



Originally published as:

Wieland, M., Torres, Y., Pittore, M., Benito, B. (2016): Object-based urban structure type pattern recognition from Landsat TM with a Support Vector Machine. - *International Journal of Remote Sensing*, 37, 17, pp. 4059—4083.

DOI: <http://doi.org/10.1080/01431161.2016.1207261>

Object-based Urban Structure Type Pattern Recognition from Landsat TM with a Support Vector Machine

Marc WIELAND¹, Yolanda TORRES², Massimiliano PITTORE¹, Belen BENITO²

¹ Centre for Early Warning Systems, GFZ German Research Centre for Geosciences, Telegrafenberg, 14473 Potsdam, Germany

² Earthquake Engineering Research Group, Technical University of Madrid. Ctra. Valencia, km 7.5, 28031 Madrid, Spain

Corresponding author: Marc Wieland, mwieland@gfz-potsdam.de

Object-based Urban Structure Type Pattern Recognition from Landsat TM with a Support Vector Machine

Abstract

This study evaluates the potential of object-based image analysis in combination with supervised machine learning to identify urban structure type patterns from Landsat Thematic Mapper (TM) images. The main aim is to assess the influence of several critical choices commonly made during the training stage of a learning machine on the classification performance and to give recommendations for classifier-dependent intelligent training. Particular emphasis is given to assess the influence of size and class distribution of the training data, the approach of training data sampling (user-guided or random) and the type of training samples (squares or segments) on the classification performance of a Support Vector Machine (SVM). Different feature selection algorithms are compared and segmentation and classifier parameters are dynamically tuned for the specific image scene, classification task and training data. The performance of the classifier is measured against a set of reference datasets from manual image interpretation and furthermore compared on the basis of landscape metrics to a very high-resolution reference classification derived from Light Detection And Ranging (LiDAR) measurements. The study highlights the importance of a careful design of the training stage and dynamically tuned classifier parameters, especially when dealing with noisy data and small training datasets. For the given experimental set-up the study concludes that given optimized feature space and classifier parameters, training a SVM with segment-shaped samples that were sampled in a guided manner and are balanced between the classes provided the best classification results. If square-shaped samples are used, a random sampling provided better results than a guided selection. Equally balanced sample distributions outperformed unbalanced training sets.

Keywords: machine learning; object-based image analysis; urban structure types; Landsat; Haiti.

1. Introduction

The research presented in this paper is part of an overall seismic risk study in Haiti and focuses on the optimization of the study area stratification from remote sensing (Figure 1). Seismic risk assessment usually involves the characterization and analysis of an exposed building stock according to vulnerability-relevant attributes and integrates it with the outcomes of a seismic hazard analysis of the same area to compute expected damage distributions. The collection of exposure and vulnerability data frequently turns into a challenging task. Typically used by risk analysts, cadastral and census databases are not always available or updated. Then, field surveys are often needed to collect data. However, as the campaigns are time and resource consuming, they are rather used for sampling purposes. Consequently, a stratification of exposed assets over large areas is needed in order to efficiently prioritize field surveys. Often, regular grids or administrative units are used in risk computation as strata to map exposure and vulnerability. However, such units are artificially created and do not necessarily relate to the actual distribution of exposed assets. They can therefore be meaningless or at least inefficient for data disaggregation or aggregation. In this context, urban structure types can provide a valuable concept for the stratification of a city into meaningful spatially defined entities that better reflect the actual composition of exposed assets (Herold, Liu, and Clarke 2003). A useful stratification is defined, based on the hypothesis that the building inventory composition follows Tobler's law of spatial autocorrelation (Tobler 1970). In this regard, the hypothesis proposes that neighbouring buildings have several characteristics in common (e.g., typology, occupancy, etc.) and can therefore be clustered together. In this study urban structure type patterns are extracted directly from Landsat Thematic Mapper (TM) data at an aggregated neighbourhood scale ranging from one to several blocks. The minimum unit size is defined by image statistics and outlined by unsupervised image segmentation. Labelling of segments is performed by a Support Vector Machine (SVM). Urban structure types are, therefore, in the following defined as spatial units of the built environment at an aggregated neighbourhood scale, which are relatively homogeneous with respect to the overall population in terms of their physical appearance (landcover) and usage (landuse).

Figure 1: Overview of the method.

Several studies exist that use machine learning for urban landcover / landuse mapping from Landsat data at the per-pixel level (Schneider 2012), (Griffiths et al. 2010). Only recently studies explored the potentials of machine learning in the context of an object-based image analysis (Peña et al. 2014). In object-based image analysis, the feature vector is calculated using all the pixels inside a segment. The image segmentation process does not yield perfectly homogeneous objects; hence the noise introduced in the training phase might lead to less effective classification models. Besides the computational unit, several other choices are usually being made during the training stage of a learning machine that can significantly influence its performance. Amongst the most important are training data sample size (Li et al. 2014), class distribution (Weiss and Provost 2003) and sample approach (Foody and Mathur 2004). Therefore, Pal and Foody (2012) suggest to design classifier-specific intelligent training data acquisition activities for accurate classification from small training datasets. Previous work showed that also feature selection (Löw et al. 2013), segmentation (Wieland and Pittore 2014) and classifier parameter tuning (Qian et al. 2014) can have a significant effect on the performance that may be even larger than the selection of a particular learning machine. Given these observations, it becomes imperative that learning machines are dynamically tuned for each classification task, image scene and training dataset.

The aim of this work is to assess the influence of several critical choices commonly made during the training stage of a learning machine on the classification performance. The specific research questions tackled within this study include:

- (1) How does the choice of feature selection algorithms influence the classification performance?
- (2) How does the size and class distribution of training samples influence the classification performance?
- (3) How does the sampling strategy of training instances (user-guided, random) affect the classification performance?
- (4) How should the sampling area in which to calculate the feature vector for training be defined (square or segment)?

With respect to previous work in this direction, the study at hand covers characteristics that are intrinsic to object-based image analysis on multi-spectral satellite images at medium resolution. We propose the use of homogeneous sampling areas that are independent on the segmentation, and analyse whether it affects the classification results. Furthermore, the classification task can be considered rather complex with respect to the input data, the level of thematic detail and the high cost of training sample selection.

2. Study area and data

2.1 Study area

Port-au-Prince, the capital city of Haiti, was selected as study area given the confluence of several factors. Firstly, the city and its metropolitan area is a good example of rapidly changing urban area, where new construction is not always recorded in official cadastre. Hence the interest of testing an exposure and vulnerability assessment technique based on remote sensing, that is able to cope with the rapid development of the urban environment. Secondly, the large amount of spatial data of Haiti made available after the 2010 earthquake has enabled this study. Of relevant importance for this study are the building database compiled by the Ministry of Public Works, as well as the Light Detection And Ranging (LiDAR) points and very high resolution aerial images collected by the Rochester Institute of Technology. Finally, recent studies (Benito et al. 2012), (Torres et al. in preparation), (Calais et al. 2010), (Frankel et al. 2011) have proven that seismic hazard and risk still remain high in the region, where the population exposed to this risk reaches 2.4 million (“Population of Haiti” 2015). Therefore, new initiatives aimed at earthquake risk estimation are being promoted, such as (“Plan de Prévention Séisme Pour Le Grand Nord” 2015), for which this study would be a support.

Figure 2: Study area Port-au-Prince and surroundings, Haiti. Landsat TM image of the study area, extracted built-up area and distribution of reference data. The zoom shows a subset of the segmentation result superimposed on the input Landsat image (top) and on a very high-resolution aerial image (bottom).

The study area covers the larger urban agglomeration of Port-au-Prince (Figure 2). The centre and south-west of the study area are mainly occupied by the urban areas of Port-au-Prince

and Carrefour; whereas rural areas are found to the north (Marin) and east (Croix-des-Bouquets). Petionville is a residential area located to the south.

2.2 Landsat imagery

In this study we use a Landsat TM satellite image of 26 November 2009 (previous to the 12 January 2010 earthquake) from the United States Geological Survey (USGS) satellite image repository (“USGS Earthexplorer” 2015). The image is cloud-free within the analysis mask. The image has a geometrical resolution of 30 m in six spectral bands (0.45-12.5 μm) and 120 m in the thermal infra-red band, with a radiometric resolution of 8 bits. The image scene was cut to a subset of 1,000 pixels \times 780 pixels size. Since only a single image scene is used in the study, no atmospheric correction and ortho-rectification were performed in order to minimize the modification of the image intensity values (Song et al. 2001). A nearest neighbor transformation has been performed on the thermal infra-red band to resample it to the 30 m pixel-resolution of the other multi-spectral bands. The image is in a Universal Transverse Mercator (UTM) projection with World Geodetic System 1984 (WGS84) datum.

