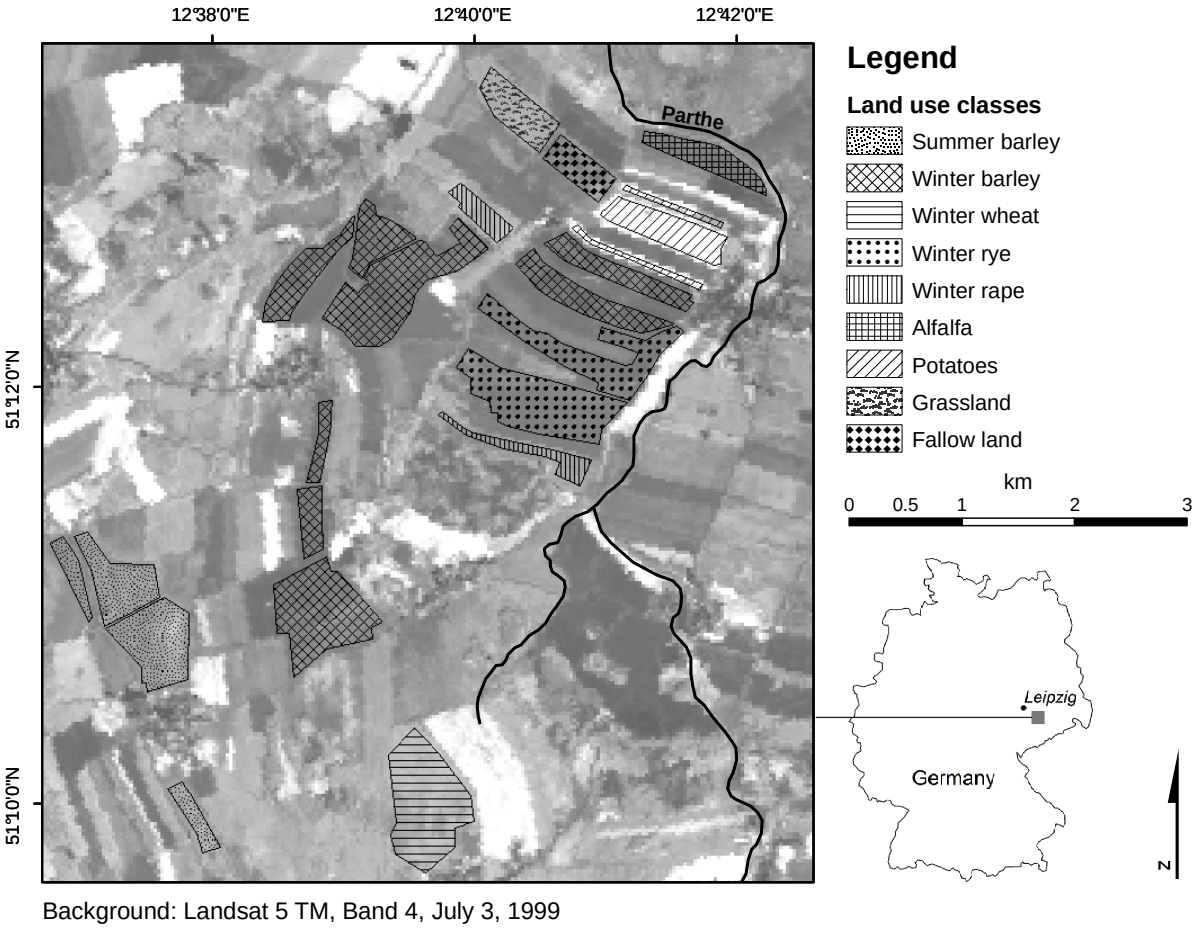
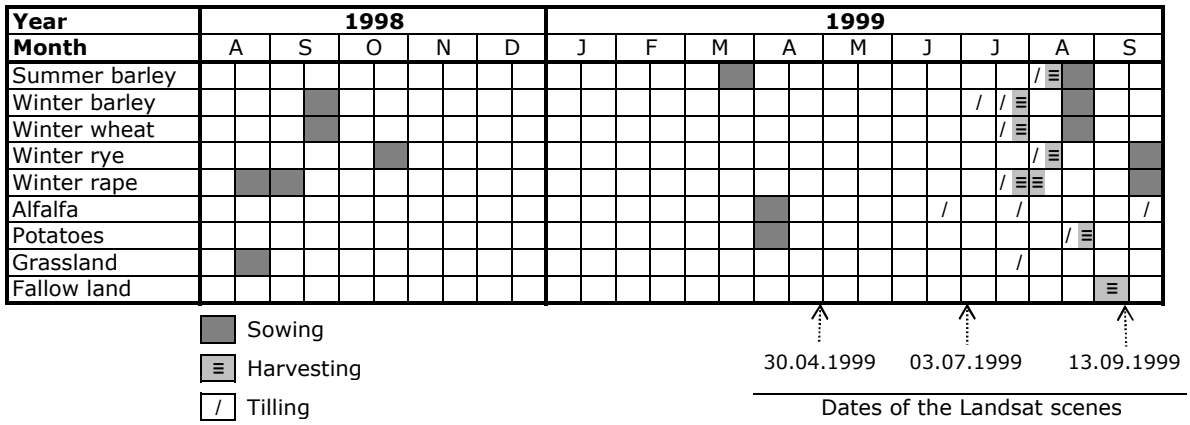


Table 1. Summary of the land use classes and its distribution within the test area.

No.	Land use type	Number of plots	Area [ha]	% Total
1	Summer barley	4	81.5	14.2
2	Winter barley	8	219.9	38.1
3	Winter wheat	2	97.9	17.0
4	Winter rye	1	60.4	10.5
5	Winter rape	2	22.4	3.9
6	Alfalfa	1	19.8	3.4
7	Potatoes	3	36.6	6.3
8	Grassland	1	22.1	3.8
9	Fallow land	1	15.9	2.8
Total		23	576.5	100.0

Figure 2. Chronogram of the vegetation periods and management activities. The dates of the landsat scenes are also depicted.



