International Journal of Geographical Information Science
Vol. 00, No. 00, Month 200x, 1–51

RESEARCH ARTICLE

Incorporating Spatial Association into Statistical Classifiers: Local Pattern-based Prior Tuning

HEXIANG BAI[†], FENG CAO[†], ATKINSON M. PETER[‡], QIAN CHEN[†], JINFENG WANG[§] and YONG GE[*§]

[†]*School of Computer and Information Technology, Shanxi University, Taiyuan, 030006 Shanxi, China*; [‡]*Faculty of Science and Technology, Engineering Building, Lancaster University, Lancaster LA1 4YR, UK.*; [§]*State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China*

(*xxxx.xx.xx*)

This paper proposes a new classification method for spatial data by adjusting prior class probabilities according to local spatial patterns. First, the proposed method uses a classical statistical classifier to model training data. Second, the prior class probabilities are estimated according to the local spatial pattern and the classifier for each unseen object is adapted using the estimated prior probability. Finally, each unseen object is classified using its adapted classifier. Because the new method can be coupled with both generative and discriminant statistical classifiers, it performs generally more accurately than other methods for a variety of different spatial datasets. Experimental results show that this method has a lower prediction error than statistical classifiers that take no spatial information into account. Moreover, in the experiments, the new method also outperforms spatial auto-logistic regression and Markov random field-based methods when an appropriate estimate of local prior class distribution is used.

Keywords: Spatial Pattern; Statistical Classifier; Spatial Auto-logistic Regression; Spatial Data

1. Introduction

As modern Earth observation devices and programmes have been developed, the classification of spatial data has been increasingly used to discover knowledge from very large datasets. For example, periodic changes in the distribution of different types of vege-

*Corresponding author. Email: gey@lreis.ac.cn

ISSN: 1365-8816 print/ISSN 1362-3087 online

DOI: 10.1080/1365881YYxxxxxxxx
http://www.informaworld.com

tation can be obtained through the classification of satellite sensor image time-series. Furthermore, classification results are information sources for further interpretation of related geographical phenomena. For example, land cover type has been used to study the urban heat island phenomenon (Gartland 2010), analyze the run-off dynamics of river basins (Khare *et al.* 2015) and plan biosphere reserves (Evans 2017). Accordingly, classification techniques have attracted much attention.

Statistical classifiers are one of the most popular types of classifier in the geosciences. Various kinds of statistical classifiers exist, such as the naive Bayesian, maximum likelihood, k-nearest neighbor, support vector machine and logistic/softmax regression classifiers (Duda *et al.* 2001, Bishop 2006, Murphy 2012). These classifiers model the feature space of geographical objects from different perspectives. Generally, some classifiers try to find the best hyperplane between classes directly in the feature space, while others attempt to model the distribution of feature values for different classes. No matter which classifier is used, the distribution of different categories in space is generally neglected during modelling and reasoning.

It is important to model spatial patterns as well as the characteristics of other features when modelling geographical phenomena and processes. Geographical phenomena may have some patterns in their distribution in space. For example, spatial auto-correlation is a simple measurable spatial pattern. For datasets with spatial auto-correlation, ignoring spatial auto-correlation may lead to spatially auto-correlated regression residuals (Anselin 1988, Haining 1990). Meanwhile, compared with the situation that sampling units are independent from each other, fewer samples may be needed to infer the parameters; for example, the mean of the population given the same allowed error when using spatial sampling methods (Goovaerts 1997, Wang *et al.* 2002).

Undoubtedly, the modelling of spatial patterns is a necessary step in the classification of spatial data. There have been many related studies about incorporating spatial pattern information into the classification process. These studies can be divided roughly into three types. The first type creates new features that consider spatial pattern information (Ding *et al.* 2009, Vainer *et al.* 2009), for example, the kernel-based image smoothing method. These methods are not classification methods. The features created can be coupled with any classifiers to increase classification accuracy. The disadvantage of this type of method is that spatial feature creation is often application-specific and time-consuming (Jiang and Shekhar 2017). The second type uses data fusing technologies, such as consensus theory, Dempster–Shafer theory, Bayes updating and many other fusion techniques (Strebelle 2000, Ge and Bai 2010, 2011, Zhang and Prasad 2016, Joshi *et al.* 2016). However, data fusion methods will inevitably introduce new assumptions about the data and some new parameters. These will complicate the modelling process and introduce some additional unnecessary uncertainties.

The final approach is to develop new statistical classifiers. Currently, there are two major approaches for incorporating spatial patterns into statistical classification models: Markov random field (MRF) models (Jeon and Landgrebe 1992, Solberg *et al.* 1996) and spatial auto-regressive (SAR) models (Griffith 1989). These two models have been used widely in classification and have attracted the attention of many researchers (Ni *et al.* 2014, Xia *et al.* 2015, Yu *et al.* 2016, White *et al.* 2017, Hughes *et al.* 2011, Chun and Griffith 2013, Liu *et al.* 2017a,b). However, an MRF needs a large amount of computing resource, especially when high-order neighbors of objects are involved in finding the minimum value of the energy function, although this method has been shown to be effective (Austad and Tjelmeland 2017). SAR needs less computational resource than MRF to estimate parameters (Sherman *et al.* 2006, Paciorek 2007). However, it requires the

data to match many restrictive assumptions about the probability distributions of feature values as well as the class boundaries (Shekhar *et al.* 2002). Accordingly, SAR alone is insufficient to classify all kinds of spatial datasets, which may have different specific feature space distributions and class boundary characteristics. Some studies have tried to incorporate spatial pattern information into other classifiers. For example, Goovaerts (2002) used indicator kriging to estimate the prior probability for the maximum likelihood classifier to increase the classification accuracy of hyperspectral data.

Besides the above mentioned classifiers, there are many other statistical classifiers in the field of machine learning. These classifiers model the data from a variety of perspectives using different assumptions. For spatial data that do not match some of the assumptions of MRF and SAR, these statistical classifiers provide alternative approaches. However, these classifiers lack appropriate modelling of spatial patterns. Accordingly, it is imperative to develop these statistical classifiers by incorporating spatial patterns into the framework of statistical classification.

The aim of this paper was to develop a unified framework for traditional statistical classifiers to introduce spatial pattern information into the modelling process for spatial data. This framework first uses sample data to train traditional classifiers. Secondly, it introduces the assumption that the prior probabilities of different classes are the marginal probabilities of different classes in the sample data. Finally, it modifies the prediction model by replacing the prior distribution of different classes with the local prior class probabilities for each unseen object. In the experiment section, we compare the proposed framework with traditional statistical classifiers. Moreover, the SAR and MRF models are compared with the proposed framework. The comparison shows that the proposed framework has an advantage over MRF and SAR when the classes have statistically significant spatial association and the local prior distribution of different categories is accurately estimated.

The rest of the paper first reviews the basis of statistical classifier systems. Next, a new classification algorithm for spatial data is proposed through incorporating spatial pattern information using a unified framework. Finally, a series of experiments on four real-life spatial datasets is performed and the proposed algorithm is compared with other commonly used classifiers to validate the effectiveness of the proposed method. The final section concludes the paper.

2. Background

We take the classification of a remotely sensed image as an example to illustrate a typical scenario. Assuming there is a remotely sensed image which has N pixels and J bands, each pixel of the remotely sensed image can be represented using its values on all J bands. A pixel u can be represented using a feature vector $\vec{x}(u) = [x_1(u), x_2(u), \cdots, x_J(u)]^T$, in which $x_i(u)$ is the gray value of band $1 \leq i \leq J$. Generally, some pixels are used for training classifiers. The categories of these pixels have already been assigned to one of K land cover types (categories) $\{y_1, y_2, \cdots y_K\}$. The remaining pixels are unseen objects to be classified. These pixels are described only using a feature vector and have no pre-assigned categories.

The aim of classifying a remotely sensed image is to infer which land cover type should be assigned to each unseen pixel according to its feature vector. For a statistical classifier, the purpose is to find the most probable land cover type for pixel u, that is, $c(u) = \arg\max\{P(y_i|\vec{x}(u))|i \in 1 \cdots K\}$, where $c(u)$ is the estimated category of u and $P(y_i|\vec{x}(u))$

is the probability of observing land cover type y_i given feature vector $\vec{x}(u)$. Obviously, estimating $P(y_i|\vec{x}(u))$ is the key issue in designing a statistical classifier. In terms of the approach to $P(y_i|\vec{x}(u))$ estimation, statistical classification can be divided into two types (Bishop 2006). In the following, $\vec{x}(u)$ is simplified to $\vec{x}$.

The first type of statistical classifier is called a generative model. For each land cover type y_i, these models first learn $P(\vec{x}|y_i)$ and then infer $P(y_i|\vec{x})$ using Bayes' theorem with the help of the prior probability $P(y_i)$, i.e.,

$$P(y_i|\vec{x}) = \frac{P(\vec{x}|y_i)P(y_i)}{\sum_{j=1}^{K} P(\vec{x}|y_j)P(y_j)} \tag{1}$$

Representative generative models for classification include the naive Bayesian classifier, maximum likelihood classifier, and hidden Markov model.

Another type of statistical classifier is called the discriminant model. A discriminant model learns $P(y_i|\vec{x})$ directly from the sample dataset and then subsequently uses this distribution to make optimal decisions. Generally, it first defines a parametric model $P(y_i|\vec{x}, \boldsymbol{\theta})$ in terms of the characteristics of the data and some necessary assumptions. Then, it estimates parameter $\boldsymbol{\theta}$ using statistical parameter estimation methods. Finally, given any input feature vector, it can generate the corresponding y_i (sometimes with the help of a sigmoid or softmax function). Representative discriminant models for classification include logistic regression, neural networks, k-nearest neighbor, and relevance vector machine classifiers.

For both models, the relation between $P(\vec{x}|y_i)$ and $P(y_i|\vec{x})$ is

$$P(\vec{x}|y_i) = \frac{P(\vec{x}) \times P(y_i|\vec{x})}{P(y_i)}, \tag{2}$$

where a reasonable assumption of $P(y_i)$ is the marginal probability of y_i in the sample data when no prior knowledge of the probability distribution of different classes is given (Duda *et al.* 2001).

3. Method

The lack of consideration of spatial patterns in statistical classification decreases prediction accuracy. The traditional statistical classifier uses only the feature set, which is represented using $\vec{x} = [x_1, x_2, \cdots, x_J]^T$, of a geographical object to learn the conditional probability $P(y_i|\vec{x})$ of observing category y_i given $\vec{x}$. When the model is used to predict unseen objects, the same model is used without considering the locations of the target objects. For example, the maximum likelihood classifier (MLC) can be used to classify a remotely sensed image. It generally assumes that the bands of the remotely sensed image have a multi-Gaussian distribution. Then, MLC learns the parameters from sample data and constructs the posterior distribution $P(y_i|\vec{x})$. Finally, each un-labelled pixel is classified to the category that has the maximum posterior probability. The whole classification process neglects completely the spatial pattern that prevails in the remotely sensed imagery.

However, according to the first law of geography (Tobler 1970), objects within an area where some category prevails are more likely to belong to that category than another category. For example, two pixels a and b have the same spectral values. a lies in an area

full of grasses and b is located in forest. a is more likely to belong to grassland and b is more likely to belong to forest. As different categories of geographical objects are always unevenly distributed in space, traditional statistical classifiers inevitably underestimate the probability of $P(y_i|\vec{x})$ in the region in which y_i prevails and overestimate the probability of $P(y_i|\vec{x})$ in the region where y_i is rare. This is the over-generalization of the classifier to unseen objects in the regions that are not suitable for the trained classier, which leads to a lower classification accuracy than expected.

More specifically, Figure 1 presents an example of the above task. When all objects of the study area are used as training data, the decision plane corresponds to the black line. However, if only the left (right) part of the study area is used to train a classifier, the red (blue) line is the appropriate decision plane. Obviously, when classifying a new object in the left (right) part of the study area, the red (blue) decision plane is more appropriate than the black one.

[Figure 1 about here.]

To overcome the above deficiency, a natural idea is to adjust the learned probabilistic model $P(y_i|\vec{x})$ through modelling the local spatial pattern for each unseen object. We, thus, propose a new local pattern-based prior tuning statistical classifier (LPPT), which is illustrated in Figure 2. This classifier proceeds as follows.

[Figure 2 about here.]

First, the probabilistic model $P(y_i|\vec{x})$ can be learned from the sample data. The training process is the same as that of the selected traditional statistical classifier. One should choose the classifier that is suitable for the application need. The selected traditional statistical classifier is called the original classifier. For example, an MLC model can be learned from the sample data of a remotely sensed image.

Second, local patterns are used to infer the prior probability of different categories for each unseen object. Generally, the local pattern of the unseen object is modelled using its neighboring geographical objects. Let u denote an unseen object, for example, a pixel with no land cover type in a remotely sensed image, and $\mathcal{N}(u)$ be the neighbor of u, for example pixel, classifications in the local neighborhood of u. Information in $\mathcal{N}(u)$ can be used to estimate the probability with which u belongs to y_i. The probability of a given category y_i to which u belongs in terms of the neighboring objects is referred to as $P_{\mathcal{N}(u)}(y_i)$.

There are many different approaches to calculate the neighbors of a geographical object u. One approach is to use the first to kth order adjacency neighbors. The first order neighbors can be constructed using any connectivity algorithm (Fortin and Dale 2005), and the kth order neighbors can be established using the concept of relation composition (Bai *et al.* 2016). Another possible approach is to select objects within a given distance to u. For simplicity, this distance is denoted as the neighboring distance. Users can choose the most suitable method in terms of applications and the characteristics of the underlying geographical phenomena. For example, when the area of different objects differs greatly, the latter approach may be more suitable. In all the following experiments, one object is another object's neighbor if its distance is less than the predefined neighboring distance.

The simplest way of estimating $P_{\mathcal{N}(u)}(y_i)$ is to use the proportions of the categories of neighboring geographical objects to estimate the prior class probabilities for this object, if $P(y_i)$ in the neighborhood of u is stationary. This is called the local observed frequency (LOF) estimator. All the neighboring objects of u are used as samples. Meanwhile, the

neighbors labelled y_i form a set $\mathcal{N}_{y_i}(u)$. Then

$$P_{\mathcal{N}(u)}(y_i) = \frac{|\mathcal{N}_{y_i}(u)|}{|\mathcal{N}(u)|}, \tag{3}$$

where $|\mathcal{N}_{y_i}(u)|$ is the number of neighboring objects of category y_i and $|\mathcal{N}(u)|$ is the number of neighboring objects in the training data. Taking a remotely sensed image as an example, each pixel's $P_{\mathcal{N}(u)}(y_i)$ could be learned from its neighboring training pixels. The percentage of each category of all neighboring training pixels can be used as the prior class distribution.

Alternative approaches exist for modelling spatial patterns under the assumption of stationarity. For example, indicator kriging can be used to model the prior probabilities of different categories for each unseen object. When the geographical phenomenon or process has complex spatial structures, multiple point statistics (e.g., SNESIM), can be used to model the prior probabilities of different categories.

