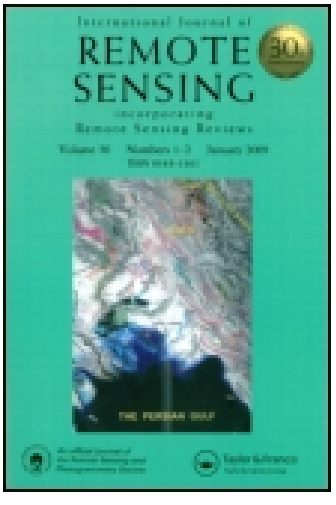

International Journal of Remote Sensing

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/tres20

Assessing the impact of training sample selection on accuracy of an urban classification: a case study in Denver, Colorado

Huiran Jin[a], Stephen V. Stehman[b] & Giorgos Mountrakis[a]

[a] Department of Environmental Resources Engineering, State University of New York College of Environmental Science and Forestry, Syracuse, NY 13210, USA

[b] Department of Forest and Natural Resources Management, State University of New York College of Environmental Science and Forestry, Syracuse, NY 13210, USA

Published online: 06 Mar 2014.

To cite this article: Huiran Jin, Stephen V. Stehman & Giorgos Mountrakis (2014) Assessing the impact of training sample selection on accuracy of an urban classification: a case study in Denver, Colorado, International Journal of Remote Sensing, 35:6, 2067-2081

To link to this article: http://dx.doi.org/10.1080/01431161.2014.885152

International Journal of Remote Sensing, 2014
Vol. 35, No. 6, 2067–2081, http://dx.doi.org/10.1080/01431161.2014.885152

Assessing the impact of training sample selection on accuracy of an urban classification: a case study in Denver, Colorado

Huiran Jin[a]*, Stephen V. Stehman[b], and Giorgos Mountrakis[a]

[a]*Department of Environmental Resources Engineering, State University of New York College of Environmental Science and Forestry, Syracuse, NY 13210, USA;* [b]*Department of Forest and Natural Resources Management, State University of New York College of Environmental Science and Forestry, Syracuse, NY 13210, USA*

(*Received 12 October 2013; accepted 3 January 2014*)

Understanding the factors that influence the performance of classifications over urban areas is of considerable importance to applications of remote-sensing-derived products in urban design and planning. We examined the impact of training sample selection on a binary classification of urban and nonurban for the Denver, Colorado, metropolitan area. Complete coverage reference data for urban and nonurban cover were available for the year 1997, which allowed us to examine variability in accuracy of the classification over multiple repetitions of the training sample selection and classification process. Four sampling designs for selecting training data were evaluated. These designs represented two options for stratification (spatial and class-specific) and two options for sample allocation (proportional to area and equal allocation). The binary urban and nonurban classification was obtained by employing a decision tree classifier with Landsat imagery. The decision tree classifier was applied to 1000 training samples selected by each of the four training data sampling designs, and accuracy for each classification was derived using the complete coverage reference data. The allocation of sample size to the two classes had a greater effect on classifier performance than the spatial distribution of the training data. The choice of proportional or equal allocation depends on which accuracy objectives have higher priority for a given application. For example, proportionally allocating the training sample to urban and nonurban classes favoured user's accuracy of urban whereas equally allocating the training sample to the two classes favoured producer's accuracy of urban. Although this study focused on urban and nonurban classes, the results and conclusions likely generalize to any binary classification in which the two classes represent disproportionate areas.

1. Introduction

Mapping urban land in a timely and accurate manner is indispensable to a multitude of planning applications. Remote sensing is considered an essential technology in urban-related environmental and socioeconomic studies due to its ability to provide spatially detailed and temporally efficient information on urban land cover. Over the past two decades, researchers have become increasingly interested in using remotely sensed imagery to address urban and suburban problems (Jacquin, Misaova, and Gay 2008). A number of algorithms have been proposed to automatically extract urban land uses and changes (e.g. Bauer, Loeffelholz, and Wilson 2007; Weng, Hu, and Liu 2009; Jin and Mountrakis 2013), and to model urban sprawl patterns (e.g. White, Engelen, and Uljee 1997; Pijanowski et al. 2005; Sesnie et al. 2008; Huang, Zhang, and Wu 2009). Detailed

*Corresponding author. Email: hjin02@syr.edu

reviews of urban monitoring classifiers and urban growth prediction models are provided by Weng (2012) and Triantakonstantis and Mountrakis (2012), respectively.

The value of thematic maps constructed from remotely sensed data is clearly a function of classification accuracy (Foody 2002). The quality of a supervised classification is impacted by the sample of reference data selected to train the classifier. Although much effort has been devoted to investigating sampling designs and estimation protocols for accuracy assessment of the completed classification (e.g. Stehman 2009), less emphasis has been placed on investigating the impact of the sample selection protocol for obtaining training data. For example, Edwards et al. (2006) explored the advantages of implementing a probability sampling design relative to a non-probabilistic, purposive selection of training data in modelling and predicting species distributions, and Zhen et al. (2013) examined the effects of spatial dependence between training and validation samples collected from reference polygons on classification accuracy and accuracy estimates.

Sample size is another attribute of a training sample that may greatly affect classification accuracy (Foody et al. 2006). For example, for conventional statistical classifiers such as the maximum likelihood classification algorithm, the suggested number of training samples per class should be at least 10–30 times the number of wavebands used (Piper 1992; Mather 1999). Critically, however, this heuristic rule is often enforced in determining the per-class sample size, irrespective of the characteristics of the study site or the aim of the classification analysis (van Niel, McVicar, and Datt 2005; Foody et al. 2006). It is, therefore, an attractive proposition to investigate the allocation of training sample size to classes such that a more representative training set can be obtained with desired information for target objectives.