3.2. Variable Definition

For the calibration and validation phases of this study, several disjoint sets were sampled without replacement from the images mentioned above. The sample size of the validation set was fixed ad hoc to 270 observations distributed equally among all nine classes (i.e., $l = 9$). Additionally, six calibration sets were also sampled to perform a sensitivity analysis of the impact of the sample size on the efficiency of the classifier, thus, the number of observations per class were fixed as $n_h \in \{5, 10, 15, 20, 25, 30\}$. For all these calibration sets, the variables listed in Table 2 were extracted from the original images and then tabulated as indicated in Section 1.

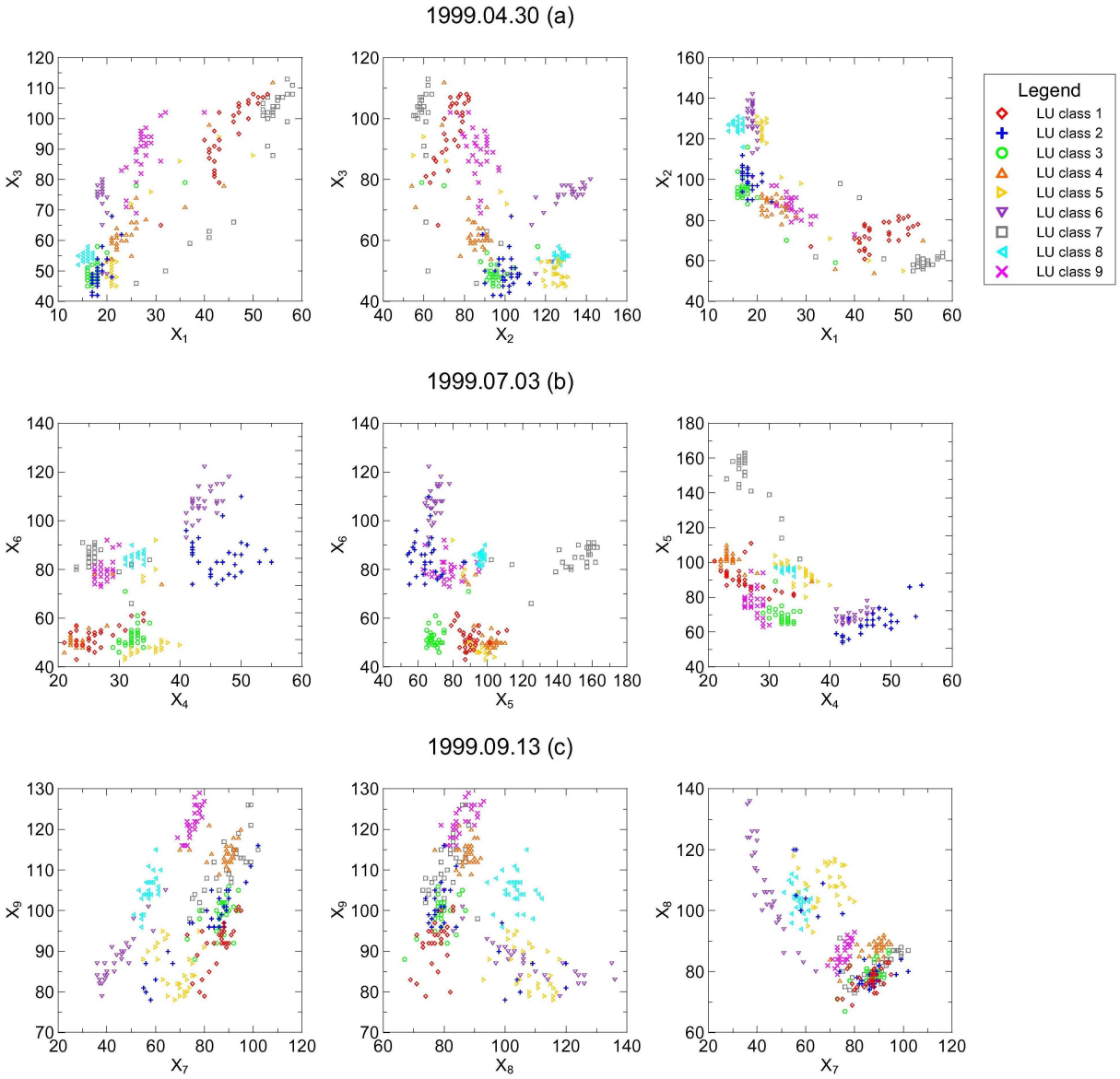
Table 2. Definition and notation of the selected predictors at different points in time.

	Mono-temporal		Multi-temporal	
Variable Name	Landsat Band	Date	Landsat Band	Date (dd.mm.yyyy)
x_1	1		3	
x_2	2		4	30.04.1999
x_3	3		5	
x_4	4	30.04.1999	3	
x_5	5		4	03.07.1999
x_6	7		5	
x_7			3	
x_8			4	13.09.1999
x_9			5	

It is convenient to visualize the relationships that might exist between predictors (see Figure 3) before the calibration starts, as the main task of a supervised classifier is finding patterns and rules within the predictor's space aiming at forming disjoint classes. In fact, as it is shown at the panel (c) of this figure, these relationships can be so intertwined that most classifiers produce a high number of false positives. In this example, for instance, classes 1, 2, 3, 4, 7, and 9 are heavily clustered as can be seen in each

pairwise scatterplot of variables x_7, x_8, and x_9, respectively. This highlights the main motivation behind this paper stated in Section 2. to find an embedding space $\mathbf{u}$ where close points in the $\mathbf{x}$ space correspond to a similar class.

Figure 3. Scatterplots depicting the location of the land use classes in the predictor space $\mathbf{x}$ at three different points in time as indicated at each panel. In this case, the reflectance measurements represent bands 3-4-5 of the training set whose sample size is $n = 270$.