Finally, $P_{\mathcal{N}(u)}(y_i)$ is used to replace the marginal probability $P(y_i)$ in Equation (1), that is,

$$P_{\mathcal{N}(u)}(y_i|\vec{x}) = \frac{P(\vec{x}|y_i)P_{\mathcal{N}(u)}(y_i)}{\sum_{j=1}^{K} P(\vec{x}|y_j)P_{\mathcal{N}(u)}(y_j)} \tag{4}$$

to calculate the final adjusted $P_{\mathcal{N}(u)}(y_i|\vec{x})$ for u. From the decision plane $P(y_i|\vec{x})$ learned by the original statistical classifier, $P(\vec{x}|y_i) = P(\vec{x}) \times P(y_i|\vec{x})/P(y_i)$ (see Equation (2)). Accordingly,

$$\begin{aligned}
&\frac{P(\vec{x}|y_i)P_{\mathcal{N}(u)}(y_i)}{\sum_{j=1}^{K} P(\vec{x}|y_j)P_{\mathcal{N}(u)}(y_j)} \\
=&\frac{P(\vec{x}) \times P(y_i|\vec{x})/P(y_i) \times P_{\mathcal{N}(u)}(y_i)}{\sum_{j=1}^{K} P(\vec{x}) \times P(y_i|\vec{x})/P(y_j) \times P_{\mathcal{N}(u)}(y_j)} \\
=&\frac{P(y_i|\vec{x})P_{\mathcal{N}(u)}(y_i)/P(y_i)}{\sum_{j=1}^{K} P(y_i|\vec{x})/P(y_j)P_{\mathcal{N}(u)}(y_j)}
\end{aligned}$$

For simplicity,

$$P_{\mathcal{N}(u)}(y_i|\vec{x}) \propto \frac{P(y_i|\vec{x})}{P(y_i)} \times P_{\mathcal{N}(u)}(y_i). \tag{5}$$

Replacing of the prior distribution with a local one is an effective and commonly used strategy for measuring local spatial association, for example, as used in the Local Indicators of Spatial Association (LISA) (Anselin 1995) and the Local Indicators of Categorical Data (LICD) (Boots 2003) approaches. In both methods, spatial association is measured in the neighborhoods of each geographical object.

Based on the general procedure presented, a new classification algorithm based on the LOF estimator (see Equation (3)) was developed. The algorithm is divided into two stages, training and classification. The training stage starts with the selection of the appropriate traditional statistical classifier. Next, the traditional classifier is trained for spatial data $\mathcal{D} = \{< \vec{x}(u), y(u) >: u = 1, 2, \cdots, N\}$, where $\vec{x}(u)$ and $y(u)$ are the feature vector and

category of the ith object in the training data, respectively. The traditional classification model learned from the training data is $\{P(y_i|\vec{x}) : i = 1, 2, \cdots, K\}$.

The classification stage involves three main steps: finding the neighboring objects $\mathcal{N}(u)$ of the unseen object u, estimating $P_{\mathcal{N}(u)}(y_i)$ in terms of the LOF estimator, and calculating $P_{\mathcal{N}(u)}(y_i|\vec{x})$ in terms of Equation (5). During the estimation of $P_{\mathcal{N}(u)}(y_i)$, some unseen objects may have very few or even no neighboring objects with decision values. We add a smoothing factor to Equation (3) to solve this issue:

$$P_{\mathcal{N}(u)}(y_i) = \frac{|\mathcal{N}_{y_i}(u)| + P(y_i)}{|\mathcal{N}(u)| + 1}. \tag{6}$$

When there are no neighboring objects with decision values, $P_{\mathcal{N}(u)}(y_i)$ falls back to the marginal probability of category y_i.

Compared with traditional statistical classifiers, additional time is needed to estimate $P_{\mathcal{N}(u)}(y_i)$. Generally, the number of neighboring objects is very small compared to the total number of objects, hence the time spent in counting the number of different categories is short.

When the classes have statistically significant spatial associations, the prior probabilities of different categories estimated by the LOF reflect the real situation more accurately than the marginal probabilities for most unseen objects. As illustrated in the experiment in Section 5.1.1.2, the more accurate prior gave the opportunity to generate more accurate classifiers. More accurate classifiers generally produce more accurate classification results. Accordingly, the successful application of LPPT relies on the existence of statistically significant spatial associations in the classes and the correct estimation of the local prior distribution of different categories for unseen objects.

4. Experiments

A series of experiments on simulated datasets and three real datasets are presented in this section. The first real dataset consisted of a Landsat Thematic Mapper (TM) remotely sensed image and a Gaofen-2 remotely sensed image. The second real dataset concerned neural tube birth defects (NTD) in Heshun, Shanxi, China. The third real dataset focused on poverty-stricken villages in Yunyang, Hubei, China. The three real-life datasets were collected to serve different objectives. The first real dataset showed that LPPT was effective for multi-category cases and can increase the classification accuracy greatly when there are strong spatial associations in the classes. The second real dataset was used to evaluate LPPT when there are statistically significant, but relatively weak spatial associations. The third real dataset was used as an example to analyze the performance of LPPT when there was no spatial auto-correlation.

4.1. *Classification of simulated data*

Different simulated datasets were used to evaluate the performance of LPPT under different scenarios. First, simple two-category simulated data were used to show the effectiveness of the proposed method. Figure 3(a) shows a simulated two-category spatial dataset. This dataset was simulated using sequential indicator simulation (SIS). The proportions of the "Black" and "White" categories were both set to 0.5. Both x and y axes ranged from 0 to 50. The exponential variogram model was used, and the sill and range were set

to 1 and 7 units for the simulation. For simplicity, each object had only one feature. The feature values that corresponded to the "Black" and "White" categories obeyed Gaussian distributions $N(3,1)$ and $N(4,1)$, respectively.

[Figure 3 about here.]

Furthermore, simulated datasets with multiple categories and attributes were generated to analyze the performance of LPPT under different scenarios. Figure 3(b) shows a simulated six-category dataset. SIS was used to simulate multiple category datasets. Similarly, the proportions of different categories were all set to $1/K$, where K is the number of categories. Both x and y axes ranged from 0 to 50. The exponential variogram model was used, and the sill and range were set to 1 and 7 units for each category during the simulation. The values of different features that corresponded to different categories obeyed multi-Gaussian distributions. Sensitivity analysis was conducted for the neighboring distance and the number of categories. Finally, the LPPT method was compared with other classifiers on 1,000 simulated datasets with a random number of categories (from two to six) and random number of attributes (from one to six).

4.2. *Recognition of vegetation types from remotely sensed imagery*

This dataset was used to illustrate the typical scenario of applying LPPT: there were multiple categories in the study area and the classes had statistically significant spatial associations. Two remotely sensed images were used for classification to further validate the effectiveness of the proposed classification model. The first image was clipped from the Landsat TM image with product id LT05_L1TP_125034_20100923_20161013_01_T1. The study area was located in the south-east of Taiyuan, Shanxi Province. The TM image size was $1,000 \times 1,000$ pixels. The image contained seven spectral bands. The spatial resolution of band six was 120 m. The spatial resolution of all other bands was 30 m. The upper-left latitude and longitude coordinates of this image were $113°19'32.41''E$ and $37°59'43.34''N$, and its lower-right latitude and longitude coordinates were $113°39'25.96''E$ and $37°43'8.49''N$, respectively. Figure 4(a) shows the 5, 4, 3-band pseudo-colour composite image. The second remotely sensed image used was the Gaofen-2 image with product id GF2_PMS2__L1A0001708261-MSS2. The image size was 7299×6999 pixels and it contained four spectral bands with a spatial resolution of 4 m. This image and its label is available at http://captain.whu.edu.cn/GID/. Please refer to Tong *et al.* (2020) for a detailed description of the Gaofen-2 image. Figure 4(b) shows the red, green, blue color composite image of the Gaofen-2 image.

[Figure 4 about here.]

Both the TM image and Gaofen-2 image were segmented using the multi-resolution segmentation function of eCognition 8.9 before classification. Both images were classified following the object based image analysis framework. During the segmentation, "the weight of color criterium" was set to 0.5, and "the maximum standard deviation of the homogeneity in regard to the weighted image layers" was set to five for the TM image and 200 for the Gaofen-2 image. The segmentation result of the TM image is shown in Figure 4(a). Each object of the segmentation result of the TM image and Gaofen-2 image had seven features and four features, respectively. Each feature corresponded to a spectral band and its value was the algorithmic mean of all the pixels in the object. For the TM image, the land cover type of each object was obtained from the Global Land Cover Map (GLOBCOVER) 2009 (Bicheron *et al.* 2008), which had a spatial resolution of 300 m.

Each object in the segmentation result intersected with the pixels of the GLOBCOVER 2009 map. The category with the largest area was used as the object's category. For the Gaofen-2 image, the land cover types were labeled by Tong *et al.* (2020).

When these category values were assumed as the true categories of each object, new uncertainties were introduced. For example, the intersection between the segmented result and the pixels introduced polygon overlay uncertainties (Smith and Campbell 1989) in the result and the GLOBCOVER 2009 map also contained system or random errors (Bicheron *et al.* 2008). These uncertainties led to bias in the estimation of the parameters of the classifiers. If the classification results were used in other tasks, these uncertainties would influence the consequent analysis. Accordingly, it is necessary to provide uncertainty analysis in real life applications using, for example, uncertainty propagation methods (Rajabi 2019, Chen *et al.* 2019). To alleviate any such influence on the evaluation of LPPT, the training data and validation data both used the assumed true categories.

Although, there were 10 types of vegetation in the TM image, some vegetation types had very few instances. Accordingly, six categories with more than 100 instances were selected to perform classification. All the other four categories were merged into one category. In the following, "1" to "7" were used to represent the "Rainfed croplands," "Others," the "Mosaic Cropland (50-70%) / Vegetation (grassland, shrubland, forest) (20-50%)," the "Closed (>40%) needleleaved evergreen forest (>5m)," the "Closed to open (>15%) mixed broadleaved and needleleaved forest (>5m)," the "Mosaic Forest / Shrubland (50-70%) / Grassland (20-50%)," and "Artificial surfaces and associated areas (urban areas >50%)" classes, respectively. The Gaofen-2 image had five types of vegetation. In the following "a" to "e" were used to represent the "Built-up", "Others", "Farmland", "Forest", and "Meadow" classes, respectively. Figure 5 shows the distribution of different vegetation types. Table 1 and 2 shows the degree of spatial association calculated using NCP[1] (Bai *et al.* 2016) for different neighboring distances.

[Figure 5 about here.]

[Table 1 about here.]

[Table 2 about here.]

The GLOBCOVER map had a coarse spatial resolution, such that some fine resolution details of the patterns of different categories might be missed, which might increase the degree of spatial association for target categories in real life applications. In our experiment, this dataset had a higher degree of overall spatial association than the other two real-life datasets.

The main purpose of this dataset was to evaluate the effectiveness of LPPT in the situation where there is strong spatial association in multiple classes. To strengthen the multi-category and strong spatial association characteristics, some less important factors for the validation of LPPT were ignored. This is instructive for exploring the performance of LPPT in such situations.

4.3. *Prediction of NTD occurrences*

This dataset was used to test the performance of LPPT when there are statistically significant, but relatively weak spatial associations in classes. NTD data have been collected

[1] $NCP > 0$, $NCP = 0$ and $NCP < 0$ indicate positive, no and negative spatial associations, respectively. The larger the NCP is, the stronger the spatial association.

over several years and investigated in many previous studies (Wu *et al.* 2004, Liao *et al.* 2009a,b, Wang *et al.* 2010, Bai *et al.* 2010, 2016). In the study area, there were 322 villages and one town. The locations of the 322 villages were determined using a geographical information system. The data were collected by a field survey. This research project was approved by the Ministry of Science and Technology of the People's Republic of China. The study used only local statistical data. There were no experimental or ethical issues. As there were no boundaries defined for the villages, they were drawn for each village using Voronoi polygons (see Figure 6). Meanwhile, the villages that did not have new births from 1998 to 2003 were not included in the figure and the experiment.

[Figure 6 about here.]

Each village had 14 conditional attributes and one decision attribute. Nine continuous-valued attributes were used in the experiment, including gross domestic product (GDP) *per capita*, fertilizer used in the area, access to a doctor, production of fruit, production of vegetables, elevation, distance to rivers, distance to roads, and distance to fault lines. All the maps of the attributes can be found in Wang *et al.* (2010) and Bai *et al.* (2010). A detailed description of the NTD data can be found in Wang *et al.* (2010). The decision attribute was whether there were NTD instances in a village. If there were NTD instances, then the village was labeled as "Yes"; otherwise, the village was labeled "No". Table 3 shows the degree of spatial association calculated using NCP for different neighboring distances.

[Table 3 about here.]

4.4. *Identification of poverty in villages*

This dataset was used to show the effectiveness of LPPT when there are no statistically significant spatial associations in classes. Poverty data in the village level for Yunyang, Hubei, China were collected using a local government field survey. The poverty headcount was identified using the income poverty line of China (2736 Chinese yuan in 2013). If the proportion of poor people in a village was larger than 2%, then the village was considered a poverty-stricken village (Hubei provincial government 2018). There were 85 poverty-stricken villages in the study area which consisted of 340 villages. The locations of these 340 villages were determined using a geographical information system (see Figure 7).

[Figure 7 about here.]

Each village had six conditional attributes and one decision attribute, including the population of working ages, the proportion of migrant workers, the number of participants in the new rural cooperative medical system, cement road mileage, and the number of households with broadcast and television fiber optic cables. The decision attribute was whether the village is in poverty. Poverty-stricken villages were labelled as "Yes"; otherwise, the village was labelled "No". Table 4 shows the degree of spatial association calculated using NCP for different neighboring distances.

[Table 4 about here.]

4.5. *Classifier comparison process*

For each dataset, the comparison between the LPPT-based and other methods was performed following the process shown in Figure 8. First, the dataset was divided into train-

ing and validation data using simple random sampling. Second, different classifiers were trained using the training data. Next, the validation data were classified using all trained classifiers. Finally, different accuracy assessment indices were computed for all classification results to compare the effectiveness of the different classifiers. In the experiments, the entire process was repeated 1,000 times.

[Figure 8 about here.]

In the sampling step, some geographical objects were selected randomly as training data and all the remaining objects were used as validation data. When predicting the occurrence of NTD instances and identifying poverty-stricken villages, half the villages were drawn randomly as training data to feed sufficient samples into the classifier. Regarding recognizing vegetation types and experiments on simulated datasets, there were more than 2900 objects, and 10% of objects were drawn randomly as training data.

During the training stage, four traditional statistical classifiers were selected as the original classifiers and were trained using the training data: naive Bayes (NB), k-nearest neighbors (kNN), relevance vector machines (RVM) and logistic regression (LR). In these classifiers, NB is a generative model-based classifier and the other three classifiers are discriminant model-based classifiers. The scikit-learn (Pedregosa *et al.* 2011) and scikit-rvm (Ritchie 2017) packages in Python were used to implement these algorithms. In the NB classifier, the Gaussian distribution was used to model each attribute. The kNN algorithm used five nearest neighbors. The pseudo-random number generator seed of LR was set to 42. All other parameters of these algorithms were set to their default values.

Additionally, two spatial classification models, SAR and the model which incorporated neighboring object features, were trained using the training data. The SAR was solved using the pseudo-likelihood method (Sherman *et al.* 2006) and was also implemented in Python. The second approach, NeighFea for simplicity, included 10 neighboring object feature values in the current pixel's feature vector for classification. The training and classification steps were the same as those of the classical statistical classifiers. The NeighFea models that corresponded to the above four classical statistical classifiers were denoted by NeighFea_NB, NeighFea_kNN, NeighFea_RVM, and NeighFea_LR.

In the classification step, only the validation data were used to assess classification accuracy. The classical statistical classifier, SAR, and NeighFea models were used directly to classify the validation data. Meanwhile, the local prior was calculated for each geographical object using neighboring objects in the training data to update the corresponding original classifiers. The LPPT-based methods are also implemented using Python. All four classical statistical classifiers were coupled with LPPT, and the corresponding classifiers were denoted by NE_NB, NE_kNN, NE_RVM, and NE_LR.