2.3 Built-up area mask

The analysis presented in this study has been performed on the built-up area only. In order to distinguish built-up from non-built-up areas, a binary classification of the Landsat TM image has been carried out using an object-based image analysis with SVM, as is described in greater detail in (Wieland and Pittore 2014). A learning machine has been tuned and trained on the basis of several image scenes (not including the Port-au-Prince scene) as part of the previous study for the detection of built-up areas from Landsat TM images. The trained learning machine has been applied in this study, therefore rendering the classification process for the study area fully automated.

2.4 Urban structure types

Based on expert knowledge of the study area, acquired after several visits to the country, five classes of urban structure type patterns could be identified, namely residential, urban regular, urban irregular, rural and industrial (Table 1).

Residential patterns (class 1) are characterized by isolated medium-to-big size family houses, with garden and green open spaces, as well as broad and straight paved streets. Residential areas are mainly located in the metropolitan ring of larger cities, and are normally of planned urbanism and well-constructed buildings.

Urban patterns (classes 2 and 3) present higher building density and smaller buildings than residential patterns. Class 2 is characterized by regular street network and some green spaces, unlike irregular pattern, where constructions are stuck to each other and unpaved streets present a chaotic distribution. Irregular urban areas are found in the outskirts of main cities and are associated to low income level and brittle buildings (class 3).

Rural patterns (class 4) correspond to the country side areas and small towns surrounding Port-au-Prince. The vegetation density is higher than in any other classes and buildings are sparse. Some regular shaped structures are present, such as fences or walls dividing family properties or crop fields. Small houses are generally low-quality, whereas bigger buildings indicate better construction technique and materials.

Industrial patterns (class 5) are clearly identified by their large dimensions and regular shape. The immediate surroundings are normally not vegetated and bare soil dominates.

2.5 Reference data

2.5.1 Visual image interpretation

To train and test the SVM classifier under varying conditions, eight reference datasets have been compiled for the study area. Table 2 summarizes their main characteristics. R50-SQ-RD and R50-SE-RD are composed of 50 samples per class (R50) defined over a square (SQ) or segment (SE) shape and selected randomly (RD). The data were sub-sampled from a larger reference dataset that follows the natural distribution of the classes in the study area as a result of a simple random sampling. From these reference datasets smaller subsets with 15 samples per class (R15-SQ-RD and R15-SE-RD) have been derived for comparison with guided and unbalanced random sampling approaches. R15-SQ-GD and R15-SE-GD have 15 samples per class (R15) defined over a square (SQ) or segment (SE) shape and were selected

manually by a human operator in a guided manner (GD). RUB-SQ-RD and RUB-SE-RD follow the natural (unbalanced) class distribution (RUB) and were derived by sub-sampling the larger reference dataset with natural class distribution but without constraining the number of samples per class. For each sampling approach two different sampling types, namely segmentation objects and regular 200 m squares (Table 1) are provided. The square and segment samples refer always to the same sample location and only the sample type, meaning the shape of the area covered by the particular sample, varies between a regular square or the respective segment at this location. Labelling of the samples has been done manually by visual image interpretation under consideration of ground-truth knowledge and comparison with aerial imagery of the study area. Class assignment to samples is based on approximate area coverage, where the class that covers the largest area within a sample is chosen. Moreover, an independent testing dataset of 100 samples that are not included in the above mentioned reference datasets has been compiled for an assessment of the final map accuracies.

2.5.2 LiDAR classification

For the 100 samples of the independent test dataset very high-resolution LiDAR data was available (“Open Topography” 2015). The LiDAR dataset was collected by the Center for Imaging Science at Rochester Institute of Technology, Kucera International, and ImageCat, Inc. right after the 2010 earthquake, between 21 and 27 January 2010. The point density is 3.4 points/m² and the coordinates are given in UTM Zone 18 N WGS84. A cloud of 975 million points was needed to cover the whole study area. The LiDAR point analysis has been done with the MDTopX software, which incorporates a powerful algorithm to classify the points (Arranz 2013). The classes of interest for this study were ground, buildings, and vegetation (low-, medium-, and high-rise). The algorithm is able to distinguish between these classes following rules based mainly on geometric features derived from a Delaunay triangulation, such as elevation, area and slope of the triangles. Once the points falling in the 100 samples were selected and classified, a very high resolution raster image was created (0.5 m spatial resolution) upon which to landscape metrics (Section 3.4) were computed for the different urban patterns (Table 3).

3. Methods

3.1 Image segmentation

An efficient graph-based image segmentation algorithm (Felzenszwalb and Huttenlocher 2004) is used within this study to cluster the original image pixels into segments. The algorithm measures the evidence for a boundary between two segments by comparing the feature vectors across the boundary with those between neighbouring pixels within each segment. The brightness values of the image pixels in all spectral bands of the satellite image are used to define the input feature space for segmentation, and the L2 Euclidean distance is used to measure the dissimilarity. A function of the segment size m and a runtime parameter k allow to effectively set a scale of analysis in that a larger value for k causes a preference for larger segments. Segmentation parameters have been tuned in this study based on a comparison of intra-segment homogeneity and inter-segment heterogeneity (Dey, Zhang, and Zhong 2010). A global Moran's I has been utilized to quantify inter-segment heterogeneity and the standard deviation of the brightness values of the input image bands, weighted by each segment's size and summed over all the segments in the image scene has been used for assessing intra-segment homogeneity (Johnson and Xie 2011). In a set of segmentations, the segmentation that minimizes intra-segment variability and maximizes inter-segment separability is defined as being optimal in terms of providing the most appropriate analysis scale for a given image scene.

3.2 Support Vector Machine (SVM)

SVM has been selected as most promising classifier to be used within this study for the task of urban structure type pattern recognition within built-up areas. The classifier choice is based on a comprehensive study that was previously carried out to assess the performance of different learning machines to distinguish built-up from non-built-up areas (Wieland and Pittore 2014). SVM is a non-parametric classifier derived from statistical learning theory and originally developed by Vladimir Vapnik (Vapnik 2000). SVM utilize kernel functions to project non-linearly separable classes into higher dimensional feature space, where non-linearly separable classes can be separated by a linear hyperplane. The optimal separating

hyperplane between two classes is chosen by maximizing the margin between the separating hyperplane and the closest feature vectors. Therefore, only the closest training samples (support vectors) to the edge of the class distribution are used, which is why potentially SVM can deal well with small training datasets given that they are well selected (Foody and Mathur 2004). A soft-margin parameter is introduced in SVM to deal with outliers in the data. This allows some data points to violate the separation through the hyperplane without affecting the final result. Although SVM was originally developed to solve binary classification tasks it has been extended to solve multi-class problems by applying a one-against-one scheme. In this study we use the scikit-learn (“Scikit-Learn” 2015) Python interface to the well-known libsvm (Chang and Lin 2011). Optimal SVM parameters (kernel function Φ , kernel coefficient γ and penalty or regularization parameter C) are tuned for each classification according to a ten-fold cross-validation method during the training phase of the classifier. An optimal parameter selection is reached when the cross-validation estimate of the test samples error is minimal.

3.3 Feature space and feature selection

The feature space can be grouped into (a) spectral features (29), (b) spectral band indices (2), (c) textural features (56) derived from Grey-Level Co-occurrence Matrix (GLCM) (Haralick, Shanmugam, and Dinstein 1973) and (d) geometrical features (10). The feature values were normalized to zero mean and unit variance because the SVM classifier is dependent on a distance measure and its discriminating power may be negatively affected by not normalized data. For all the experiments making use of the square-shaped reference data, geometrical features were excluded from the feature vector since the regular and equal geometry of the square-shaped samples would not provide any significant information to the classification. A comprehensive list of the features is provided together with references and definitions in Wieland and Pittore (2014). Three feature selection methods are tested and compared in this study to select for each training-testing-scenario the most significant image features.

A recursive feature selection algorithm as proposed by Guyon et al. (2002) is used in this study. Given an external classifier that assigns weights to features, recursive feature selection

considers iteratively smaller and smaller sets of features. With each iteration the features are used to train a classifier and are assigned weights according to their discriminating power. The features with the smallest weights are eliminated from the feature set for the next iteration. The feature space that maximizes a scoring value is selected. The feature selection has been performed in a five-fold cross-validation loop using classification accuracy as scoring value.

Furthermore, a ReliefF feature ranking algorithm as described by Kononenko (1994) and implemented in the Weka software (“WEKA3 Data Mining with Open Source Machine Learning Software” 2012) is used in this study. ReliefF uses iterative random sampling to select multiple times an instance R and its k nearest neighbours H from the same class and k nearest neighbours M from the other classes. The feature values of R are compared to H and M and a quality score is adjusted for each feature depending on how well it separates the classes. After a defined number of iterations the scores are ordered to form the final feature rank.