The main objective of this research is to determine the impact of the sampling design used to select training data on the accuracy of the classification. Our investigation takes advantage of a complete coverage reference database of urban change between 1977 and 1997 for the Denver, Colorado (USA), metropolitan area. Four sampling designs representing two options for stratification (spatial and class-specific) and two options for sample allocation (proportional to area and equal allocation) were implemented for selecting training data. The evaluation of the different training sample selection protocols was based on the accuracy of the urban/nonurban classification for 1997, obtained by applying the classification algorithm to 1000 different training samples selected by each of the four sampling protocols. By repeating the training sample selection and classification process many times, we were able to observe the variation in accuracy over the set of potential training samples that might arise. In practice, only one sample would be used for training the classifier, but the reality is that different training samples will yield different classifications with different accuracies and our results provide a rare glimpse into the magnitude of this variability. The four sampling designs evaluated are all probability sampling designs (Stehman 2009). It is strongly recommended to use a probability sampling design for estimating accuracy because rigorous design-based inference can be established if such a design has been implemented (Stehman 2000). In contrast, it is not a requirement that the training sample be selected via a probability sampling protocol. However, in some applications it may be necessary to select the training sample and validation sample concurrently. In such cases, a practical approach is to select a sample via a probability sampling protocol and then randomly split that sample into independent training and validation samples. This would yield a validation sample that satisfied the probability sampling criterion required to invoke design-based inference for the accuracy estimates obtained from the validation sample.

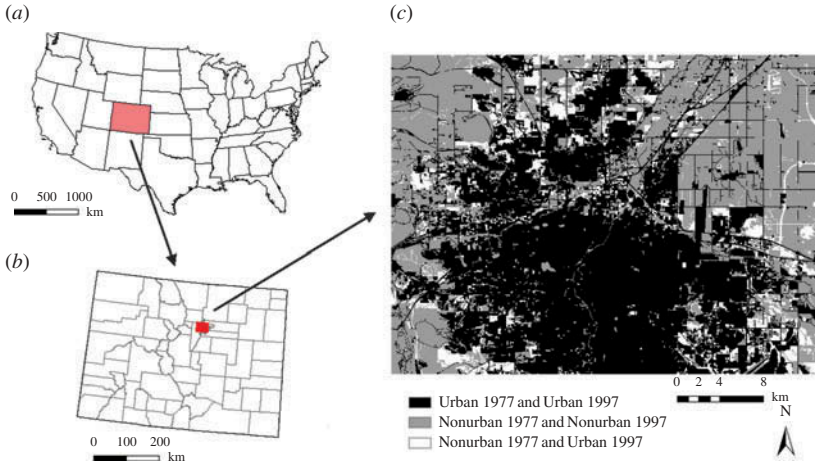

Figure 1. Location of the study area. Maps of (*a*) USA, (*b*) State of Colorado and (*c*) urban change (white area) in the Denver metropolitan area from 1977 to 1997.

2. Study area and data

The area selected for this study covers the Denver, Colorado, metropolitan area (Figure 1). The study area is centred at 39° 48′ 30" N latitude and 104° 59′ 5" W longitude with a ground extent of 39.6 km × 29.7 km (approximately 1176 km^2). According to land-use maps provided by the US Geological Survey Rocky Mountain Mapping Center (USGS 2003), urban land use reached 62% of the study area in 1997 compared with 51% in 1977, reflecting the rapid urban growth that Denver experienced in the late twentieth century.

Figure 1 displays urban growth in the Denver metropolitan area from 1977 to 1997, where urban growth was derived by comparing the bitemporal land-use data sets produced from manually digitized aerial photographs over the entire site. To focus on relatively large-scale patterns, all data were rasterized to a spatial resolution of 30 m × 30 m. If a pixel overlapped with any digitized data relating to urban development (i.e. residential areas, commercial/light industries, institutions, communication and utilities, heavy industries, entertainment/recreation, and roads and other transportation infrastructure), it was then assigned to the urban class; otherwise, the pixel was considered as nonurban. Urban areas in 1977 (black pixels in Figure 1) were excluded so that we could assess the performance of the classifier to map more recently developed urban areas (i.e. areas that had transitioned to urban within 20 years prior to the 1997 classification) and to maintain a greater disparity between the area proportions of the two classes. Consequently, white (defined as *urban*) and grey (defined as *nonurban*) pixels constitute the population of reference data on which the experiments were conducted. The reference set is assumed to be a 100% accurate representation of reality with a total of 638,204 pixels, 22.2% of which were identified as urban and the remaining 77.8% as nonurban.

A Level 1T Landsat 5 Thematic Mapper (TM) image acquired on 26 June 1997 was downloaded from the USGS Earth Resources Observation and Science Center (EROS) website (source: http://glovis.usgs.gov/), and a subscene with the same ground coverage as shown in Figure 1 was used to provide spectral information (Figure 2). The image consists of the six reflective bands (bands 1−5 and 7) and 1320 × 990 pixels with a constant pixel size of 30 m × 30 m.

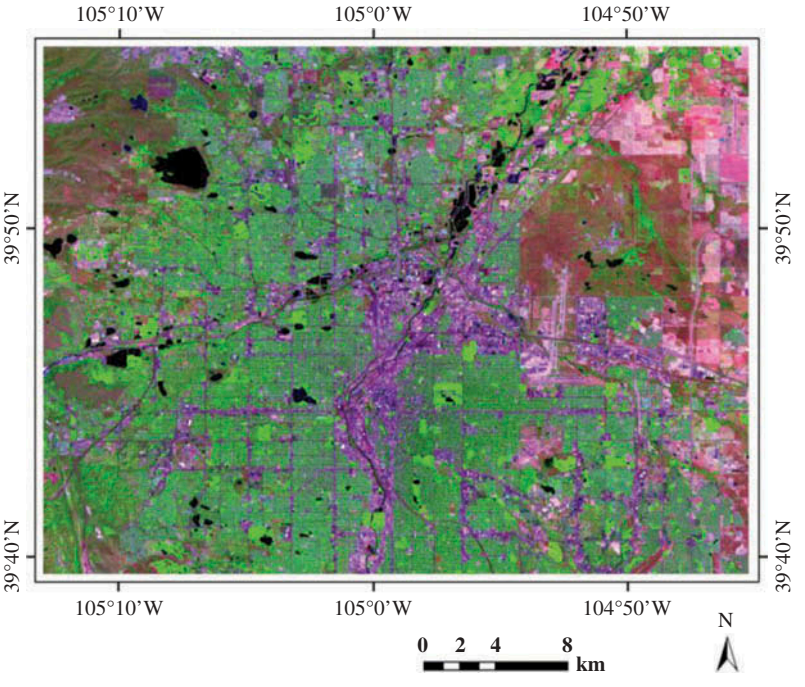

Figure 2. Landsat-5 TM false-colour image (RGB bands 7, 4, and 3) of the study area from 26 June 1997.