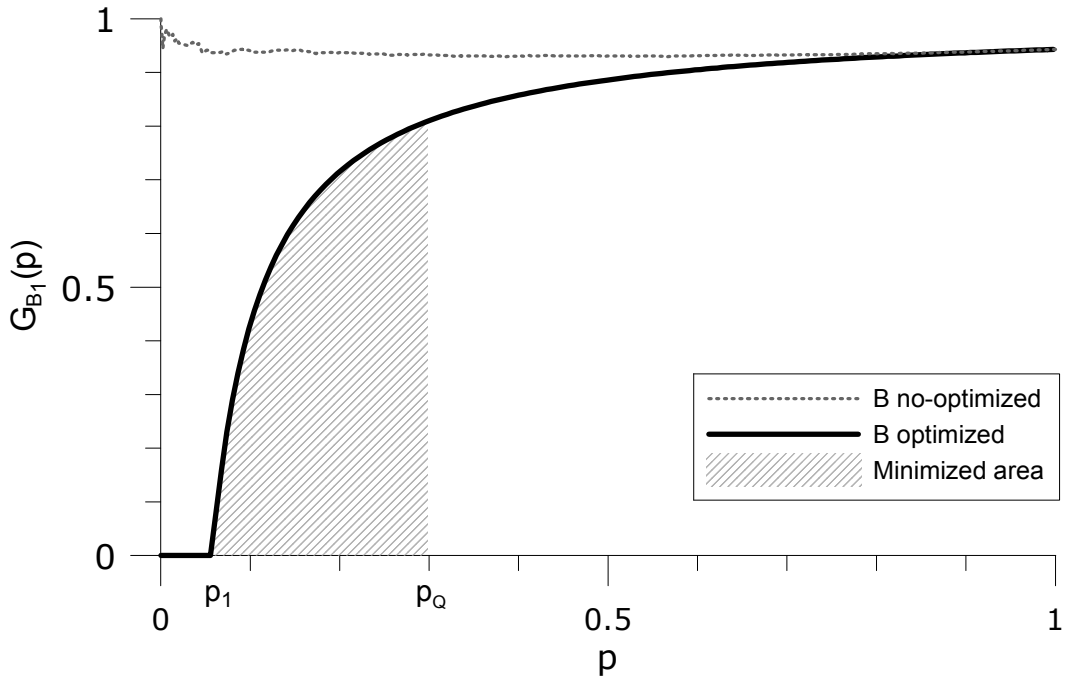


4. Results

To find the embedding spaces $\mathbf{u}_h$, $h = 1, \ldots, l$, the algorithm presented in subsection 2.3. was sequentially applied. In each optimization, the maximum limit of the integration was $p_Q = 0.30$ (see Equation 9). The result of the optimization can be seen in Figure 4. This graph shows that the variance function G exhibits a completely different behavior depending on whether the distance is calculated in the original predictors' space $\mathbf{x}$ or in the embedding space $\mathbf{u}$. In the former, the variance of the (p) closest neighbors remains almost constant for different values of p, whereas in the latter an

abrupt change occurs at $p_1 \approx 0.057$. This threshold represents the ratio between the number of pairs within class 1 and the total number of pairs of the boundary of the set $\mathcal{C}_1$ and has already been defined in Equation 10. Therefore, as a result of the optimization, the variance of the "outputs" $\mathbf{e}_i$ has been reduced to zero, i.e., approximately the 6% of the closest neighbors belong to class 1 in the transformed space $\mathbf{u_1}$. This can also be visualized in Figure 5, where it can be seen that class 1 (encircled by a continuous line) was clearly pulled apart from the rest.

Figure 4. Effect of the optimization on the variance function G_{B_1}. Here $h = 1$, i.e., the variance function of class 1 for the training set $\mathcal{T}$.



The robustness and the reliability of the MNN method was estimated with the standard confusion matrix approach [8]. Based on this sort of contingency table, several efficiency measures were evaluated, including: the probability of detection or producer accuracy, the user accuracy which is the complement of the false positive rate, and the overall accuracy OA $\in [0, 1]$ where $0 \equiv$ worst, and $1 \equiv$ best. Figure 6 depicts the robustness of the proposed approach compared with both ML and the k-NN using the standard Euclidian distance in the predictors' space ($\mathbf{B}$ not optimized). The 90% confidence interval of the statistic OA is almost independent with respect to the number of nearest neighbors N, which is not the case in the non-optimized case. For the estimation of empirical confidence intervals shown in Figure 6, 100 independent training sets ($n = 135$, $n_h = 15$ $\forall h$) were randomly sampled without replacement. The simulation results show that the average OA obtained for MNN is even greater than the upper 95%-confidence limit of the results obtained with ML. The standard k-NN (without any transformation) performs poorly in all these independent samples.

Several sensitivity analysis were carried out to determine the behavior of MNN with respect to variable composition as well to several key parameters. First, we investigated how the selection of the variables influence the efficiency of the different methods. Based on Figure 7, it can be concluded that the multi-temporal supervised classifications provide, in general, better results than those of the mono-temporal ones. In the present case, however, k-NN with data from the 30th of April, 1999, is an

exception. This indicates that k-NN might be sensitive to artifacts present (e.g., related with climate, vegetative cycle) in the data that may severely affect the efficiency of the classifier. This shortcoming has not been observed in the several hundreds of tests carried out with MNN. Another results derived from these experiments is that MNN performs much better than ML when the sample size per class (n_h) is small (see Figure 7). Furthermore, influence of n_h on the efficiency is small. This is an advantage because it reduces the cost of data collection without jeopardizing the quality of the results.

Figure 5. Scatterplots depicting the location of the land use classes in the embedded space $\mathbf{u}_1$. Class 1 (encircled by a continuous line) is completely isolated from the others. Sample size $n = 270$.