Finally, a linear discriminant analysis (Huberty 1994) is carried out. Linear discriminant analysis is a multivariate statistical technique that generates a set of discriminant functions for classification on the basis of linear combinations of those predictor variables with higher discriminant power. The discriminant power of the predictor variables (or features) is computed by comparison of the intra- and inter-groups variances through the Snedecor’s F and Wilks’ lambda indices. Variables with low intra-group variance and high inter-group variance have high discriminant power, and are ranked accordingly. To generate the discriminant functions, the ranked features are included one-by-one, in a stepwise procedure, until the inclusion of a new feature does not improve the classification performance of the functions.

3.4 Performance measures

To assess the classification performance on the selected subsets of the feature space for different training-testing scenarios, a set of accuracy measures has been used. Standard accuracy measures that have been derived from the error matrix through comparison with

ground-truth labels include accuracy, sensitivity, specificity, precision and F_1 score (Fawcett 2006). Cross-validation has been used in order to increase their reliability by reducing the bias resulting from specific training-testing datasets. Performance measures are therefore reported as average over the results of the cross-validation iterations. Moreover, in case of multi-class classification problems, the weighted average of the performance measures of each class are provided.

Receiver Operating Characteristics (ROC) curves are used to visualize the performance of a classifier against ground-truth labels as its discrimination threshold is varied. An ROC curve is created by plotting the true positive rate against the false positive rate at various thresholds. Compared to accuracy, F_1 score and precision, ROC curves are solely based on true positive and false positive rate, and thus are insensitive to changes in class distribution of the test dataset (Fawcett 2006). The Area Under the Curve (AUC) is used as a single scalar value to describe the classifier performance as derived from an ROC curve.

Learning curves are used to evaluate the influence of the training data size on the classification performance and to identify whether the classifier is more affected by a bias or variance error. A learning curve shows the training and validation score of a classifier for varying sizes of training samples. The training dataset size is iteratively increased and a five-fold cross-validation is used for each iteration to derive the validation and training scores. The curves provide the mean score and the range of scores as derived from the cross-validation for each iteration. Accuracy is used as score in this study.

To further quantify the spatial characteristics of the land-use / land-cover inside the segments, landscape metrics were computed from a very high-resolution classification of LiDAR point-clouds. Landscape metrics were found to provide important information for differentiating urban structural patterns (Herold, Liu, and Clarke 2003). In this study, we compare three well known metrics (patch density, Euclidean nearest neighbour distance and shape index) across the urban structure type classes. Patch density (PD) equals the number of patches of a specific land cover class (buildings) divided by the total segment area. Euclidean Nearest Neighbour distance (ENN) provides the shortest edge-to-edge distance from cell centre to cell centre

between each patch of a class and its nearest neighbour patch. It can be interpreted as a measure of patch isolation per segment. The complexity of a patch shape compared to a square shape of the same size is measured by the shape index (Shape). The index is computed at the patch (per-building) level and the area weighted mean is reported per segment. The landscape metrics were calculated using the public domain software FRAGSTATS Version 4.2 (McGarigal, Cushman, and Banzhaf 2012).

Assuming that these metrics can provide a description of the urban structure type classes at higher thematic detail than ground-truth labels, evaluation is performed using the metrics themselves. Silhouette coefficients are computed to quantify and compare the quality of classifications based on how well they separate the feature space spanned by the landscape metrics (Rousseeuw 1987). Silhouette coefficients compare the mean distance between a sample and all other points in the same class, with the mean distance between a sample and all other points in the next nearest cluster. A higher silhouette coefficient score relates to a model with more dense and well separated clusters.

4. Results

For the Port-au-Prince scene, a total built-up area of 158 km² has been extracted using the above mentioned classification procedure. An accuracy assessment carried out with an independent reference dataset (100 stratified random samples) indicates a very good classification result with an accuracy of 0.92, a precision of 0.93 and a F_1 score of 0.92. The built-up area has been further segmented by using the previously mentioned segmentation algorithm and parameter tuning method. For the built-up area of the image scene 1,736 segments have been produced with a parameter combination of $k=20$ and $m=20$ that provided the best overall quality scores of 384 segmentation runs. Superimposing the segmentation result onto an aerial image of the same area and from the same acquisition year, allowed also for a qualitative assessment of the segmentation result by visual inspection. Urban areas of different complexity, which appear homogeneous in terms of their spectral response in the Landsat image, could be outlined properly (Figure 1). It can be observed that they correspond to clusters of relatively homogeneous urban structure types. In the following the results of

several experiments related to the training stage of the classifier for the detection of urban structure type patterns are described.

4.1 Feature space

This experiment aims at identifying the most relevant features for class separability and provides indication on how sensitive the learning machine is to the size and composition of the feature vector. Figure 3 shows a comparison of the three different feature selection methods (fs-disc: Discriminant Analysis; fs-relief: ReliefF; fs-rec: Recursive Feature Selection) applied to (a) the squares reference dataset (R50-SQ-RD) and (b) the segments dataset (R50-SE-RD) separately. Additionally, also the full feature space (fs-all) has been used as input to a ROC analysis with five-fold cross-validation. The computed micro-average ROC curves and corresponding AUC values for the different feature subsets indicate a general improvement of the classification performance through feature selection. The choice of the feature selection method, however, seems to not strongly influence the performance of the classifier. The feature compositions derived by the different methods are also presented in Figure 3. It can be seen that the feature spaces identified by Recursive Feature Selection and Discriminant Analysis are significantly smaller than the ones derived through ReliefF. The feature composition from Recursive Feature Selection is dominated by spectral features (R50-SQ-RD: 55%; R50-SE-RD: 52%), followed by textural (R50-SQ-RD: 27%; R50-SE-RD: 33%) and spectral band index features (R50-SQ-RD: 18%; R50-SE-RD: 10%). Also the Discriminant Analysis identifies a similar distribution of feature classes with spectral (R50-SQ-RD: 55%; R50-SE-RD: 43%), textural (R50-SQ-RD: 27%; R50-SE-RD: 25%) and spectral band index features (R50-SQ-RD: 18%; R50-SE-RD: 16%) being most important for class separability. The ReliefF method gives more weight to geometrical features for segments while showing a similar importance distribution for the squares with spectral (R50-SQ-RD: 53%; R50-SE-RD: 45%), textural (R50-SQ-RD: 40%; R50-SE-RD: 30%), spectral band index features (R50-SQ-RD: 7%; R50-SE-RD: 6%) and geometrical features (R50-SE-RD: 19%). The most important spectral features for class separability over all methods appear to be standard deviation of band 1 (blue), mean of band 4 (near infra-red) and the thermal infra-red band as well as the minimum pixel value of the thermal infra-red band. GLCM

mean of band 4 (near infra-red) and band 1 (blue) as well as GLCM homogeneity of band 3 (red) appear to be most important textural features. Mean and standard deviation of the Normalized Difference Vegetation Index (NDVI) are identified by all methods to be important features. For the geometrical features compactness, perimeter and convexity of the segments seems to play an important role.

Figure 3: Micro-average ROC curves derived for a) the squares reference dataset (R50-SQ-RD) and b) the segments reference dataset (R50-SE-RD) under consideration of different feature subsets.

In the following experiments the Recursive Feature Selection is used as the underlying feature selection algorithm, due its slightly better performance under consideration of small feature subsets (especially compared to ReliefF). The Discriminant Analysis performs equally well but is relatively complex to run in an automatized manner.

4.2 Sample size

Figure 4 shows the learning curves derived for the squares (R50-SQ-RD) and segments (R50-SE-RD) reference datasets. The large gap between validation and training score at small training sample size indicates a tendency of the classifier to over-fit the data. Increasing the training sample size can significantly improve the validation score, lower the variance and therefore increase the generalization ability of the classifier. In the case of square samples a minimum of 100 samples is needed to achieve an acceptable validation score of 0.6. In case of segment samples this number increases to around 140 samples. However, even when the maximum amount of training data is reached adding more data could potentially further increase the generalization ability of the classifier on both the squares and segments datasets. Other possible strategies to reduce variance include using less features or increasing the regularization of the classifier. We already optimize both parts automatically in the analysis chain through feature selection and tuning of the SVM parameters (including kernel function Φ and regularization parameter C). A linear kernel function with $C=1.0$ and a reduced feature space of 11 (squares) and 21 (segments) dimensions underline the fact that the implemented analysis chain is reacting to a limited training sample size.

Figure 4: Learning curves derived for a) the squares (R50-SQ-RD) and b) the segments reference datasets (R50-SE-RD) under consideration of the selected feature subsets from recursive feature selection.