3. Methods

3.1. *Sampling designs for training data selection*

Training data should be representative of the study area and of the classes in the classification scheme. Because urban is often a relatively rare class covering only a small proportion of the landscape, stratified sampling is an obvious consideration to ensure that a specified sample size is obtained for each rare class (Biging, Colby, and Congalton 1998). The four sampling designs investigated in this study all incorporate stratification, and a pixel is chosen as the sampling unit.

All four of the stratified designs investigated used the urban and nonurban classes to form the strata. Two different allocations of sample size to strata were investigated. One allocation was proportional to the stratum area (the design labelled as '*Prop*') and the other allocation was equal (the design labelled as '*Eq*'), in which the sample size was evenly split between the urban and nonurban classes. Proportional allocation provides no assurance that the rare urban class would be adequately represented in the training sample because proportional allocation is an equal probability design in which all pixels have the same probability ($\pi_h = n_h/N_h$) of being included in the sample, where n_h and N_h are the sample and population size in stratum h, respectively (because the sample size n_h of stratum h must be a whole number, it may not be possible to choose n_h so that exact proportionality occurs). Equal allocation in which n_h is the same for all strata can be used to increase the sample size for the rare urban class, so it leads to unequal inclusion probabilities for pixels from different strata. That is, for a pixel in stratum h, the inclusion probability is $\pi_h = n_h/N_h = n/2N_h$ and thus π_h varies with the stratum size, N_h. Although for equal allocation the sample size in each stratum is the same, $n_h = n/2$, the inclusion

probabilities for the two strata will differ if the N_h differ. For both designs, *Prop* and *Eq*, simple random sampling was applied to select the sample of pixels from each stratum.

The other two sampling designs evaluated included a second level of stratification in addition to the stratification by the urban and nonurban classes. The second level of stratification was spatial stratification implemented by partitioning the region of interest into equal-area, square tessellation cells, and then distributing the total sample size equally among each of these pre-defined spatial strata (i.e. the sample size n_i in spatial stratum i was n/H, where n = total sample size for the entire study region, and H = number of spatial strata). Within each spatial stratum, a stratified random sample (using the urban and nonurban strata) was selected. We evaluated both proportional and equal allocation for the sample size allocation to the urban and nonurban strata. That is, for the design labelled '*SpatialProp*', the sample size was allocated to each class proportional to the areal coverage in the reference set, with the constraint that each spatial stratum received an equal total sample size. For example, if the urban and nonurban classes comprised 20% and 80% of the area of the entire region, respectively, the sample allocation in each spatial stratum would be 20% urban and 80% nonurban. For the design labelled '*SpatialEq*', the sample size in each spatial stratum was equally allocated to the urban and nonurban strata. Similar to the *Eq* design, the *SpatialEq* design increases the sample size from the rare urban class. Compared with the *Prop* and *Eq* sampling designs, each sample obtained by the *SpatialProp* and *SpatialEq* designs is guaranteed to be spatially well distributed. For both spatially stratified designs, the general form for the inclusion probability of a pixel from stratum h (where $h = 1$ for the nonurban stratum and $h = 2$ for the urban stratum) in the spatial stratum i is $\pi_{hi} = n_{hi}/N_{hi}$, where n_{hi} is the number of pixels sampled per class h in spatial stratum i and N_{hi} is the number of pixels per class h in spatial stratum i. Because $n_{hi} = nP_h/H$, the inclusion probability of a pixel with class h can be expressed as $\pi_{hi} = n_{hi}/N_{hi} = (nP_h/H)/N_{hi}$, where H is the number of spatial strata, and P_h indicates the class allocation technique adopted (e.g. for our study, $P_h = 0.78$ for the nonurban class and $P_h = 0.22$ for the urban class in the *SpatialProp* design, and $P_h = 0.50$ for both classes in *SpatialEq*). Therefore, π_{hi} varies spatially over the region. Higher inclusion probabilities are associated with more isolated pixels (i.e. spatial strata with smaller N_{hi}). The construction of the spatially stratified designs was motivated by the objective of testing whether classification performance can be improved with a spatially well-distributed training sample compared to the use of a sample that may yield groups or clusters of pixels in a few locations.

A simplified scenario shown in Figure 3 illustrates the four sampling designs for a hypothetical situation in which nonurban occupies 75% of the area and urban occupies 25%. Given a total sample size of 16, 12 nonurban pixels and 4 urban pixels will be selected following the designs of *Prop* and *SpatialProp* (proportional allocation) and 8 pixels will be sampled in each class by designs *Eq* and *SpatialEq* (equal allocation). Figures 3(*c*) and (*d*) illustrate the intuitively appealing spatial balance feature of the spatially stratified designs. In contrast, the stratified designs *Prop* and *Eq* are more likely to yield large unsampled gaps as observed in Figures 3(*a*) and (*b*).

In practice, it is necessary to determine the total training sample size along with an appropriate spatial extent for applying geographic stratification. In this study, we first used a large but reasonable size of 5000 pixels to ensure reliable classification in accordance with a previous experiment conducted by Jin and Mountrakis (2013), who tested a series of training sample sizes from 100 to 5000 pixels. A smaller training sample size of 1000 was also examined in this study, reflecting the circumstances in which limited time/effort can be devoted to training data selection. Meanwhile, the study area was divided into 48 non-overlapping 5 km × 5 km blocks (a block is also a spatial stratum in *SpatialEq* and

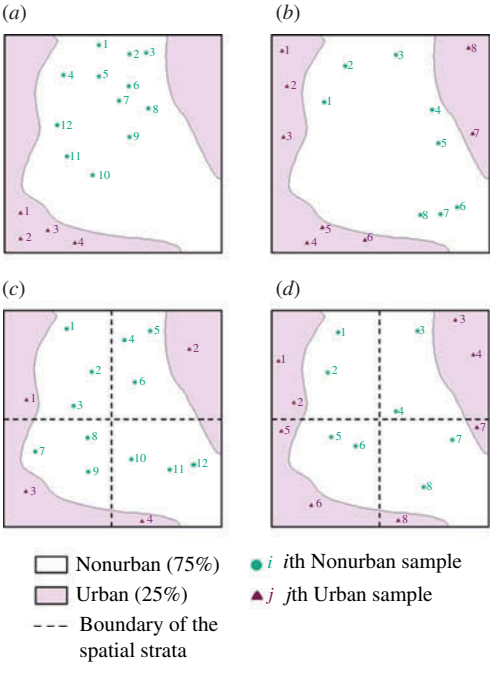

Figure 3. Example realizations of the four sampling designs (n = 16) for (*a*) *Prop*, (*b*) *Eq*, (*c*) *SpatialProp*, and (*d*) *SpatialEq*.