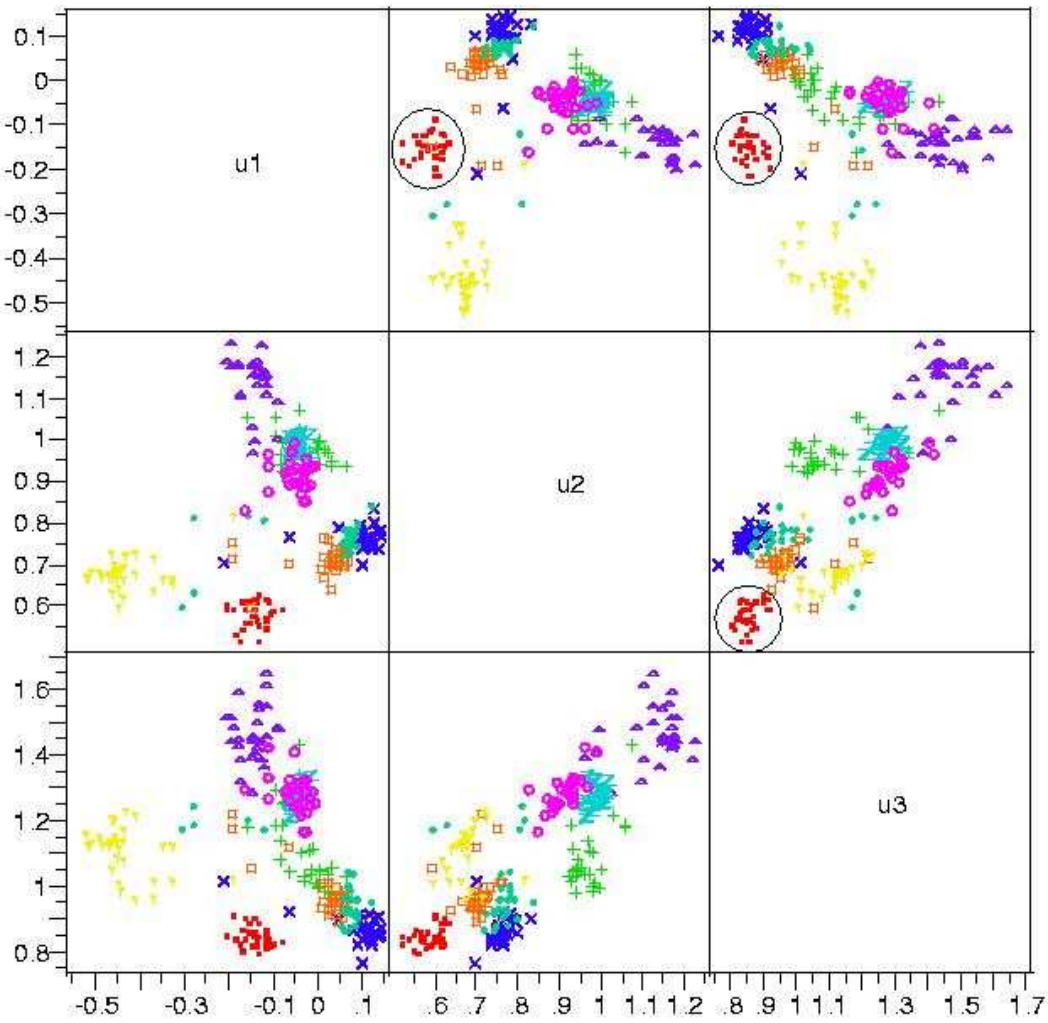
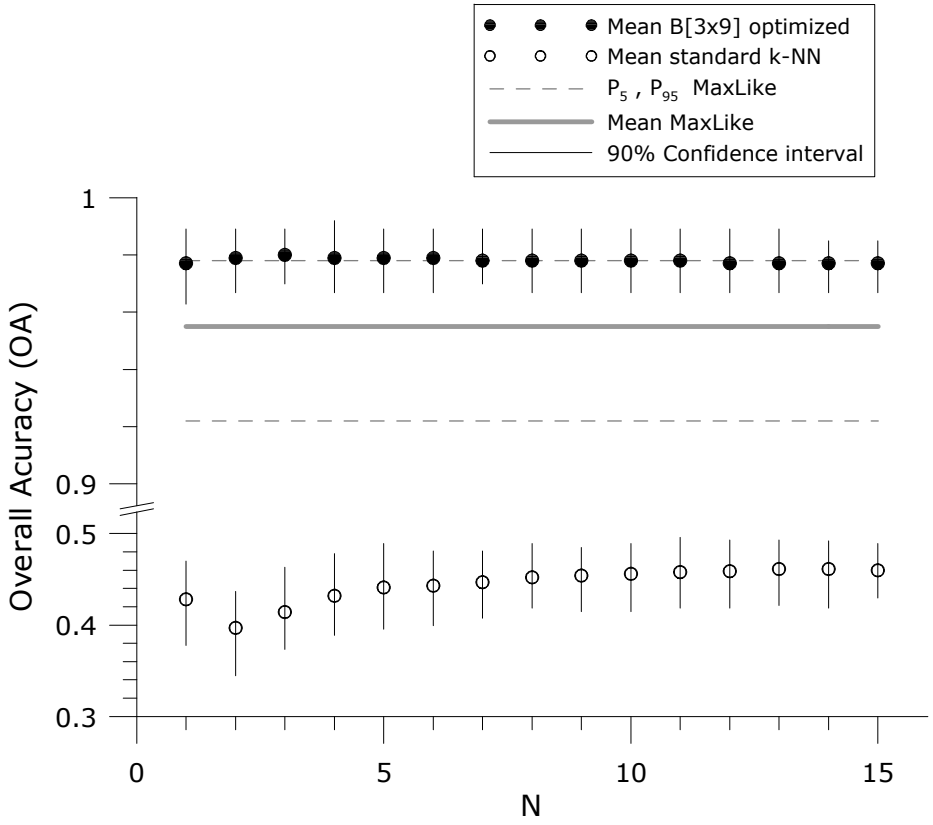


Figure 6. Graph showing the variation of the 90% confidence interval depending on the number of nearest neighbors. As reference the confidence intervals for ML and k-NN ($\mathbf{B}$ not optimized) are also shown.



The selection of the dimension of the embedding space k plays a very important role on the efficiency of MNN as can be seen in Figure 8. Because there are no rules for the determination of k the trial-and-error approach was followed. This figure shows that embedding a large dimensional space into an scalar (i.e., $k = 1$) is not a good solution. As k approaches m, the same effect is observed, i.e. large values of k do not perform much better than that of the $k-1$-case. It should be noted that increasing k by one increases the number of degrees of freedom l times. In the present case the best results were obtained for $k = 3$ in the mono-temporal experiment. A modification of the Mallows-CP [35] statistic, however, was suggested in [22] to help determine the most efficient embedding dimension, which can also be used in this context.