The comparison also used MRF-based classifiers (Solberg *et al.* 1996). The difference between the four MRF-based models used was their energy functions. All the energy functions contained two parts: a spatial contextual term E_{sp}, and a class-conditional term E_{data}. E_{sp} was modelled in the same manner used in (Solberg *et al.* 1996). E_{data} was generally modelled using $ln(P(\vec{x}|y_i))$. Different traditional statistical classifiers learned different $P(\vec{x}|y_i)$, and generated four different E_{data} models. Corresponding to the four traditional statistical classifiers used, four MRF models, MRF_NB, MRF_kNN, MRF_RVM, and MRF_LR, were used. The weights of E_{data} and E_{sp} were both set to one. The iterated conditional model (ICM) algorithm was used to determine the local minimum energy. Details of the algorithm can be found in (Solberg *et al.* 1996).

Two methods were used to evaluate the classification accuracy. Confusion matrices (Cohen 1960, Powers 2011) are commonly used to evaluate classifications. Three indices from

the confusion matrix were selected to perform the accuracy assessment. The overall accuracy refers to the proportion of correctly classified objects in the test data. For each category, the precision is the ratio of the number of correctly classified objects to the number of all the objects that are allocated to this category by the classifier. The recall is the fraction of the objects of a category that are correctly classified. The precision and recall have many alternative names, for example producer's accuracy and user's accuracy.

Although these three indices measure the classification accuracy when the discriminant threshold of probability is 0.5, they neglect the chance level performance when different discriminant thresholds are used (Powers 2011). The area under the curve (AUC) can measure the uncertainty of a classifier when different discriminant thresholds are used in classifications. Accordingly, the AUC metric (Fawcett 2006, Theodoridis and Koutroumbas 2009) was also selected to assess the accuracy of classification to provide a richer comprehension of the accuracy.

5. Results and discussion

5.1. *Comparison with traditional statistical classifiers*

The first objective of the experiments was to test whether the proposed LPPT method outperformed the corresponding traditional statistical classifiers. All simulated datasets and three real life datasets were used for comparison.

5.1.1. Simulated datasets

5.1.1.1. Performance of LPPT. In the following, the performance of LPPT is compared with traditional statistical classifiers from the perspective of neighboring distance, different number of categories, different number of attributes. A total of 10% of the data was used as training data and the NB classifier was used by default.

Figure 9(a) shows the increment of different average accuracy measures using LPPT when neighboring distances were set from one to 10 using the simulated two category datasets (see Figure 3). Clearly, regardless of which neighboring distance was used, all average accuracies increased. Furthermore, when the neighboring distance increased, the accuracy increment increased firstly and then decreased and gradually approached zero. The Bonferroni corrected[1] Student's t-test showed that there was a statistically significant difference between the average accuracy of LPPT and classical classifiers in all cases, when the significance level was 0.05. The Cohen's d of the difference was larger than 0.9 in most cases, which meant a relatively large increase in accuracy.

[Figure 9 about here.]

The classical classifier and LPPT were also compared for simulated datasets with different numbers of categories. Four new datasets with three to six categories were generated. All datasets had only one feature for simplicity. The feature's variance for each category was set to one and the feature's means for the six categories were set to three to eight, respectively. Figure 9(b) shows that the average overall accuracy increased using LPPT for datasets with different numbers of categories. Although the average overall accuracy decreased when the number of classes increased, the accuracy increment using

[1]The total number of comparisons for each dataset was one plus the triple of the number of categories in the dataset. The number of the comparison with true null hypothesis is used in the Bonferroni corrections.

LPPT became larger when the number of categories increased. The Bonferroni corrected Student's t-test showed that there was a statistically significant difference between the average accuracy of LPPT and classical classifiers in all cases, when the significance level was 0.05. The Cohen's d of the difference was larger than 0.9 in most cases, which meant a relatively large increase in accuracy. Meanwhile, the greater the number of categories, the larger the Cohen's d.

Finally, Figure 9(c) shows that the average overall accuracy increased when the original classifiers were NB, kNN, RVM, and LR. This was the average result of classification on 1,000 different simulated datasets. The number of categories of these datasets ranged from two to six, and the number of features of these datasets ranged from one to six. Clearly, regardless of which original classifier was used, the average overall accuracy increased in most cases for most simulated datasets. The Bonferroni corrected Student's t-test showed that there was a statistically significant difference between the average overall accuracy of LPPT and classical classifiers, when the significance level was 0.05. The Cohen's d was larger than 0.45 for the NB, kNN and LR classifiers, which meant a medium increase in overall accuracy. The Cohen's d for the RVM was 0.2, which meant a relatively small increase in overall accuracy. These experiments showed that LPPT was superior to the corresponding original classifiers and LPPT was effective in processing spatial data.

5.1.1.2. Case by case examples of LPPT. To further inspect the proposed algorithm, we selected some representative correctly and incorrectly rectified classifications of unseen objects using LPPT. Figure 10 shows five such examples. In Figure 10(a), the unseen object was surrounded by four neighboring objects with the label "Black." Meanwhile, $P(White|x = 2.48) = 0.92$ and $P(Black|x = 2.48) = 0.08$ using NB. When LPPT was used, $\mathcal{N}(u) = \mathcal{N}_{Black}(u) = 4$ and $\mathcal{N}_{White}(u) = 0$. Then, $P_{\mathcal{N}(u)}(White) = 0.69/5$ and $P_{\mathcal{N}(u)}(Black) = 4.31/5$. Therefore, $P_{\mathcal{N}(u)}(White|x = 2.48) \propto 0.92/0.69 * 0.69/5 = 0.184$ and $P_{\mathcal{N}(u)}(Black|x = 2.48) \propto 0.08/0.31 * (4.31)/5 = 0.2225$. Finally, the probability of being classified as "White" and "Black" categories using NE_NB was approximately 0.45 and 0.55, respectively. Then, the incorrectly classified object was correctly rectified to "Black." In the same manner, Figure 10(b) and (c) show two correctly rectified examples.

[Figure 10 about here.]

In most cases, unseen objects were surrounded by objects of the same category. Compared with classical classifiers, LPPT changed the decision hyper-plane according to the local prior distribution of different categories for each unseen object, which, in turn, rectified some originally misclassified objects and increased the classification accuracy.

LPPT might also incorrectly rectify the classification results of unseen objects surrounded by opposite categories. Consider Figure 10(d) as an example. The unseen object, whose true category was "White", was surrounded by four objects with the label "Black." $P(White|x = 1.59) = 0.90$ and $P(Black|x = 1.59) = 0.10$ using NB. Then, $P_{\mathcal{N}(u)}(White|x = 1.59) \propto 0.905/0.57 * 0.57/5 = 0.18$ and $P_{\mathcal{N}(u)}(Black|x = 1.59) \propto 0.095/0.43 * 4.43/5 = 0.206$. Accordingly, the probabilities of being classified as "White" and "Black" categories using NE_NB were approximately 0.47 and 0.53, respectively. Clearly, NE_NB incorrectly classified the unseen object, whereas NB correctly classified it. Figure 10(e) shows another failed example.

The false rectifications using LPPT shared the same characteristic; that is, the majority of the neighbors of the unseen objects were of the opposite category to their true category. Generally, neighboring objects tended to have the same category as the current object in

spatial data. If most objects were surrounded by objects with different labels, then there were negative spatial associations. LPPT is not suitable for such datasets.

Furthermore, LPPT can be adapted to give more consideration to such isolated objects by the modification of Equation 6 to

$$P_{\mathcal{N}(u)}(y_i) = \frac{|\mathcal{N}_{y_i}(u)| + f \times P(y_i)}{|\mathcal{N}(u)| + f} \tag{7}$$

where f is the weight of the smoothing factor. Equation 6 is the special case of Equation 7 when $f = 1$. Consider Figure 10(d) as an example. If f was set to 1.2, then $P_{\mathcal{N}(u)}(White|x = 1.59) \propto 0.90/0.57 * (1.2 * 0.57)/5.2 \approx 0.21$ and $P_{\mathcal{N}(u)}(Black|x = 1.59) \propto 0.10/0.43 * (4 + 0.43 * 1.2)/5.2 = 0.20$. Then the probabilities of being classified as "White" and "Black" categories using NE_NB were approximately 0.51 and 0.49, respectively. The classification result for the object was still "White."

The above method is an approach to alleviate the influence of the induction bias (Hsu 2003) of LPPT, which assumes that neighboring objects tend to be of the same categories in terms of the first law of geography. Besides the above approach, there are many candidate methods. For example, the local prior was set to the marginal probability for objects classified to a category with high probability by the original classifier. Or f can be set to different values for different regions in the study area. We will investigate this issue in our future work.

5.1.2. Real datasets

5.1.2.1. Strong spatial association. In the first real dataset experiment, although substantially fewer training data (only 10%) were used than that in the second and third real dataset experiments, the classification accuracy increased remarkably. Figure 11(a) shows the increment of the average overall accuracy when the neighboring distance was 2 km using different original classifiers for the TM image. The LPPT-based model's overall accuracy increased by at least 4% compared with the original classifiers. Furthermore, with the exception of the average recall of "1" using NE_NB, and average recalls of "2" and "4" using NE_kNN, which decreased slightly, the accuracies of all the other indices increased significantly. This indicates that LPPT outperformed the corresponding original classifiers. The Bonferroni corrected Student's t-test showed that there was a statistically significant difference between the average accuracy of LPPT and classical classifiers, when the significance level was 0.05, with the exception of the average precision of "2" of NE_NB, the average recall of "1" and "3" of NE_RVM, and the average recall of "1" and "2" of NE_LR. The Cohen's ds of most accuracy measures for the NB, kNN, and LR were larger than 0.9, which indicated a relatively large increase in accuracy. For the RVM classifier, the Cohen's ds of almost all indices were larger than 0.2, which meant that NE_RVM increased the classification accuracy slightly.

[Figure 11 about here.]

Figure 11(b) shows the increment of the average overall accuracy when the neighboring distance was 1 km using different original classifiers for the Gaofen-2 image. The LPPT-based model's overall accuracy increased by more than 2% compared with the original classifiers except for the RVM. Furthermore, For each classifier, the Bonferroni corrected Student's t-test showed that there was a statistically significant difference between the average overall accuracy of LPPT and classical classifiers when the significance level was 0.05. The Cohen's ds of the average overall accuracy for the NB, kNN, and LR were

larger than 0.9, which meant a relatively large increase in accuracy. With respect to the RVM classifier, the Cohen's d of the average overall accuracy was smaller than 0.2, which meant a slight increase in classification accuracy.

To further validate the performance of LPPT, the TM and Gaofen-2 remotely sensed image was divided into four equal subregions: top left, top right, bottom left and bottom right. In each sub-region, 20% objects were used as training data and all the remaining objects were used as validation data. The experiment followed the same procedure shown in Figure 8. Figure 11(a & b) shows that the average overall accuracy of all sub-regions increased when LPPT was used, regardless of which original classifiers were used. The Bonferroni corrected Student's t-test showed that there was a statistically significant difference between the average overall accuracy of LPPT and that of the original classifiers, except for the RVM for the Gaofen-2 image, when the significance level was 0.05. The Cohen's ds of all the average overall accuracies of the NB, kNN, and LR for the Gaofen-2 image were all larger than 0.68, among which 75% Cohen's ds were larger than 0.9. This indicates a relatively large difference in the accuracy.

5.1.2.2. Weak but significant spatial association. Consider the prediction of the occurrence of NTD instances as an example, Figure 11(c) shows the increment of the average overall accuracy, average precision, average recall, and average AUC of 1,000 classifications when the neighboring distance was 2 km. Although the average recall of "No" decreased slightly when kNN or LR was used, all the other average indices of LPPT-based models were larger than or equal to those of the original classifiers when kNN, NB, RVM, or LR were used. With respect to the NE_RVM, the AUC increased by almost 0.2 compared with the original RVM classifier, although it did not change the hard classification result. This means that NE_RVM increased the chance level performance of the RVM. The Bonferroni corrected Student's t-test showed that there was a statistically significant difference between the average accuracy of LPPT and that of classical classifiers when the significance level was 0.05, with the exception of all the average recalls of NE_NB, the average recall of "Yes" of NE_LR and the average recall, average precision, and average overall accuracy of NE_RVM. The Cohen's ds of most accuracy measures were between 0.2 and 0.7, which meant a medium difference in the classification accuracy.

5.1.2.3. No spatial association. Figure 11(d) shows the increment of the average overall accuracy, average precision, average recall, and average AUC of 1,000 classifications when the neighboring distance was 2 km. As there was no statistically significant overall spatial auto-correlation in the dataset, LPPT had an adverse effect on the classification result in most cases. The accuracy of most of the LPPT results decreased or was not significantly different to those of the classical classifiers.

The above three experiments investigated the performance of LPPT under the situation that there was strong, weak but significant, or no spatial associations amongst the classes. When there was significant spatial association, LPPT could effectively increase the classification accuracy no matter which original classifier was used. Meanwhile, the greater the spatial association, the more the prediction accuracy increased. On the contrary, the experiments on the third dataset showed that it is only beneficial to use LPPT when spatial correlation exists; otherwise, LPPT might lead to adverse effects because of the inaccurate estimation of the prior using limited samples.

5.2. *Sensitivity analysis of the neighboring distance*

The neighboring distance is an important parameter that influences the performance of LOF-based LPPT. If it is too small, then there are too few samples from neighboring objects to estimate $P_{\mathcal{N}(u)}(y_j)$. If it is too large, then the local prior distribution of different categories approaches the marginal probability distribution.

This issue is essential when classifying any spatial data. The classification of spatial objects should be conditioned by the spatial extent chosen. If the spatial extent was extended to one side, or the area was doubled, then because of natural spatial variation, the classifier parameters were expected to change slightly, or even a great deal. As a consequence, the precision of prediction or overall accuracy will probably change too. This naturally leads to the suggestion that one should consider the spatial extent carefully when classifying. And further than that, it leads to the suggestion that local classification approaches may hold benefits over global ones (i.e., restrict the spatial extent arbitrarily). There are many ways that this could be done, including geographically weighted classification type approaches (Fotheringham *et al.* 1998, Fotheringham and Brunsdon 1999, Atkinson and Naser 2010, Tang *et al.* 2016, Wu *et al.* 2019) where the model is fitted locally. Instead of fitting the model locally, the solution proposed modifies locally the prior. However, it is still greatly influenced by the spatial extent chosen.

In all the real-life experiments, the comparison between LPPT and its corresponding original classifier was performed using a neighboring distance from 1 km to 8 km. Figure 12(a) and (b) shows the tendency of the increment of the average overall accuracy of the NTD prediction using a neighboring distance from 1 km to 8 km. When the neighboring distance was 1 km, the Bonferroni corrected Student's t-test showed that the average overall accuracy and AUC using LPPT were not statistically significantly different from those using the original classifiers when the significance level was 0.05 regardless of which original classifier was used, with the exception of NE_LR. Meanwhile, the Cohen's ds of most accuracy indices were smaller than 0.2, which meant small differences between LPPT and the corresponding original classifiers. This was because that there was weak spatial auto-correlation (see Table 3) and too few neighboring objects to correctly estimate $P_{\mathcal{N}(u)}(y_i)$. When the neighboring distance was 3 km, more neighboring objects were used to estimate $P_{\mathcal{N}(u)}(y_i)$. As a consequence, the overall accuracy and AUC approached a maximum for most original classifiers.

[Figure 12 about here.]