4.3 Sample approach

This experiment aims at checking the influence of different sampling approaches for training the classifier on the classification performance. In particular, a stratified random sampling with balanced class distribution (R15-SE-RD, R15-SQ-RD) is compared to a simple random sampling with unbalanced class distribution (RUB-SE-RD, RUB-SQ-RD) and a user-guided selection of training samples (R15-SE-GD, R15-SQ-GD). All approaches are applied to both the segment and squares sample types and the balanced reference datasets are reduced to 75 samples (15 per class). Given the results of the sample size experiment, this is clearly under-sampling the training data size. However, a very small training data size is largely common in remote sensing applications, especially when using an object-based image analysis where segments and not single pixels represent the computational unit. Moreover, such scenario can give further indications on whether a well-tuned SVM can adequately cope with under-sampled and noisy training data.

Figure 5: ROC curves and performance measures derived for a) the squares and b) the segments reference datasets under consideration of different sample size, types and approaches.

After performing recursive feature selection for the six sampling approaches, ROC curves and performance measures derived from aggregated error matrices have been computed from five-fold cross-validation. It can be seen from Figure 5 that acceptable to good performance can be measured for all the scenarios with largest difficulties arising from the separation of classes 1 (residential) and 2 (urban regular), which show smallest AUC values. Through guided sample selection, however, the receiver operator characteristics for classes 1 and 2 slightly improved. With respect to the mean performance measures, on the squares dataset no significant difference can be observed between training with samples that are randomly selected or chosen in a guided manner. When training and testing on the segments dataset, however, an improvement of the classification performance can be observed when using a

guided training dataset with respect to a randomly selected one. The results of the classifier trained on the reduced datasets show comparable mean performance values with respect to the large training datasets (R50-SQ-RD and R50-SE-RD). The standard deviation of the performance measures, however, is clearly larger on the reduced training datasets than on the large dataset. This indicates that it is possible to achieve good classification performance with clearly under-sampled training data in single cases, but with a clear increase of the variance and therefore a reduction of the generalization ability of the classifier. The worst performance over all scenarios can be observed by the unbalanced datasets, where training samples were selected by a simple random sampling.

4.4 Sample type

Comparing square with segment sample types almost similar performance can be measured on the large reference datasets (Figure 5). When using the under-sampled reference datasets a better performance with the segment sample type can be observed for both balanced random and guided sampling, where the best mean performance is given by the balanced guided sampling (R15-SE-GD). In the unbalanced case, squares perform better than segments.

So far reference datasets have been treated independently, meaning that training and testing were performed by cross-validation on the same datasets – squares were trained and tested with square samples; segments were trained and tested with segment samples. This can give information about the performance of a classifier under different levels of noise, where segments can be considered to be noisier than squares. In a real-world application, however, square samples may be used to train a classifier, but the classification is done on the segment basis. For this reason an independent testing dataset has been compiled on the segment basis, against which the different sample types and training approaches are evaluated. This aims at evaluating the influence of different sample types and approaches on the performance of classifying segments. Figure 6 shows the classifications of the whole study area derived under consideration of the different sample approaches and types. It also shows the error matrices along with the derived performance measures for the different solutions. The best performance when classifying and testing on the segment basis can be observed for the

guided sampling approach with segment sample type (R15-SE-GD). This is in line with the previous test results, as well as a better performance with the segment sample type for both balanced random and guided sampling.

Figure 6: Classifications of the study area with error matrices and average (avg) performance measures under consideration of different sample types and approaches.

4.5 Comparison with very high-resolution data

Until now, we compared the classification results with visual image interpretation by a human operator assigning a discrete class label at the segment level. However, we should further consider the actual composition of the segments with respect to parameters of interest that can be observed at higher spatial resolution. In this context, a good classification should show high intra-class homogeneity and inter-class heterogeneity with respect to interpretable metrics that can describe the different urban structure type classes based on an independent very high-resolution dataset. As described above (Section 3.4), landscape metrics were computed for the 100 segments of the independent testing dataset to quantitatively describe the urban structure types classes according to the results of an independent very high-resolution classification of LiDAR point-clouds.

Figure 7: Comparison of the Landsat classifications with a very high-resolution classification from LiDAR data.

Figure 7 gives an overview of the distribution of the three landscape metrics and the area covered by the LiDAR classes (buildings, ground and vegetation) per urban structure type class for each of the classification results. Looking at the manually labelled testing dataset (Ref label), it can be seen that class 2 (urban regular) shows the highest Patch Density (PD) with small distances between the patches measured by the Euclidean Nearest-Neighbour distance (ENN). Classes 3 (urban irregular) and 1 (residential) also show high density and low distance values. The lowest density and highest distance values can be observed for classes 4 (rural) and 5 (industrial / commercial). The most complex shapes (high values for Shape) of built-up areas can be observed for class 3 (urban irregular), whereas class 4 (rural) shows rather simple shapes (low values for Shape). Also from the composition of the relative

area coverage of LiDAR classes (Comp) clear differences can be observed between the urban structure type classes.

Table 4 shows the silhouette coefficients computed for each of the urban structure types classifications from Landsat, using the euclidean distance on the feature space spanned by the three landscape metrics as measure for intra-class homogeneity and inter-class heterogeneity. The coefficients show values close to 0 for all the classifications including the manually labelled independent testing data. This indicates that classes are not densely clustered and overlapping. The negative sign shows that some of the samples might have been assigned to the wrong class. The urban structure types classification that shows the highest coefficient and therefore produces the best partition of the landscape metrics feature space is the balanced guided sampling with segment sample type (R15-SE-GD). This classification shows even better results than the manually labelled testing dataset (Ref label). The classification based on a balanced random sampling with segment sample type (R15-SE-RD) shows still a performance that is comparable with the manually labelled classification, whereas the other classifications perform worse.

5. Discussion and conclusions

In this study, we evaluated different aspects of the training stage of a SVM classifier with respect to its performance for the task of classifying urban structure type patterns from Landsat TM images. Within several experiments, emphasis was given to assess the influence of feature selection, the training data sample size, sample approach and sample type on the classification performance in the context of an object-based image analysis. The main findings can be summarized as follows:

- (1) *Feature space*: Even though SVM discard as part of the training phase irrelevant features, performing separate feature selection prior to training the classifier could significantly improve the classification performance. Amongst the tested feature selection algorithms recursive feature selection and discriminant analysis proved to be the most efficient. Especially when training data acquisition comes at a high cost and the Hughes phenomenon (Hughes 1968) is relevant, the ability to select few well

defined and significant image features becomes essential. Landsat TM images could provide a feature space that allows to sufficiently distinguish between different classes of the built-environment. In this context, spectral and textural features showed the highest importance for class separability. Geometrical features can be largely discarded. The most significant image bands that were identified by all the feature selection algorithms include the blue (0.45-0.52 μm), the near infra-red (0.76-0.90 μm) and the thermal infra-red (10.40-12.50 μm) bands. The importance of the thermal infra-red band is particularly interesting in this context, especially since it has often been ignored in pixel-based land-cover classifications, due to its low geometrical resolution (120 m). In the context of an object-based image analysis, however, the lower ground sampling distance of the thermal infra-red band did not negatively influence the geometric outline of the objects of interest, but added additional spectral content when included in the classification of the segments.

(2) *Training sample size*: Applying machine learning to remote sensing applications and especially to object-based image analysis stills comes at a high cost with respect to acquiring the needed samples for training the classifiers. In this context SVM show large potentials in that they are able to deal well also with small training sample sizes due to the fact that they make use of support vectors only. This is not a new conclusion but could be confirmed within this study also for object-based image analysis. Moreover, this study highlighted the importance of feature selection and classifier parameter tuning (especially regularization) to get the most out of the training data. A problem that, however, remains in any case when potentially under-sampling the training data is that in the single classification case good performance may be achieved but the learned model will lack generalization ability and will be highly sensitive to changes in the data.

(3) *Training sample approach*: From the experiments carried out in this study a performance improvement can be observed for a guided selection of training samples with respect to a random sampling given that the samples are balanced across the classes. This becomes particularly evident when using segments as sample type.

Guided sampling has been performed by a human operator with experience in image interpretation who selected representative samples based on profound knowledge of the study area. The performance improvement over a random sampling comes therefore at the cost of having profound ground-truth knowledge and the additional time effort that is needed to search for the actual samples. Class balance seems to play a role and better performance was observed on balanced datasets. However, in order to better understand the effect of class-balance on the classification performance larger reference datasets would be needed.

(4) *Training sample type*: The choice of the sample type is a point that is particular to object-based image analysis. The use of square-shaped samples could ease up the creation of an independent reference database, since the samples are less noisy with respect to segment samples, geometrically clearly defined and their class assignment is easier to interpret. The difference in classification performance between using segments or squares is not significant. This indicates that SVM are potentially capable of dealing well with noisy data. However, when trained on squares and applied to the segments, which is the usual real-world application case, the so trained learning machine performs slightly worse than the one trained on segments. Therefore, it seems that introducing noise into the training phase can be beneficial. However, given the small difference in performance, more evidence should be collected to underline this conclusion.

