ON THE EFFECT OF SPATIALLY NON-DISJOINT TRAINING AND TEST SAMPLES ON ESTIMATED MODEL GENERALIZATION CAPABILITIES IN SUPERVISED CLASSIFICATION WITH SPATIAL FEATURES

Christian Geiß[1], *Member, IEEE*, Patrick Aravena Pelizari[1], Henrik Schrade[1], Alexander Brenning[2], and Hannes Taubenböck[1]

[1] German Remote Sensing Data Center (DFD), German Aerospace Center (DLR), 82234 Wessling-Oberpfaffenhofen, Germany; christian.geiss@dlr.de, patrick.aravenapelizari@dlr.de, henrik.schrade@dlr.de, hannes.taubenboeck@dlr.de

[2] Department of Geography, Friedrich Schiller University Jena, 07743 Jena, Germany; alexander.brenning@uni-jena.de

Abstract— **In this letter, we establish two sampling schemes to select training and test sets for supervised classification. We do this in order to investigate whether estimated generalization capabilities of learned models can be positively biased from the usage of spatial features. Numerous spatial features impose homogeneity constraints on the image data, whereby a spatially connected set of image elements is attributed identical feature values. In addition to a frequent occurrence of intrinsic spatial autocorrelation, this leads to extrinsic spatial autocorrelation with respect to the image data. The first sampling scheme follows a spatially random partitioning into training and test sets. In contrast to that, the second strategy implements a spatially disjoint partitioning, which considers in particular topological constraints that arise from the deployment of spatial features. Experimental results are obtained from multi- and hyperspectral acquisitions over urban environments. They underline that a large share of the differences between estimated generalization capabilities obtained with the spatially disjoint and non-disjoint sampling strategy can be attributed to the use of spatial features, whereby differences increase with an increasing size of the spatial neighborhood considered for computing a spatial feature. This stresses the necessity of a proper spatial sampling scheme for model evaluation to avoid overoptimistic model assessments.**

***Index Terms*— Supervised Classification, Spatial Features, Morphological Profiles, Random Forests, Model Generalization Capability, Multispectral Images, Hyperspectral Images.**

I. INTRODUCTION

The development of methods for the derivation of thematic information such as land use / land cover (LULC) classes from remote sensing imagery has been a major research subject of the remote sensing community in the past decades. Thereby, varying ground sampling distances of individual sensors induced the development of diverse methodological approaches. In this work, we focus on situations where the ground sampling distance is much smaller than the objects of interest of a scene. This situation can occur in various remote sensing data, depending on the relation of ground sampling distance and corresponding size of the objects of interest. Nowadays, especially data from sensors with a very high spatial resolution such as WorldView I-III, or GeoEye, among others, feature this situation. Thereby, the high spatial resolution can induce high intra-class and low inter-class variability in particular in heterogeneous environments such as urban areas. This can decrease accuracy of the classification model and induce the well-known salt and pepper effect [1].

One of the most prominent ways to cope with this problem and ensure coherent spatial regularization is to *compute features which account for the neighborhood of an individual image element*, i.e., spatial features. Examples of such kinds of features are morphological profiles (MPs) (i.e., morphological transformations of the image data based on the sequential application of a structuring element (SE) with increasing size [2], [3]), texture filters [4], variation indices [5], and multi-level object-based image analysis approaches [6], [7], among others. Thereby, a considerable number of spatial features impose homogeneity constraints on the image data and attribute identical feature values to image elements in close spatial vicinity. Popular examples are MPs, which assign minimum or maximum values within a defined neighborhood to an individual image element.

Subsequently, those spatial features are fed to a learning machine (e.g., Support Vector Machine or Random Forest). There, a popular strategy is to learn the model and also optimize its hyperparameters based on a *training set* using labeled samples of relevant thematic classes (relying for instance on a *k*-fold cross-validation for hyperparameter tuning), and estimate the generalization capabilities of a learned model for unseen data based on an independent *test set* (i.e., holdout) [8]. Thereby, numerous studies do not strictly take topological (neighborhood) relations of the image elements (here pixels) of training and test set into account. In this letter, we investigate whether this can lead to substantially biased estimates of the generalization capabilities of learned models – especially when relying on spatial features for classification. This can be related to the fact that nearby image elements tend to show a high degree of similarity in the feature space not only because of the frequent presence of *intrinsic spatial autocorrelation* (i.e., image elements nearby tend to be more similar than image elements farer apart [9]) but also heavily due to the aforementioned homogeneity constraints, which are imposed on the image by certain spatial features. This can be interpreted as *extrinsic spatial autocorrelation*, where a spatially connected set of image elements is attributed identical feature values.