When there was strong spatial auto-correlation, the LPPT increased the classification accuracy even if only a few neighboring objects were used to estimate $P_{\mathcal{N}(u)}(y_i)$. Figure 12(c & d) shows the tendency of the average overall accuracy of vegetation-type recognition using a neighboring distance from 1 km to 8 km for the TM image and Gaofen-2 image, respectively. When the neighboring distance was 1 km, the local prior probability was estimated using only two to three neighboring objects. However, there was strong spatial auto-correlation in the study area when the neighboring distance was 1 km, as is shown in Table 1. Accordingly, the Bonferroni corrected Student's t-test showed that the average overall accuracy from LPPT was statistically significantly larger than those from the original classifiers when the significance level was 0.05, regardless of which original classifier was used. The Cohen's ds of all the average overall accuracies were larger than 1.5 except the RVM for Gaofen-2.

Another tendency was that the prediction accuracy decreased gradually, when the neighboring distance continually increased. When the neighboring distance was too large, more global information about the category distribution was introduced to the process of

estimation $P_{\mathcal{N}(u)}(y_i)$. As a result, $P_{\mathcal{N}(u)}(y_i)$ approached the marginal probability $P(y_i)$. Therefore, $P_{\mathcal{N}(u)}(y_i|\vec{x})$ approached $P(y_i|\vec{x})$ and the prediction accuracy of the LPPT-based classifier approached that of its original classifier.

Considering the prediction of NTD occurrences as an example, when the neighboring distance was larger than 3 km, the increment of the overall accuracy and AUC decreased gradually and the prediction accuracy of LPPT approached that of its original classifier as the neighboring distance increased. The same tendency could also be seen in the vegetation recognition examples. As is shown in Figure 12(c & d), the increment of the average LPPT prediction accuracy decreased gradually and approached zero when the neighboring distance increased.

To determine the optimal neighboring distance for different types of data, users should balance between the degree of spatial auto-correlation and the number of neighboring training objects. If there was only weak spatial auto-correlation, then a large neighboring distance could be set; otherwise, a small neighboring distance was sufficient for LPPT. Meanwhile, if there were only few training data, a relatively larger neighboring distance was necessary. For example, the dashed line in Figure 9(a) shows the overall NCP of the simulated two-category data. Although there was strong spatial auto-correlation, there was only one neighboring object for most unseen objects when the neighboring distance was one unit distance. Accordingly, it was necessary to increase the neighboring distance to estimate $P_{\mathcal{N}(u)}(y_i)$ more accurately. The optimal neighboring distance was three unit distances.

LPPT can use other estimators in addition to the LOF estimator, such as the kriging method. Some of these methods use a large neighboring distance and assign different weights to different neighbors of the current object. For distant objects, they are assigned very small weights. The same strategy can be incorporated into LOF as an alternative to selecting the optimal neighboring distance.

5.3. *Comparison with other spatial data oriented classifier*

This section compares LPPT with three commonly used classifiers for spatial data: SAR, MRF and NeighFea. In the comparison, the MRF-based method and NeighFea methods were compared with the LPPT-based method using NB, kNN, RVM, or LR as the original classifier. For fairness, the experiment compares LPPT based on LR with SAR.

5.3.1. Comparison with SAR

Figure 13(a) shows the increment of the average overall accuracy of SAR for predicting Heshun NTD occurrence and recognizing vegetation types. Clearly, SAR was superior to LR when predicting Heshun NTD occurrence. To inspect the influence of the neighboring distance on the accuracy difference between NE_LR and SAR, Figure 13(b) shows the increment of the average overall accuracy, average precision, average recall, and average AUC of NE_LR compared with SAR when the neighboring distance was set to 1 km, 2 km and 8 km. When the neighboring distance was 1 km, these two methods had almost the same accuracy. The Bonferroni corrected Student's t-test showed that the average overall accuracy and average AUC of NE_LR were not statistically significantly different from those of SAR when the significance level was 0.05.

As the neighboring distance increased, the NE_LR was more accurate than LR. When the neighboring distance was 2 km, except the recall of "No" and precision of "Yes", the Bonferroni corrected Student's t-test showed that the average overall accuracy of NE_LR was statistically significantly larger than those of SAR when the significance level was

0.05. When the neighboring distance was 8 km, the Bonferroni corrected Student's t-test showed that the average accuracy of NE_LR was statistically significantly larger than that of SAR when the significance level was 0.05, with the exception of the recall for "No.". For both neighboring distances, the Cohen's ds for the significant average accuracy measures were larger than 0.2, which meant a small increase in accuracy.

[Figure 13 about here.]

In the vegetation-type recognition experiment, it was obvious that the overall accuracy of SAR was significantly smaller than that of LR. However, this did not mean that SAR was definitely inferior to NE_LR. There were seven categories in the study area and only 10% of the data was used to train the classifiers. When SAR was used, there were many additional parameters to learn. In this case, the training data might be insufficient to effectively learn all the parameters. This hindered the correct classification of unseen objects and influenced the correct evaluation of the proposed method.

From the above comparison, the results showed that NE_LR with an appropriate neighboring distance was superior to SAR. However, when the neighboring distance was small, the classification accuracy of NE_LR had no advantage over SAR. Moreover, progressively increasing the neighboring distance caused these two methodsâĂŹ classification accuracies to approach that of LR. Accordingly, it is not possible to conclude that one method is definitely superior to the other from the perspective of accuracy assessment. However, the main advantage of LPPT is that it can be used with different types of statistical classifiers. SAR may not be suitable for the classification of some real-life data, whereas some other statistical classifier may be appropriate. For example, nominal variables may be used to describe spatial data, or the distribution of the attribute value may obey some specific distribution instead of the Gaussian distribution. In these scenarios, some other statistical classifiers may be more suitable than the SAR, or the SAR may not be applicable at all. In these circumstances, LPPT is an effective method to account for spatial patterns in the target objects for the classification of spatial data and is superior to SAR.

5.3.2. Comparison with MRF-based methods

With respect to the MRF-based methods, Figures 14 (a-c) show the increment of the average overall accuracies of the LPPT-based models (NE_kNN, NE_NB, NE_RVM, and NE_LR) compared with those of the MRF models (MRF_kNN, MRF_NB, MRF_RVM, and MRF_LR) when the neighboring distance was set to 1 km to 8 km, respectively. Clearly, the MRF-based model was inferior to the LPPT-based model in most cases. The Bonferroni corrected Student's t-test showed that the average overall accuracy of results from the LPPT-based models was statistically significantly larger than that of the results from the MRF-based models when the significance level was 0.05, except when RVM was used as the original classifier to predict Heshun NTD occurrences. For the Heshun NTD data, the Cohen's d for the significant average accuracy measures was larger than 0.2, which meant a small increase in accuracy. For the vegetation recognition, the Cohen's d for the average overall accuracy was larger than 0.9, which meant a large increase in accuracy.

[Figure 14 about here.]

In the vegetation-type recognition experiment, the MRF-based models were even inferior to the classical models. The MRF models performed poorly because the MRF model used in the experiment (Solberg *et al.* 1996) took the relation between only two objects

into account. It neglected the complex relation between multiple objects. However, it is difficult to construct appropriate cliques for irregular lattice data. Even if there is a perfect solution to construct appropriate cliques, MRF-based models need much more computational resources than LPPT-based models. Taking vegetation-type recognition from remotely sensed imagery as an example, when the neighboring distance was 4 km, in addition to the time required to train traditional statistical classifiers, MRF-based methods required 42 seconds, whereas LPPT-based methods required two seconds.

5.3.3. Comparison with NeighFea approach

Finally LPPT was compared with the NeighFea approach. NeighFea included neighborhood information, that is, neighboring objects' feature values, in the vector of each geographical object during the training and classification steps. This method was easy to implement and could increase the classification accuracy. However, NeighFea used many more features than the classical classifiers. It required more training data to estimate the classifier's parameters. When there were insufficient training data, NeighFea performed poorly.

Figure 15 shows the average overall accuracy of NB, NE_NB, and NeighFea_NB using LPPT and NeighFea for the simulated data, which had two to six categories or one to six features. When the number of categories was two and the number of features was one, the average overall accuracy of NeighFea_NB was larger than those of NB and NE_NB. Clearly, the NeighFea approach was superior to the classical classifier and LPPT method in terms of classification accuracy. However, when the number of categories increased, NeighFea was inferior to LPPT, because more categories introduced more parameters and the training data size was unchanged. Similarly, increasing the number of attributes increased the number of parameters to be estimated with the same training data, which led to inaccurate estimated parameters. Accordingly, increasing of the number of attributes decreased the average overall accuracy of NeighFea.

[Figure 15 about here.]

In the experiment on real-life data, NeighFea was inferior to LPPT because many features were used to describe geographical objects. For example, Figure 16(a) and (b) show the increment of the average overall accuracy and average AUC of NeighFea when NB, kNN, RVM and LR were used for the prediction of NTD occurrence. Clearly, the accuracy of NeighFea was less than that of the classical classifier, with the exception of the average AUC of RVM. For the recognition of vegetation types, NeighFea was also inferior to the classical classifiers. Figure 16(c) and (d) shows the increment of the average overall accuracy of NeighFea when NB, kNN, RVM and LR were used for the recognition of vegetation types. The Bonferroni corrected Student's t-test showed that there were statistically significant differences between NeighFea and the classical classifiers when the confidence level was 0.05 except for the RVM of Gaofen-2. All the Cohen's ds of all statistically significant accuracy measures were smaller than zero, which meant a decreases in classification accuracy.

[Figure 16 about here.]

NeighFea had the potential to increase classification accuracy for spatial data. It introduced many neighboring objects' features to help to improve the classification of unseen objects. However, there was a limited number of instances in the training data compared with the number of parameters to be estimated in some real-life applications. Accordingly, it might perform poorly in such scenarios. Different from NeighFea, the LOF estimator

in LPPT was independent of the training of the original classifiers. It did not introduce new parameters in the training of classical classifiers. Therefore, it is more effective in processing data with many features.

To summarize, LPPT performed more accurately than the three benchmark methods in the experiments when there is statistically significant spatial association and local spatial patterns were modelled appropriately in some situations. In particular, LPPT can be adapted readily to real-life applications in which SAR is inappropriate through selecting an appropriate original classifier. Compared with MRF, LPPT could easily estimate local prior distributions for irregular lattices and spend much less computing resources. With respect to the NeighFea, LPPT suffered less from the curse of dimensionality.

5.4. *Influence of sampling methods*

To test whether LPPT was only effective when the simple random sampling method is used, the clustering sampling method (Wang *et al.* 2012) was used to perform comparisons between LPPT and the classical statistical classifiers.

The clustering sampling method first divided the population into separate groups. Then, a group was randomly selected. In the selected group, a random sample was then selected from the group. This process iterated until sufficient samples were obtained. In our experiment, the datasets were divided into 10 groups using k-means (Murphy 2012). For simplicity, the comparison was performed on the vegetation recognition dataset.

[Figure 17 about here.]

Figure 17 shows the increment of the average overall accuracy for the vegetation type recognition using different sampling methods when the neighboring distance is 2 km. When the NB or kNN were used as the original classifier, the average overall accuracy corresponding to that of the clustering sampling method increased less than that using simple random sampling. Nevertheless, when the clustering sampling method is used, the average overall accuracy also increased at least 4%. This showed that LPPT was not limited to the simple random sampling method.

The key to increasing classification accuracy using LPPT is not which sampling method is used, but the correct estimation of the prior distribution of different classes for each unseen object. According to Equation 6, if there are no samples in the neighborhood of unseen objects, both $|\mathcal{N}(u)|$ and $|\mathcal{N}_{y_i}(u)|$ were zero and $P_{\mathcal{N}(u)}(y_i) = P(y_i)$, then LPPT degrades to its original classifier. However, if a correct prior distribution is provided to update the prior for each unseen object, LPPT can help increase classification accuracy even when there are no samples in the neighborhood of unseen objects.

To validate this, the vegetation recognition of TM, the vegetation recognition of Gaofen-2, and Heshun NTD dataset were divided into left and right parts. The left part was used as training data and the right part as validation data. The LR was used as the original classifier. When classifying the right part for each dataset, no objects from the sample were used in the classification.

Correct and incorrect prior distributions of classes were assigned to each unseen object during classification, respectively. In the experiment, the prior distribution was obtained using the neighboring objects' real labels. In real applications, the correct prior distribution can be obtained from experts or inferred from some secondary data. The incorrect prior distribution is set through subtracting the correct prior probability from 1. Results shows that almost all the classification accuracy measures increased by at least 5% when the correct prior probability was used to update the local prior for each object. However,

when an incorrect prior probability is used, the overall accuracy of LPPT decreased by at least 4%.

Therefore, when the prior distribution of different classes can be correctly estimated, LPPT is preferred and superior to classical classifiers. However, when no information is available to estimate the prior information, it is recommended to use classical classifiers because LPPT generates the same result as the original result. In future, we plan to explore methods for estimating the prior distribution under different scenarios.

5.5. *Incorporating spatial pattern information*

In SAR, the categories of neighboring objects $\vec{y}_{\mathcal{N}(u)}$ in the training data are modelled as an independent variable. If the training data have strong spatial associations, the weights of $\vec{y}_{\mathcal{N}(u)}$ will be much larger than the weights for other attributes. When such a model is used in prediction or classification, the result will depend greatly on the neighbor and neglect other attributes. In MRF, the weights of the spatial pattern information are assigned arbitrarily in the potential function. It is hard for the user to predetermine this weight to make the best use of the spatial information and attribute information simultaneously.

There exist additional methods for incorporating spatial pattern information into the classification process in the framework of statistics besides SAR and MRF. These methods all need additional assumptions to fuse spatial pattern information into the classification besides the assumptions of the statistical classifier and the correct modelling of spatial pattern information. A commonly used strategy is giving appropriate weight to spatial pattern information (Atkinson and Naser 2010, Ge and Bai 2011, Tang *et al.* 2016). However, like MRF, these methods need to pre-define the weight of spatial pattern information. No matter which type (a constant or a function) of weight is used, the user must define it at the risk of personal induction bias. Bayesian updating (Journel 2002) is another way of considering both spatial pattern information and feature space information. It needs the assumption that the contribution of the spatial pattern information is the same before and after the feature space information is used. However, the suitability of this assumption needs to be reconsidered for different datasets.

Post classification processing is another commonly used method to increase classification accuracy. A simple post classification method is moving a kernel across each object and using the mode to smooth the classification result. An experiment was performed to compare this post classification method and LPPT using the Gaofen-2 image. The NB classifier was used as the original classifier and the neighboring distance was set to 2 km. When the post classification method was used, although the average overall accuracy increased by 1%, the average precision and recall of the category "a" with strong spatial association increased at the cost of decreasing of the accuracy of all the other categories. The average recall of "a" increased more than 30%. However, for the categories, "d" and "e", which had weak spatial association, the precision and recall decreased more then 30%.

When this method was used for the detection of the Heshun NTD occurrence, which had weak spatial association, all accuracy measures decreased. For the vegetation recognition using the TM image, this simple post-classification strategy was insufficient because many different types of vegetation in the image mixed together and only two categories among all seven categories had large spatial associations. As a result, the average overall classification accuracy decreased. Accordingly, this post classification method is only suitable for categories with strong spatial association. To further increase the classification accuracy, a complex rule set for smoothing the classification should be constructed through

considering the pattern for different categories. LPPT simplified this complex rule set construction process to the updating of the prior distribution of different categories, and needed no post classification processing.