SpatialProp), with each block a square of 165 × 165 30 m pixels. A 5 km × 5 km block was chosen because it was large enough to ensure that in each block both urban and nonurban pixels were substantially present for training data selection irrespective of the proposed sampling design utilized and the sample size chosen. Consequently, there were 104–105 training samples per block for n = 5000 and 20–21 training samples per block for n = 1000 in the spatially stratified designs. Small variations in sample size resulted from the slightly inconsistent spatial allocation when splitting 5000 or 1000 samples into 48 blocks. In each block, training samples were allocated to the nonurban and urban strata proportionally by *SpatialProp* (i.e. 77.8% nonurban vs. 22.2% urban) and equally by *SpatialEq* (i.e. 50% nonurban vs. 50% urban). For designs *Prop* and *Eq*, however, proportional allocation and equal allocation were carried out throughout the entire study area without any spatial constraints.

To capture the range of outcomes that could result from different randomly selected training samples, 1000 independent replications of the process of selecting the training sample, implementing the classification algorithm, and assessing the accuracy of the classification were obtained for each sample size and sampling design. By repeating the full classification process many times, the potential variation in accuracy of the classification can be assessed by taking into account variation attributable to the randomization implemented in selecting the training sample.

3.2. *Classification*

A binary classification of urban and nonurban was performed based on the digital numbers of the six TM reflective bands as input attributes. From a practical perspective,

this is equivalent to an update of the existing 1977 urban land-use map in the Denver metropolitan area by identifying locations of new urban development using spectral information provided by the single-date 1997 TM imagery. Only one classifier was employed, and this classifier was applied to each of the 1000 training samples of each design to examine the variation in classification accuracy over different training samples. Because of the availability of the complete coverage reference data, we were able to evaluate this feature of a classification (variation over training samples), which is typically not possible to assess otherwise.

A decision (classification) tree methodology based on dichotomous partitioning (Breiman et al. 1984) was adopted because this classifier can be trained and executed efficiently (Pal and Mather 2003) with no implicit assumptions on data distribution (Hansen, Dubayah, and DeFries 1996). The basic procedure of tree construction is recursively splitting a training sample into smaller subdivisions (leaves) according to the rules defined at each branch. The algorithm was developed in the Matlab environment, followed by a 10-fold cross-validation and a pruning process cutting off the initially constructed tree to the level that produces the smallest tree that is within one standard error of the minimum-cost subtree (Breiman et al. 1984).

For any particular training sample, the outcome of the classification will vary due to the randomization incorporated in the 10-fold cross-validation implemented in the pruning process of the decision tree classifier. Thus two sources of variation in classifier performance are present, variation among training samples, and variation among realizations of the classifier applied to the same training sample. To assess the relative magnitudes of these different sources of variation, we conducted a small experiment in which 25 of the 1000 training samples were randomly chosen for the *Prop* sampling design and $n = 5000$. The classification process was repeated 1000 times for each of these 25 training samples, and an analysis of variance (ANOVA) was conducted using overall accuracy of each classification as the variable of interest. The variation among training samples was found to be statistically significant (an F-statistic of 2202 and a p-value <0.001 based on 24 and 24,975 degrees of freedom), indicating that the dominant source of variation in overall accuracy was attributable to different training samples. The variation among training samples accounted for an estimated 70% of the total variation in overall accuracy, and the within-training sample variation among replications of the classifier accounted for an estimated 30% of the total variation. The results reported (Section 4) include both sources of variation, so the variation in the accuracy measures observed is probably higher than would be the case for a classifier that produced a deterministic result for a given training sample.

3.3. *Accuracy assessment*

Assessing the accuracy of classification products is critical to determining the quality and potential utility of the information derived from remotely sensed data. Defined as the degree to which a thematic map agrees with reality or conforms to the 'truth' (Janssen and van der Wel 1994; Smits, Dellepiane, and Schowengerdt 1999), accuracy is a typical measure of the correctness of a classification (Foody 2002). Because a complete reference classification is available in this study, all metrics computed are parameters of the population rather than sample-based estimators. The complication of estimating classification accuracy from sample locations is therefore circumvented, along with the uncertainty attributable to sampling variability of validation data.

Table 1. Population error matrix for site-specific accuracy assessments.

		Reference class		
		Nonurban	Urban	Total
Map Class	Nonurban	p_{11}	p_{12}	p_{1+}
	Urban	p_{21}	p_{22}	p_{2+}
	Total	p_{+1}	p_{+2}	

Note: p_{ij} is the proportion of area in mapped class i and reference class j, and $p_{i+} = \sum_{j=1}^{2} p_{ij}$ and $p_{+j} = \sum_{i=1}^{2} p_{ij}$ denote the proportion of area mapped as class i and the true proportion of area of class j, respectively.

To evaluate an aggregate feature of accuracy of a classification, a nonsite-specific assessment was first conducted at the spatial scale of the 5 km × 5 km blocks, comparing the proportion of area classified as urban to the corresponding true proportion derived from the reference data. This assessment is motivated in the context of urban planning, as policy makers may use the percentage composition of urban mapped in each block as input into a model or statistical analysis to capture urban sprawl patterns at coarser resolutions. We chose the root mean square error (RMSE) as a descriptor of accuracy, given the quantitative and continuous character of urban proportions:

$$\mathrm{RMSE} = \sqrt{(1/H)\sum_{i=1}^{H}(x_i - y_i)^2}, \tag{1}$$

where x_i and y_i denote the classified and true percentage of urban pixels within the ith block, and H is the total number of blocks forming the partition of the study area ($H = 48$ in our case).