Although numerous studies establish efficient sample selection strategies, recent attempts aim to specifically account for spatial autocorrelation in accuracy assessment for supervised classification. Brenning [10] proposes spatial cross-validation and bootstrap to obtain performance estimates that are not biased by spatial autocorrelation. In its presence, an overfitted model cannot be distinguished from a model with

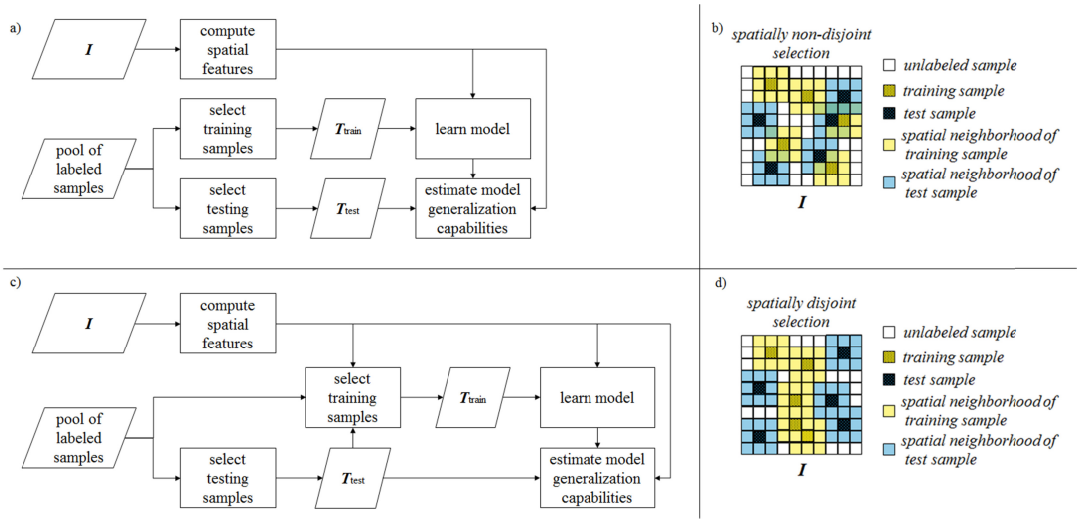

Fig. 1. Two different strategies for compiling training set and test set. (a) Spatially non-disjoint strategy for estimation of model generalization capabilities. The only constraint is that samples of the training set must not be included in the test set (b). (c) Spatially disjoint strategy for estimation of model generalization capabilities, where it is ensured that the spatial neighborhoods of the features of the samples of training set and test set do not overlap (d).

high generalization capabilities if the test set is not spatially independent from the training set (i.e., the accuracy estimates of the overfitted model would also indicate high generalization capabilities when assessed with a test set which does not feature inherent spatial independence).

In this letter, we aim to uniquely *investigate and quantify the effect of intrinsic and extrinsic spatial autocorrelation on estimated generalization capabilities of supervised classification models* (i.e. we investigate how substantial are the differences of estimated accuracies). This is done by quantifying model accuracies as related to two different strategies for partitioning the data into training and test sets. The first strategy follows a spatially random selection of training and test sets. In contrast to that, the second strategy implements a spatially disjoint partitioning into training and test regions considering also topological constraints arising from the application of spatial features (i.e., MPs).

The remainder of the letter is organized as follows. Section II documents the two different strategies for establishing training and test set for model learning. Section III is used to describe data sets and parameterization of methods. Results of the actual experiments are reported in section IV. We give concluding remarks in section V.

