remote sensing

MDPI

Article

An Explorative Study on Estimating Local Accuracies in Land-Cover Information Using Logistic Regression and Class-Heterogeneity-Stratified Data

Jingxiong Zhang [1,2,3,*], Wenjing Yang [1,3,*], Wangle Zhang [1,3], Yu Wang [1,3], Di Liu [2,3] and Yingchang Xiu [4]

1 School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China; zhangwl@whu.edu.cn (W.Z.); wangyuchn@whu.edu.cn (Y.W.)
2 School of Geodesy and Geomatics, Wuhan University, Wuhan 430079, China; alliu0815@whu.edu.cn
3 Collaborative Innovation Center of Geospatial Technology, Wuhan University, Wuhan 430079, China
4 College of Agronomy, Liaocheng University, Liaocheng 252059, China; xiuyingchang@163.com
* Correspondence: jxzhang@whu.edu.cn (J.Z.); yangwj@whu.edu.cn (W.Y.)

Received: 10 August 2018; Accepted: 26 September 2018; Published: 1 October 2018

Abstract: It is increasingly recognized that classification accuracy should be characterized locally at the level of individual pixels to depict its spatial variability to better inform users and producers of land-cover information than by conventional error-matrix-based methods. Local or per-pixel accuracy is usually estimated through empirical modelling, such as logistic regression, which often proceeds in a class-aggregated or a class-stratified way, with the latter being generally more accurate due to its accommodation for between-class inhomogeneity in accuracy-context relations. As an extension to class-stratified modelling, class-heterogeneity-stratified modelling, in which logistic models are built separately for contextually heterogeneous vs. homogeneous sub-strata in individual strata of map classes, is proposed in this paper for proper handling of within-class inhomogeneity in accuracy-context relations to increase accuracy of estimation. Unlike in existing literature where sampling is usually approached separately, the double-stratification method is also adopted in sampling design so that more sample data are likely allocated to heterogeneous sub-strata (which are more prone to misclassifications than homogeneous ones). This class-heterogeneity-stratified method furnished for sampling and modelling jointly thus constitutes an integrative framework for accuracy estimation and information refinement. As the first step in building up such a framework, this paper investigates the proposed double-stratification method's performance and sensitivity to sample size regarding local accuracy estimation in comparison with those of existing methods through a case study concerning Globeland30 2010 land cover over Wuhan, China. A detailed review of existing methods for analyses, estimation, and use of local accuracy was provided, helping to put the proposed research in a broader context. Candidate explanatory variables for logistic regression included sample pixels' map classes, positions, and contextual features that were computed in different-sized moving windows. Relative performances of these methods were evaluated based on an independent reference sample, with all methods found reliable. It was confirmed that the proposed method is in general the most accurate, as observed with varying sample sizes. The proposed method's competitive performance is thus proved, reinforcing its potential for information refinement. Extensions to and uncertainty aspects of the proposed method were discussed, with further research proposed.

Keywords: local accuracy; land cover; spatial heterogeneity; strata and sub-strata; validation sample data; class occurrence pattern indices; sampling

1. Introduction

Land-cover information is important for resource management and environmental modelling. A variety of land-cover information products (static and dynamic, crisp and soft) are generated from different sensor datasets at regional and global scales [1–6]. This research focuses on static land-cover information coded with discrete class labels rather than percent covers (or fractional covers or class proportions). However, land-cover information is always inaccurate to some extent. This is because information about land-cover status and dynamics is not directly measurable but results from complex processes of image and data analyses, interpretation, and reasoning, which are subject to various forms of uncertainty. There are increasing research efforts directed towards describing, quantifying, and analyzing accuracies (or misclassification errors) in land-cover information [7–13].

Conventionally, classification accuracy is assessed based on error matrices constructed from certain reference or validation sample data. Various accuracy measures, such as percent correctly classified (PCC) pixels (also termed overall accuracy), producer's accuracy, and user's accuracy, can be computed from error matrices [14,15]. On the other hand, as increasingly recognized, local (per-pixel) accuracy should be analyzed and estimated so that users can better understand how misclassifications are related to characteristics of the landscapes being mapped and producers may pursue classifier improvements and information refinement. Spatial analyses, modelling, estimation, and applications concerning local accuracies in land-cover information are discussed by various authors [16–34], as reviewed below.

Research on local accuracy has focused on two major inter-related aspects: (1) local accuracy characterization through spatial and statistical analyses of accuracy-context associations, and (2) local accuracy estimation which is usually based on sample data and empirically built accuracy models. Here, context, as a broadly defined term, includes map class labels, locations, and indices quantifying patterns of class occurrences, as is the case in this paper. It may also be defined in image data and feature space [16]. Classes can refer to static land-cover types or their changes (e.g., forest loss and urban gain, as in [17]), although this paper concerns the former case. We review related work on these two aspects below.

Research on analyses of accuracy-context relationships has found that informative contextual features (for explaining spatial variations in classification accuracy) include spatial heterogeneity, patch size, and other landscape pattern indices [18–20]. Heterogeneity indicates textural complexity of land-cover classes occurring in certain neighborhoods and generally includes compositional (the number and proportions of different classes) and configurational (the spatial arrangement of classes) types [21]. A few examples are as follows.

It was found that land-cover heterogeneity and patch size were important factors determining local accuracy for the United States National Land-Cover Data (NLCD) land-cover product [18,19]. Van Oort et al. established relationships between classification error and landscape characteristics, showing that the probability of correct classification decreases with higher focal heterogeneity (in a neighborhood of 3 by 3 pixels) and smaller patch size [20]. Lechner et al. developed a statistical simulation model to test the effects of patch size and shape, classification threshold, and grid location on classification accuracy of small and linear features. They found that the patch size was an important factor affecting classification accuracy [22]. Chen et al. analyzed and examined the relationships between accuracies of crop classification and area estimation and spatial heterogeneities, in particular, sample pixel impurity and landscape heterogeneity, and found that the impact of configurational heterogeneity on the area estimation was more significant than that of the compositional heterogeneity [21]. As reviewed above, complex landscapes (as indicated by increased heterogeneity, decreased dominance, and smaller patch sizes, etc.) likely lead to more misclassifications. Clearly, misclassifications are also more likely with blurred remote-sensing images and lack of class separability in feature space (e.g., [16]). Logistic models were usually used for describing statistical relationships between local accuracies and contextual/landscape patterns [18–20]. There is also increasing research on local accuracy estimation (or prediction) as follows.

Various methods were explored for estimating local accuracies. These include empirical modelling (e.g., logistic regression [17,23,24]), interpolation with inverse distance weighting (after computing accuracy measures based on locally constrained error matrices) [25,26], kernel functions [24], estimation based on local error matrices that are constructed by geographic weightings [27], kriging [28], and logistic-regression-kriging [29,30]. Maps displaying estimated per-pixel accuracies, such as probabilities of correct classification (or misclassification), user's accuracies (commission errors), and producer's accuracies (omission errors), were also generated (see [17,23] for examples).

The aforementioned methods are mostly employed in the spatial domain (i.e., user's domain), while some of them are also applicable in the spectral domain (i.e., producer's domain) (e.g., kernel functions and logistic regression, as described in [24]). A useful method for local accuracy estimation in the spectral domain is the so-called calibration method that seeks to transform various classification certainty measures, such as maximum posterior probabilities, which are computed as intermediate results prior to output of end results, to accuracy indicators [31]. There was also research on local accuracy estimation in combined spectral and spatial domains. For example, Steele et al. [28] formulated a concept of misclassification probability and present a resampling-based method of estimating misclassification probabilities at training sample locations, from which misclassification probability estimates are then interpolated to a lattice of points via kriging. Additional examples of combined spectral and spatial methods [16,29], in which spectral data and spectrally—derived soft class probabilities were used as the basis for modelling local accuracies, respectively.

As this research is oriented to local accuracy estimation in spatial domains, we elaborate on relevant (spatial—domain) methods, though most of them are mentioned above. A useful method is to compare map and reference class labels at certain sample locations so that a map of misclassifications can be created, helping to analyze their occurrences in the map being assessed. However, such an error location map does not show complete-coverage misclassifications over the problem domain. For mapping per-pixel accuracies, Foody [25] proposed a method based on interpolating accuracy measures computed from locally constructed error matrices. This method relies heavily on availability of relatively dense sample data to work well (a sampling intensity of about 6% was employed [25]). However, sampling intensities say 2.5% which are definitely affordable for small areas (e.g., [24]) will become prohibitive for large-area assessments (e.g., [17]). Developments on this method are reflected on geographical weighting and other extensions in local construction of error matrices [23,27,32]. In addition to such methods making use of only locational information contained in the sample data, logistic modelling using contextual information (in addition to locational information) for estimating local accuracy may be usefully explored (as in this paper), given the observation that per-pixel probability of correct classification is closely related to contextual features characterizing patterns of map class occurrences in the neighborhoods [18–20]. In fact, logistic regression was implemented in both geographic space (e.g., locations) [24] and contextual feature space (e.g., contextual information about class occurrence patterns and landscape characteristics) [17]. The fitted logistic models can then be used to estimate the per-pixel probabilities of correct classification and hence generate maps showing spatially varying accuracies. See the work by Wickham et al. [17], Khatami et al. [24] and Zhang and Mei [30] for examples of using logistic models built on sample data and land-cover map data to estimate local accuracies.

Having provided some relatively solid justification for local accuracy estimation based on logistic modelling (which this research adopts), we consider issues of sampling (for collecting reference sample data), in particular, coupling of sampling designs and modelling approaches, below. The coupling of modelling and sampling facilitates integration of accuracy estimation and information refinement, with the latter using information about local accuracies in fusion of map and reference data for enhancing quality of fused maps [33,34]. This integrative framework actually represents the paper's major contribution to the literature, as is seen below.

Like in error-matrix-based accuracy assessment, reference or validation sample data consisting of reference class labels (from which binary data indicating correct or incorrect classifications at sample pixels are obtained) are necessary for model-building. As understandable, models empirically built and model predictions are conditional to specific sample data employed (for model training), which are collected following certain sampling designs. It is thus important to reflect on how logistic modelling was implemented in combination with sampling in the past. The review below aims to provide a general indication to the largely loose coupling between modelling and sampling in existing literature, though it is by no mean comprehensive or detailed.