Based on the results of the sensitivity analysis, the final classification of the whole area of the test sites was carried out using multi-temporal data (bands 3-4-5 form the three scenes) embedded into a three dimensional space, i.e. $\mathbf{B}[3 \times 9]$ and using $N = 5$ neighbors. As a result, the land cover map depicted in Figure 9 was obtained. Additionally, the ambiguity index $b(\mathbf{y}_0)$ proposed in Equation (13) was estimated. With the help of this index, it is possible to determine those places where the classification is not clear (say $b(\mathbf{y}_0) > 0.5$). Consequently, one should interpret the results in those areas carefully and try to determine why the classifier performs ambiguously on these areas (e.g., a very mixed class or errors in the determination in the ground truths). Furthermore, by plotting the density distribution of the ambiguity index and estimating statistics like the mean ambiguity index or the portion of pixels exceeding a given

threshold, one can have a very reliable measure of the uncertainty of the classification (see Figure 10). Taking, for example, a threshold value of $b = 0.2$ we observe that a training set consisting of only $n_h = 5$ observations for each class will have a large number of false positives, whereas with $n_h = 25$ we only obtain a very small number indicating a more reliable classification. The slight increase of the classification uncertainty between $n_h = 25$ and $n_h = 30$ suggest that large training sets might be less efficient due to possible data artifacts introduced during the sampling process. We here find an optimum sample size of $n_h = 25$ observation per class.

Figure 7. Graph showing the influence of the sample size per class n_h on the overall accuracy of the classification. The dimension of the transformation matrix $\mathbf{B}$ in the mono- and multi-temporal classifications is $[3 \times 6]$ and $[3 \times 9]$ respectively.

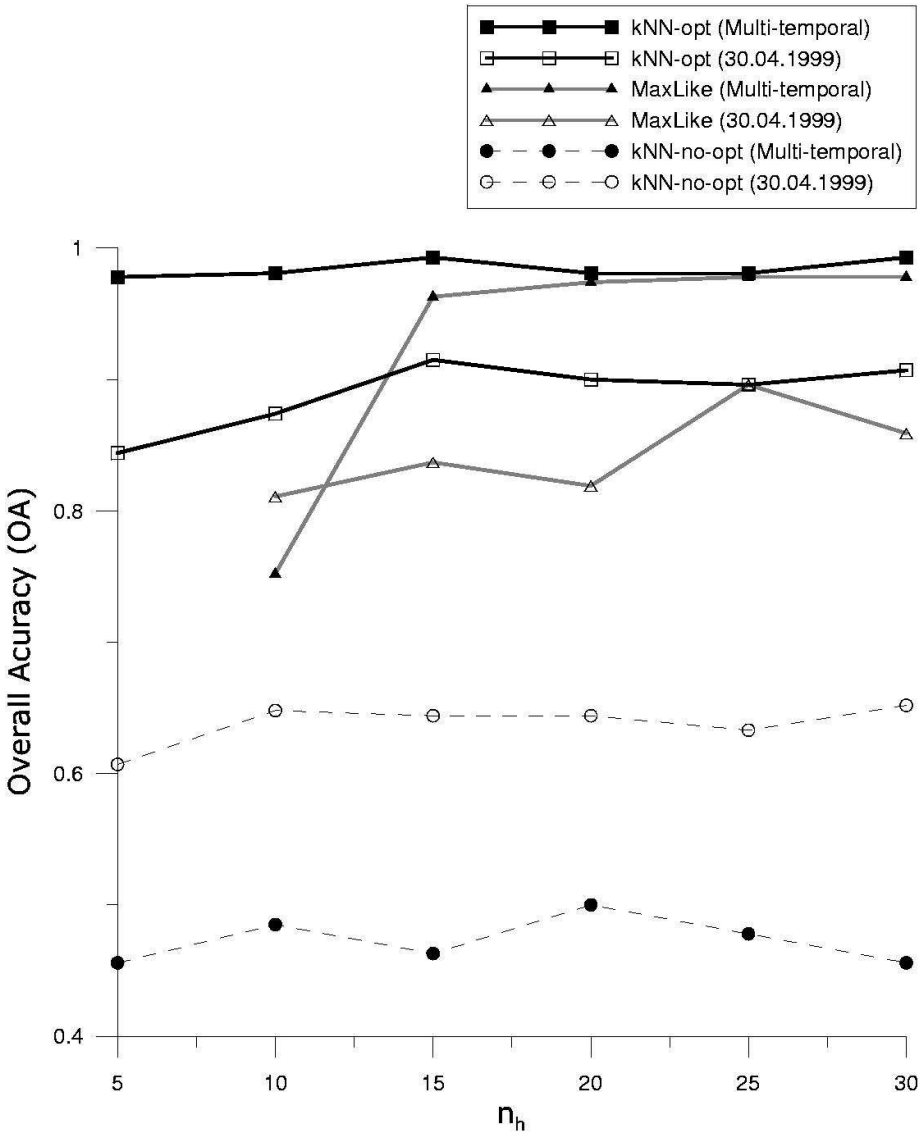


Figure 8. Graph showing the influence of the type of transformation and the number of nearest neighbors on the overall accuracy of the classification.