An error matrix approach (Story and Congalton 1986) was used to convey location-specific accuracy information. Table 1 represents a population error matrix for a binary classification of nonurban and urban created by a census of the reference classification over the entire study area. The main diagonal cells indicate correctly allocated pixels, and the off-diagonal cells indicate classification errors. Parameters such as overall, producer's, and user's accuracies, and the bias of classification for urban pixels were then calculated from the error matrix to quantify classification performance and to characterize errors. Bias of the urban class was defined as the difference between the map proportion of urban and the reference proportion of urban, $p_{2+} - p_{+2} = p_{21} - p_{12}$. Unlike the aforementioned RMSE, overall, user's, and producer's accuracies are characterized as 'site-specific' because the comparison between the reference and map classifications is conducted on a location-by-location or pixel-by-pixel basis (Stehman and Foody 2009).

4. Results

4.1. *Nonsite-specific assessment*

For each sampling design, 1000 training samples of each sample size (5000 and 1000) were selected and a binary classification of urban and nonurban was derived from each training sample. Results are first presented for accuracy characterized by agreement between the map percentage urban area and reference percentage urban area at the support

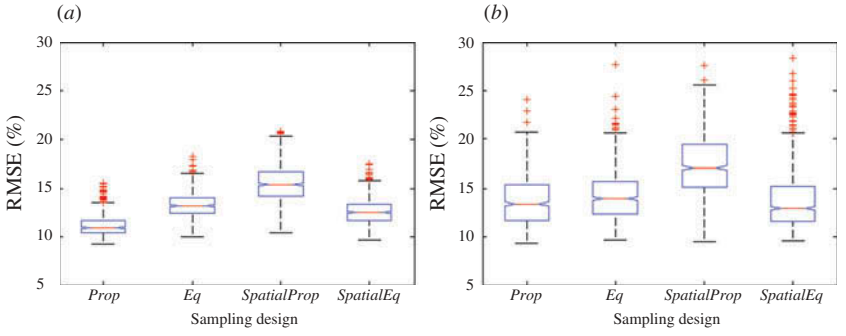

Figure 4. RMSE for percentage area of urban at the 5 km × 5 km support using different sampling designs and training sample sizes of (*a*) 5000 and (*b*) 1000.

of a 5 km × 5 km block. The RMSE for the set of 1000 classifications of each sampling design is graphically displayed by the boxplots in Figure 4.

Training samples selected by *Prop* with a sample size of 5000 resulted in the smallest interquartile range (IQR) and the lowest median RMSE (Figure 4 and Table 2) for the classifications aggregated to the 5 km × 5 km support. The two equally allocated sampling designs, *Eq* and *SpatialEq* had almost identical RMSE distributions to *Prop* for the n = 1000 sample size and a slightly higher median RMSE relative to *Prop* for the n = 5000 sample size. *SpatialProp* led to the highest RMSE and greatest variability (IQR) among the four training sampling designs, whereas the two equally allocated designs appeared more prone to RMSE outliers for the smaller sample size. The results confirm that a larger training sample size improves classifier performance because the median RMSE for n = 5000 was smaller than the median RMSE for n = 1000. Additionally, the IQR of RMSE was smaller for n = 5000, indicating that classifier performance varied less at the larger training sample size (Table 2). The relative performance of the four designs remained consistent for the two sample sizes evaluated.

4.2. *Site-specific assessment*

The results for the site-specific assessment of classifier performance for different training sampling designs are summarized by boxplots (Figures 5 and 6) of overall accuracy, bias, and class-specific producer's and user's accuracies. Generally, the sample sizes allocated to the urban and nonurban strata had a greater effect on classifier performance than the

Table 2. Median and interquartile range (IQR) of the distribution of RMSE for percentage urban (%) at the 5 km × 5 km support over the 1000 repetitions of the classification (training sample sizes of n = 5000 and n = 1000).

	Median RMSE		IQR	
Sampling Design	5000	1000	5000	1000
Prop	10.9	13.4	1.2	3.6
Eq	13.2	14.0	1.7	3.3
SpatialProp	15.4	17.1	2.5	4.3
SpatialEq	12.5	12.9	1.7	3.6

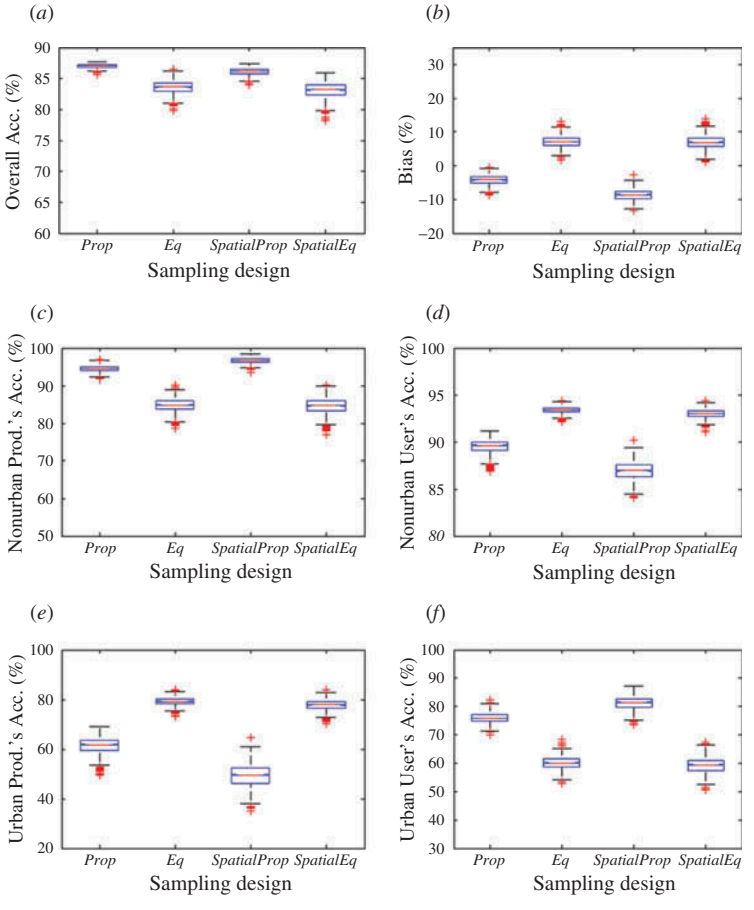

Figure 5. (*a*) Overall accuracy, (*b*) bias of area of urban, and (*c*) nonurban producer's accuracy, (*d*) nonurban user's accuracy, (*e*) urban producer's accuracy, and (*f*) urban user's accuracy for classifications developed from different training sampling designs with a sample size of 5000 (vertical axes chosen to match Figure 6).