II. Supervised Classification with Spatial Features and Different Sampling Strategies

Let us consider an image (e.g., acquired by a multi- or hyperspectral sensor) I_F constituted by a single or multiple spectral bands F. The spectral information allows deriving spatial features using θ_W-parametrized functions $\phi_{\theta_W}(\cdot)$, which consider a spatial neighborhood W of each image element $\mathbf{x}_i \in \mathbb{R}^F$. Those functions map an image element into the feature space of a filter determined by θ_W [11]. For an exhaustive description of an image, a set of spatial features φ is derived, consisting of d features $\varphi = \left\{\phi_{\theta_{W_j}}\right\}_{j=1}^{d}$. For classification, a stacked matrix can be compiled, i.e., $\Phi_{F,\varphi} \in \mathbb{R}^{I \times F,d}$, containing the spectral bands F and d spatial features in φ computed for all elements of I. Following a supervised approach, a pool of labeled samples $S = \{X, Y\}$ is given, where $X = \{\mathbf{x}_l\}_{l=1}^{n} \in \Phi_{F,\varphi}$ is a subset of $\Phi_{F,\varphi}$ made up of n labeled samples, and associated labels $Y = \{\mathrm{y}_l\}_{l=1}^{n} \in \{1, \ldots, C\}$ of C classes. Subsequently, a model is learned to assign a class label to all unlabeled instances $X^* = \{\mathbf{x}_u^*\}_{u=1}^{m} \in \Phi_{F,\varphi}$, i.e., $f(\mathbf{x}_u^*) = sign(\cdot)$.

A. Spatially Non-disjoint Training and Test Sets

The first strategy for establishing training set and test set is referred to as *spatially non-disjoint* [Fig. 1a]. Thereby, the pool of labeled samples is split in training set $T_{train} = \{\mathbf{x}_l, y_l\}_{l=1}^{j} \in S$ and test set $T_{test} = \{\mathbf{x}_k, y_k\}_{k=j+1}^{n} \in S$. The only constraint is that samples of the training set must not be included in the test set, i.e., $T_{train} \cap T_{test} = \emptyset$ [Fig. 1b]. Preferably, S is drawn from the complete image data to avoid a possible shift of covariance (i.e., induce a bias related to domain adaption). Consequently, this strategy does not account for autocorrelation and, thus, is biased towards over-optimistic estimates of generalization capabilities. This is particularly the case for a progressive increase of the size of the considered spatial neighborhood of a spatial filter.

B. Spatially Disjoint Training and Test Sets

The second strategy for establishing training set and test set is referred to as *spatially disjoint* [Fig. 1c]. Labeled samples are drawn from the complete image data, whereby feature vectors of instances of training and test set are not attributed identical feature values given the application of a spatial filter. Thus, it is ensured that the spatial neighborhoods of the features of the samples of training and test set do not show any overlap [Fig. 1d]. We implemented this constraint with a random compilation of T_{test} from the complete image data and compilation of T_{train} from the residual areas. In order to ensure comparability with results from the spatially non-disjoint sampling strategy, training samples are drawn from within these residual areas without any spatial constraints. This spatial partitioning strategy is believed to reveal less

biased estimates compared to the spatially non-disjoint strategy since both intrinsic and extrinsic spatial autocorrelation is consistently considered.

III. Data and Experimental Setup

A. Data Sets

We consider three data sets from multispectral, and hyperspectral acquisitions.

1) Munich: The image was acquired by the multispectral WorldView-II sensor. The subset from the panchromatic band has a size of 1000 × 1001 pixels with a resampled spatial resolution of 2 m, using nearest neighbor interpolation, to reduce the computational burden for the experiments. Labeled samples are available for the thematic classes "impervious" surfaces, "vegetation", and "shadow" areas [Fig. 2a]. Those labeled samples were determined based on photointerpretation analysis under consideration of aerial imagery and cadastral maps.

2) Cologne: The data set considered in the experiments is a subset of a multispectral image of 500 × 500 pixels acquired by the QuickBird sensor with a spatial resolution of 0.65 m. However, also here we resampled the image data to a spatial resolution of 2 m to reduce the computational burden [12]. Labeled samples are available for the same thematic classes as for the previous data set [Fig. 2b].