Smith et al. implemented logistic regression for characterizing local accuracy in the NLCD datasets in the eastern US encompassing four regions across 21 states with 5020 sample pixels (presumably with a region-stratified random sampling design) by a class-aggregated modelling strategy [18], with models built for individual regions separately. Then, Smith et al. carried out logistic modelling of local accuracies by a (map) class-stratified modelling strategy (stratifications with map classes at both Levels I and II), using the same sample set (5020 sample pixels) [19]. Based on a class-aggregated modelling strategy, Van Oort et al. used a sample set of 1161 grid cells (collected with a kind of near-systematic sub-sampling) to model and analyze the classification accuracy of agricultural crops in the Dutch national land-cover database [20]. Based on a simple random sample data collected at a sampling intensity of about 5%, Zhang and Mei integrated logistic regression and geostatistics for local accuracy characterization in land-cover change information via class-aggregated modelling [30]. With stratified random sample data collected at intensities of 0.5% and 2.5%, Khatami et al. compared logistic modelling with other modelling approaches for estimating local accuracies in classified remote-sensing images, with both class-aggregated and class-specific (i.e., class-stratified) modelling approaches considered in the spatial domain or spectral domain [24]. It was confirmed that class-specific modelling provides more accurate estimation of local accuracies than class-aggregated modelling, as investigated in [18,19,24].

As reviewed above, with reference sample data collected, logistic modelling can be performed in a (map) class-aggregated or class-stratified way. The latter is well suited to accommodating between-class inhomogeneity in accuracy-context relations, as demonstrated in [19], and has been confirmed to be more accurate than the former [24]. In addition to systematic sampling and random sampling (simple or stratified), which are among the commonly used sampling designs, sampling adaptive to local class heterogeneity (e.g., class impurity in a focal neighborhood of 3 by 3 pixels) was also explored for accuracy assessment [35]. This is motivated by the observation that boundary areas (i.e., edge pixels) are more likely misclassified than inner areas (i.e., interior pixels), as amply demonstrated in the literature on local accuracy estimation [35]. Based on sample data in which edge pixels and interior pixels were treated separately, accuracy assessment was carried out, showing large differences between classification accuracies in segments of edge pixels and those of interior pixels [36].

Similar to the aforementioned error-matrix-based accuracy assessment, models of local accuracies may be built separately for contextually heterogeneous vs. homogeneous pixel segments (sub-strata) in individual strata of map classes, hopefully increasing accuracy in resultant model estimation. In other words, as an extension to class-stratified modelling, class-heterogeneity-stratified modelling can be usefully explored for proper handling of within-strata inhomogeneity in accuracy-context relations. This double-stratified method should also be considered for sampling pertaining to reference sample data collection so that sampling and modelling are well coupled with each other. More importantly, with this double-stratified method applied in sampling designs, heterogeneous sub-strata (which usually are more prone to misclassification than homogeneous sub-strata) are likely sampled at greater sampling intensities than with other designs without considering sub-stratification by heterogeneity. The increased number of sample pixels in error-prone locations will, in turn, enable detailed studies of misclassification patterns and facilitate direct correction of misclassification errors for refinement of land-cover information through fusion of map data and reference sample data. This helps to broaden usability of sample data for not only local accuracy estimation but also

information refinement. Therefore, the aforementioned class-heterogeneity-stratified method for sampling and logistic regression modelling constitutes this paper's major contribution to the literature. The main features and values of the proposed double-stratification method include a combined perspective of sampling and modelling (which were seldom treated coherently in the past) and an integrative construct for local accuracy characterization and information refinement.

As the first step towards building up the aforementioned integrative framework, this paper investigates performances of the proposed double-stratified method (featuring class-heterogeneity-stratification in both logistic modelling and sampling) in comparison with those of alternative methods (i.e., logistic regression modelling and sampling that are not class-heterogeneity-stratified). This is important as the proposed method needs to be proved competitive in terms of performance for local accuracy estimation at the first place to be worthy of being pursued further for information refinement. In addition to comparing the proposed and alternative methods' performances based on a separate model-testing sample, these methods' sensitivities to sample sizes were also analyzed, with their robustness to varying sample sizes examined. This (sensitivity analysis) actually represents another contribution of this research to the literature, as it was rarely considered in similar research. As shown in the case study, the proposed class-heterogeneity-stratified method generates significantly more accurate estimation of local accuracies than alternative methods including a double-stratification method with sub-stratification by edge vs. interior pixels (as described in [36]), according to results of statistical testing and sensitivity analyses.

The remainder of the article is as follows. In Section 2, the study area and data used in the research are described first, followed by descriptions of methods for sampling and logistic regression modelling, in particular, those with double stratifications by class and heterogeneity. Section 3 describes the experiment carried out and the results obtained, aiming to test the proposed method in comparison with alternative methods. Finally, Section 5 concludes the paper after discussing some issues in Section 4.

2. Materials and Methods

2.1. The Study Area and Experimental Data

GlobeLand30 2010 land-cover dataset for Wuhan city was used for the study in this paper, as shown in Figure 1. As a global fine-resolution land-cover information product, GlobeLand30 (for 2000 and 2010), which was produced by the National Geomatics Center of China (NGCC) in 2014, has ten land-cover classes (http://www.globallandcover.com). The city of Wuhan (Lat 29°58′–31°22′ N, Long 113°41′–115°05′ E) is about 8495 km^2 in areal extent, located in the middle and lower reaches of the Yangtze, and is the provincial capital of China's Hubei province, as shown in Figure 1 (the inset map of China, lower right corner). For Wuhan, there are seven classes (Table 1, except for shrub, tundra, and permanent snow and ice), as shown in Figure 1. In particular, the dominate class is cultivated land, occupying about 60 percent of Wuhan's areal extent, followed by water, forest, and artificial surface, which account for 15 percent, 12 percent, and 7 percent of the total area of Wuhan, respectively. Grassland, wetland, and bare land together take about 6 percent of Wuhan's areal extent.

Reference data recording reference class labels are required for local accuracy estimation. Reference class labeling is defined as the best available assessment of the ground conditions. Collecting reference data is a time-consuming and costly procedure. In the study, the reference classes at sample pixels (sampling will be described in the next subsection) were obtained using visual interpretation of high spatial resolution images (i.e., Google Earth images). Interpretation was undertaken according to the standards consistent with GlobeLand30 classification system (Table 1). In most cases, Google Earth images were used for interpretation. When such images were not available, actual ground visits and Landsat TM image flown in temporal proximity of corresponding GlobeLand30 2010 maps were used as sources to obtain reference class labels. A set of reference data were used as (model) training

data, with another as testing data for performance evaluation (which is to be described in Section 3.4). The training data and testing data were independent. For the training data, further information about sampling design and resultant sample data collected is provided in Sections 2.2 and 3.1, respectively.

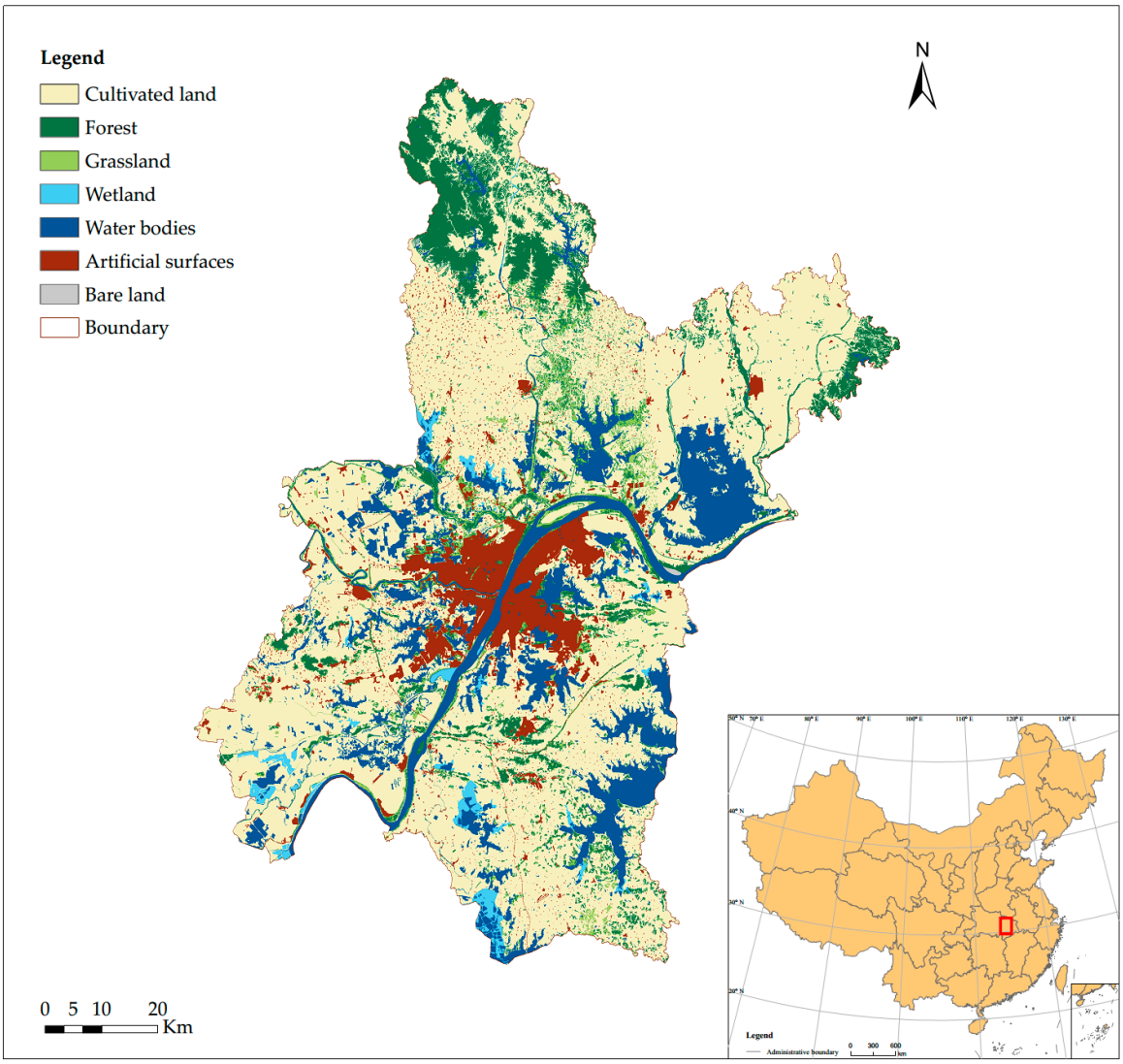

Figure 1. GlobeLand30 2010 land cover for Wuhan, China.