(5) *Comparison with very high resolution LiDAR classification*: When comparing the different Landsat classifications with the distribution of landscape metrics that describe the urban structure type classes at a higher spatial resolution, it can be seen that the classes can be well distinguished. However, overlaps between ranges of landscape metrics for the different classes are observable for all classifications including the manual reference. This is resulting in low silhouette coefficients for all classifications. A possible reason for the generally low coefficients even for the manually labelled reference data could lie in the LiDAR classification quality (see Table 3) which does not always provide perfectly delineated building footprints and

may cause a certain degree of overlap of the landscape metrics between the classes. Therefore, looking only at the relative distribution of the silhouette coefficients the best stratification is achieved by a guided sampling at the segment level. This stratification even outperforms the manually labelled reference data. Therefore, a strong component of personal judgement in both the training and testing phases of a typical remote sensing classification work flow becomes apparent. This means that even for a skilled human operator it is often difficult to clearly distinguish different landuse / landcover classes and thus manually labelled classifications are not always necessarily providing the best stratification.

In summary, the study could highlight the strong influence of the training stage on the classification performance. Especially when tackling a rather difficult classification task on noisy data with high training costs, feature selection and parameter tuning are crucial to improve the results. For the given experimental setup it can be concluded that given optimized feature space and classifier parameters, training a SVM with segment-shaped samples that were sampled in a guided manner and are balanced between the classes provided the best classification results (Precision: 0.68; Recall: 0.65; F_1 : 0.65). If square-shaped samples are used, a random sampling provided better results (Precision: 0.67; Recall: 0.58; F_1 : 0.56) than a guided selection. Equally balanced sample distributions always outperformed unbalanced training sets. A comparison with a very high resolution LiDAR classification could underline these findings and highlighted that manually labelled classifications are not always necessarily providing the best stratification.

The results of this study will be used as base-level stratification to further characterize the exposed building stock (Wieland et al. 2015) and to infer its vulnerability (Pittore and Wieland 2013) as input for a seismic risk assessment of the larger urban agglomeration of Port-au-Prince. Traditionally, the seismic vulnerability assessment is performed within samples which are selected by the risk analyst according to their knowledge of the study area. Then, the building stock of these samples is classified into model building types, to which a vulnerability function is allocated. Finally, the vulnerability model is extrapolated to the whole city in order to estimate the earthquake risk. The fact that the samples of different

urban patterns are selected by the analyst introduces a component of subjectivity which may influence the results. With the approach we present here, the different urban patterns are delimited using machine learning procedures, what restricts the participation of the risk analyst. This way the subsequent sampling becomes an informed, less subjective process. Hence, a better sampling process implies a better inference of the vulnerability of the population (the whole study area). Moreover, given the semi-automated character of the procedure, the elaboration of exposure databases becomes a faster task that is able to cope with the rapid urban sprawl of expanding cities. Finally, the exposure and vulnerability analysis carried out on the basis of remotely sensed data will enable to evaluate the seismic vulnerability even in cities where cadastre is not available. This fact opens the door to an ideal global earthquake risk model monitoring and mitigation.

Acknowledgements

The authors would like to thank the editors and the anonymous reviewers for suggestions which helped to improve this paper and K. Fleming for English language revision. This study has been supported by the SENSUM project (Grant Agreement Number 312972) and a research scholarship jointly funded by the Technical University of Madrid (UPM) and the German Academic Exchange Programme (DAAD).

References

- Arranz, J. J. 2013. "Diseño, Optimización Y Análisis de Sistemas Basados En Técnicas Láser, Para El Modelado Geométrico, Registro Y Documentación, Aplicados a Entidades de Interés Patrimonial." Phd Thesis, Madrid: Technical University of Madrid.
- Benito, B., D. Belizaire, Y. Torres Fernández, J.J. Martínez Díaz, V. Huérfano, E. Polanco, R. Garcia, P. González-Crende, A.R. Serna Martínez, and F. Zevallos. 2012. "An Evaluation of Seismic Hazard in La Hispaniola, after the 2010 Haiti Earthquake." In: *XXXIII General Assembly European Seismological Commission*, Moscow, Russia, <http://oa.upm.es/21850/>.

- Calais, E., A. Freed, G. Mattioli, F. Amelung, S. Jónsson, P. Jansma, S. Hong, T. Dixon, C. Prépetit, and R. Momplaisir. 2010. "Transpressional Rupture of an Unmapped Fault during the 2010 Haiti Earthquake." *Nature Geoscience* 3 (11): 794–99. doi:10.1038/ngeo992.
- Chang, C., and C. Lin. 2011. "LIBSVM: A Library for Support Vector Machines." *ACM Transactions on Intelligent Systems and Technology (TIST)* 2 (3): 27.
- Dey, V., Y. Zhang, and M. Zhong. 2010. "A Review on Image Segmentation Techniques with Remote Sensing Perspective." In: Wagner W., Székely, B. (eds.): *ISPRS TC VII Symposium – 100 Years ISPRS*, Vienna, Austria.
- Fawcett, T. 2006. "An Introduction to ROC Analysis." *Pattern Recognition Letters* 27 (8): 861–74. doi:10.1016/j.patrec.2005.10.010.
- Felzenszwalb, P. and D. Huttenlocher. 2004. "Efficient Graph-Based Image Segmentation." *International Journal of Computer Vision* 59 (2): 167–81.
- Foody, G.M., and A. Mathur. 2004. "Toward Intelligent Training of Supervised Image Classifications: Directing Training Data Acquisition for SVM Classification." *Remote Sensing of Environment* 93 (1-2): 107–17. doi:10.1016/j.rse.2004.06.017.
- Frankel, A., S. Harmsen, C. Mueller, E. Calais, and J. Haase. 2011. "Seismic Hazard Maps for Haiti." *Earthquake Spectra* 27 (S1): S23–41.
- Griffiths, P., P. Hostert, O. Gruebner, and S. van der Linden. 2010. "Mapping Megacity Growth with Multi-Sensor Data." *Remote Sensing of Environment* 114 (2): 426–39. doi:10.1016/j.rse.2009.09.012.
- Guyon, I., J. Weston, S. Barnhill, and V. Vapnik. 2002. "Gene Selection for Cancer Classification Using Support Vector Machines." *Machine Learning* 46 (1-3): 389–422.
- Haralick, R., K. Shanmugam, and I. Dinstein. 1973. "Textural Features for Image Classification." *IEEE Transactions on System, Man and Cybernetics* 3 (6): 610–621.
- Herold, M., X. Liu, and K.C. Clarke. 2003. "Spatial Metrics and Image Texture for Mapping Urban Land Use." *Photogrammetric Engineering & Remote Sensing* 69 (9): 991–1001.
- Huberty, C. J. 1994. *Applied Discriminant Analysis*. New York: Wiley.

- Hughes, G. 1968. "On the Mean Accuracy of Statistical Pattern Recognizers." *IEEE Transactions on Information Theory* 14 (1): 55–63. doi:10.1109/TIT.1968.1054102.
- Johnson, B., and Z. Xie. 2011. "Unsupervised Image Segmentation Evaluation and Refinement Using a Multi-Scale Approach." *ISPRS Journal of Photogrammetry and Remote Sensing* 66 (4): 473–483.
- Kononenko, I. 1994. "Estimating Attributes: Analysis and Extensions of Relief." In: *Lecture Notes in Computer Science* 784: 171–182.
- Li, C., J. Wang, L. Wang, Luanyun Hu, and Peng Gong. 2014. "Comparison of Classification Algorithms and Training Sample Sizes in Urban Land Classification with Landsat Thematic Mapper Imagery." *Remote Sensing* 6 (2): 964–83. doi:10.3390/rs6020964.
- Löw, F., U. Michel, S. Dech, and C. Conrad. 2013. "Impact of Feature Selection on the Accuracy and Spatial Uncertainty of per-Field Crop Classification Using Support Vector Machines." *ISPRS Journal of Photogrammetry and Remote Sensing* 85: 102–19. doi:10.1016/j.isprsjprs.2013.08.007.
- McGarigal, K., S. Cushman, and E. Banzhaf. 2012. *FRAGSTATS: Spatial Pattern Analysis Program for Categorical and Continuous Maps* (version 4). <http://www.umass.edu/landeco/research/fragstats/fragstats.html>.
- "Open Topography." 2015. Accessed December 7. http://www.opentopography.org/index.php/blog/detail/haiti_lidar_imagery_in_google_earth.
- Pal, M., and G.M. Foody. 2012. "Evaluation of SVM, RVM and SMLR for Accurate Image Classification with Limited Ground Data." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 5 (5): 1344–55. doi:10.1109/JSTARS.2012.2215310.
- Peña, J., P. Gutiérrez, C. Hervás-Martínez, J. Six, R. Plant, and F. López-Granados. 2014. "Object-Based Image Classification of Summer Crops with Machine Learning Methods." *Remote Sensing* 6 (6): 5019–41. doi:10.3390/rs6065019.