3) Pavia: The third data set is the well-known hyperspectral acquisition of the center of Pavia from the Digital Airborne Imaging Spectrometer (DAIS). It features an extent of 1096 × 1096 pixels with a spatial resolution of 1 m. Ground truth information is provided by the University of Pavia and is available for detailed urban LULC classes including water, trees, asphalt, parking lots, bitumen, brick roofs, meadows, bare soil, and shadows. However, labeled samples were aggregated to three thematic classes in order to being able to compile training sets of sufficient size, even when many test samples are jointly considered with a large neighborhood for the spatial features (i.e., strongly limiting the residual areas for compilation of training sets) [Fig. 2c].

B. Experimental Setup

The panchromatic band and first principal component were deployed from the multi- and hyperspectral acquisitions, respectively, to compute the spatial features [Fig. 2]. This is done to establish a situation with limited spectral resolution, which specifically encourages the use of spatial features. The spatial features considered for the experiments consist of MPs regarding the panchromatic imagery and an extended MP [13] regarding the hyperspectral imagery using opening and closing operations [14]. Thereby, a square-shaped SE of linear increasing size $B = \{3,5,\dots,21\}$ was used, since this parameterization showed viable performance properties in comparable settings previously [7], [12]. In the subsequent section, experimental results are presented as a function of an increasing size of B. Thereby, it was made sure that a feature vector obtained with a comparatively smaller SE is part of a feature vector derived with a comparatively larger SE. Consequently, the feature vector obtained for B_{max} has 21 dimensions. To underline the validity of the presented experimental setup, we computed the Moran's I as global

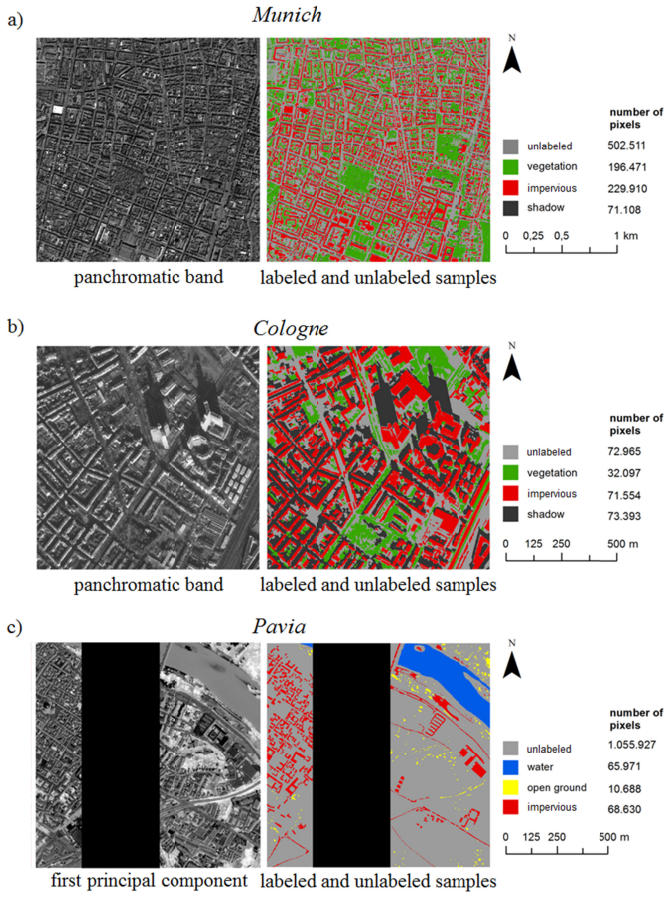

Fig. 2. Imagery for (a) Munich, (b) Cologne, and (c) Pavia with affiliated unlabeled samples and labeled samples of different thematic classes.

measure for spatial autocorrelation [15] for the closing operations of the Munich data set. This analysis revealed a strictly monotonic function I = {0.81,0.84,0.86,0.88,0.90,0.92,0.93,0.94,0.95,0.96}, which confirms an increasing level of *extrinsic spatial autocorrelation* with respect to an increasing size of the considered spatial neighborhood.