Table 1. GlobeLand30 2010 land-cover classes.

Class Name (Abbreviation)	Definition
Cultivated land (Cultivt)	Land used for agriculture, horticulture, and gardens, including paddy fields, irrigated and dry farmland, vegetable and fruit gardens, etc.
Forest	Land covered by trees, vegetation covers over 30%, including deciduous and coniferous forests, and sparse woodland with cover 10–30%, etc.
Grassland (Grass)	Land covered by natural grass with cover over 10%, etc.
Shrub	Land covered by shrubs with cover over 30%, including deciduous and evergreen shrubs, and desert steppe with cover over 10%, etc.
Wetland	Land covered by wetland plants and water bodies, including inland marsh, lake marsh, river floodplain wetland, forest/shrub wetland, peat bogs, mangrove, and salt marsh, etc.
Water bodies (Water)	Water bodies in land area, including river, lake, reservoir, fish pond, etc.

Table 1. *Cont.*

Class Name (Abbreviation)	Definition
Tundra	Land covered by lichen, moss, hardy perennial herbs and shrubs in the polar regions, including shrub tundra, herbaceous tundra, wet tundra, and barren tundra, etc.
Artificial surfaces (Artfct)	Land modified by human activities, including all kinds of habitation, industrial and mining area, transportation facilities, and interior urban green zones and water bodies, etc.
Bare land (Bare)	Land with vegetation cover lower than 10%, including desert, sandy fields, Gobi, bare rocks, saline and alkaline land, etc.
Permanent snow and ice	Lands covered by permanent snow, glacier, and icecap.

2.2. Sampling Design and Sample Allocation for Reference Data

In simple random sampling (SRS), n (sample) units are selected out of N units in the population such that every one of the distinct samples has an equal chance of being drawn. By a stratified random sampling (StRS) that may be based on geographic regions or map classes [12–14], independent and random sample units are drawn from individual strata. In this study, SRS and StRS were employed for comparative study about the proposed class-heterogeneity-stratified sampling design in the context of local accuracy estimation. They were sampling designs well suited for class-aggregated (CA) and class-stratified (CS) modelling, respectively, thus identified also as CA and CS, respectively, in the remainder of the paper.

As mentioned previously, for the proposed method, stratification is first based on map class and then heterogeneity/homogeneity. Here, homogeneity is defined as the number of pixels with the same class label as that of the center pixel in a focal neighborhood of 3 by 3 pixels. The homogeneity value can be viewed as the patch size of the center pixel in the focal neighborhood. A homogeneity value of 4 is chosen to be the threshold value to determine if the center pixel lies in a homogeneous sub-stratum or a heterogeneous one within a stratum of a certain map class. In contrast, a threshold of 8 was used for determining if the center pixel is interior (i.e., it belongs to a homogeneous sub-stratum) in [36]. The former sampling design (stratified by map classes and sub-stratified by pixels being at the centers of homogenous vs. heterogeneous focal neighborhoods) is labeled EO, while the latter (stratified by class and sub-stratified by edge vs. interior pixels) EI.

For sample allocation, Neyman allocation method was used in this study. This is because Neyman allocation considering stratum size or proportion and the degree of variation of each stratum will greatly improve variance or standard error of estimation. For StRS, in particular, when sample size is fixed, variance or standard error of estimation can be minimized by Neyman allocation. For example, Neyman allocation method was used for sample allocation among different strata of map classes [35]. By Neyman allocation method, the sample size for a stratum (but sub-stratum for EI or EO) is calculated as:

$$n_h = n \frac{W_h S_h}{\sum_{h=1}^{L} W_h S_h} \quad (1)$$

where n_h is the number of sample units in stratum h, n is total sample size, W_h indicates the stratum's area proportion, and S_h represents the stratum's standard deviation.

In this paper, for CS sampling and class-heterogeneity-stratified sampling (EO and EI), the "strata" when using Equation (1) were map classes and combinations of map classes and heterogeneity/homogeneity sub-classes, respectively. When conducting Neyman allocation, the standard deviation (or variance) of each stratum should be known for calculating the sample size of each stratum. Standard deviation of each stratum was estimated based on initial sample data, as in [13,35]. For stratum h, its standard deviation (S_h) is calculated according to Equation (5.55) in [37].

The initial sample data were collected using stratified sampling design with proper stratification plan (e.g., class-stratification for Method CS, class-heterogeneity-stratification for Methods EI and EO).

2.3. Logistic Regression Modelling

Logistic regression models are usually used to describe relationships between a binary response variable $I(x)$ and one or more explanatory variables $Z_k(x)$ ($k = 1, \ldots, K$) at pixel x. For mapping local accuracies, the response variable is an indicator $I(x)$ for classification correctness (or agreement between the map and reference class labels) at pixel x, and is coded as 1 if pixel x was correctly classified and 0 otherwise. The explanatory variables are pattern indices for map class occurrences within multi-scale neighborhoods, as discussed previously. Model predictions are probabilities of individual pixels being correctly classified. A logistic model is

$$\log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \sum_{k=1}^{K} \beta_k Z_k(x) \quad (2)$$

where $p(x)$ is the probability of pixel x being correctly classified, $\beta = (\beta_0, \beta_1, \beta_2 \ldots, \beta_k)$ represents the parameters to be estimated. Relations between $p(x)$ and $I(x)$ are further discussed in the Discussion, so is possible extension to logistic-regression-based estimation.

Logistic regression analyses are preferably approached in a way well coupled with the model-training sample data available. In this study, logistic modelling was performed in different ways: for Methods EO and EI, models were built for individual sub-strata separately; models were built for individual strata in CS; for CA, a single model was built for the study areas as a whole. Resultant logistic models built for EO and EI should be applied to their corresponding sub-strata in the land-cover map concerned, and those for CS to corresponding strata. When modelling is performed separately per land-cover class, as in EO, EI, and CS methods, logistic models can be applied to commission errors [17], although only per-pixel probabilities of correct classification were considered in this study.

Class occurrence pattern indices including homogeneity, heterogeneity, dominance, entropy, and contagion were used as candidate explanatory variables in this study (Table 2). These pattern indices were quantified in different-sized moving windows. In the study, due to computational limitation, moving window sizes were 3 by 3, 5 by 5, to 39 by 39 pixels at the maximum.

Definitions for the candidate explanatory variables shown in Table 2 are as follows. Class is represented by six binary variables, as there are seven land classes occurring in the study area. Homogeneity (Hom) refers to the number of pixels with the same map class label as the center pixel in a neighborhood (or moving window). Heterogeneity (Het) indicates the number of different classes in the moving window. If the value of heterogeneity equals 1, it indicates that the neighborhood is pure or homogeneous with the same class. Entropy (Ent) indicates the average uncertainty of class occurrences. When the probability of each class being present in a given neighborhood is roughly the same, entropy value reaches its maximum, and when only one class dominates, entropy value is zero. Dominance is the difference between the maximum possible diversity of the neighborhood or moving window being considered (measured by entropy, $\ln(K)$, K being the number of classes occurring in the moving window) and computed diversity in the moving window. Thus, dominance measures the extent to which one or a few cover types dominate the landscape (the moving window, to be more precise). A higher value indicates that the neighborhood is dominated by one or a few land-cover classes, and a lower value indicates that land-cover types have nearly equal proportions [38], given the same number of classes in the area (moving window) (as the maximum diversity value is determined by the number of classes in an area). Clearly, dominance is related to entropy, although their relation is modulated by K, which is itself varied rather than fixed across the landscape where moving windows fall in. Contagion index describes the degree of clumping of land cover, which is computed from the frequencies by which different pairs of cover types occur as adjacent pixels within a moving window [39]. Because the index contains spatial information, it is one of the most important

landscape pattern indices. In general, high contagion values show that certain landscape patch types form good connectivity; on the contrary, low contagion values indicate that the landscape is highly fragmented [20,39].

The −2LogLikelihood statistics (−2LL) is often used to test whether all regression parameters in a model are simultaneously zero, furthermore, the difference between the -2LL of the two models follows a chi-square distribution with degrees freedom being equal to the number of the extra variables to those shared by the two models [40]. A χ^2-test then can be used to test whether the addition of these extra variables can significantly improve model-fitting.

Table 2. Class occurrence pattern indices.

Variable	Abbreviation	Description
Land cover	Class	the class labels of the sample pixel
Homogeneity	Hom	the number of pixels with the same label as the sample pixel in its neighborhood
Heterogeneity	Het	the number of land-cover class in the window centered on the sample pixel
Entropy	Ent	the weighted mean of the entropy associated with the single class, expressed in terms of probability of various class types
Dominance	Dom	the degree to which one or few land-cover types predominate the landscape in terms of proportion
Contagion	Con	the extent to which classes in a patch (moving window centered at the pixel being considered) are clumped

3. Experiment, Results, and Analyses

The experiment procedures consisted of reference sample data collection (for model-training and testing, see Sections 3.1 and 3.4, respectively), model-building, predictive mapping of local accuracies using models fitted with training data, and performance evaluation based on testing data. The performances of different methods for local accuracy estimation were compared, so were their sensitivities to sample size.