spatial distribution of the training samples. That is, the differences between the performance of *Prop* and *SpatialProp* and between *Eq* and *SpatialEq* were smaller relative to the more substantial differences observed between *Prop* and *Eq* and between *SpatialProp* and *SpatialEq*. The differences in classifier performance were associated with the distribution of sample pixels to the two classes in the training data. For the proportional allocation designs, approximately 22% of the training sample was from the reference urban class, and for the equal allocation designs approximately 50% of the training sample was from the reference urban class.

Overall accuracies for the classifications based on *Prop* were higher by 3–5% than those achieved by *Eq* (Figures 5(*a*) and 6(*a*), Table 3). In contrast, producer's and user's accuracies (e.g. Figures 5(*c*)–(*f*)) differed more substantially among sampling designs, particularly for the rare urban class. For example, there was an increase of 17% in the median producer's accuracy of urban from 62% (*Prop*) to 79% (*Eq*), indicating an advantage to equal allocation on the criterion of urban producer's accuracy (Figure 5(*e*)).

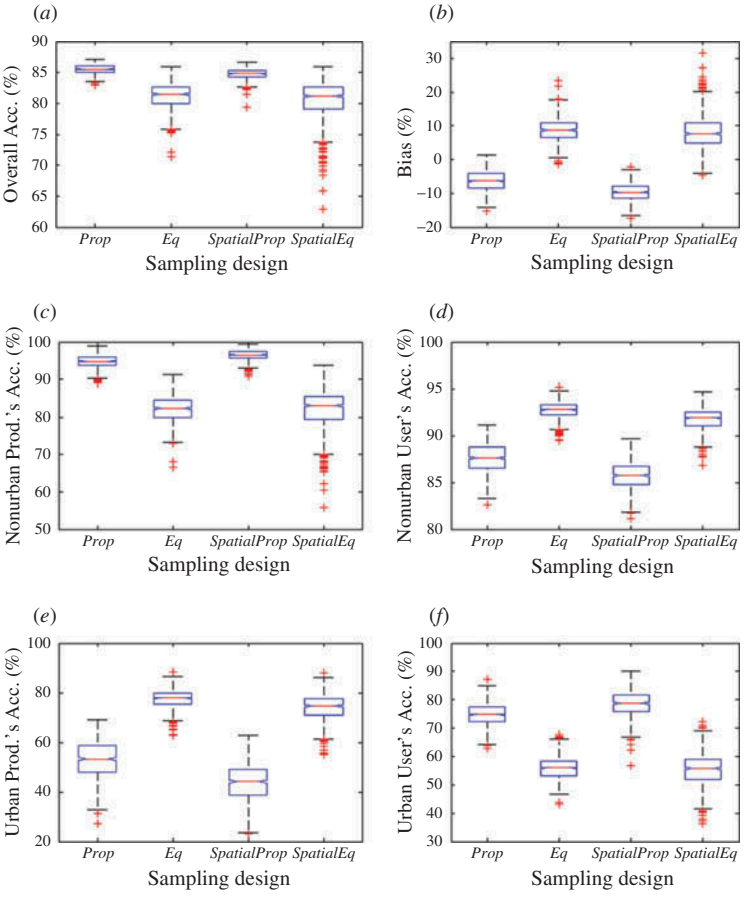

Figure 6. (*a*) Overall accuracy, (*b*) bias of area of urban, (*c*) nonurban producer's accuracy, (*d*) nonurban user's accuracy, (*e*) urban producer's accuracy, and (*f*) urban user's accuracy for classifications derived from 1000 repetitions of each of the four training sampling designs with a sample size of 1000.

Table 3. Median of site-specific accuracy measures (%) extracted from 1000 replications of each sampling design and training sample size (n = 5000 and n = 1000).

	Sample size (n)							
	5000				1000			
	Prop	*Eq*	*SpatialProp*	*SpatialEq*	*Prop*	*Eq*	*SpatialProp*	*SpatialEq*
Overall	87	84	86	83	86	81	85	81
Bias	−4.2	7.1	−8.7	6.9	−6.4	8.7	−9.7	7.6
Prod.'s acc. (nonurban)	94	85	97	85	95	82	97	83
User's acc. (nonurban)	90	94	87	93	88	93	86	92
Prod.'s acc. (urban)	62	79	50	78	53	78	44	75
User's acc. (urban)	76	60	81	59	75	56	79	56

The median of the bias (Figure 5(*b*)) varied from –8.7% (*SpatialProp*) to 7.1% (*Eq*). Selecting the training sample using one of the proportionally allocated designs (*Prop* and *SpatialProp*) resulted in an underestimate of the urban proportion (bias < 0), whereas *Eq* and *SpatialEq* resulted in an overestimate (bias > 0). Clearly the proportion of urban classified by this algorithm depended on the proportion of urban in the training sample.

Taking advantage of the larger sample size of nonurban training pixels, classifications developed from *Prop* had markedly higher producer's accuracies of nonurban (Figure 5(*c*)). Conversely, to obtain higher producer's accuracies of urban it would be preferable to employ *Eq* (Figure 5(*e*)), in which half of the training sample is guaranteed to be collected from the rare urban class for a better representation of urban signatures in the subsequent classification tree construction. As the complement to producer's accuracy, omission error of urban associated with *Eq* was relatively low, suggesting that fewer urban pixels were classified erroneously to the nonurban class (i.e. p_{12} is small in Table 1). Therefore, it was more likely that p_{12} comprised a small part of the row marginal of p_{1+}, leading to lower commission errors and higher user's accuracies of nonurban, as can be seen in Figure 5(*d*). Similar characteristics existed for producer's accuracy of nonurban (Figure 5(*c*)) and user's accuracy of urban (Figure 5(*f*)), as p_{21} was conditioned on the column and row marginal proportions, respectively. The contrasting patterns between producer's and user's accuracies for a given class are attributable to the description of classification performance from two perspectives – reference and map – with the use of different off-diagonal elements in an error matrix.