- Pittore, M., and M. Wieland. 2013. "Toward a Rapid Probabilistic Seismic Vulnerability Assessment Using Satellite and Ground-Based Remote Sensing." *Natural Hazards* 68 (1): 115–45. doi:10.1007/s11069-012-0475-z.
- "Plan de Prévention Séisme Pour Le Grand Nord." 2015. Accessed December 6. http://www.ht.undp.org/content/haiti/fr/home/operations/projects/crisis_prevention_and_recovery/plan-de-prevention-seisme-pour-le-grand-nord-dhaiti.html.
- "Population of Haiti." 2015. Accessed December 6. <http://www.citypopulation.de/Haiti.html>.
- Qian, Y., W. Zhou, J. Yan, W. Li, and L. Han. 2014. "Comparing Machine Learning Classifiers for Object-Based Land Cover Classification Using Very High Resolution Imagery." *Remote Sensing* 7 (1): 153–68. doi:10.3390/rs70100153.
- Rousseuw, P.J. 1987. "Silhouettes: A Graphical Aid to the Interpretation and Validation." *Journal of Computational and Applied Mathematics* 20: 53–65.
- Schneider, A. 2012. "Monitoring Land Cover Change in Urban and Peri-Urban Areas Using Dense Time Stacks of Landsat Satellite Data and a Data Mining Approach." *Remote Sensing of Environment* 124: 689–704. doi:10.1016/j.rse.2012.06.006.
- "Scikit-Learn." 2015. Accessed January 8. <http://scikit-learn.org>.
- Song, C., C. E. Woodcock, K. C. Seto, M. P. Lenney and S.A. Macomber. 2001. "Classification and change detection using Landsat TM data: when and how to correct atmospheric effects?" *Remote sensing of Environment* 75 (2): 230-244.
- Tobler, W. 1970. "A Computer Movie Simulating Urban Growth in the Detroit Region." *Economic Geography* 46 (2): 234–40.
- Torres, Y., S. Molina, S. Martínez-Cuevas, M. Navarro, J. Martínez-Díaz, B. Benito, J. Galiana, and D. Belizaire. In preparation. "A First Approach to the Earthquake Loss Estimation in Haiti. Advises to Minimize the Seismic Risk"
- "USGS Earthexplorer." 2015. Accessed January 5. <http://earthexplorer.usgs.gov/>.
- Vapnik, V. 2000. *The Nature of Statistical Learning Theory*. New York: Springer.
- Weiss, G.M., and F. Provost. 2003. "Learning When Training Data Are Costly: The Effect of Class Distribution on Tree Induction." *Journal of Artificial Intelligence Research* 19: 315–354.
- "WEKA3 Data Mining with Open Source Machine Learning Software." 2012. Accessed October 29. <http://www.cs.waikato.ac.nz/ml/weka/>.

- Wieland, M., and M. Pittore. 2014. "Performance Evaluation of Machine Learning Algorithms for Urban Pattern Recognition." *Remote Sensing* 6 (4): 2912–39.
- Wieland, M., M. Pittore, S. Parolai, U. Begaliev, P. Yasunov, J. Niyazov, S. Tyagunov, et al. 2015. "Towards a Cross-Border Exposure Model for the Earthquake Model Central Asia." *Annals of Geophysics* 58 (1): S0106. doi:10.4401/ag-6663.

Table 1: Urban structure type patterns identified for the study area with sample examples superimposed on the input Landsat image and a very high-resolution aerial image. The square-shaped samples are 200 m x 200 m.

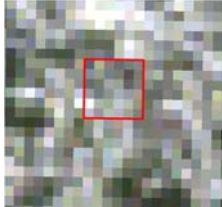



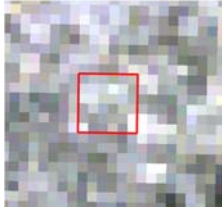

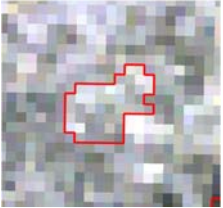

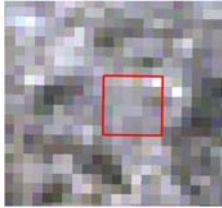

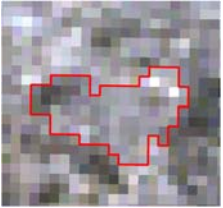









UST class	Square-shaped sample		Segment-shaped sample	
1 – Residential				
2 – Urban regular				
3 – Urban irregular				
4 – Rural				
5 – Industrial / Commercial				

Table 2: References datasets and their characteristics.

Reference dataset	R50-SQ-RD	R50-SE-RD	R15-SQ-RD	R15-SE-RD	RUB-SQ-RD	RUB-SE-RD	R15-SQ-GD	R15-SE-GD
Sample size	250	250	75	75	75	75	75	75
Sample type	Square	Segment	Square	Segment	Square	Segment	Square	Segment
Sample approach	Random (balanced)	Random (balanced)	Random (balanced)	Random (balanced)	Random	Random	Guided (balanced)	Guided (balanced)

Table 3: Examples of LiDAR classifications of urban structure type patterns superimposed on a very high-resolution aerial image.

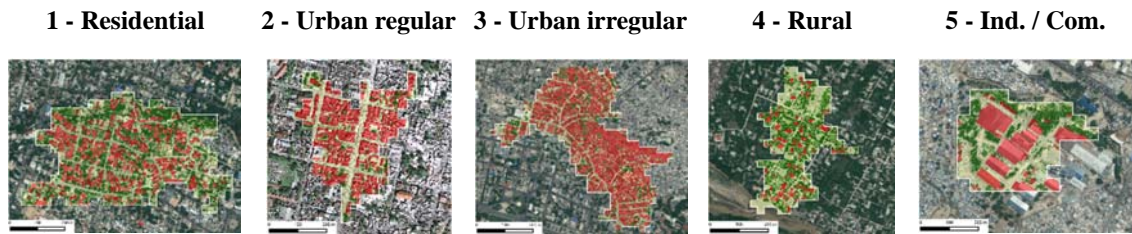


Table 4: Mean silhouette coefficients computed for the different Landsat classifications, based on landscape metrics derived from a very high-resolution LiDAR classification.

Classification	Ref label	R15-SQ-RD	R15-SE-RD	RUB-SQ-RD	RUB-SE-RD	R15-SQ-GD	R15-SE-GD
Silhouette coefficient	-0.04	-0.07	-0.04	-0.06	-0.06	-0.12	-0.01

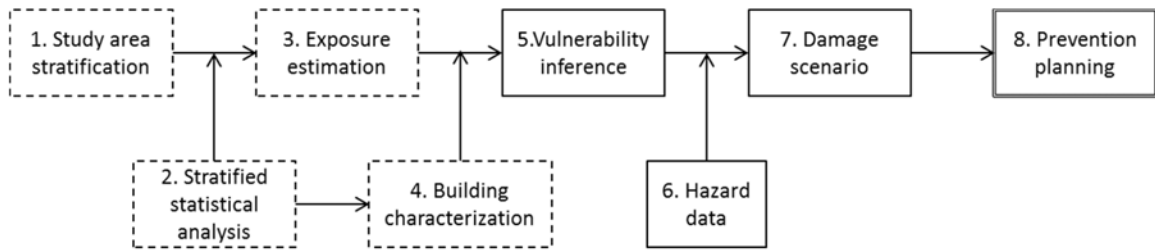


Figure 1: Overview of the method.