3.1. Training Sample Data Collection

As mentioned in Section 2.2, the sampling designs tested were: (1) SRS (CA); (2) StRS (CS), (3) EI (a sampling design with stratification by map classes and sub-stratification by edge/interior pixels), and (4) EO (the proposed class-heterogeneity-stratified random sampling). For all methods being examined, the sample size was 3000 pixels (for model-training). The rationales are that a sample size of 3000 pixels is about the minimum to enable wetland and bare land (rare classes) to be allocated minimum numbers of sample pixels necessary for model-building and that the sample size is to be reduced to 1000 pixels at the minimum for sensitivity analysis as in Section 3.5. Sample allocations for the aforementioned sampling designs are shown in Table 3. For sample allocation, the Neyman method was used for all methods except CA (where SRS was adopted). For EO and EI, the homogenous and heterogonous areas of each class were considered as different strata during sampling.

As shown in Table 3 (upper part, strata of classes), sampling intensities are about the same across different classes (strata) for Method CA as SRS was employed therein, while variations among different classes are apparently increasing in CS, EI, and EO. As shown in Table 3 (lower part), sampling intensities at heterogeneous sub-strata are obviously higher than at homogeneous sub-strata, as evaluated relative to the respective population sub-strata (e.g., about 3.93% for Wetland_E, wetland class, heterogeneous sub-stratum, but only 0.10% for Wetland_O, wetland class, homogeneous sub-stratum). Such a remarkable difference is seen to be narrower in the cases of EI, even reversed in CS and CA (Wetland_E vs. Wetland_O, for example).

Table 3. Sample allocations for different sampling designs (number of sample pixels belonging to individual strata or sub-strata, while sampling intensities (in percentages) shown in parentheses are with respect to the total number of pixels (N_{strata}) belonging to specific strata or sub-strata in the land-cover map).

Strata	Sample Designs				N_{strata}
	SRS (CA)	StRS (CS)	EI	EO	
Cultivt	1812(0.03)	1795(0.03)	1570(0.03)	1215(0.02)	5,795,924
Forest	363(0.03)	360(0.03)	340(0.03)	420(0.04)	1,135,422
Grass	104(0.03)	110(0.03)	175(0.05)	290(0.09)	319,278
Wetland	57(0.04)	90(0.07)	165(0.12)	220(0.16)	135,768
Water	459(0.03)	370(0.03)	375(0.03)	385(0.03)	1,418,938
Artfct	201(0.03)	195(0.03)	215(0.03)	300(0.04)	719,111
Bare	4(0.04)	80(0.75)	160(1.50)	170(1.60)	10,652
Strata and Sub-Strata (E-Heterogeneous O-Homogeneous)					
Cultivt_E	17(0.03)	20(0.04)	21(0.04)	120(0.21)	56,721
Cultivt_O	1795(0.03)	1775(0.03)	1549(0.03)	1095(0.02)	5,739,203
Forest_E	42(0.03)	46(0.03)	44(0.03)	140(0.10)	133,655
Forest_O	321(0.03)	314(0.03)	296(0.03)	280(0.03)	1,001,767
Grass_E	32(0.05)	24(0.03)	29(0.04)	120(0.17)	70,366
Grass_O	72(0.03)	86(0.03)	146(0.06)	170(0.07)	248,912
Wetland_E	0(0)	1(0.05)	6(0.30)	80(3.93)	2033
Wetland_O	57(0.04)	89(0.07)	159(0.12)	140(0.10)	133,735
Water_E	2(0.01)	2(0.01)	7(0.03)	100(0.42)	23,895
Water_O	457(0.03)	368(0.03)	368(0.03)	285(0.02)	1,395,043
Artfct_E	6(0.03)	10(0.05)	11(0.06)	100(0.52)	19,324
Artfct_O	195(0.03)	185(0.03)	204(0.03)	200(0.03)	699,787
Bare_E	0(0)	22(0.60)	41(1.12)	80(2.19)	3658
Bare_O	4(0.06)	58(0.83)	119(1.70)	90(1.29)	6994

3.2. Model Selection

For model selection, an exhaustive procedure was applied to find the optimal model containing the largest number of significant explanatory variables based on a particular model-training sample, as in [20]. There were 98 candidate explanatory variables (i.e., map class (1), pattern indices computed in different-sized windows (95), and sample pixel's coordinates (2)) to choose from. The pattern indices in different windows were written as the combination of abbreviations and numbers, where numbers represented the sizes of the window. For example, Hom3 indicated the homogeneity of the sample data in a 3 by 3 pixels' window. Individual candidate explanatory variables were tested using chi-square statistics with respect to their statistical significance in model-fitting. This was to test if adding a candidate variable to a model already selected (i.e., a simpler model) significantly improves model-fitting (i.e., leading to a significant decrease in model deviance) (at a significance level (α) of 0.05). The significance of interaction terms for every two significant explanatory variables already selected in logistic regression was also assessed. Optimal models with the largest number of significant explanatory variables were identified when it was confirmed that adding variables further leads to insignificant reduction in model deviances. This means that a simpler model (with less significant explanatory variables) is preferred between two models with the same level of goodness of fit. This procedure was implemented on the R software system.

Results of model selection are shown in Table 4. "E" and "I" (Method EI) represent edge (more heterogeneous) and interior (more homogeneous) pixels, respectively, in Table 4. Thus, Forest_I indicates interior pixels sub-stratum for Forest class, and corresponds (though not one-to-one) to Forest_O (for Method EO) of homogenous sub-stratum of Forest class.

Table 4. Optimal models with significant explanatory variables for different methods.

Methods	Strata or Sub-Strata	Significant Explanatory Variables
CA		Ent9 Class Con3 Ent39 Con37 Ent5
CS	Cultivt	Hom5 Ent33 Con33 Dom15
	Forest	Hom3 Con39
	Grass	Dom19 Hom35 Het3
	Wetland	Het27 Dom33 Ent25 Y
	Water	Y
	Artfct	X Hom3
	Bare	Y Ent3 Dom13 Con3 Het31
EI	Cultivt_E	Y Hom5
	Cultivt_I	Hom39 X Y
	Forest_E	Hom5 Y Ent39 Dom15
	Forest_I	Con31
	Grass_E	Hom3 Hom37
	Grass_I	Hom5 Het27
	Wetland_E	Hom19 Con39 Con3 Ent5
	Wetland_I	X Hom39 Y
	Water_E	X
	Water_I	Y Hom5 Dom7
	Artfct_E	Het5
	Artfct_I	Hom39
	Bare_E	Het5 Het15
	Bare_I	Con11 Dom35
EO	Cultivt_E	Dom37 Het11
	Cultivt_O	Het7 Het3 Hom39
	Forest_E	Con5 Hom5
	Forest_O	Het25 Hom5
	Grass_E	Ent3 Dom5 Het13
	Grass_O	Dom33
	Wetland_E	Con3 Y
	Wetland_O	Dom13
	Water_E	X Con5
	Water_O	Y
	Artfct_E	Y
	Artfct_O	Het3
	Bare_E	Het5
	Bare_O	X Con17 Het39

Unsurprisingly, there are apparent differences in the optimal models for individual strata or sub-strata, as shown in Table 4. In other words, individual strata or sub-strata have their own optimal models with unique significant explanatory variables of their own. This (non-uniformity of logistic models of local accuracies across map strata and heterogeneity sub-strata) confirms the need for modelling local accuracies for individual map classes and heterogeneous/homogeneous sub-strata (of each class) separately. This also necessitates sensitivity analyses (to be described in Section 3.5) for methods being tested with respect to different sample sizes.

Lastly, regarding relative significance of locational vs. contextual explanatory variables in logistic modelling, the latter seem to be more informative in explaining observed classification correctness. This is supported by the results shown in Table 4, where accuracy models have solely contextual features (e.g., CS, Forest) as explanatory variables in 21 cases out of a total of 36 cases, while there are only four cases when explanatory variables are locations alone (e.g., CS, Water).

3.3. Mapping Per-Pixel Probabilities of Correct Classification and Classification Correctness at Sample Locations

With models of local accuracies obtained for the four methods (i.e., CA, CS, EI, and EO) as in Table 4, maps of per-pixel probabilities of correct classification were generated using these models, as shown in Figure 2a–d, respectively. The differences between the maps of estimated local accuracies shown in Figure 2 are somehow appreciable. Also, we can check their differences quantitatively using model-testing sample data, as shown later.