For learning the actual classification model, a Random Forest (RF) approach [16] was deployed. RF is a decision-tree-based ensemble learning method for classification and regression. In concordance with previous studies [12], this nonparametric approach was used to account for a certain level of redundancy shown by the MPs (induced by the application of consecutively increasing sizes of the SE), which can be critical for the estimation of statistics in parametric approaches. The hyperparameters that need to be determined for generating a RF model consist of the number of classification trees to be grown n_{tree} and the number of features m_{try} used at each node. To establish a reliable error estimation and simultaneously maintain computation times in reasonable ranges, we selected an n_{tree} value of 500. A value for $m_{\text{try}} = \sqrt{p}$, with p denoting the number of input features, yields near optimum results [16], and we parametrized the models accordingly.

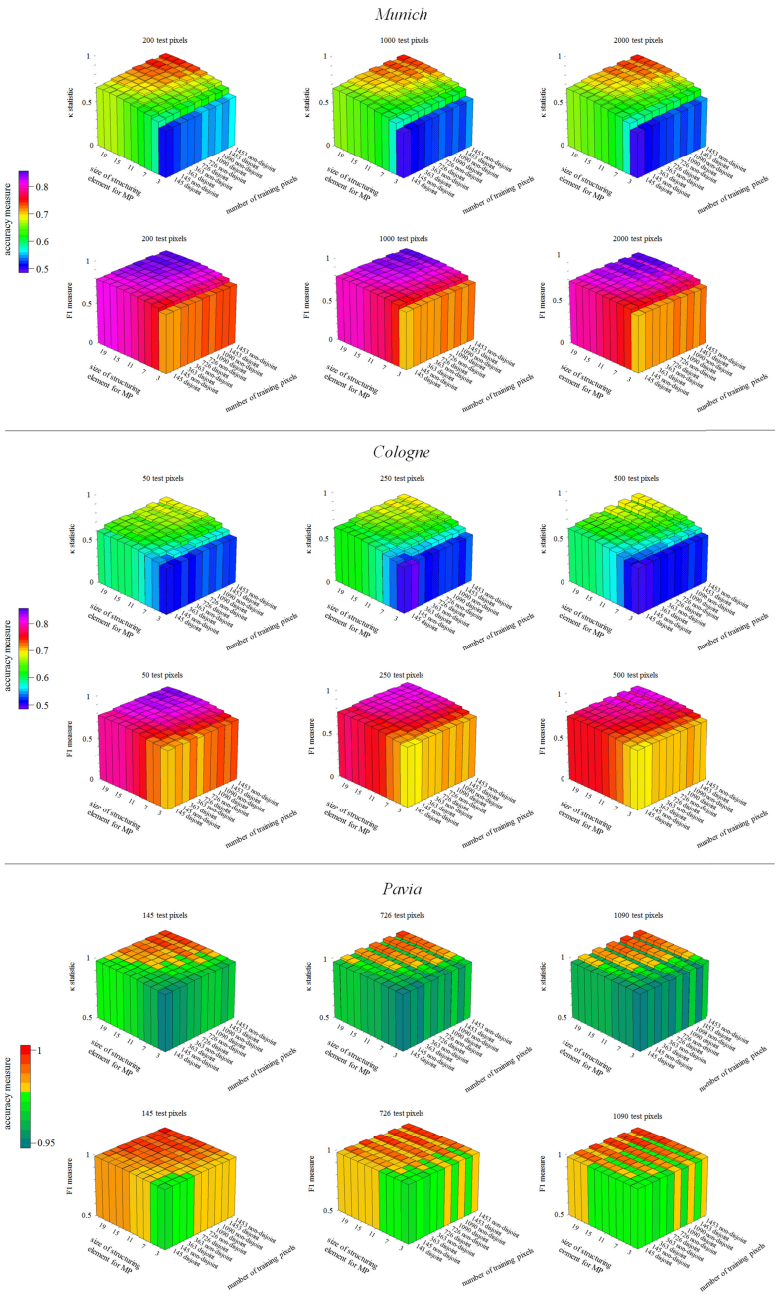

Fig. 3. Results for the image data of Munich, Cologne, and Pavia. Model generalization capabilities are quantified by means of κ statistic and $\bar{F}$ measure as a function of the size of the SE, and number of labeled samples of training set and test set for the spatially disjoint and non-disjoint sampling strategy.