The designs of *Prop* and *SpatialProp* – where the sample allocation was proportional to the corresponding population size – were sensitive to the spatial location of training samples, whereas accuracies from *Eq* and *SpatialEq* were weakly influenced by the implementation of geographic stratification. The improvement achieved by the spatially balanced sampling of *SpatialProp* over *Prop* was more apparent in the user's accuracies of urban, with a near 6% increase (Figure 5(*f*)), indicating a higher probability that a pixel labelled as urban by the classifier was indeed found to be urban on the ground. However, such improvement was gained at the expense of ability to correctly detect urban pixels in the reference data, as median producer's accuracy of urban for *SpatialProp* declined by more than 12% relative to producer's accuracy of urban for *Prop* (Figure 5(*e*)).

The general patterns of the comparisons between different training sampling designs remained the same for the smaller sample size of $n = 1000$ (Figure 6) relative to the results for $n = 5000$ (Figure 5). The one noticeable difference in results for the different sample sizes was that the IQR of each accuracy measure was considerably larger for the smaller sample size, indicating the expected greater variability in classification accuracy over different realizations of the training samples selected. Training samples yielding low accuracy and high bias were much more prevalent at the smaller sample size as illustrated by the high number of outliers visible in the boxplots (Figure 6), with the *SpatialEq* design being the most prone to these inaccurate classification outcomes.

Table 3 summarizes the median (over the set of 1000 repetitions) of every accuracy measure per sample size and sampling design. For both sample sizes, *Prop* performed better than the other designs in terms of median overall accuracy and median bias of urban proportion. The highest median values of user's accuracy of nonurban and producer's accuracy of urban were achieved by *Eq*, while *SpatialProp* yielded the highest median producer's accuracy of nonurban and the highest median user's accuracy of urban. Generally, there was an approximate 1–3% decrease in overall accuracy when the training sample size was smaller. Together with the more 'repeatable' accuracy (i.e. smaller IQR)

of the classifiers for n = 5000 (Figure 5), the benefits of devoting more time/effort to collecting sufficient training samples was evident.

5. Discussion and conclusions

The quality of a map constructed from remotely sensed data is of considerable importance to its applications in scientific investigations and policy decisions. Understanding the factors that influence classification accuracies may guide the decision of which sampling design to implement for training given the objectives specified by the map producer. This study provides insight into the potential impact of training sample selection on the performance of classifications over urban areas that are subject to rapid expansion, as exhibited in Denver, Colorado. Although our study focused on an urban/nonurban classification, it is reasonable to assume that the results observed are driven by the relative proportions of each class and that differences in classifier performance associated with choice of training sample are not specific to just an urban/nonurban classification.

Descriptions of accuracy derived from the four proposed sampling designs and two training sample sizes encompassed a 5 km × 5 km block-oriented assessment of urban composition (nonsite-specific) and the conventional per-pixel assessment of accuracy (site-specific). The uncertainty of inferring from a sample to the population when conducting a sample-based accuracy assessment was circumvented because complete coverage reference data were available. Results were obtained from a large number of repetitions of the classification process (i.e. selecting the training sample and implementing the classification algorithm independently for each training sample). The results for repeated application of the classification are thus more representative of the performance of the classifiers than would be obtained from a single realization of the process, and the distributions of the accuracy measures (as represented by the boxplots of Figures 5 and 6) provide a novel view of the potential variability in the outcome of the classification depending on which training sample is selected.

The choice of an appropriate sampling design for training data selection is generally an application-specific decision that will depend on which accuracy features are the highest priority for that application. For example, if the priority objective is to increase overall accuracy or to achieve small RMSE of within-block percentage area of urban, the proportionally allocated stratified design (*Prop*) is the preferred option. The *Eq* and *SpatialEq* designs increased the sample size of urban in the training sample, which led to an improvement in producer's accuracy of the rare urban class. Spatial stratification with proportional class allocation (*SpatialProp*) becomes appealing if the objective is to obtain high user's accuracy of urban. As shown in Figures 5 and 6, trade-offs between class-specific producer's and user's accuracies may exert a strong influence on the decision as to which training sampling design to implement.

Our sampling designs were established given that archived reference data (i.e. aerial photography in both 1977 and 1997) were available for the entire region of interest. In practice, however, complete coverage reference data will usually not be available and it will not be possible to stratify by the true urban or nonurban condition of a pixel. The results we report thus evaluate an ideal application of stratification in which the true class of each pixel is known. Because proportional allocation is an equal probability sampling design, it is possible to obtain the equal probability feature of *Prop* using simple random or systematic sampling. Whereas proportional allocation guarantees that any sample selected will have the sample size in each stratum proportional to the area of that stratum, simple random and systematic sampling ensure this proportional sample representation in

an average sense (averaging over all possible sample realizations). However, simple random sampling would be a reasonable approximation to the *Prop* design and systematic sampling would approximate the performance of the *SpatialProp* design if these proportional allocation designs are desired for a practical application. Equal allocation would be more problematic to implement because it is necessary to assign pixels to urban and nonurban strata on the basis of whatever *a priori* information existed to determine the stratum assignments or areas.

The existence of complete coverage reference urban information provided a novel opportunity to evaluate different sampling schemes for selecting training data. Although the results are based on a single site and single date, we would expect that results for other study sites and dates would be qualitatively similar to our findings. However, additional investigation is needed to allow broader generalization of the quantitative differences in accuracy that will result from different training sample selection protocols.

Acknowledgements

We thank two anonymous reviewers and the editor for their constructive comments that helped improve the manuscript. The research is solely the responsibility of the authors and does not necessarily represent the official views of the USGS.

Funding

Stehman was supported by the United States Geological Survey [Grant/Cooperative Agreement Number G12AC20221]. Jin and Mountrakis were supported by the National Aeronautics and Space Administration Biodiversity Program [grant number NNX09AK16G].