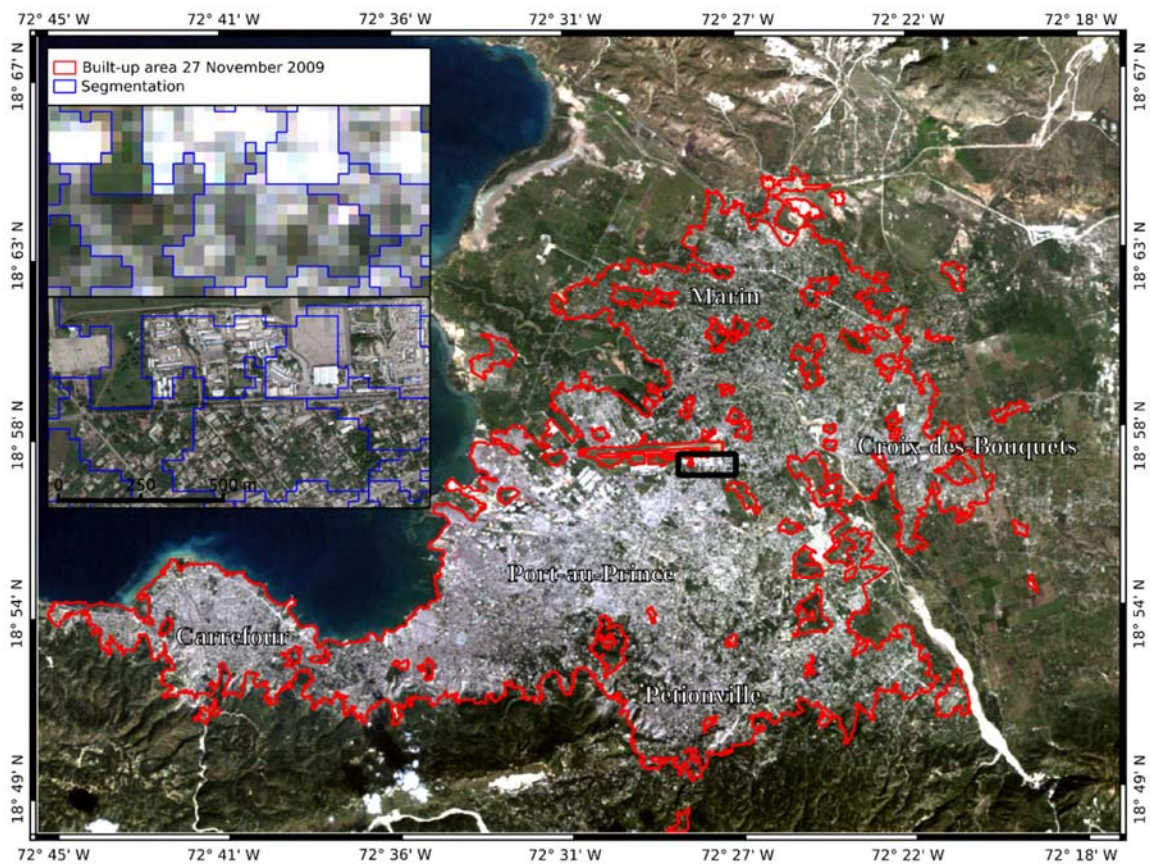


Figure 2: Study area Port-au-Prince and surroundings, Haiti. Landsat TM image of the study area, extracted built-up area, and distribution of reference data. The zoom shows a subset of the segmentation result superimposed on the input Landsat image (top) and on a very high resolution aerial image (bottom).

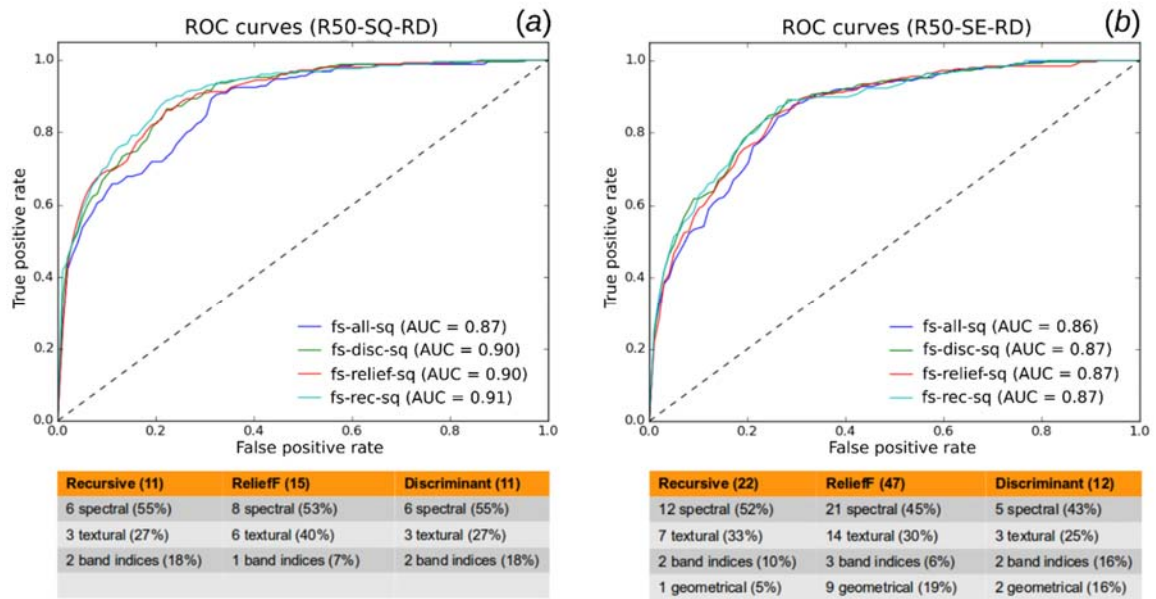


Figure 3: Micro-average ROC curves derived for (a) the squares reference data set (R50-SQ-RD) and (b) the segments reference data set (R50-SE-RD) under consideration of different feature subsets.

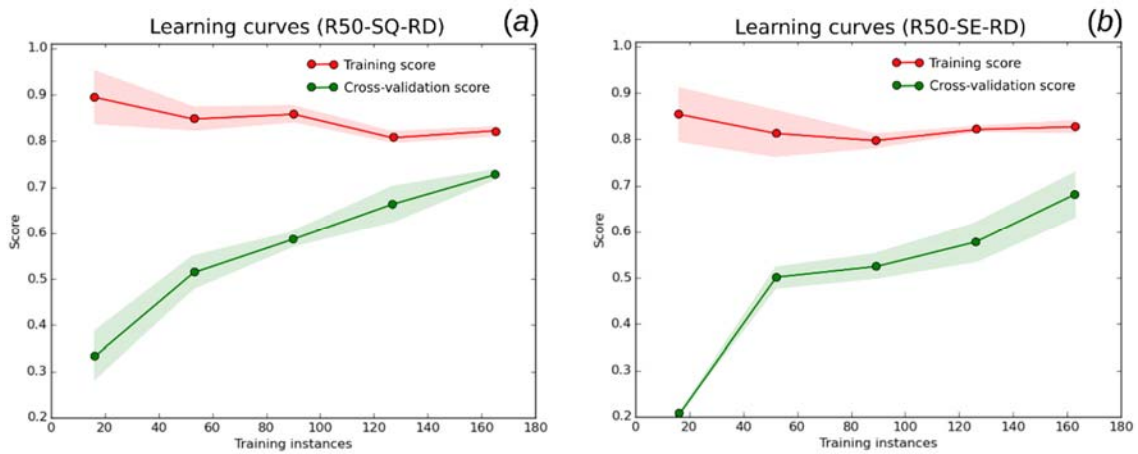


Figure 4: Learning curves derived for (a) the squares (R50-SQ-RD) and (b) the segments reference data sets (R50-SE-RD) under consideration of the selected feature subsets from recursive feature selection.