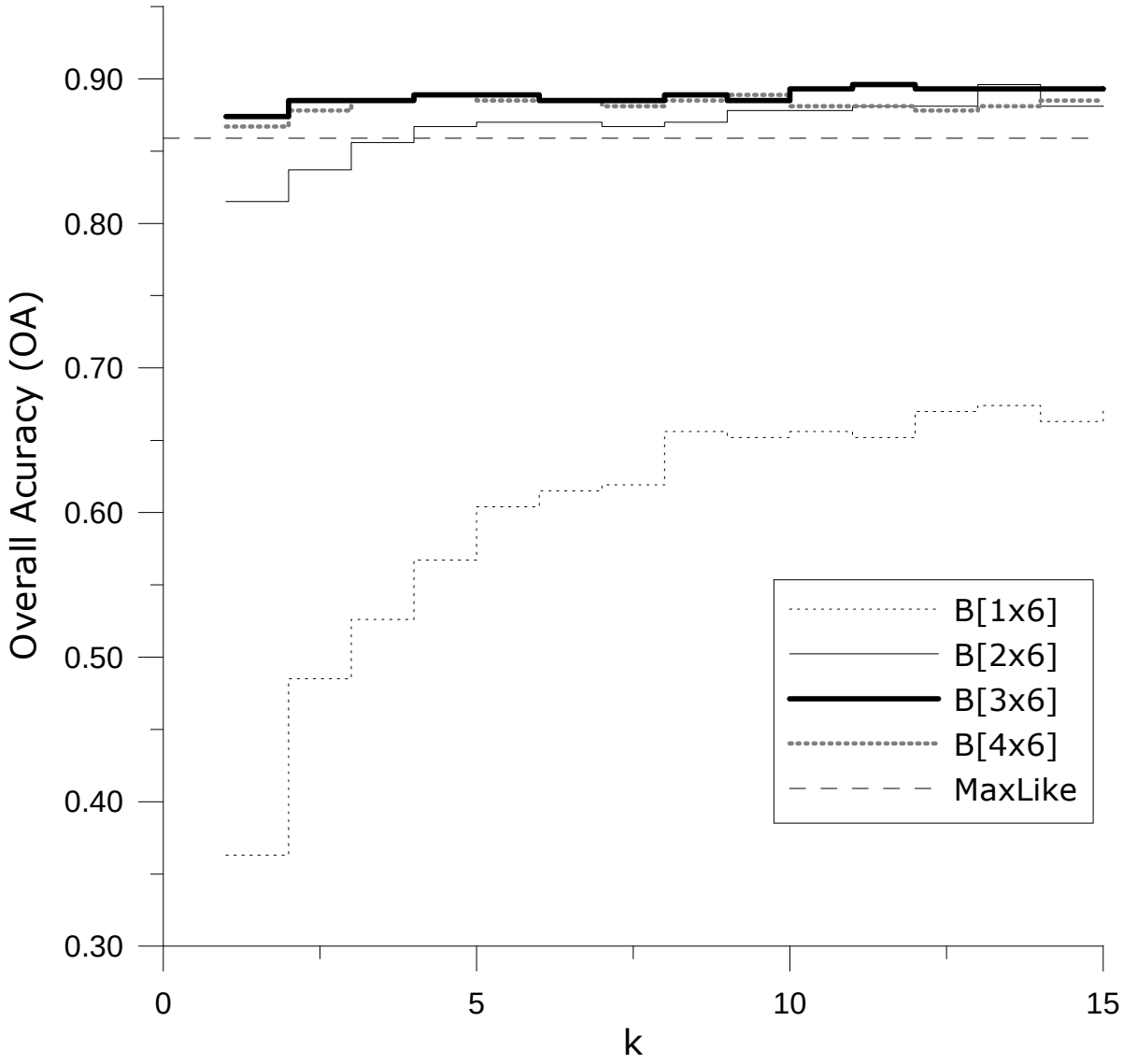


Figure 9. Ambiguity index [panel (a)] and land cover map obtained with the MNN classifier using $N = 5$ neighbors [panel (b)].

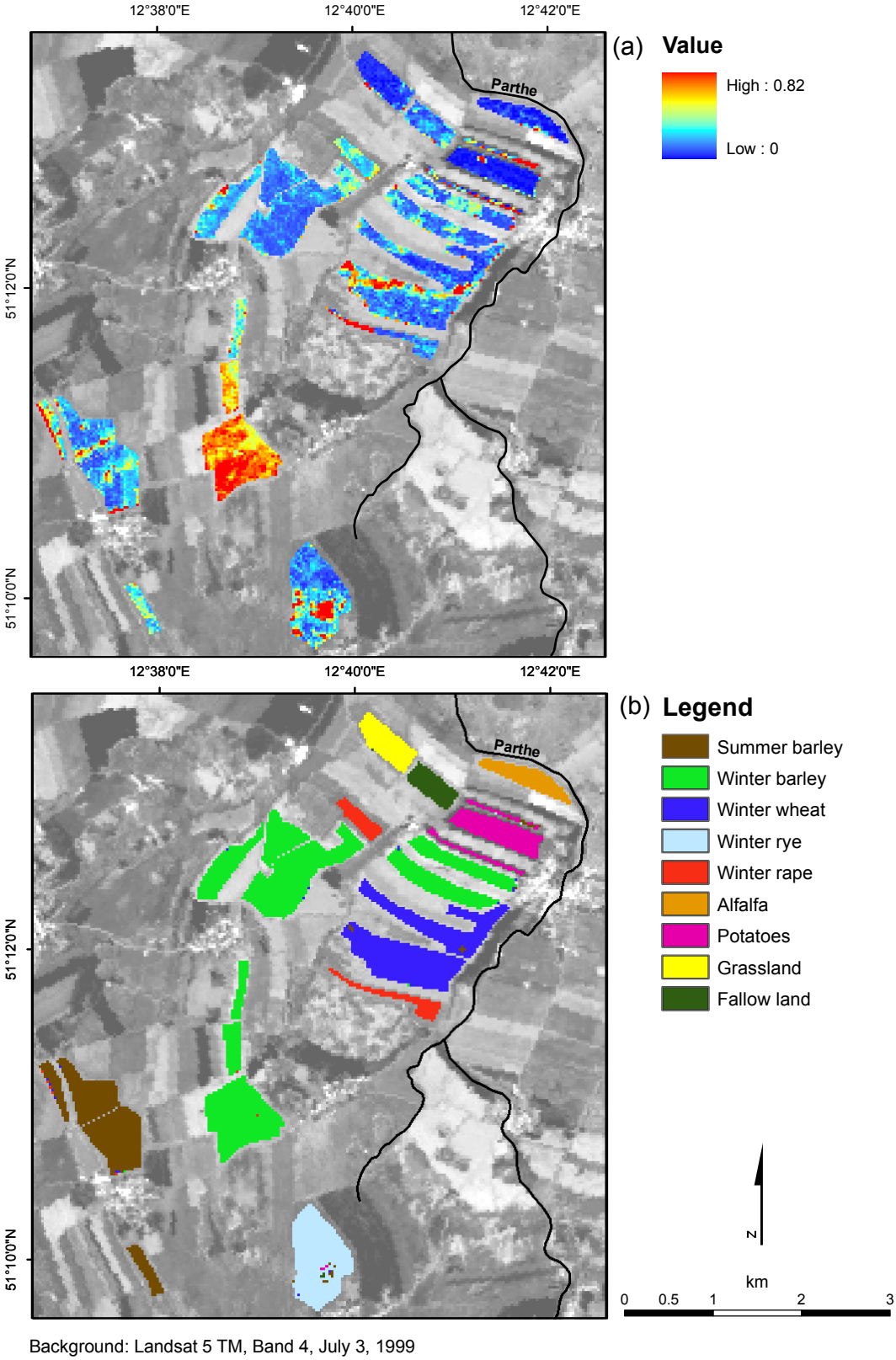
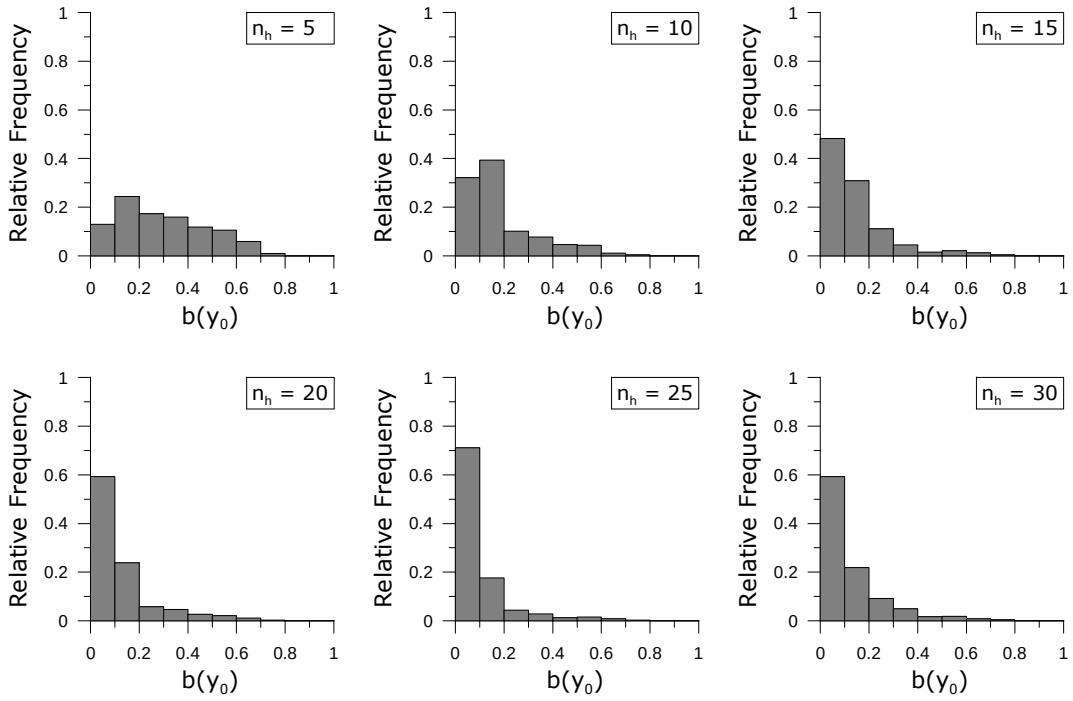