Compared with the above methods, LPPT updates the prior probability distribution of each category according to the spatial pattern information for each unseen object. If a generative model is used as the original classifier, LPPT does not introduce any more assumptions besides those of the statistical classifier and the correct modelling of spatial pattern information. For example, incorporating indicator kriging into maximum likelihood classification through prior updating is effective for classifying hyperspectral data (Goovaerts 2002). If a discriminant model is used, as traditional statistical classifiers do not use spatial pattern information, it is appropriate to assume that $P(y_i)$ is equal to its marginal probability. Meanwhile, this assumption is easy to check for a statistical classifier through synthetic experiments by changing the proportion of objects of different categories. This assumption is related only to the original classifier used and has nothing to do with the data at hand and user induction bias. Accordingly, when using LPPT, users need not set appropriate weights for spatial pattern information or consider the contribution of the spatial pattern information to be the same before and after the feature space information is used. Rather, users can focus on the effective and efficient modelling of spatial pattern information.

Although LPPT provides a new way to incorporate spatial pattern information, as with other spatial classifiers, users should also consider the possibility of over-generalization of the spatial pattern information in practical use. The spatial pattern information is learned inevitably using some models or assumptions. For example, the LOF estimator-based LPPT uses the first law of geography and sample objects surrounding the unseen object to infer the prior probability of different categories. If the spatial pattern prior of a category is very large, it may suppress the feature space information.

6. Conclusion

This paper proposed a novel classification model for spatial data based on traditional statistical classifiers. In this model, a traditional statistical classifier is first trained. Then, each unseen geographical object's local prior class distribution is estimated and used to replace the marginal class distribution in the prediction model of the traditional statistical classifier. The experiment results show that the classifier using the proposed method outperformed its corresponding original classifier. Meanwhile, the proposed model's classification accuracy was larger than those of logistic spatial autoregression and MRF based models when there was statistically significant spatial association and $P_{\mathcal{N}(u)}(y_i)$ was correctly estimated in the experiments.

Because LPPT can be coupled readily with traditional statistical classifiers, in addition to its effectiveness, the main advantage of LPPT is its wide applicability. Even a discriminant model can be coupled with LPPT under the assumption that the prior distribution is the marginal classes' distribution. With the help of LPPT, a traditional statistical classifier can take the spatial distribution pattern of classes into account. Accordingly, the use of LPPT can increase classification accuracy for spatial data. The experimental results show that LPPT does increase classification accuracy, even for discriminant statistical classifiers. Therefore, as well as effectively modelling the feature space using a traditional statistical classifier and properly estimating the local prior class probability, LPPT can be applied to generate classification results with higher accuracy than that of its original

classifier.

The current LPPT model may be regarded as a generic modelling approach. It may need further reinforcement to adapt it to various kinds of real-life spatial data. In particular, future research is required to extend this model in the following five ways:

(1). Applying LPPT to additional types of spatial data. This paper shows three real-life applications for the proposed classifier. There exist many different types of spatial data. These spatial data may have different forms, such as points, lines, polygons, or networks. Some new methods, for example geostatistics and multiple point simulation, can be used to estimate $P_{\mathcal{N}(u)}(y_i)$ to help adapt LPPT for different spatial data.

(2). Determining how to estimate $P_{\mathcal{N}(u)}(y_i)$ for objects with no or very few surrounding sample data. For example, sample data and unseen objects may not be in the same study area, or there are too few samples in the study area. The proposed method is not efficient in both cases. However, if some unseen objects can be pre-classified using some methods, there will be more neighboring objects to estimate $P_{\mathcal{N}(u)}(y_i)$.

(3). Determining how to consider multiple spatial scale structure information. Geographical phenomena generally have spatial characteristics at different scales. Under the framework of LPPT, the classification accuracy can be increased by making the best use of multiple scale structure information, as long as the estimation of $P_{\mathcal{N}(u)}(y_i)$ takes it into account.

(4). Generating finer spatial resolution classifications. This method has the potential to be used in sub-pixel mapping. Finer spatial resolution spatial information could be acquired in many different ways (Tatem *et al.* 2001, Ge *et al.* 2009, Zhong and Zhang 2012, Wang *et al.* 2014, Li *et al.* 2017). This information can then be coupled with soft classification results from statistical classifiers using LPPT. The LPPT can then adjust the category at the sub-pixel level in terms of both spatial information and spectral information.

(5). Further exploring LPPT using Bayes decision theory. In addition to the experiments, effort should also be made to explain why LPPT can classify spatial data more accurately than the original classifiers, for example, whether there are any properties of LPPT that demonstrate its advantage over classical statistical classifiers when used on spatial data. Bayesian decision theory is a powerful candidate tool for further exploring this question.

7. Data and codes availability statement

The data and codes that support the findings of this study are available in "figshare.com" with the identifier ":doi:10.6084/m9.figshare.11876490.v1". The Heshun NTD dataset and the poverty-stricken village dataset of Yunyang is owned by two local governments. Their websites are http://www.chinacdc.cn and http://www.shiyan.gov.cn, respectively. We are not authorized to publish these datasets. Accordingly, these two datasets are not included and are replaced with mock data. Meanwhile, The Gaofen-2 remotely sensed image and its label are provided by Tong *et al.* (2020) and are available at http://captain.whu.edu.cn/GID/.

Acknowledgments

The work is supported by the Strategic Priority Research Program of the Chinese Academy of Science under Grant XDA19040501, the National Natural Science Foundation for Distinguished Young Scholars of China under Grant 41725006, the National Natural Science Foundation of China (Nos. 41871286, 61672331) and the Natural Science Foundation of Shanxi Province, China (No. 201701D121055).

References

Anselin, L., 1988. *Spatial econometrics: methods and models*. Studies in operational regional science Dordrecht, Netherlands: Kluwer Academic Publishers.

Anselin, L., 1995. Local indicators of spatial association–LISA. *Geographical Analysis*, 27 (2), 93–115.

Atkinson, P.M. and Naser, D.K., 2010. A Geostatistically Weighted k-NN Classifier for Remotely Sensed Imagery. *Geographical Analysis*, 42 (2), 204–225.

Austad, H.M. and Tjelmeland, H., 2017. Approximate computations for binary Markov random fields and their use in Bayesian models. *Statistics and Computing*, 27 (5), 1271–1292.

Bai, H., *et al.*, 2010. Using rough set theory to identify villages affected by birth defects: the example of Heshun, Shanxi, China.. *International Journal of Geographical Information Science*, 24 (4), 559–576.

Bai, H., *et al.*, 2016. Detecting nominal variables' spatial associations using conditional probabilities of neighboring surface objects' categories. *Information Sciences*, 329 (Supplement C), 701 – 718 Special issue on Discovery Science.

Bicheron, P., *et al.*, 2008. *GlobCover: products description and validation report*. Technical report.

Bishop, C.M., 2006. *Pattern Recognition and Machine Learning*. Information Science and Statistics Springer.

Boots, B., 2003. Developing local measures of spatial association for categorical data. *Journal of Geographical Systems*, 5 (2), 139–160.

Chen, X., *et al.*, 2019. A Novel Method for Inverse Uncertainty Propagation. *In*: E. Minisci, M. Vasile, J. Periaux, N.R. Gauger, K.C. Giannakoglou and D. Quagliarella, eds. *Advances in Evolutionary and Deterministic Methods for Design, Optimization and Control in Engineering and Sciences*. Cham: Springer International Publishing, 353–370.

Chun, Y. and Griffith, D., 2013. *Spatial Statistics and Geostatistics: Theory and Applications for Geographic Information Science and Technology*. Thousand Oaks: Sage.

Cohen, J., 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20 (1), 37–46.

Ding, W., Stepinski, T., and Salazar, J., 2009. Discovery of Geospatial Discriminating Patterns from Remote Sensing Datasets. Proceedings, *In*: C. Apte, H. Park, K. Wang and M.J. Zaki, eds. *Proceedings of the 2009 SIAM International Conference on Data Mining*. DOI: 10.1137/1.9781611972795.37 Society for Industrial and Applied Mathematics, 425–436.

Duda, R.O., Hart, P.E., and Stork, D.G., 2001. *Pattern Classification*. 2nd New York, NY, USA: Wiley & Sons, Inc.

Evans, S.W., 2017. An assessment of land cover change as a source of information for con-

servation planning in the Vhembe Biosphere Reserve. *Applied Geography*, 82 (Supplement C), 35–47.

Fawcett, T., 2006. An introduction to ROC analysis. *Pattern Recognition Letters*, 27 (8), 861–874.

Fortin, M. and Dale, M., 2005. *Spatial Analysis: A Guide for Ecologists*. Cambridge, United Kingdom: Cambridge University Press.

Fotheringham, A.S., Charlton, M.E., and Brunsdon, C., 1998. Geographically Weighted Regression: A Natural Evolution of the Expansion Method for Spatial Data Analysis. *Environment and Planning A: Economy and Space*, 30 (11), 1905–1927.

Fotheringham, A.S. and Brunsdon, C., 1999. Local Forms of Spatial Analysis. *Geographical Analysis*, 31 (4), 340–358.

Gartland, L., 2010. *Heat Islands: Understanding and Mitigating Heat in Urban Areas*. Abingdon, UK and New York, USA: Routledge.

Ge, Y., Li, S., and Lakhan, V.C., 2009. Development and Testing of a Subpixel Mapping Algorithm. *IEEE Transactions on Geoscience and Remote Sensing*, 47 (7), 2155–2164.

Ge, Y. and Bai, H., 2010. MPS-based information extraction method for remotely sensed imagery: a comparison of fusion methods. *Canadian Journal of Remote Sensing*, 36, 763–779.

Ge, Y. and Bai, H., 2011. Multiple-point simulation-based method for extraction of objects with spatial structure from remotely sensed imagery. *International Journal of Remote Sensing*, 32, 2311–2335.

Goovaerts, P., 2002. Geostatistical incorporation of spatial coordinates into supervised classification of hyperspectral data. *Journal of Geographical Systems*, 4 (1), 99–111.

Goovaerts, P., 1997. *Geostatistics for natural resources evaluation*. Oxford, United Kingdom: Oxford University Press.

Griffith, D.A., 1989. Spatial Econometrics: Methods and Models. *Economic Geography*, 65 (2), 160–162.

Haining, R.P., 1990. *Spatial data analysis in the social and environmental sciences*. Cambridge, United Kingdom: Cambridge University Press.

Hsu, W.H., 2003. Data Mining. *In*: J. Wang, ed. . Hershey, PA, USA: IGI Global, chap. Control of Inductive Bias in Supervised Learning Using Evolutionary Computation: A Wrapper-based Approach, 27–54.

Hubei provincial government, 2018. Compilation of Precise Poverty Alleviation Policies in Hubei Province. [online] [2019.06.18].

Hughes, J., Haran, M., and Caragea, P., 2011. Autologistic models for binary data on a lattice. *Environmetrics*, 22 (7), 857–871.

Jeon, B. and Landgrebe, D.A., 1992. Classification with spatio-temporal interpixel class dependency contexts. *IEEE Transactions on Geoscience and Remote Sensing*, 30 (4), 663–672.

Jiang, Z. and Shekhar, S., 2017. *Spatial Big Data Science Classification Techniques for Earth Observation Imagery*. Cham, Switzerland: Springer International Publishing AG.

Joshi, N., *et al.*, 2016. A Review of the Application of Optical and Radar Remote Sensing Data Fusion to Land Use Mapping and Monitoring. *Remote Sensing*, 8 (1), 70.

Journel, A.G., 2002. Combining Knowledge from Diverse Sources: An Alternative to Traditional Data Independence Hypotheses. *Mathematical Geology*, 34 (5), 573–596.

Khare, D., *et al.*, 2015. Impact of landuse/land cover change on run-off in a catchment of Narmada river in India. *Applied Geomatics*, 7 (1), 23–35.

Li, X., *et al.*, 2017. Generating a series of fine spatial and temporal resolution land cover maps by fusing coarse spatial resolution remotely sensed images and fine spatial resolution land cover maps. *Remote Sensing of Environment*, 196, 293–311.

Liao, Y., *et al.*, 2009a. Risk assessment of human neural tube defects using a Bayesian belief network. *Stochastic Environmental Research and Risk Assessment*, 24 (1), 93–100.

Liao, Y., *et al.*, 2009b. Identifying environmental risk factors for human neural tube defects before and after folic acid supplementation. *BMC Public Health*, 9, 391.

Liu, J., *et al.*, 2017a. Autologistic models for benchmark risk or vulnerability assessment of urban terrorism outcomes. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*.

Liu, W., Fowler, J.E., and Zhao, C., 2017b. Spatial Logistic Regression for Support-Vector Classification of Hyperspectral Imagery. *IEEE Geoscience and Remote Sensing Letters*, 14 (3), 439–443.

Murphy, K.P., 2012. *Machine Learning: A Probabilistic Perspective*. Cambridge, MA, USA: The MIT Press.

Ni, L., *et al.*, 2014. Edge-constrained Markov random field classification by integrating hyperspectral image with LiDAR data over urban areas. *Journal of Applied Remote Sensing*, 8 (1), 085089.

Paciorek, C.J., 2007. Computational techniques for spatial logistic regression with large data sets. *Computational Statistics & Data Analysis*, 51 (8), 3631 – 3653.

Pedregosa, F., *et al.*, 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

Powers, D.M.W., 2011. Evaluation: From Precision, Recall and F-Measure To Roc, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies*, 2 (1), 37–63.

Rajabi, M.M., 2019. Review and comparison of two meta-model-based uncertainty propagation analysis methods in groundwater applications: polynomial chaos expansion and Gaussian process emulation. *Stochastic Environmental Research and Risk Assessment*, 33 (2), 607–631.

Ritchie, J., 2017. Relevance Vector Machine implementation using the scikit-learn API. [online] [Online; accessed 15-Novenber-2017] [2019.01.07].

Shekhar, S., *et al.*, 2002. Spatial contextual classification and prediction models for mining geospatial data. *IEEE Transactions on Multimedia*, 4 (2), 174–188.

Sherman, M., Apanasovich, T.V., and Carroll, R.J., 2006. On estimation in binary autologistic spatial models. *Journal of Statistical Computation and Simulation*, 76 (2), 167–179.

Smith, J.W.F. and Campbell, I.A., 1989. Error in polygon overlay processing of geomorphic data. *Earth Surface Processes and Landforms*, 14, 703–717.

Solberg, A.H.S., Taxt, T., and Jain, A.K., 1996. A Markov random field model for classification of multisource satellite imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 34 (1), 100–113.

Strebelle, S., 2000. Sequential Simulation Drawing Structures from Training Images. Thesis (PhD). Stanford University, Stanford, CA, USA.

Tang, Y., *et al.*, 2016. A multiple-point spatially weighted k-NN classifier for remote sensing. *International Journal of Remote Sensing*, 37 (18), 4441–4459.

Tatem, A.J., *et al.*, 2001. Super-resolution target identification from remotely sensed images using a Hopfield neural network. *IEEE Transactions on Geoscience and Remote Sensing*, 39 (4), 781–796.

Theodoridis, S. and Koutroumbas, K., 2009. *Pattern Recognition (Fourth Edition)*. Fourth Edition Boston: Academic Press.

Tobler, W., 1970. A computer movie simulating urban growth in the Detroit Region. *Economic Geology*, 46 (2), 234–240.

Tong, X.Y., *et al.*, 2020. Land-cover classification with high-resolution remote sensing images using transferable deep models. *Remote Sensing of Environment*, 237, 111322.

Vainer, I., *et al.*, 2009. Scalable Classification in Large Scale Spatiotemporal Domains Applied to Voltage-Sensitive Dye Imaging. *In: 2009 Ninth IEEE International Conference on Data Mining*, Dec.., 543–551.