Generalization capabilities of learned models are evaluated based on global accuracy measures, which allow considering both omission and commission errors. Results are reported as the average of 50 independent trials. In particular, κ statistic, and weighted mean $\bar{F}$ of the F-measures (weighted by the cardinals of the thematic classes) were computed. These measures were chosen, since they are less governed by class imbalance compared to other global measures such as overall accuracy. Experimental results are presented as a function of the size of the training and test set. It was made sure that samples contained in one set are also contained in the affiliated set with a larger number of samples to allow for quantifying the effect of training and test set size on prediction accuracy. Thereby, test samples were drawn randomly according to the prior probability of the classes, whereby training samples were selected with an equal number of samples per class to avoid possible problems related to class imbalance. Finally, it is important to note that labeled samples of all test sets of the spatially disjoint selection strategy were drawn according to B_{max} to enable a consistent comparative quantification of model accuracy with respect to an increasing size of B regarding a spatially non-disjoint selection (i.e.,

ensure that *differences* of the spatially disjoint and non-disjoint selection strategy are not influenced by varying sizes of the considered spatial neighborhood).

IV. Experimental Results

Experimental results obtained for the image data of *Munich*, *Cologne*, and *Pavia* are visualized in Fig. 3 as a function of the number of labeled samples of training and test set, and the size of the spatial neighborhood considered for both the spatially disjoint and non-disjoint sampling strategy. First of all, it can be noted that overall accuracy levels vary between the different data sets. The highest levels with the smallest variations in terms of numeric values of the considered accuracy measures are obtained from the *Pavia* data set, whereas obtaining the correct thematic classes is most challenging for the *Cologne* data set.

Within an individual plot (i.e., for a fixed number of test samples), generally, an increasing number of training samples allows for an increasing level of accuracy. This is an intuitive result since more prior knowledge is encoded in the models. However, estimated generalization capabilities are distinctively higher when partitioning non-spatially compared to the spatially disjoint strategy. Additionally we find those differences increase with an increasing number of training samples. This can be related to an increased number of samples lying within an area affected by spatial autocorrelation. Likewise, differences between the spatially disjoint and non-disjoint sampling strategy also increase with an increasing size of the SE (i.e., size of the spatial neighborhood considered for a spatial feature). This observation can be attributed to enlarged areas, which induce spatial autocorrelation and underline the significant influence of the use of spatial features. In particular, we observe for our data sets, that those differences can reach up to 5.2 percentage points (p.p.) in κ statistic and 3.2 p.p. in $\bar{F}$ measure. Thereby, it can be noted that a slight decrease of accuracy especially for the spatially disjoint strategy is observable between individual plots (i.e., for an increasing number of test samples) for a data set. This can be related to the circumstance that few areas for selecting training samples are left when a large number of test samples is drawn in relation to the size of the image data. This can lead to undersampling if the image data is highly heterogeneous. Nevertheless, this is for instance hardly observable for the *Munich* data set, which is the most homogeneous data set considered. There, differences in κ and $\bar{F}$ reach up to 4.2 p.p. and 2.6 p.p., respectively, which unambiguously underlines the substantial influence of spatial autocorrelation especially when using spatial features.

V. Conclusion

In this letter, we investigated whether estimated generalization capabilities of supervised classification models are positively biased without a proper spatial sampling scheme that considers topological relations in establishing training and test sets. Particular emphasis was on the use of spatial features (i.e., MPs) for classification. We reasoned that those spatial features induce *extrinsic spatial autocorrelation* in addition to *intrinsic spatial autocorrelation* due to homogeneity constraints, which are imposed on the image data, and the corresponding attribution of identical feature values to multiple image elements in close spatial proximity. To test this conjecture, we followed two different strategies for partitioning into training and test sets. The first strategy established a spatially random selection, whereas the second strategy implements a spatially disjoint selection considering also topological constraints arising from the application of spatial features.