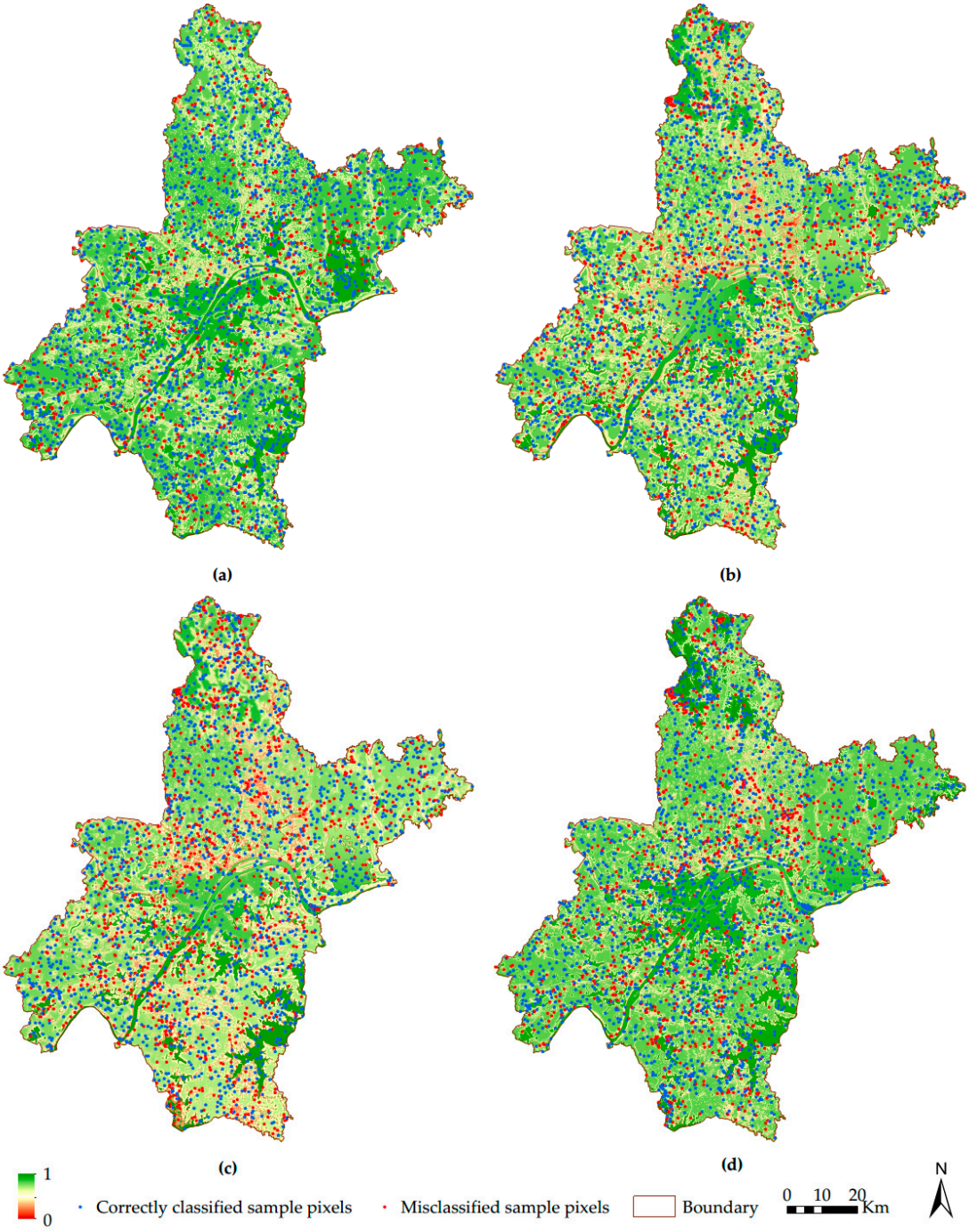

Figure 2. Maps of per-pixel probabilities of correct classification estimated by different methods (overlaid with observed classification correctness indicators at respective model-training sample pixels): (**a**) CA, (**b**) CS, (**c**) EI, and (**d**) EO.

Also shown in Figure 2 are maps of observed classification correctness indicators (simply termed maps of misclassifications) at model-training sample pixels. These maps of misclassifications were obtained by comparing map and reference class labels, with 1 indicating correct classification, and 0 for misclassification, shown as blue and red points in subfigures of Figure 2. They are overlaid on their corresponding maps of estimated per-pixel probabilities of correct classification. Differences in error locations appear more appreciable than those in estimated per-pixel accuracies, due to use of different training samples.

As shown in Figure 2, locations of misclassifications (in red) are more likely found near locations with smaller estimated probabilities of correct classification (from orange to red). On the other hand, locations of correct classifications (in blue) are more likely found near locations with larger estimated probabilities of correct classification (from green to blue).

The closeness between estimated probabilities and actual indicators of correct classification can be assessed quantitatively using the area under the receiver operating characteristic curve (AUC) [41–44], as in the next subsection for performance evaluation based on independent testing sample data. AUC is a commonly used metric for assessing performances (discriminatory power) of models constructed to predict binary outcomes. AUC values can theoretically range from 0 to 1, with larger value indicating greater accuracy in predictions [42,43]. Using training sample data (3000 sample pixels but different designs for different methods) as references, AUC values for the four methods tested were computed using an R package pROC [45]. Their AUC values were estimated to be 0.6810, 0.7198, 0.7349, and 0.7625, for methods CA, CS, EI, and EO, respectively, indicating their increasing accuracy in predicting per-pixel probabilities of correct classification. Nevertheless, it should be noted that assessing models' accuracy based on training data is not recommendable. Rather, we should use independent sample data to assess these models' performances, as in the next subsection.

3.4. Performance Evaluation

As mentioned previously, AUC was used to evaluate performances of different methods for local accuracy estimation. *T*-test was performed on AUC values to examine whether there exist significant differences between the aforementioned different methods.

For this, an independent set of 1020 model-testing sample pixels were acquired (with a simple random design) from visual interpretation of high-resolution satellite images, as described in Section 2.1. Different sets of local accuracy estimations (obtained with samples collected with different sampling designs) were compared to their corresponding reference indicator data, with AUCs computed also using the R package pROC [45]. AUC values obtained by these different methods are shown in Table 5.

Table 5. AUC values for different methods.

Method	CA	CS	EI	EO
AUC	0.7516	0.7724	0.7667	0.7968

As mentioned previously, the range of AUC is from 0.0 to 1.0. Models are graded as excellent, fair, and poor. Specifically, models providing excellent predictions have AUC values higher than 0.9, fair models have AUC values between 0.7 and 0.9, and models are considered as poor when their AUC values are below 0.7 [41,43]. The AUC values shown in Table 5 (based on testing sample data) are in the range between 0.75 and 0.8. Thus, all methods should be considered as fair. In comparison, as shown in Table 5, the performances of predictions obtained with EO, CS, and EI are better than that with CA in terms of AUC values, with EO method getting the greatest AUC value of 0.7968. AUC values shown in Table 5 are greater than their respective AUC values obtained with training sample data. This is perhaps due to use of different sample designs by them and a much smaller sample size used in the latter. After all, AUC values obtained from sample data were estimates and can be assessed with respect to their variability, as examined in the next subsection.

T-test was performed to determine statistical significances of differences between these AUC values, leading to results shown in Table 6. R package pROC [45] was used also for testing the significances of differences in AUC values obtained from alternative methods. pROC has a built-in function for t-test, assessing the significance of differences in AUC values, with variance and covariance of AUC estimates for a pair of methods computed by the package before *p*-values are output.

Table 6. *T*-test results of statistical significance in differences between AUC values shown in Table 5 (*p*-value shown, * for significant at $\alpha = 0.10$, ** for significant at $\alpha = 0.05$).

***T*-test**	**EO vs. EI**	**EO vs. CS**	**EO vs. CA**	**EI vs. CS**	**EI vs. CA**	**CS vs. CA**
p-value	0.005 **	0.028 **	0.0002 **	0.664	0.119	0.069 *

As shown in Table 6, EO method is significantly more accurate ($\alpha = 0.05$) than EI, CS, and CA methods, while CS is significantly more accurate than SRS ($\alpha = 0.10$). Comparing EI and CS methods, there was no significant difference between them.

3.5. Sensitivity Analysis

As model predictions depend on model-training sample data characteristics (i.e., sampling design and sample size), it is important to undertake sensitivity analysis for the methods being compared in the study. This was done in the following steps:

- The original sample size (3000 pixels) decreased at an equal step of 200 (pixels), down to 1000 pixels at the minimum, leading to ten reduced sample sizes (N_d $d = 1, 2, \ldots, 10$) for ten new samples. Sample allocations to strata and sub-strata were done proportionately to the reduced sample sizes N_d for new samples, according to sample allocation for the original sample shown in Table 3.
- The ten new sample sets (of reduced sizes) for each method were collected from the original sample by SRS from the original sample (CA), individual strata (CS), or sub-strata (EO and EI).
- Logistic regression modelling was done using samples of reduced sizes, for the four methods. For this, model selection was carried out for all methods with all new samples with reduced sizes (refer to Section 3.2 for technical detail). This resulted in different models with different optimal explanatory variables based on different new samples for different methods.
- Local accuracies were estimated for different methods using their respective new samples of reduced sizes. The resultant accuracy estimations were assessed using the test sample (1020 pixels) mentioned in Section 3.4, leading to AUC values for different methods with new samples of reduced sizes.
- Steps 2 through 4 were repeated 20 times, giving rises to 20 AUC values for each of the method with a specific reduced sample size N_d Means and standard deviation were computed based on these AUC values. The mean AUC values for different methods with different reduced sample sizes are shown in Figure 3.

As shown in Figure 3, for all methods, mean AUC values decrease as sample sizes decrease. EO method has the highest mean AUC values except for the sample with 1200 pixels when CS method performs slightly better. When the sample size is greater than 1800 pixels, the order of performances from good to poor in terms of mean AUC values is Methods EO, CS, EI, and CA. As sample size decreases from 1800 to 1000, the order of performances for Methods CS, EI, and CA fluctuates, with Method EI performing worse than Method CA. Two more observations are outstanding from Figure 3. First, CA method's quality of predictions (as quantified by AUC values) does not increase as sample size increases after 1400. Second, even with much smaller sample size, class-specific methods, especially EO method, can perform better than or the same as CA. For example, AUC of EO with 1400 samples is very close to AUC of CA with 3000 samples. From a practical point of view, this is important as it shows a simple modification in modelling can save costs via decreasing sample size.

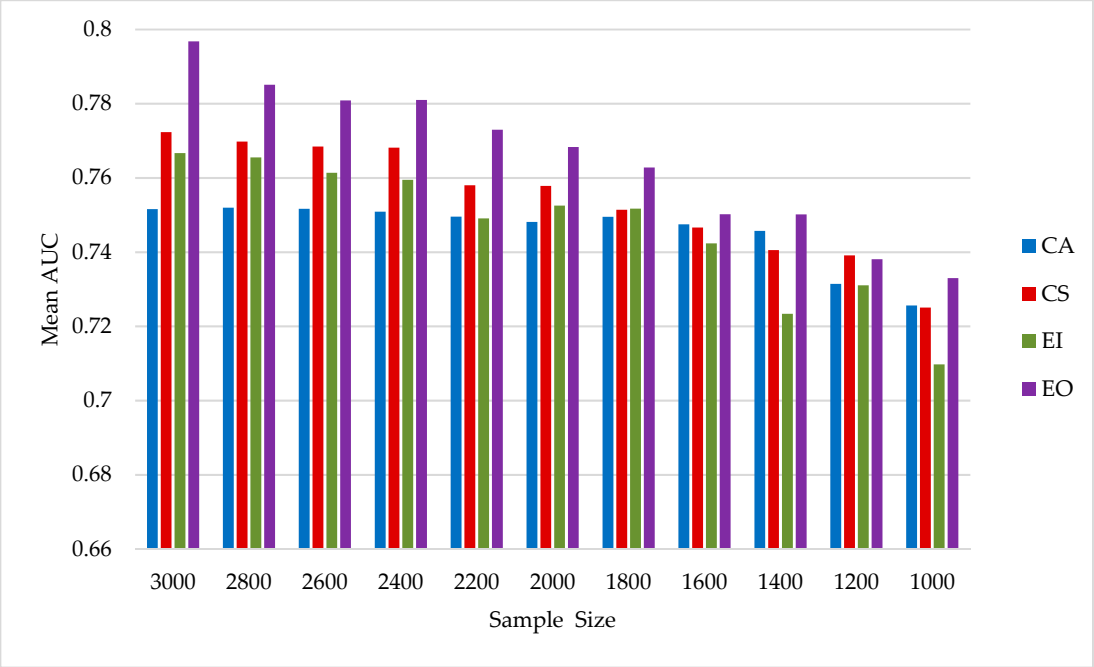

Figure 3. Mean AUC values for different methods with varying sample sizes.