Wang, J., *et al.*, 2002. Spatial sampling design for monitoring the area of cultivated land. *International Journal of Remote Sensing*, 23 (2), 263–284.

Wang, J.F., *et al.*, 2012. A review of spatial sampling. *Spatial Statistics*, 2, 1 – 14.

Wang, J., *et al.*, 2010. Assessing local determinants of neural tube defects in the Heshun Region, Shanxi Province, China. *BMC Public Health*, 10, 52.

Wang, Q., Shi, W., and Wang, L., 2014. Allocating Classes for Soft-Then-Hard Subpixel Mapping Algorithms in Units of Class. *IEEE Transactions on Geoscience and Remote Sensing*, 52 (5), 2940–2959.

White, P., Gelfand, A., and Utlaut, T., 2017. Prediction and model comparison for areal unit data. *Spatial Statistics*, 22 (Part 1), 89–106.

Wu, C., *et al.*, 2019. Multiscale geographically and temporally weighted regression: exploring the spatiotemporal determinants of housing prices. *International Journal of Geographical Information Science*, 33 (3), 489–511.

Wu, J., *et al.*, 2004. Exploratory spatial data analysis for the identification of risk factors to birth defects. *BMC Public Health*, 4, 23.

Xia, J., *et al.*, 2015. Spectral-Spatial Classification for Hyperspectral Data Using Rotation Forests With Local Feature Extraction and Markov Random Fields. *IEEE Transactions on Geoscience and Remote Sensing*, 53 (5), 2532–2546.

Yu, H., *et al.*, 2016. Spectral-Spatial Hyperspectral Image Classification Using Subspace-Based Support Vector Machines and Adaptive Markov Random Fields. *Remote Sensing*, 8 (4), 355.

Zhang, Y. and Prasad, S., 2016. Multisource Geospatial Data Fusion via Local Joint Sparse Representation. *IEEE Transactions on Geoscience and Remote Sensing*, 54 (6), 3265–3276.

Zhong, Y. and Zhang, L., 2012. Remote Sensing Image Subpixel Mapping Based on Adaptive Differential Evolution. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 42 (5), 1306–1329.

Table 1. Overall NCP and NCP of different vegetation types in the TM image using different neighboring distances

Neighboring Distance	Overall	"1"	"2"	"3"	"4"	"5"	"6"	"7"
1km	0.25	0.32	0.09	0.18	0.10	0.18	0.11	0.53
2km	0.16	0.19	0.04	0.11	0.05	0.13	0.15	0.39
3km	0.11	0.11	0.01	0.17	0.04	0.09	0.04	0.30
4km	0.08	0.05	0.00[a]	0.04	0.03	0.07	0.03	0.23

[a] failed to pass the permutation test.

Table 2. Overall NCP and NCP of different vegetation types in the Gaofen-2 image using different neighboring distances

Neighboring Distance	Overall	"a"	"b"	"c"	"d"	"e"
1km	0.20	0.31	0.18	0.23	0.03	0.09
2km	0.12	0.16	0.11	0.15	0.01	0.05
3km	0.07	0.09	0.06	0.10	0.01	0.03
4km	0.05	0.05	0.04	0.07	0.00[a]	0.02

[a] failed to pass the permutation test.

Table 3. Overall NCP and NCP of different categories of NTD occurrences in Heshun, Shanxi, China, using different neighboring distances.

Neighboring Distance	Overall	"Yes"	"No"
1km	0.12	0.10	0.13
2km	0.12	0.12	0.13
3km	0.11	0.11	0.11
4km	0.09	0.09	0.09

Table 4. Overall NCP and NCP of different categories of poverty-stricken villages in Yunyang, Hubei, China, using different neighboring distances

Neighboring Distance	Overall	"Yes"	"No"
1km	0.08[a]	0.10	0.02[a]
2km	0.06[a]	0.09	0.04[a]
3km	0.06[a]	0.09	0.02[a]
4km	0.04[a]	0.07	0.02[a]

[a] failed to pass the permutation test.

List of Figures

represents the precision of "Yes," RYes represents the recall of "Yes," PNo represents the precision of "No," and RNo represents the recall of "No."

Figure 12. Tendency of the increment of the average accuracy of NTD prediction and vegetation type recognition using a neighboring distance from 1 km to 8 km when the original classifiers were NB, kNN, RVM, and LR, respectively: (a) the average overall accuracy increment tendency for NTD prediction; (b) the average AUC increment tendency for NTD prediction; (c) the average overall accuracy increment tendency for vegetation type recognition (TM); (d) the average overall accuracy increment tendency for vegetation type recognition (Gaofen-2).

Figure 13. Comparison between LPPT and SAR: (a) increment of the average overall accuracy of SAR for predicting Heshun NTD occurrence and recognizing vegetation types when the neighboring distance was from 1 km to 8 km; (b) increment of the average overall accuracy, average precision, average recall and average AUC of NE_LR compared with SAR for predicting Heshun NTD occurrence when the neighboring distance was set to 1 km, 2 km and 8 km. PYes represents the precision of "Yes," RYes represents the recall of "Yes," PNo represents the precision of "No," and RNo represents the recall of "No."

Figure 14. Increment of the average overall accuracy of the LPPT models (NE_kNN, NE_NB, NE_RVM, and NE_LR) compared with those of the MRF models (MRF_kNN, MRF_NB, MRF_RVM, and MRF_LR) for (a) predicting NTD instances, (b) vegetation recognition (TM) and (c) vegetation recognition (Gaofen-2) when the neighboring distance was set to 1 to 8 km, respectively.

Figure 15. Average overall accuracy of NB, NE_NB, and NeighFea_NB using LPPT and NeighFea for the simulated data, which had two to six categories or one to six features when the neighboring distance was two unit distances.

Figure 16. Increment of the average accuracy of NeighFea when NB, kNN, RVM, and LR were used: (a) average overall accuracy for the prediction of NTD occurrence; (b) average AUC for the prediction of NTD occurrence; (c) average overall accuracy for the vegetation type recognition (TM); (d) average overall accuracy for the vegetation type recognition (Gaofen-2)

Figure 17. The increment of the average overall accuracy for the vegetation type recognition using different sampling methods when the neighboring distance is 2km.

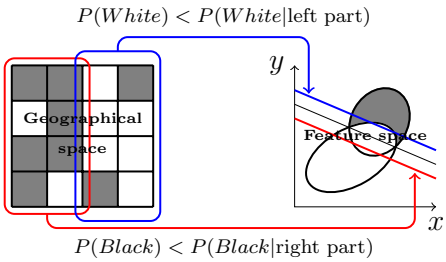


Figure 1. Spatial distribution and statistical classifier. $P(Black)$ ($P(White)$) is the marginal probability of 'Black'('White'). $P(Black|\text{right part})$ ($P(White|\text{left part})$) is the probabilities of observing 'Black'('White') in the right part (left part).

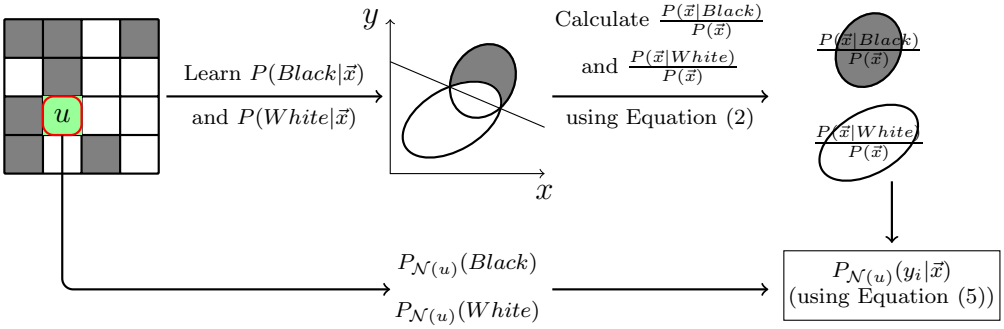
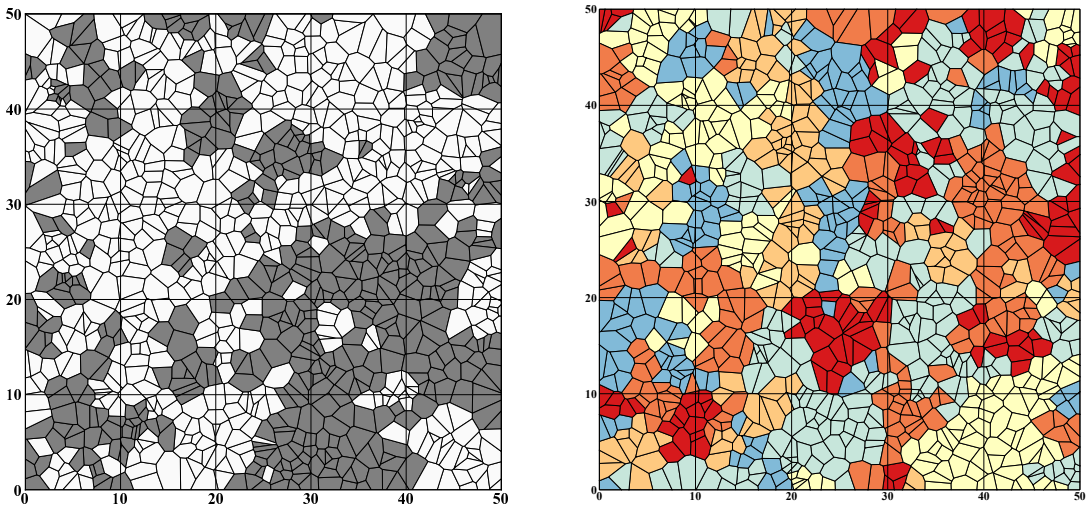


Figure 2. Local pattern-based prior tuning statistical classifier.



(a) (b)

Figure 3. Simulated datasets: (a) two categories; (b) six categories.

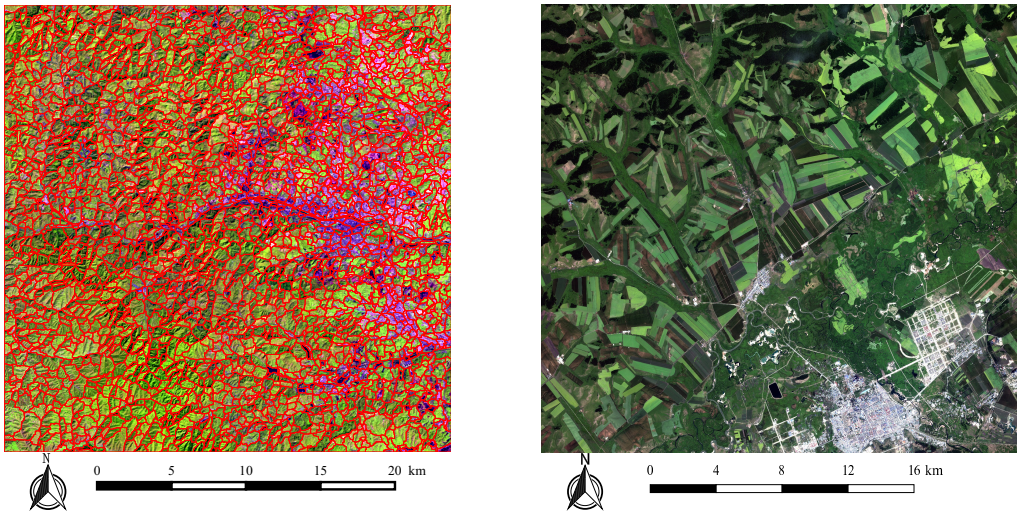
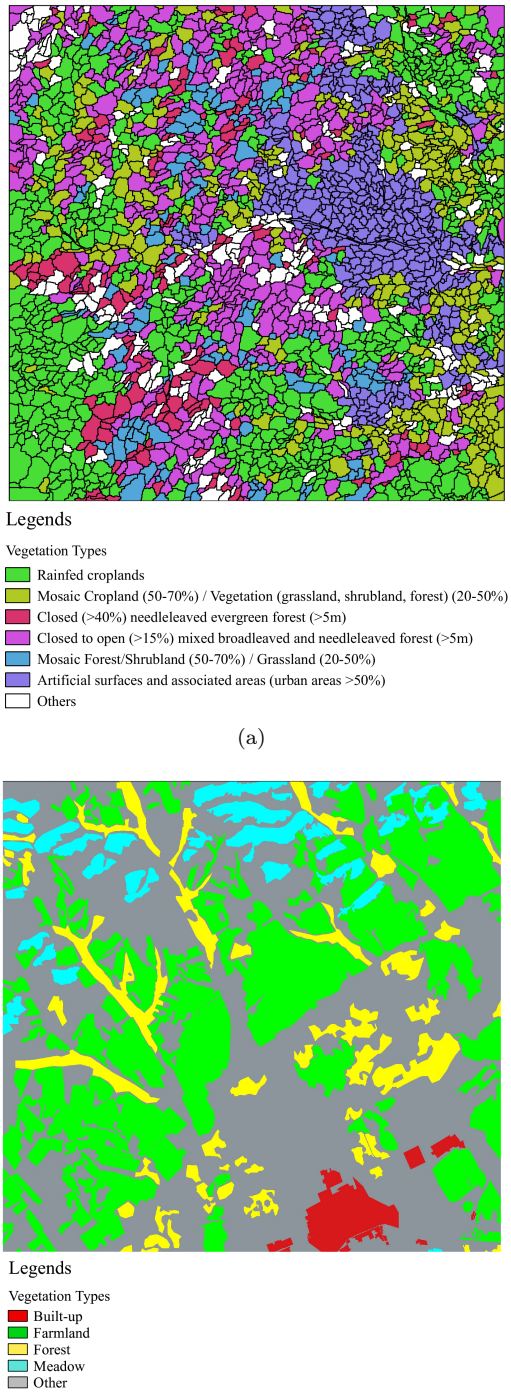


Figure 4. The segmentation and pseudo-color composites of the TM and Gaofen-2 image: (a) the segmentation and pseudo-color composite of the TM image; (b) the true color composite image of the Gaofen-2 image.



(a)

(b)

Figure 5. Vegetation type maps of Figure 4: (a) vegetation type map of the TM image; (b) vegetation type map of the Gaofen-2 image.

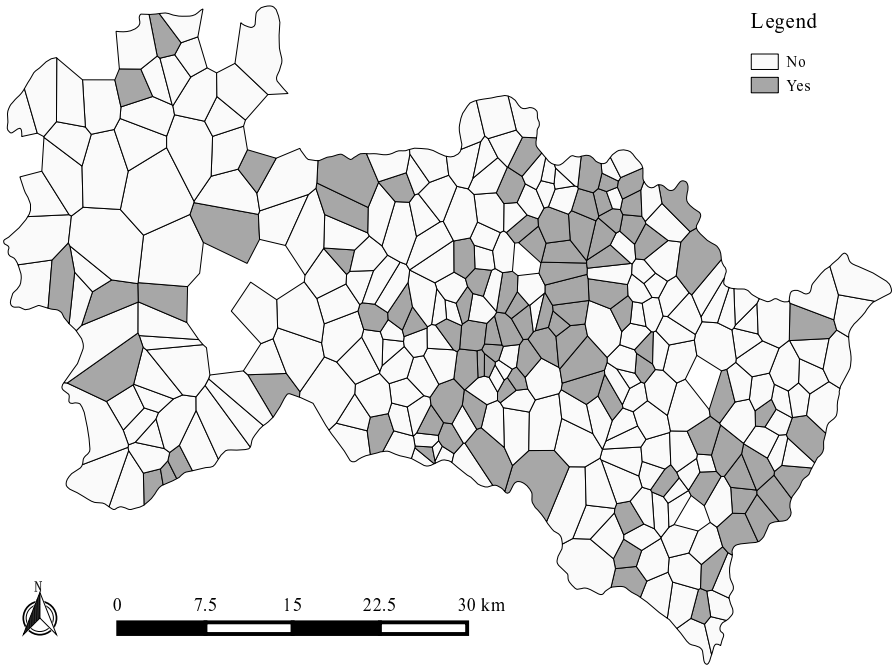


Figure 6. Map of NTD occurrences in Heshun, Shanxi, China.