Figure 10. Histograms depicting the ambiguity index $b(\mathbf{y}_0)$ as a function of the sample size of the training set n_h. (All histograms were obtained with the MNN classifier using $N = 5$ neighbors.)



5. Summary and Conclusions

Here, we have introduced an improved k-NN technique based on a combination of a local variance reducing technique and a linear embedding of the observation space into an appropriate Euclidian space (MNN) and applied this method to a (supervised) land use and cover (LULC) classification example using mono- and multi-temporal Landsat imagery. Results show that MNN performs significantly better than both Maximum Likelihood (ML) and the standard k-NN method for all cases considered. The advantage of using multi-temporal data was also confirmed for the MNN method.

While the efficiency of the ML is dependent on a sufficient large sample size, a further strength of MNN is its small sensitivity to the cardinality of the training set leading to a more robust and cost-effective calibration of the classification problem. The formulation of an ambiguity index and its distribution provides a good measure of the uncertainty associated with a classification. Analysis of its spatial extent and characteristic will allow critical areas and LULC types within the classification process to be identified so that additional effort in ground-truth and calibration activities can be organized in a more focussed and efficient way.

Given these advantages of MNN compared with standard methods, it will be of particular interest for any of the current and future global or regional observation activities, where large areas have to be (frequently) explored under limited resources. Good examples are the Global Observation of Forest and Land Cover Dynamics (GOFC-GOLD) or the Monitoring Agriculture by Remote Sensing and Supporting Agricultural Policy (MARS-CAP) projects (among many others).

However, before becoming operational, some further aspects of the method still have to be investigated. So far, we have limited our analysis to the classification of agricultural land use types using Landsat images with a spatial resolution of 30m as presented here in this paper and to hydrologically related application considering 3 different land use classes (forest, pervious and impervious). This needs to be extended considering a broader spectrum of LULC types as well as different levels of informational aggregation. Data from different sensors and platforms need to be analyzed in order to explore the sensitivity and efficiency of our method to different spatial and spectral resolutions. However, given the results presented here and in our previous application [23], MNN is expected to also outperform other classification methods under these conditions.

Analyzing in detail the robustness and transferability of the optimized transformation of the observation space to other geographic locations will help to keep the cost for calibration to a minimum. Also, information on the scale and temporal invariance of the transformation will be of great importance. The MNN technique might be further improved by considering non-linear transformations and/or some appropriate formulation for its variations within the observation space. While being beyond the scope of this paper, these aspects will direct our future research activities and will be presented in the near future.

6. Acknowledgement

The authors would like to thank two anonymous reviewers for their helpful comments. They would also like to thank K. Hannemann for the preparation of the data.