Experimental results were obtained from multi- and hyperspectral acquisitions over varying urban environments. Spatial features were computed based on the concept of MPs, and models were learned within RF architecture. Our results point out that a large share of the differences between the accuracies obtained with the spatially disjoint and non-disjoint sampling strategies can be attributed to the use of spatial features. Differences increase with an increasing size of the spatial neighborhood considered for computing a spatial feature. This work underlines the necessity of appropriate strategies for establishing training and test areas in a spatially disjoint way and, thus, learning models that are not influenced by intrinsic or extrinsic spatial autocorrelation. Since different classifiers tend to show different degrees of (over/under)fitting the training data, future research should also investigate additional classifiers in order to generalize our findings.

VI. Acknowledgements

This work was supported by the German Federal Ministry for Economic Affairs and Energy's initiative "Smart Data—innovations from data" under grant agreement: "smart data for catastrophe management (sd-kama, 01MD15008B)". The work of Christian Geiß was supported by the Helmholtz Association under the grant "pre_DICT" (PD-305). We would like to thank Paolo Gamba for provision of the hyperspectral imagery, and. acknowledge the valuable works of Henry Schubert (University of Jena) on this topic. We also want to thank the two anonymous reviewers for very helpful comments.

References

[1] T. Blaschke, "Object based image analysis for remote sensing," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 65, no. 1, pp. 2–16.
[2] M. Pesaresi, and J. A. Benediktsson, "A New Approach for the Morphological Segmentation of High-Resolution Satellite Imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 39, no. 2, pp. 309–320, 2001.
[3] D. Tuia, F. Pacifici, M. Kanevski, and W. J. Emery, "Classification of very high spatial resolution image ry using mathematical morphology and support vector machines*," IEEE Trans. Geosci. Remote Sens*., vol. 47, no. 11, pp. 3866–3879, Nov. 2009.
[4] A. Brenning, S. Long, and P. Fieguth, "Detecting rock glacier flow structures using Gabor filters and IKONOS imagery," *Remote Sensing of Environment*, vol. 125, pp. 227-237, 2012.
[5] X. Huang, Q. Lu, and L. Zhang, "A multi-index learning approach for classification of high-resolution remotely sensed images over urban areas," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 90, no. 1, pp. 36–48.
[6] L. Bruzzone, and L. Carlin, "A Multilevel Context-Based System for Classification of Very High Spatial Resolution Images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 44, no. 9, pp. 2587–2600, 2006.
[7] C. Geiß, M. Klotz, A. Schmitt, and H. Taubenböck, „Object-based Morphological Profiles for Classification of Remote Sensing Imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 10, pp. 5952–5963, 2016.
[8] C. Persello, and L. Bruzzone, "A Novel Protocol for Accuracy Assessment in Classification of Very High Resolution Images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 3, pp. 1232–1244, 2006.
[9] W. R. Tobler, "A Computer Movie Simulating Urban Growth in the Detroit Region," Econ. Geogr, vol. 46, pp. 234–240, 1970.
[10] A. Brenning, "Spatial Cross-Validation and Bootstrap for the Assessment of Prediction Rules in Remote Sensing: The R Package Sperrorest," Proceedings, 2012 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 23-27 July 2012, pp. 5372-5375, 2012.
[11] D. Tuia, R. Flamary, and N. Courty, "Multiclass feature learning for hyperspectral image classification: sparse and hierarchical solutions," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 105, pp. 272–285, 2015.
[12] C. Geiß, and H. Taubenböck, "Object-based Postclassification Relearning," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 11, 2336-2340, 2015.
[13] J.A. Benediktsson, J.A. Palmason, J.R. Sveinson, „Classification of hyperspectral data from urban areas based on extended morphological profiles," *IEEE Trans. Geosci. Remote Sens*., vol. 43, no. 3, pp. 480–491, Feb. 2005.
[14] Soille, Morphological Image Analysis: Principles and Applications 2nd ed. Berlin, Germany: Springer-Verlag, 2004.
[15] P.A.P. Moran, "Notes on Continuous Stochastic Phenomena," *Biometrika*, vol. 37, no. 1, 17-23, 1950.
[16] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, 2001.