T-test was undertaken based on the aforementioned sets of AUC values (means and standard deviation) for different methods with sample sizes. The results are shown in Table 7. Method EO is significantly more accurate than other methods, especially when the sample size is greater than 1800 pixels. Method EI is significantly less accurate than Method CS. When the sample size is no less than 2000, all kinds of stratified methods are significantly more accurate than CA. However, when sample size is less than 1800, the relativity among Methods EI, CS, and CA fluctuates.

Table 7. *T*-test results for significance in method pairwise differences based on mean AUC values with reduced sample sizes (*p*-value shown, * for significance at $\alpha = 0.10$, ** at $\alpha = 0.05$, and *** at $\alpha = 0.01$).

Sample Size	EO vs. EI	EO vs. CS	EO vs. CA	EI vs. CS	EI vs. CA	CS vs. CA
3000	—	—	—	—	—	—
2800	4.8×10^{-11} ***	1.8×10^{-6} ***	2.2×10^{-16} ***	0.927	4.6×10^{-8} ***	1.3×10^{-7} ***
2600	4.2×10^{-8} ***	2.1×10^{-5} ***	4.4×10^{-14} ***	0.991	2.0×10^{-4} ***	2.8×10^{-9} ***
2400	7.7×10^{-10} ***	4.1×10^{-5} ***	1.1×10^{-13} ***	0.999	4.0×10^{-5} ***	7.1×10^{-9} ***
2200	4.7×10^{-7} ***	2.2×10^{-4} ***	4.7×10^{-10} ***	0.974	0.550	0.008 ***
2000	1.4×10^{-4} ***	0.010 **	1.6×10^{-7} ***	0.883	0.085 *	0.006 ***
1800	0.004 ***	0.010 **	6.4×10^{-5} ***	0.473	0.242	0.323
1600	0.059 *	0.259	0.288	0.788	0.862	0.563
1400	5.9×10^{-7} ***	0.070 *	0.195	0.996	1.000	0.793
1200	0.118	0.566	0.156	0.937	0.529	0.098 *
1000	8.4×10^{-5} ***	0.101	0.123	0.992	0.991	0.521

4. Discussion

As shown in the results obtained in the study, the proposed class-heterogeneity-stratified method (applied for sampling and logistic modelling jointly) was confirmed to be the most accurate for estimating local accuracies in comparison with other methods. Sensitivity analyses also showed the proposed method's effectiveness and robustness, confirming its fair level of reliability. This study has met its goal of testing the proposed method's performance in local accuracy estimation, as the first step towards building an integrative framework for accuracy estimation and information refinement.

Below, some aspects of the work reported in the paper are reflected upon, with further work prospected briefly.

Firstly, in the paper, residuals of logistic regression predictions were not analyzed with respect to spatial correlation, nor was logistic-regression-kriging explored for mapping local accuracies as in [30]. In the logistic model in Equation (2), $p(x)$ represents the probability of correct classification (or agreement between map and reference class labels) at pixel x. $p(x)$ is actually the mean of a binary variable $I(x)$ indicating if x is correctly classified: $p(x) = E(I(x))$. Logistic-regression-kriging can thus be viewed as kriging with local means to get estimation of $I(x)$, with logistic regression predicting local means, while kriging transferring spatial information contained in residuals (i.e., $I(x) - p(x)$) from sampled locations to unsampled ones [46]. It (logistic-regression-kriging) certainly merits consideration for mapping local accuracies, especially when regression residuals are spatial correlated (hence should be incorporated for improved estimation of local accuracies). However, given the paper's future orientation to information refinement (after local accuracy estimation), it makes sense to perform kriging based on land-cover data concerned directly (rather than indicator data representing classification correctness) when pursuing data fusion in the future. Another reason for having not pursued kriging in the paper is the extra computational cost that would be incurred by implementing kriging after logistic regression, since sensitivity analysis, as a relatively novel aspect of this study, was already computationally expensive.

Secondly, in this study, double-stratified modelling in combination with double-stratified sampling was confirmed to be the most accurate for local accuracy estimation, given same sample sizes. However, it is worth exploring how sampling may be optimally configured (beyond double stratifications) with respect to information refinement [33,34] (the top priority in the future). Related to this is the issue of how we may figure out the optimal sample size for a specific study area given the budgets for reference data collection. Furthermore, it is important to devise methods for combined use of all reference data available to improve accuracy characterization and data fusion, regardless of with what designs (which may be more complex than those in this study) they were originally collected. We acknowledge that there is great room for improvements of and extensions to the work done in the paper, being aware that sampling is itself a topic of breadth and depth.

Thirdly, some technical aspects are worth further explorations. On one hand, given the facts that double-stratified modelling tends to become complicated with much more models to build than CA modelling and that sub-stratified models are very much homogenized over corresponding data sub-strata, it seems sensible to explore possible simplification of sub-stratified models without significantly compromising accuracy of estimation (e.g., using sub-stratum means). On the other hand, it is worth exploring the double-stratification method in time. For this, it is interesting to investigate how the double-stratification method may be used to characterize per-pixel accuracies in land-cover change [17,30].

Fourthly, we discuss the issue concerning threshold selection for defining heterogeneous vs. homogeneous sub-strata, which are essential for the proposed Method EO. As described in Section 2.2, the threshold for a sub-stratum being homogeneous was 4 pixels in a neighborhood of 3 by 3 pixels. This means that the class type of center pixel needs to be in majority (no less than 5/9) to claim it being in a relatively homogeneous neighborhood. Please note that homogeneity is defined as the number of pixels with the same class label as that of the center pixel in the focal neighborhood (see also Table 2). Clearly, unlike first level stratification by map classes that are fixed for a given map, sub-stratification into heterogeneous vs. homogeneous pixels segments in a stratum can be made on a more adaptive basis, as threshold selection is obviously related to and varies by the land-cover mosaic (cover types, patch shape, and landscape texture, etc.) depicted in the map being assessed. By adaptively selecting thresholds of homogeneity in sub-stratification, we can optimize sub-stratification by optimal thresholding to maximize reliability in estimated local accuracies under the constraints of sampling intensity and sample size. This issue (threshold selection) is certainly worth exploring in future research.

Fifthly, we discuss potentially useful methods for per-pixel accuracy estimation in soft (subpixel) classifications [2,5]. Soft classifications are often considered as a kind of fuzzy classifications. In other words, "fuzzy" is more general than "soft" in conceptual terms, as the former refers to the cases whereby classes themselves are vaguely defined (e.g., the severity of drought). However, for soft classifications representing subpixel proportions of candidate classes in individual pixels, their probabilistic interpretations seem to be more relevant. With this understanding, we assume numerical equivalence (or similarity, more correctly) between subpixel class proportions and fuzzy membership values without causing confusion in the following discussion. Comber et al. [32] represented one piece of pioneering work on per-pixel accuracy estimation for fuzzy/soft classifications, while Khatami et al. [47] was more recent contribution to the relevant literature. For such maps, per-pixel accuracy measures were differences between map and reference membership values (or class proportions) for a candidate class (denoted D) in [32,47] (absolute differences, $|D|$, were used in the former). In the papers by Comber et al. [32] and Khatami et al. [47], spatial interpolation was applied to generate surfaces of weighted moving window means of D's at sample locations, where weights are computed with distance-based kernels in a similar manner to geographically weighted regression (GWR). Clearly, the proposed method is not directly applicable to estimating local accuracies in fuzzy maps. To make the proposed method applicable to fuzzy classifications, two extensions are required. One concerns adaptions to regression modelling, the other is related to how heterogeneity is defined on fuzzy maps to better facilitate double-stratification on such maps. Regression modelling needs to consider the fact that accuracy measures applied to fuzzy maps are no longer probabilistic but continuous-valued D. Thus, regression analyses rather than logistic regression may be explored. See the work by Shortridge and Messina [48] for an example of analyses of continuous-valued errors in Shuttle Radar Topography Mission (SRTM) DEM and their associations with globally available topographic and land-cover variables across a wide range of landscapes in the United States, although it was not about classifications *per se*. On the other hand, definitions of heterogeneity vs. homogeneity should be reviewed in the context of fuzzy maps and their local accuracy modelling [49]. It seems that continua of heterogeneity-homogeneity are closely related to class proportions (or fuzziness in class memberships), although relations are not yet well understood. Once heterogeneity is defined with proper thresholds, double-stratification may be implemented: sub-stratifications of heterogeneity vs. homogeneity are based on the thresholds chosen while strata of prototype map classes are based on alpha-cuts [49,50] or maximum membership values (or dominant classes' proportions) [51].

Lastly but not the least, it should be recognized that there is issue of uncertainty related to estimated local accuracies. This is so because the models obtained (i.e., significant explanatory variables and model parameters) were conditional to the specific sample data given, as shown in Table 4, even to sample data with same sampling design and sample size (Table 7). Reference sample data quality [52] is also a factor, although reference data in this research were assumed to be accurate. More importantly, the explanatory variables used in this research were derived from map data that were known to be contaminated with unavoidable misclassification errors. This means that estimated local accuracies were subject to two-levels of uncertainties propagating: from map data to explanatory variables (i.e., contextual features or landscape pattern indices computed from a land-cover map being assessed) and from explanatory variables to logistic-modelling-based estimation. Local accuracy estimation (reported in the paper and elsewhere) is, thus, by no means perfect, no matter how sophisticated the methods employed are. It is important to develop and promote methods that not only depict spatially varying accuracies in land-cover information products but also support uncertainty analyses in predicted per-pixel accuracies.