References

1. Powell, S.L.; Cohen, W.B.; Yang, Z.; Pierce, J.D.; Alberti, M. Quantification of impervious surface in the Snohomish water resources inventory area of Western Washington from 1972-2006. *Remote Sens. Environ.* **2008**, *112*, 1895–1908.
2. Joshi, P.K.K.; Roy, P.S.; Singh, S.; Agrawal, S.; Yadav, D. Vegetation cover mapping in India using multi-temporal IRS Wide Field Sensor (WiFS) data. *Remote Sens. Environ.* **2006**, *103*, 190–202.
3. Potapov, P.; Hansen, M.C.; Stehman, S.V.; Loveland, T.R.; Pittman, K. Combining MODIS and Landsat imagery to estimate and map boreal forest cover loss. *Remote Sens. Environ.* **2008**, *112*, 3708–3719.
4. Dennison, P.E.; Roberts, D.A. Daytime fire detection using airborne hyperspectral data. *Remote Sens. Environ.* **2009**, *113*, 1646–1657.
5. Liu, J.Y.; Liu, M.L.; Tian, H.Q.; Zhuang, D.F.; Zhang, Z.X.; Zhang, W.; Tang, X.M.; Deng, X.Z. Spatial and temporal patterns of China's cropland during 1990-2000: an analysis based on Landsat TM data. *Remote Sens. Environ.* **2005**, *98*, 442–456.
6. Zhu, H.W.; Basir, O. An adaptive fuzzy evidential nearest neighbor formulation for classifying remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2005**, *43*, 1874–1889.
7. Muller, K.R.; Mika, S.; Ratsch, G.; Tsuda, K.; Scholkopf, B. An introduction to kernel-based learning algorithms. *IEEE Trans. Neural Networks* **2001**, *12*, 181–201.
8. Richards, J.A.; Jia, X. *Remote Sensing Digital Image Analysis: an Introduction*, 4th ed.; Springer-Verlag: Secaucus, NJ, USA, 2006.

9. Bárdossy, A.; Samaniego, L. Fuzzy rule-based classification of remotely sensed imagery. *IEEE Trans. Geosci. Remote Sens.* **2002**, *40*, 362–374.
10. Deer, P.; Eklund, P. A study of parameter values for a Mahalanobis distance fuzzy classifier. *Fuzzy Sets Syst.* **2003**, *137*, 191–213.
11. Friedman, N.; Geiger, D.; Goldszmidt, M. Bayesian network classifiers. *Mach. Learn.* **1997**, *29*, 131–163.
12. Atkinson, P.M.; Tatnall, A.R.L. Introduction neural networks in remote sensing. *Int. J. Remote Sens.* **1997**, *18*, 699–709.
13. Foody, G.M.; Mathur, A. A relative evaluation of multiclass image classification by support vector machines. *IEEE Trans. Geosci. Remote Sens.* **2004**, *42*, 1335–1343.
14. Massa, A.; Boni, A.; Donelli, M. A classification approach based on SVM for electromagnetic subsurface sensing. *IEEE Trans. Geosci. Remote Sens.* **2005**, *43*, 2084–2093.
15. Cleveland, W.; Devlin, S. Locally weighted regression: an approach to regression analysis by local fitting. *J. Amer. Statist. Assoc.* **1988**, *83*, 596–610.
16. Franco-Lopez, H.; Ek, A.; Bauer, M. Estimation and mapping of forest stand density, volume, and cover type using the k-nearest neighbors method. *Remote Sens. Environ.* **2001**, *77*, 251–274.
17. McLachlan, G. *Discriminant Analysis and Statistical Pattern Recognition*; John Wiley & Sons: New York, NY, USA, 1992.
18. Foody, G.M. Approaches for the production and evaluation of fuzzy land cover classifications from remotely-sensed data. *Int. J. Remote Sens.* **1996**, *17*, 1317–1340.
19. Poggi, G.; Scarpa, G.; Zerubia, J.B. Supervised segmentation of remote sensing images based on a tree-structured MRF model. *IEEE Trans. Geosci. Remote Sens.* **2005**, *43*, 1901–1911.
20. Bachmann, C.M.; Ainsworth, T.L.; Fusina, R.A. Exploiting manifold geometry in hyperspectral imagery. *IEEE Trans. Geosci. Remote Sens.* **2005**, *43*, 441–454.
21. Serpico, S.B.; Moser, G. Extraction of spectral channels from hyperspectral images for classification purposes. *IEEE Trans. Geosci. Remote Sens.* **2007**, *45*, 484–495.
22. Bárdossy, A.; Pegram, G.S.; Samaniego, L. Modeling data relationships with a local variance reducing technique: applications in hydrology. *Water Resour. Res.* **2005**, *41*, W08404, doi:10.1029/2004WR003851.
23. Samaniego, L.; Bardossy, A.; Schulz, K. Supervised classification of remotely sensed imagery using a modified k-nn technique. *IEEE Trans. Geosci. Remote Sens.* **2008**, *46*, 2112–2125.
24. Cover, T.M.; Hart, P.E. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* **1967**, *13*, 21–27.
25. Hastie, T.; Tibshirani, R. Discriminant adaptive nearest neighbor classification. *IEEE Trans. Pattern Anal. Mach. Intell.* **1996**, *18*, 607–615.
26. Peng, J.; Heisterkamp, D.R.; Dai, H.K. Adaptive quasiconformal kernel nearest neighbor classification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2004**, *26*, 656–661.
27. Lowe, D. Similarity metric learning for a variable-kernel classifier. *Neural Computat.* **1995**, *7*, 72–85.
28. Mahalanobis, P.C. On the generalised distance in statistics. *Proc. Natl. Inst. Sci.* **1936**, *2*, 4955.
29. Fukunaga, K. *Introduction to Statistical Pattern Recognition*, 2nd ed.; Academic Press Professional,

Inc.: San Diego, CA, USA, 1990.

30. Goodin, D.G.; Gao, J.; Henebry, G.M. The effect of solar illumination angle and sensor view angle on observed patterns of spatial structure in tallgrass prairie. *IEEE Trans. Geosci. Remote Sens.* **2004**, *42*, 154–165.
31. Isaaks, E.H.; Srivastava, R.M. *An Introduction to Applied Geostatistics*; Oxford University Press: New York, NY, USA, 1989.
32. Aarts, E.; Korst, J. *Simulated Annealing and Boltzmann Machines: a Stochastic Approach to Combinatorial Optimization and Neural Computing*; John Wiley & Sons: Chichester, UK, 1989.
33. Falkenauer, E. *Genetic Algorithms and Grouping Problems*; John Wiley & Sons: Chichester, UK, 1997.
34. Tolson, B.A.; Shoemaker, C.A. Dynamically dimensioned search algorithm for computationally efficient watershed model calibration. *Water Resour. Res.* **2007**, *43*, W01413, doi:10.1029/2005WR004723.
35. Mallows, C.L. Some comments on C_p. *Technometrics* **1973**, *15*, 661–667.