References

Bauer, M., B. Loeffelholz, and B. Wilson. 2007. "Estimating and Mapping Impervious Surface Area by Regression Analysis of Landsat Imagery." In *Remote Sensing of Impervious Surfaces*, edited by Q. Weng, 3–19. Boca Raton, FL: CRC Press.

Biging, G. S., D. R. Colby, and R. G. Congalton. 1998. "Sampling Systems for Change Detection Accuracy Assessment." In *Remote Sensing Change Detection: Environmental Monitoring Methods and Applications*, edited by R. S. Lunetta, and C. D. Elvidge, 281–308. Chelsea, Michigan: Ann Arbor Press.

Breiman, L., J. H. Friedman, R. A. Olshen, and C. Stone. 1984. *Classification and Regression Trees*. Belmont, CA: Wadsworth.

Edwards, Jr. T. C., D. R. Cutler, N. E. Zimmermann, L. Geiser, and G. G. Moisen. 2006. "Effects of Sample Survey Design on the Accuracy of Classification Tree Models in Species Distribution Models." *Ecological Modeling* 199: 132–141.

Foody, G. M. 2002. "Status of Land Cover Classification Accuracy Assessment." *Remote Sensing of Environment* 80: 185–201.

Foody, G. M., A. Mathur, C. Sanchez-Hernandez, and D. S. Boyd. 2006. "Training Set Size Requirements for the Classification of a Specific Class." *Remote Sensing of Environment* 104: 1–14.

Hansen, M., R. Dubayah, and R. DeFries 1996. "Classification Trees: An Alternative to Traditional Land Cover Classifiers." *International Journal of Remote Sensing* 17: 1075–1081.

Huang, B., L. Zhang, and B. Wu. 2009. "Spatiotemporal Analysis of Rural-Urban Land Conversion." *International Journal of Geographical Information Science* 23: 379–398.

Jacquin, A., L. Misaova, and M. Gay. 2008. "A Hybrid Object-Based Classification Approach for Mapping Urban Sprawl in Periurban Environment." *Landscape and Urban Planning* 84: 152–165.

Janssen, L. L. F., and F. J. M. van der Wel. 1994. "Accuracy Assessment of Satellite Derived Land-Cover Data: A Review." *Photogrammetric Engineering and Remote Sensing* 60: 419–426.
Jin, H, and G. Mountrakis. 2013. "Integration of Urban Growth Modelling Products with Image-Based Urban Change Analysis." *International Journal of Remote Sensing* 34: 5468–5486.
Mather, P. M. 1999. *Computer Processing of Remotely-Sensed Images: An Introduction*. 2nd ed. New York, NY: John Wiley & Sons.
Pal, M., and P. M. Mather. 2003. "An Assessment of the Effectiveness of Decision Tree Methods of Land Cover Classification." *Remote Sensing of Environment* 86: 554–565.
Pijanowski, B., S. Pithadia, B. Shellito, and A. Alexandridis. 2005. "Calibrating a Neural Network-Based Urban Change Model for Two Metropolitan Areas of the Upper Midwest of the United States." *International Journal of Geographical Information Science* 19: 197–215.
Piper, J. 1992. "Variability and Bias in Experimentally Measured Classifier Error Rates." *Pattern Recognition Letters* 13: 685–692.
Sesnie, S. E., P. E. Gessler, B. Finegan, and S. Thessler. 2008. "Integrating Landsat TM and SRTM-DEM Derived Variables with Decision Trees for Habitat Classification and Change Detection in Complex Neotropical Environments." *Remote Sensing of Environment* 112: 2145–2159.
Smits, P. C., S. G. Dellepiane, and R. A. Schowengerdt. 1999. "Quality Assessment of Image Classification Algorithms for Land-Cover Mapping: A Review and Proposal for a Cost-Based Approach." *International Journal of Remote Sensing* 20: 1461–1486.
Stehman, S. V. 2000. "Practical Implications of Design-Based Sampling Inference for Thematic Map Accuracy Assessment." *Remote Sensing of Environment* 72: 35–45.
Stehman, S. V. 2009. "Sampling Designs for Accuracy Assessment of Land Cover." *International Journal of Remote Sensing* 30: 5243–5272.
Stehman, S. V., and G. M. Foody. 2009. "Accuracy Assessment." In *The SAGE Handbook of Remote Sensing*, edited by T. A. Warner, M. D. Nellis, and G. M. Foody, 297–309. London: Sage Publications.
Story, M., and R. G. Congalton. 1986. "Accuracy Assessment: A User's Perspective." *Photogrammetric Engineering and Remote Sensing* 52: 397–399.
Triantakonstantis, D., and G. Mountrakis. 2012. "Urban Growth Prediction: A Review of Computational Models and Human Perceptions." *Journal of Geographic Information System* 4: 555–587.
USGS (US Geological Survey Rocky Mountain Mapping Center). 2003. "Front Range Infrastructure Resources (Frir) Project." Accessed May 9, 2013. http://energy.cr.usgs.gov/regional_studies/frip/.
van Niel, T. G., T. R. McVicar, and B. Datt. 2005. "On the Relationship between Training Sample Size and Data Dimensionality: Monte Carlo Analysis of Broadband Multi-Temporal Classification." *Remote Sensing of Environment* 98: 468–480.
Weng, Q. 2012. "Remote Sensing of Impervious Surfaces in the Urban Areas: Requirements, Methods, and Trends." *Remote Sensing of Environment* 117: 34–49.
Weng, Q., X. Hu, and H. Liu. 2009. "Estimating Impervious Surfaces Using Linear Spectral Mixture Analysis with Multitemporal ASTER Images." *International Journal of Remote Sensing* 30: 4807–4830.
White, R., G. Engelen, and I. Uljee. 1997. "The Use of Constrained Cellular Automata for High-Resolution Modelling of Urban Land-Use Dynamics." *Environment and Planning B: Planning and Design* 24: 323–343.
Zhen, Z., L. J. Quackenbush, S. V. Stehman, and L. Zhang. 2013. "Impact of Training and Validation Sample Selection on Classification Accuracy and Accuracy Assessment When Using Reference Polygons in Object-Based Classification." *International Journal of Remote Sensing* 34: 6914–6930.