We discuss further the aforementioned two-level uncertainties in the remainder of this section. As mentioned above, unless field-measured data that are sufficiently accurate are used [53], spatial analyses and modelling based on remote-sensing data and land-cover information estimated from them are subject to uncertainty. There is impressive literature on uncertainty in landscape pattern indices (or metrics) and analyses due to misclassification errors in land-cover maps [54–59]. As landscape

pattern indices were used as explanatory variables for logistic modelling of accuracy in this paper, existing methods in the literature listed above may be usefully explored for analyzing sensitivities of relevant pattern indices to misclassification errors.

However, we need to go further to analyze and quantify uncertainty in estimated local accuracies using map-data-derived pattern indices in future research. Relevant literature is rather limited, especially with respect to uncertainty in logistic-modelling-based accuracy estimation. Nevertheless, literature on error-in-variables in regression analysis may shed light on issues of two-level uncertainties, while simulation-based error modelling is well worth exploring as another promising methodology. Regarding regression modelling considering error-in-variables, Zhang et al. [60] and Fu et al. [61]) addressed error-in-variables issue in the context of forest inventory using linear regression analyses based on remote-sensing data that are known to suffer from errors. Literature on error-in-variables in the context of logistic regression is more relevant to furthering this research, as logistic regression is designed for binary response variables (e.g., agreement/disagreement between map and reference labels). The work by Carroll and Wand [62] and Yi et al. [63] may serve as good starting points for further research. On the other hand, simulation-based approaches (e.g., [59]) also merit consideration. We can simulate (land-cover) maps containing misclassification errors. These simulated maps can be used to generate a large sample of maps showing estimated local accuracies. Statistical summary and analyses on these maps of local accuracies can provide useful information about the effects of map inaccuracy on resultant per-pixel accuracy estimation, supporting uncertainty-informed local accuracy estimation and information refinement. Simulation-based methods are necessarily adapted to facilitate conditioning to reference sample data, to accommodate spatial correlation in misclassification errors in land-cover maps being assessed, and to promote mechanism-based uncertainty analyses [64].

5. Conclusions

Local accuracy characterization for land-cover information products is important for users and producers alike. In this paper, Method EO (a class-heterogeneity-stratified method) is proposed for sampling and modelling in the context local accuracy estimation. This method was compared with three alternative methods (Methods EI, CS, and CA) based on GlobeLand30 2010 land cover over Wuhan. Comparisons were also made under different sample sizes through sensitivity analysis. It was confirmed that Method EO generally yields the most accurate estimates of local accuracies with varying sample sizes, while Method CS performs with the second highest accuracy. The work accomplished in this paper holds great potentials for further research on local accuracy estimation and information refinement based on data fusion, given continuing proliferation of land-cover information products and growing accumulation of reference sample data for product validation.

As discussed in Section 4, although logistic modelling using class-heterogeneity-stratified data was advocated as a promising method for local accuracy estimation with a fair level of reliability confirmed, there exist issues deserving further research. The major issue concerns two-level uncertainty in Method EO and other alternatives, as explanatory variables used were actually map data and their derivatives (i.e., contextual features), which are subject to errors themselves. As conventionally understood, the reliability of resultant local accuracy estimates depends on various factors, such as sampling intensity and sample size, domains of explanatory variables (spatial vs. spectral), and strengths of empirically derived accuracy-context relationships. While these factors were analyzed or discussed to some extent in the past, potential effects of two-level uncertainty on local accuracy estimation have not received the kind of research attention they deserve. Advancements on methods for handling and incorporating two-level uncertainty in local accuracy mapping will bring our research on land-cover information validation and refinement to an elevated level of sophistication.

Author Contributions: J.Z. proposed the study, advised on the manuscript structure, and contributed to the manuscript writing and revision; W.Y. conducted the experiments and wrote the original draft; W.Z. helped with computing and programming; Y.W., D.L. and Y.X. contributed to acquisition of sample data. Constructive comments and suggestions from anonymous reviewers and external editor were received with thanks. Michael

F. Goodchild (Emeritus Professor of Geography at the University of California, Santa Barbara) and Roger Kirby (retiree from The University of Edinburgh, Scotland) have provided long-term advice about spatial uncertainty for J.Z., the principal author.

Funding: This research was funded by the National Natural Science Foundation of China (grant no. 41471375).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Friedl, M.A.; Sulla Menashe, D.; Tan, B.; Schneider, A.; Raman kutty, N.; Sibley, A.; Huang, X.M. MODIS Collection 5 global land cover: Algorithm refinements and characterization of new datasets. *Remote Sens. Environ.* **2010**, *114*, 168–182. [CrossRef]
2. DeFries, R.; Hansen, M.; Steininger, M.; Dubayeh, R.; Sohlberg, R.; Townshend, J. Subpixel forest cover in Central Africa from multisensor, multitemporal data. *Remote Sens. Environ.* **1997**, *60*, 228–246. [CrossRef]
3. Chen, J.; Ban, Y.; Li, S. Open access to Earth land-cover map. *Nature* **2015**, *514*, 434. [CrossRef]
4. Homer, C.; Dewitz, J.; Yang, L.; Jin, S.; Danielson, P.; Xian, G.; Coulston, J.; Herold, N.; Wickham, J.; Megown, K. Completion of the 2011 National Land Cover Database for the conterminous United States—Representing a decade of land cover change information. *Photogramm. Eng. Remote Sens.* **2015**, *81*, 345–354. [CrossRef]
5. Hansen, M.C.; Egorov, A.; Roy, D.P.; Potapov, P.; Ju, J.C.; Turubanova, S.; Kommareddy, I.; Loveland, T.R. Continuous fields of land cover for the conterminous United States using Landsat data: First results from the Web-Enabled Landsat Data (WELD) project. *Remote Sens. Lett.* **2011**, *2*, 279–288. [CrossRef]
6. Wickham, J.; Stehman, S.V.; Gass, L.; Dewitz, J.A.; Sorenson, D.G.; Granneman, B.J.; Poss, R.V.; Baer, L.A. Thematic accuracy assessment of the 2011 National Land Cover Database (NLCD). *Remote Sens. Environ.* **2017**, *191*, 328–341. [CrossRef]
7. Congalton, R.G. Using spatial autocorrelation analysis to explore the errors in maps generated from remotely sensed data. *Photogramm. Eng. Remote Sens.* **1988**, *54*, 587–592.
8. Wang, Y.; Zhang, J.; Liu, D.; Yang, W.; Zhang, W. Accuracy assessment of GlobeLand30 2010 land cover over China based on geographically and categorically stratified validation sample data. *Remote Sens.* **2018**, *10*, 1213. [CrossRef]
9. Foody, G.M. Status of land cover classification accuracy assessment. *Remote Sens. Environ.* **2002**, *80*, 185–201. [CrossRef]
10. Herold, M.; Mayaux, P.; Woodcock, C.E.; Baccini, A.; Schmullius, C. Some challenges in global land cover mapping: An assessment of agreement and accuracy in existing 1km datasets. *Remote Sens. Environ.* **2008**, *112*, 2538–2556. [CrossRef]
11. McRoberts, R.E. Satellite image-based maps: Scientific inference or pretty pictures? *Remote Sens. Environ.* **2011**, *115*, 715–724. [CrossRef]
12. Wickham, J.D.; Stehman, S.V.; Gass, L.; Dewitz, J.; Fry, J.A.; Wade, T.G. Accuracy assessment of NLCD 2006 land cover and impervious surface. *Remote Sens. Environ.* **2013**, *130*, 294–304. [CrossRef]
13. Olofsson, P.; Foody, G.M.; Herold, M.; Stehman, S.V.; Woodcock, C.E.; Wulder, M.A. Good practices for estimating area and assessing accuracy of land change. *Remote Sens. Environ.* **2014**, *148*, 42–57. [CrossRef]
14. Stehman, S.V.; Czaplewski, R.L. Design and analysis for thematic map accuracy assessment: Fundamental principles. *Remote Sens. Environ.* **1998**, *64*, 331–344. [CrossRef]
15. Feng, M.; Sexton, J.O.; Huang, C.; Anand, A.; Channan, S.; Song, X.P.; Song, D.X.; Kim, D.H.; Noojipady, P.; Townshend, J.R. Earth science data records of global forest cover and change: Assessment of accuracy in 1990, 2000, and 2005 epochs. *Remote Sens. Environ.* **2016**, *184*, 73–85. [CrossRef]
16. Burnicki, A.C. Modeling the probability of misclassification in a map of land cover change. *Photogramm. Eng. Remote Sens.* **2011**, *77*, 39–49. [CrossRef]
17. Wickham, J.; Stehman, S.V.; Homer, C.G. Spatial patterns of the United States National Land Cover Dataset (NLCD) land-cover change thematic accuracy (2001-2011). *Int. J. Remote Sens.* **2018**, *39*, 1729–1743. [CrossRef] [PubMed]
18. Smith, J.H.; Stehman, S.V.; Wickham, J.D. Impacts of patch size and land-cover heterogeneity on thematic image classification accuracy. *Photogramm. Eng. Remote Sens.* **2002**, *68*, 65–70.