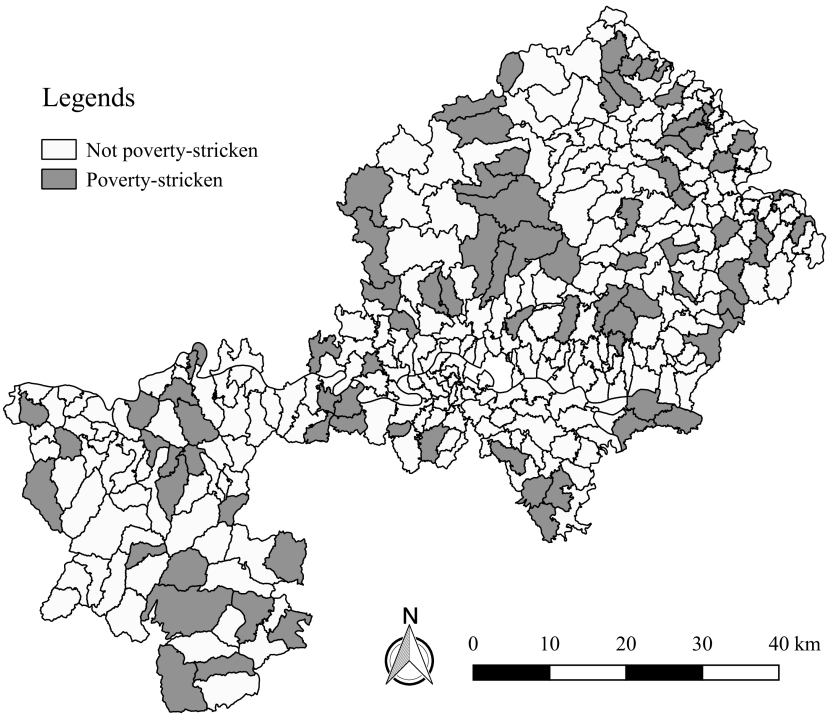


Figure 7. Map of poverty-stricken villages in Yunyang, Hubei, China.

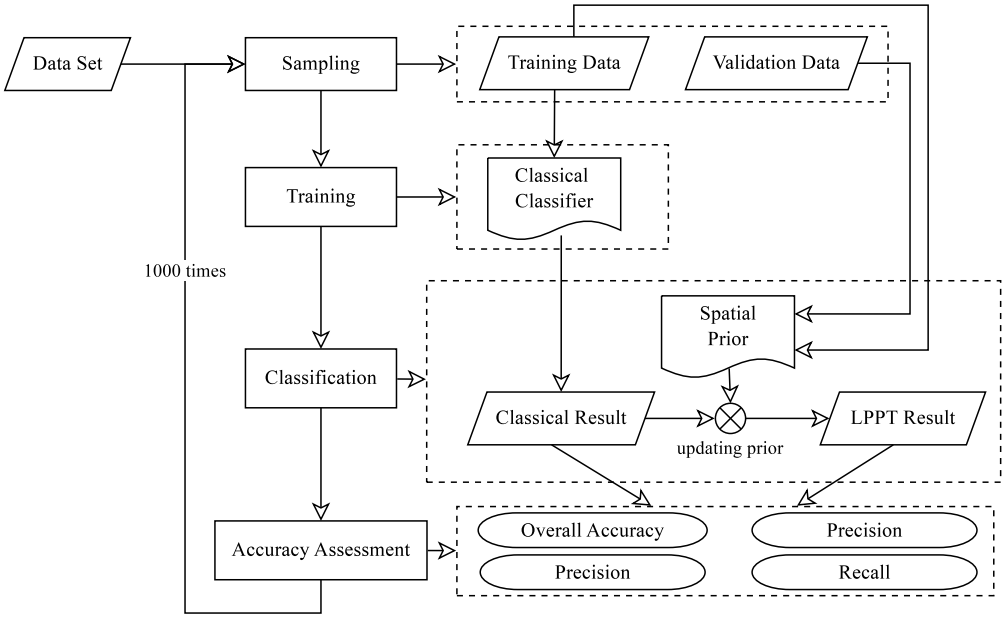


Figure 8. Classifier comparison process for all experiments.

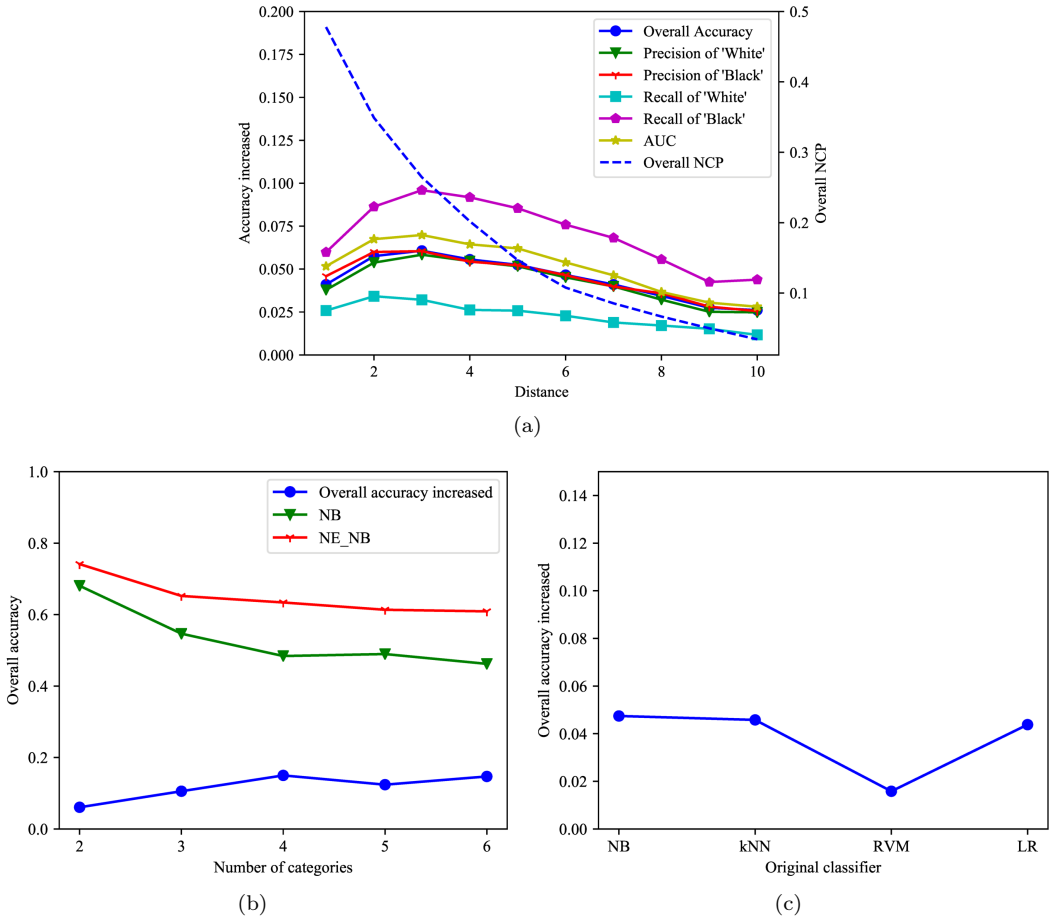


Figure 9. Average accuracy increase in the experiments on simulated datasets: (a) average accuracy increase using the neighboring distances from one to 10. The dashed line in the top right subchart is the overall spatial association calculated; (b) average overall accuracy of NB and NE_NB, and the accuracy increment of LPPT when the numbers of categories are two to six, respectively; (c) average overall accuracy increase using LPPT for 1,000 datasets with random numbers of features and random numbers of categories when the original classifiers were NB, kNN, RVM and LR, respectively.

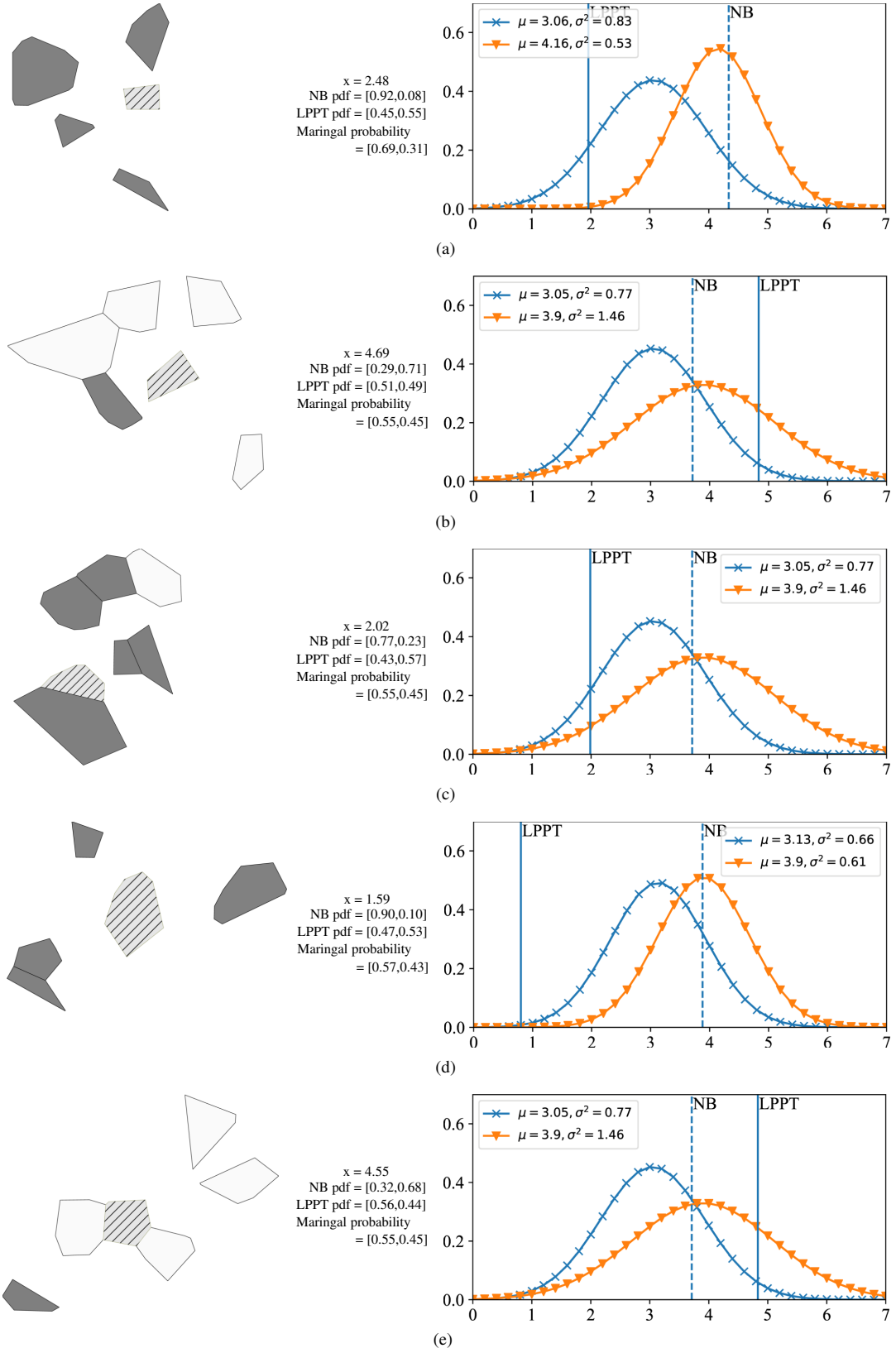
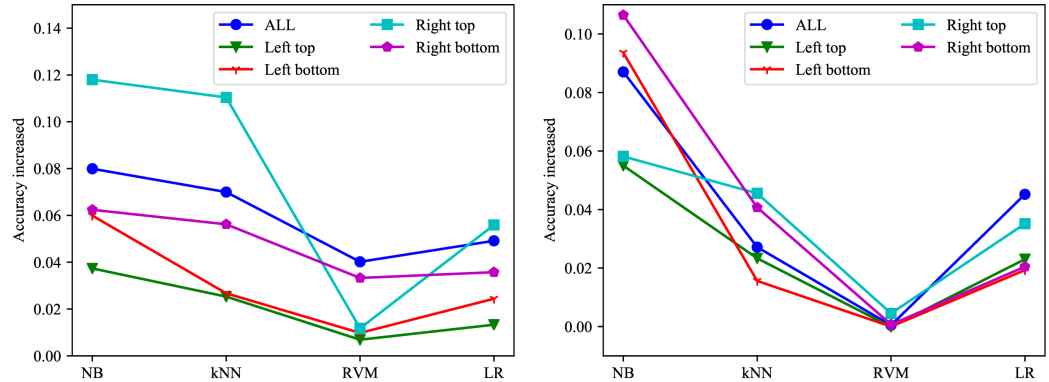
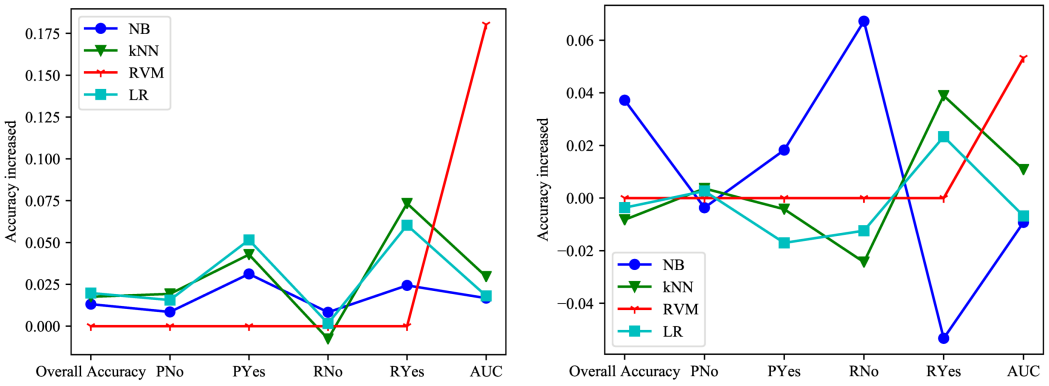


Figure 10. Examples of correctly and incorrectly rectified classification of unseen objects. Left column shows the neighboring objects found, and the right column shows how the decision hyperplane changed as a result of using the updated local prior. For each case, x denotes the feature value, NB pdf denotes the soft classification from NB, and LPPT pdf denotes the soft classification from NE_NB. The dashed and solid lines represent the decision hyperplane of NB and NE_NB, respectively.



(a) Strong spatial association (TM)

(b) Strong spatial association (Gaofen-2)



(c) Weak spatial association

(d) No spatial association

Figure 11. Increment of the average accuracy for the three real datasets: (1) The increment of the average overall accuracy for recognizing vegetation types (TM) using different original classifiers when the neighboring distance was 2 km; (2) The increment of the average overall accuracy for recognizing vegetation types (Gaofen-2) using different original classifiers when the neighboring distance was 2 km; (3) the increment of the average overall accuracy, precision, recall and AUC for the predicting NTD occurrences using different original classifiers when the neighboring distance was 2 km; (4) the increment of the average overall accuracy, precision, recall and AUC for the identification of poverty-stricken villages using different original classifiers when the neighboring distance was 2 km. PYes represents the precision of "Yes," RYes represents the recall of "Yes," PNo represents the precision of "No," and RNo represents the recall of "No."