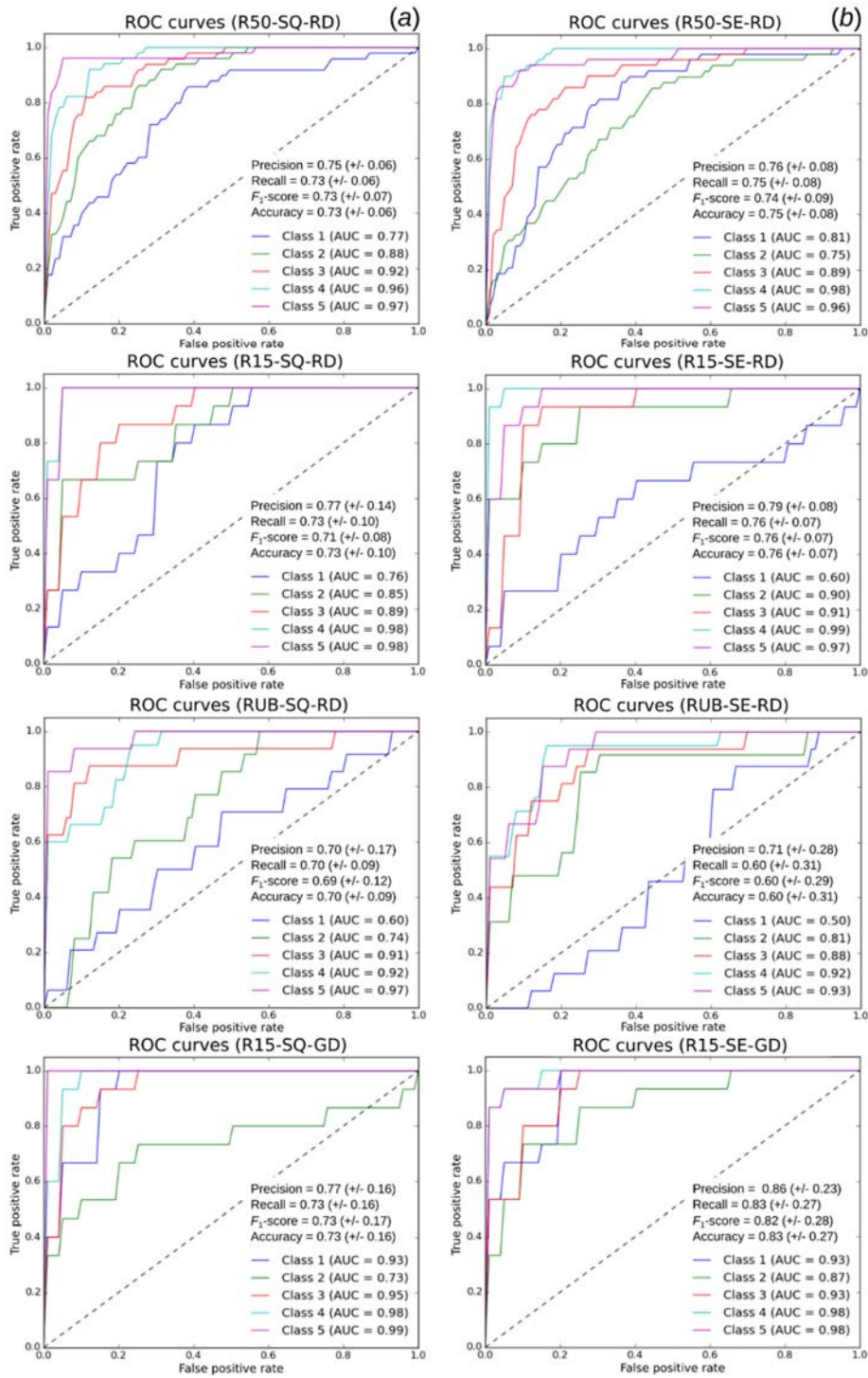


Figure 5: ROC curves and performance measures derived for (a) the squares and (b) the segments reference data sets under consideration of different sample size, types, and approaches.

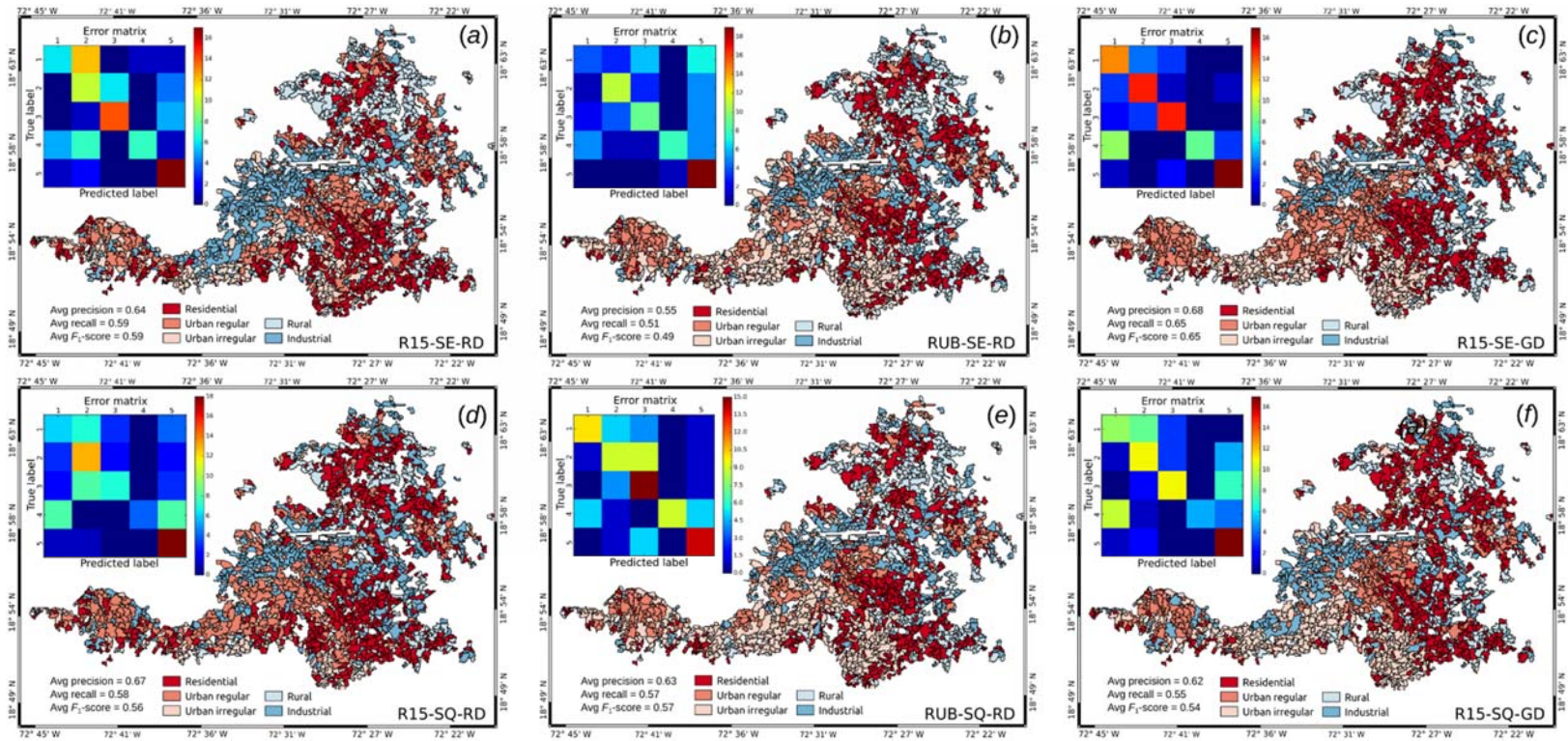


Figure 6: Classifications of the study area with error matrices and average (avg) performance measures under consideration of different sample types and approaches.

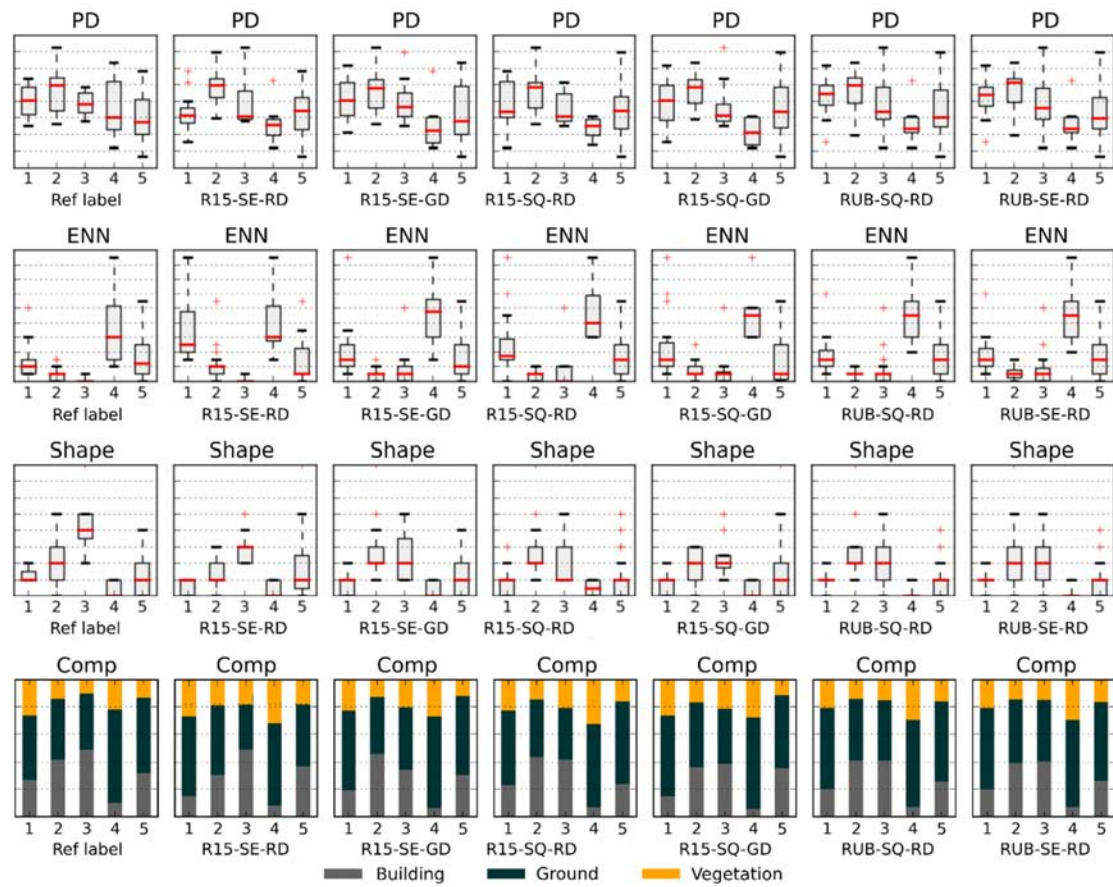


Figure 7: Comparison of the Landsat classifications with a very high resolution classification from lidar data.