19. Smith, J.H.; Stehman, S.V.; Wickham, J.D.; Yang, L. Effects of landscape characteristics on land-cover class accuracy. *Remote Sens. Environ.* **2003**, *84*, 342–349. [CrossRef]
20. Van Oort, P.A.J.; Bregt, A.K.; de Bruin, S.; de Wit, A.J.W.; Stein, A. Spatial variability in classification accuracy of agricultural crops in the Dutch national land-cover database. *Int. J. Geogr. Inf. Sci.* **2004**, *18*, 611–626. [CrossRef]
21. Chen, Y.; Song, X.; Wang, S.; Huang, J.; Mansaray, L.R. Impacts of spatial heterogeneity on crop area mapping in Canada using MODIS data. *ISPRS J. Photogramm. Remote Sens.* **2016**, *119*, 451–461. [CrossRef]
22. Lechner, A.M.; Stein, A.; Jones, S.D.; Ferwerda, J.G. Remote sensing of small and linear features: Quantifying the effects of patch size and length, grid position and detectability on land cover mapping. *Remote Sens. Environ.* **2009**, *113*, 2194–2204. [CrossRef]
23. Comber, A.J. Geographically weighted methods for estimating local surfaces of overall, user and producer accuracies. *Remote Sens. Lett.* **2013**, *4*, 373–380. [CrossRef]
24. Khatami, R.; Mountrakis, G.; Stehman, S.V. Mapping per-pixel predicted accuracy of classified remote sensing images. *Remote Sens. Environ.* **2017**, *191*, 156–167. [CrossRef]
25. Foody, G.M. Local characterization of thematic classification accuracy through spatially constrained confusion matrices. *Int. J. Remote Sens.* **2005**, *26*, 1217–1228. [CrossRef]
26. Yang, Y.K.; Xiao, P.F.; Feng, X.Z.; Li, H.X. Accuracy assessment of seven global land cover datasets over China. *ISPRS J. Photogramm. Remote Sens.* **2017**, *125*, 156–173. [CrossRef]
27. Comber, A.; Brunsdon, C.; Charlton, C.; Harris, P. Geographically weighted correspondence matrices for local error reporting and change analyses: Mapping the spatial distribution of errors and change. *Remote Sens. Lett.* **2017**, *8*, 234–243. [CrossRef]
28. Steele, B.M.; Winne, J.C.; Redmond, R.L. Estimation and Mapping of Misclassification Probabilities for Thematic Land Cover Maps. *Remote Sens. Environ.* **1998**, *66*, 192–202. [CrossRef]
29. Park, N.W.; Kyriakidis, P.C.; Hong, S.Y. Spatial estimation of classification accuracy using indicator kriging with an image-derived ambiguity index. *Remote Sens.* **2016**, *8*, 320. [CrossRef]
30. Zhang, J.; Mei, Y. Integrating logistic regression and geostatistics for user-oriented and uncertainty-informed accuracy characterization in remotely-sensed land cover change information. *ISPRS Int. J. Geo-Inf.* **2016**, *5*, 113. [CrossRef]
31. Steele, B.M. Maximum posterior probability estimators of map accuracy. *Remote Sens. Environ.* **2005**, *99*, 254–270. [CrossRef]
32. Comber, A.; Fisher, P.F.; Brunsdon, C.; Khmag, A. Spatial analysis of remote sensing image classification accuracy. *Remote Sens. Environ.* **2012**, *127*, 237–246. [CrossRef]
33. See, L.; Schepaschenko, D.; Lesiv, M.; McCallum, I.; Fritz, S.; Comber, A.; Perger, C.; Schill, C.; Zhao, Y.; Maus, V.; et al. Building a hybrid land cover map with crowdsourcing and geographically weighted regression. *ISPRS J. Photogramm. Remote Sens.* **2015**, *103*, 48–56. [CrossRef]
34. Tsendbazar, N.E.; Bruin, S.D.; Fritz, S.; Herold, M. Spatial accuracy assessment and integration of global land cover datasets. *Remote Sens.* **2015**, *7*, 15804–15821. [CrossRef]
35. Liu, M.; Cao, X.; Li, Y.; Chen, J.; Chen, X.H. Method for land cover classification accuracy assessment considering edges. *Sci. China-Earth Sci.* **2016**, *59*, 2318–2327. [CrossRef]
36. Sweeney, S.; Evans, T.P. An edge-oriented approach to thematic map error assessment. *Geocarto Int.* **2012**, *27*, 31–56. [CrossRef]
37. Cochran, W.G. *Sampling Techniques*, 3rd ed.; Wiley: New York, NY, USA, 1977; ISBN 978-0471162407.
38. O'Neill, R.V.; Krummel, J.R.; Gardner, R.H.; Sugihara, G.; Jackson, B.; DeAngelis, D.L.; Milne, B.T.; Turner, M.G.; Zygmunt, B.; Christensen, S.W.; et al. Indices of landscape pattern. *Landsc. Ecol.* **1988**, *1*, 153–162. [CrossRef]
39. Riitters, K.H.; Oneill, R.V.; Wickham, J.D.; Jones, K.B. A note on contagion indices for landscape analysis. *Landsc. Ecol.* **1996**, *11*, 197–202. [CrossRef]
40. Hosmer, D.W.; Lemeshow, S.; Sturdivant, R.X. *Applied Logistic Regression*, 3rd ed.; Wiley: New York, NY, USA, 2013; pp. 81–82. ISBN 9781-118548356.
41. Swets, K. Measuring the accuracy of diagnostic systems. *Science* **1988**, *240*, 1285–1293. [CrossRef] [PubMed]
42. Pearce, J.; Ferrier, S. Evaluating the predictive performance of habitat models developed using logistic regression. *Ecol. Model.* **2000**, *133*, 225–245. [CrossRef]

43. Luoto, M.; Marmion, M.; Hjort, J. Assessing spatial uncertainty in predictive geomorphological mapping: A multi-modelling approach. *Comput. Geosci.* **2010**, *36*, 355–361. [CrossRef]
44. Mas, J.O.; Filho, B.S.; Pontius, R.G.; Gutiérrez, M.F.; Rodrigues, H. A suite of tools for ROC analysis of spatial models. *ISPRS Int. J. Geo-Inf.* **2013**, *2*, 869–887. [CrossRef]
45. Robin, X.; Turck, N.; Hainard, A.; Tiberti, N.; Lisacek, F.; Sanchez, J.C.; Müller, M. pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinform.* **2011**, *12*, 77. [CrossRef] [PubMed]
46. Hengl, T.; Heuvelink, G.B.M.; Stein, A. A generic framework for spatial prediction of soil variables based on regression-kriging. *Geoderma* **2004**, *120*, 75–93. [CrossRef]
47. Khatami, R.; Mountrakis, G.; Stehman, S.V. Predicting individual pixel error in remote sensing soft classification. *Remote Sens. Environ.* **2017**, *199*, 401–414. [CrossRef]
48. Shortridge, A.; Messina, J. Spatial structure and landscape associations of SRTM error. *Remote Sens. Environ.* **2011**, *115*, 1576–1587. [CrossRef]
49. Arnot, C.; Fisher, P.F.; Wadsworth, R.; Wellens, J. Landscape metrics with ecotones: Pattern under uncertainty. *Landsc. Ecol.* **2004**, *19*, 181–195. [CrossRef]
50. Zhang, J.; Stuart, N. Fuzzy methods for categorical mapping with image-based land cover data. *Int. J. Geogr. Inf. Sci.* **2001**, *15*, 175–195. [CrossRef]
51. Stehman, S.V.; Arora, M.K.; Kasetkasem, T.; Varshney, P.K. Estimation of fuzzy error matrix accuracy measures under stratified random sampling. *Photogramm. Eng. Remote Sens.* **2007**, *73*, 165–173. [CrossRef]
52. Foody, G.; Pal, M.; Rocchini, D.; Garzon-Lopez, C.; Bastin, L. The sensitivity of mapping methods to reference data quality: Training supervised image classifications with imperfect reference data. *ISPRS Int. J. Geo-Inf.* **2016**, *5*, 199. [CrossRef]
53. Johnston, M.R.; Elmore, A.J.; Mokany, K.; Lisk, M.; Fitzpatrick, M.C. Field-measured variables outperform derived alternatives in Maryland stream biodiversity models. *Divers. Distrib.* **2017**, *23*, 1054–1066. [CrossRef]
54. Wickham, J.D.; O'neill, R.V.; Riitters, K.H.; Wade, T.G.; Jones, K.B. Sensitivity of selected landscape pattern metrics to land-cover misclassification and differences in land-cover composition. *Photogramm. Eng. Remote Sens.* **1997**, *63*, 397–401.
55. Hess, G.; Bay, J.M. Generating confidence intervals for composition-based landscape indexes. *Landsc. Ecol.* **1997**, *12*, 309–320. [CrossRef]
56. Hunsaker, C.T.; Goodchild, M.F.; Friedl, M.A.; Case, T.J. *Spatial Uncertainty in Ecology*; Springer: New York, NY, USA, 2001; ISBN 978-0387988894.
57. Fang, S.; Gertner, G.; Wang, G.; Anderson, A. The impact of misclassification in land use maps in the prediction of landscape dynamics. *Landsc. Ecol.* **2006**, *21*, 233–242. [CrossRef]
58. Langford, W.T.; Gergel, S.E.; Dietterich, T.G.; Cohen, W. Map misclassification can cause large errors in landscape pattern indices: Examples from habitat fragmentation. *Ecosystems* **2006**, *9*, 474–488. [CrossRef]
59. Kleindl, W.J.; Powell, S.L.; Hauer, F.R. Effect of thematic map misclassification on landscape multi-metric assessment. *Environ. Monit. Assess.* **2015**, *187*, 321. [CrossRef] [PubMed]
60. Zhang, W.; Ke, Y.; Quackenbush, L.J.; Zhang, L.J. Using error-in-variable regression to predict tree diameter and crown width from remotely sensed imagery. *Can. J. For. Res.* **2010**, *40*, 1095–1108. [CrossRef]
61. Fu, L.; Liu, Q.; Sun, H.; Wang, Q.; Li, Z.; Chen, E.; Pang, Y.; Song, X.; Wang, G. Development of a system of compatible individual tree diameter and aboveground biomass prediction models using error-in-variable regression and airborne LiDAR data. *Remote Sens* **2018**, *10*, 325. [CrossRef]
62. Carroll, R.J.; Wand, M.P. Semi-parametric estimation in logistic measurement error models. *J. R. Stat. Soc. Ser. B Methodol.* **1990**, *53*, 573–587.
63. Yi, G.Y.; Ma, Y.Y.; Spiegelman, D.; Carroll, R.J. Functional and structural methods with mixed measurement error and misclassification in covariates. *J. Am. Stat. Assoc.* **2015**, *110*, 681–696. [CrossRef] [PubMed]
64. Goodchild, M.; Zhang, J.; Kyriakidis, P. Discriminant models of uncertainty in nominal fields. *Trans. GIS* **2009**, *13*, 7–23. [CrossRef]