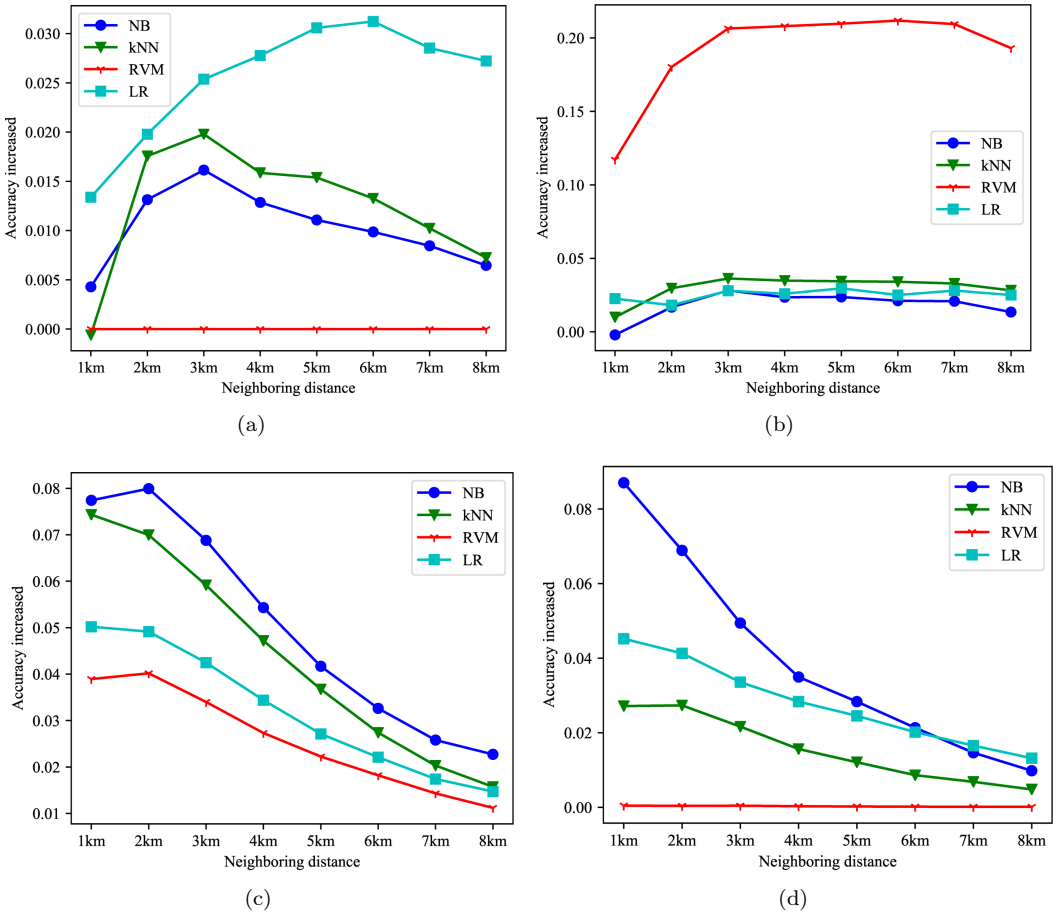
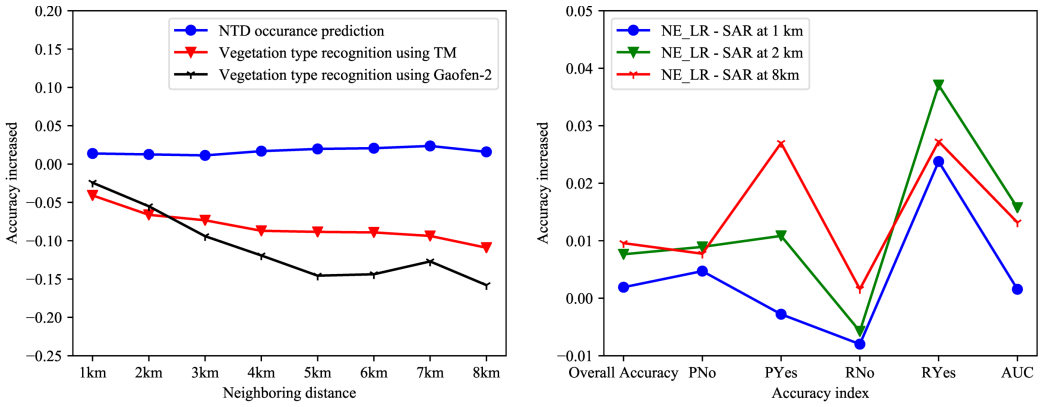


Figure 12. Tendency of the increment of the average accuracy of NTD prediction and vegetation type recognition using a neighboring distance from 1 km to 8 km when the original classifiers were NB, *k*NN, RVM, and LR, respectively: (a) the average overall accuracy increment tendency for NTD prediction; (b) the average AUC increment tendency for NTD prediction; (c) the average overall accuracy increment tendency for vegetation type recognition (TM); (d) the average overall accuracy increment tendency for vegetation type recognition (Gaofen-2).



(a) (b)

Figure 13. Comparison between LPPT and SAR: (a) increment of the average overall accuracy of SAR for predicting Heshun NTD occurrence and recognizing vegetation types when the neighboring distance was from 1 km to 8 km; (b) increment of the average overall accuracy, average precision, average recall and average AUC of NE_LR compared with SAR for predicting Heshun NTD occurrence when the neighboring distance was set to 1 km, 2 km and 8 km. PYes represents the precision of "Yes," RYes represents the recall of "Yes," PNo represents the precision of "No," and RNo represents the recall of "No."

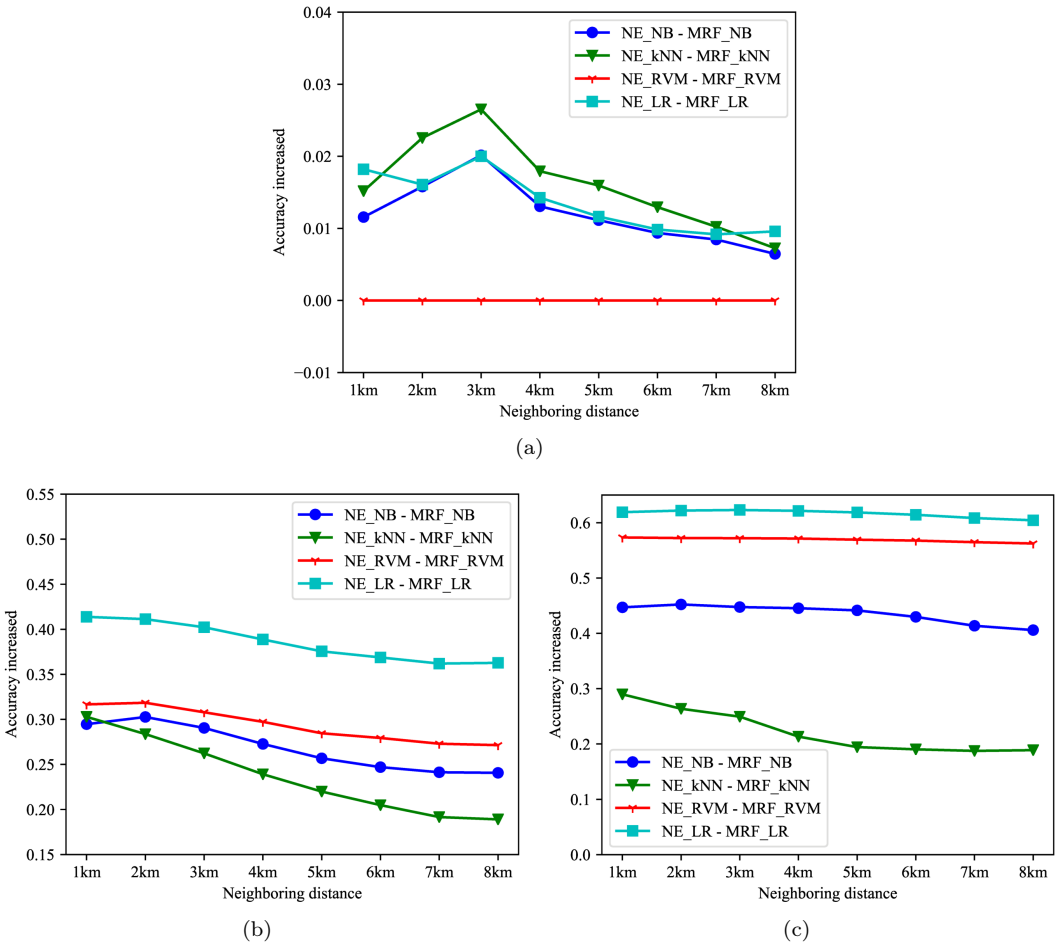
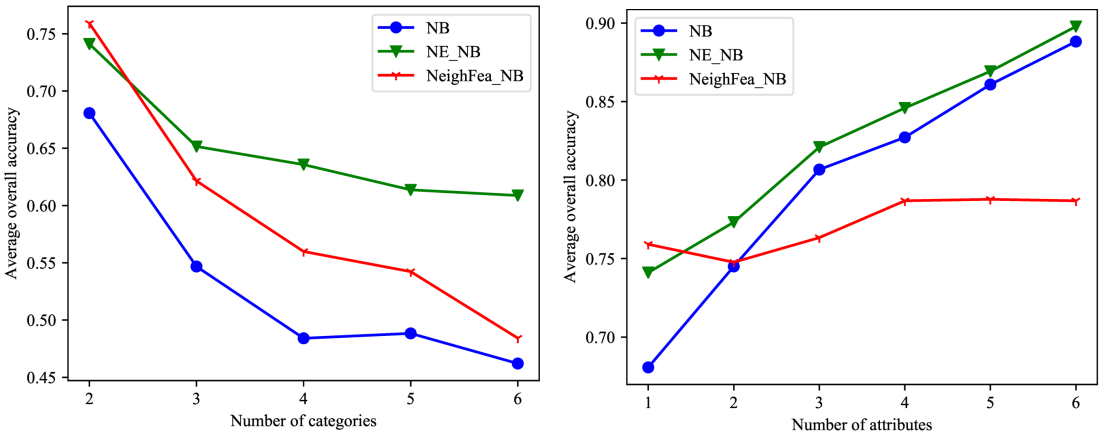


Figure 14. Increment of the average overall accuracy of the LPPT models (NE_kNN, NE_NB, NE_RVM, and NE_LR) compared with those of the MRF models (MRF_kNN, MRF_NB, MRF_RVM, and MRF_LR) for (a) predicting NTD instances, (b) vegetation recognition (TM) and (c) vegetation recognition (Gaofen-2) when the neighboring distance was set to 1 to 8 km, respectively.



(a) (b)

Figure 15. Average overall accuracy of NB, NE_NB, and NeighFea_NB using LPPT and NeighFea for the simulated data, which had two to six categories or one to six features when the neighboring distance was two unit distances.

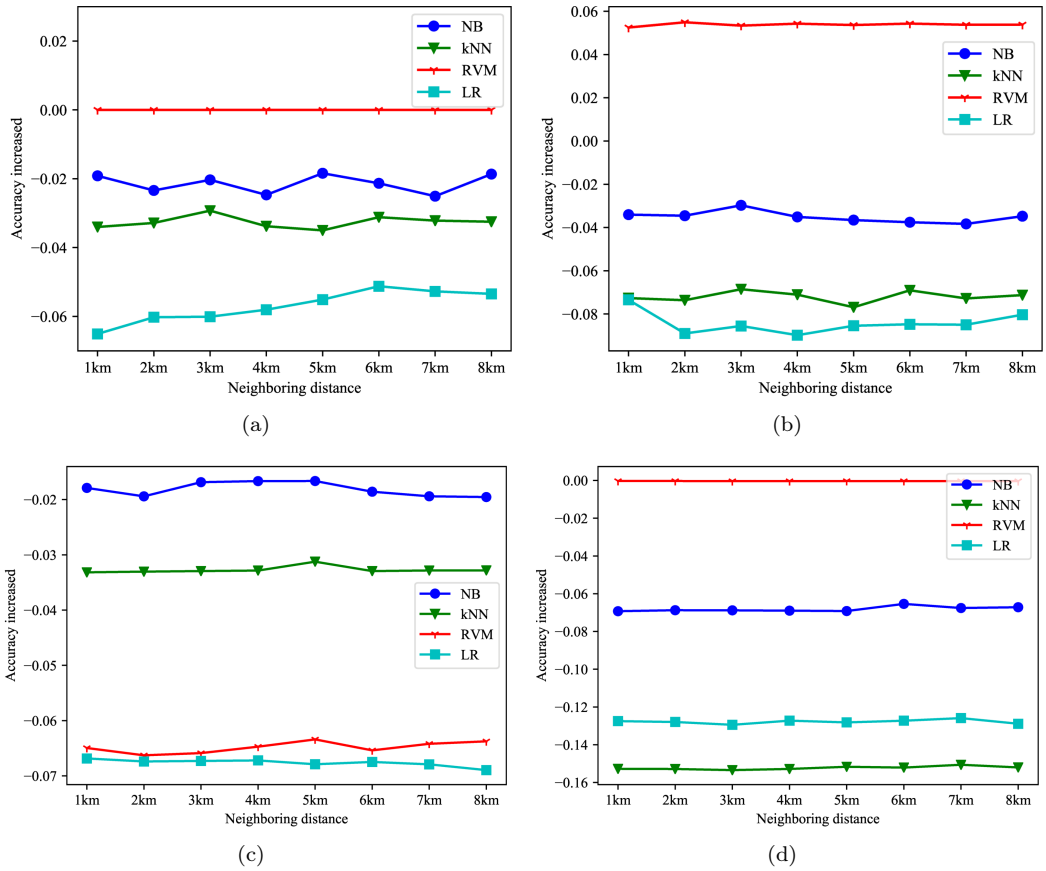


Figure 16. Increment of the average accuracy of NeighFea when NB, *k*NN, RVM, and LR were used: (a) average overall accuracy for the prediction of NTD occurrence; (b) average AUC for the prediction of NTD occurrence; (c) average overall accuracy for the vegetation type recognition (TM); (d) average overall accuracy for the vegetation type recognition (Gaofen-2)

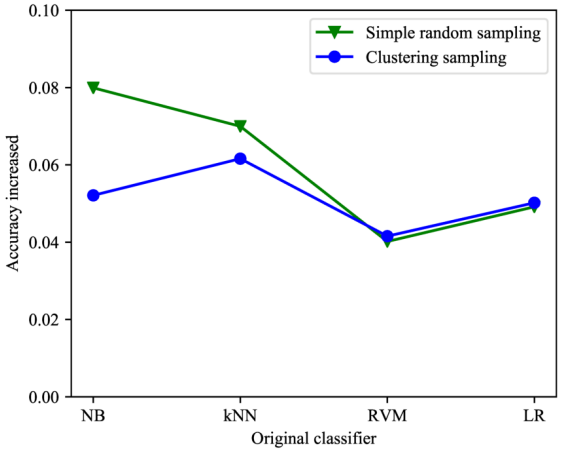


Figure 17. The increment of the average overall accuracy for the vegetation type recognition using different sampling methods when the neighboring distance is 2km.