ISPRS Journal of Photogrammetry and Remote Sensing 140 (2018) 122–132

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

Skipping the real world: Classification of PolSAR images without explicit feature extraction

Ronny Hänsch *, Olaf Hellwich

Technische Universität Berlin, Computer Vision & Remote Sensing, Germany

ARTICLE INFO

Article history:
Received 17 March 2017
Received in revised form 28 November 2017
Accepted 30 November 2017
Available online 13 December 2017

Keywords:
Random Forest
PolSAR
Classification
Feature learning

ABSTRACT

The typical processing chain for pixel-wise classification from PolSAR images starts with an optional preprocessing step (e.g. speckle reduction), continues with extracting features projecting the complex-valued data into the real domain (e.g. by polarimetric decompositions) which are then used as input for a machine-learning based classifier, and ends in an optional postprocessing (e.g. label smoothing). The extracted features are usually hand-crafted as well as preselected and represent (a somewhat arbitrary) projection from the complex to the real domain in order to fit the requirements of standard machine-learning approaches such as Support Vector Machines or Artificial Neural Networks. This paper proposes to adapt the internal node tests of Random Forests to work directly on the complex-valued PolSAR data, which makes any explicit feature extraction obsolete. This approach leads to a classification framework with a significantly decreased computation time and memory footprint since no image features have to be computed and stored beforehand. The experimental results on one fully-polarimetric and one dual-polarimetric dataset show that, despite the simpler approach, accuracy can be maintained (decreased by only less than 2% for the fully-polarimetric dataset) or even improved (increased by roughly 9% for the dual-polarimetric dataset).
© 2017 International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS). Published by Elsevier B.V. All rights reserved.

1. Introduction

Synthetic Aperture Radar (SAR) is an active air- or space-borne sensor that emits a microwave signal and measures amplitude and phase (thus a complex number) of the echo which is backscattered at the ground. As active sensor it is independent of daylight and thus contrasts optical and hyperspectral sensors. Due to the electromagnetic properties of the used microwave it is less influenced by weather conditions and can penetrate clouds, dust, and to some degree even vegetation. Polarimetric SAR (PolSAR) transmits and measures in different polarisations which enables it to provide information contained neither in single channel SAR nor in other remote sensing data. The measured echoes depend on several properties including moisture, surface roughness, as well as object geometry and are therefore highly correlated to specific object categories. Due to these advantages there are nowadays many modern sensors available that provide PolSAR data, i.e. images that contain complex-valued vectors in each pixel (see Section 2).

The increasing amount of data, but also methodological problems of manual interpretation such as loss of information during visualisation and image effects human operators are unaccustomed with, led to a dire need of automatic procedures for PolSAR image analysis. The generation of semantic maps of land use/cover by pixelwise classification is one of the most typical and most important tasks of automatic interpretation of remote sensing images. It is usually solved within a supervised machine learning framework, where the internal parameters of a generic model are adjusted based on training data which contains the desired class labels alongside with the image data. The corresponding literature can be coarsely divided into two groups: Approaches that directly work on the (Pol)SAR data by modelling its statistical properties and methods that apply general purpose classifiers to extracted features. The first group has seen a lot of success in the early years of (Pol)SAR image classification and continues to propose new models that are better able to cope with the challenges of modern data. These mostly parametric approaches include models based on distributions such as Rayleigh (Kuruoglu and Zerubia, 2004), generalized Gaussian (Moser et al., 2006), generalized Gamma (Li et al., 2011), Weibull (Taravat et al., 2014), and Fisher (Tison et al., 2004) as well as models such as Finite Mixture Models

* Corresponding author.
E-mail address: r.haensch@tu-berlin.de (R. Hänsch).

https://doi.org/10.1016/j.isprsjprs.2017.11.022
0924-2716/© 2017 International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS). Published by Elsevier B.V. All rights reserved.

(Krylov et al., 2011) and approaches based on the Mellin transform (Nicolas and Tupin, 2016) that combine and generalize different distributions. These methods usually make parametric assumptions about the underlying data distributions, which tend to fail within heterogeneous regions or with contemporary high-resolution SAR data. If the ultimate goal is to derive a classification decision, discriminative approaches have been shown to be easier trained and to be more robust and accurate as such generative models. The general approach is to extract (often hand-crafted and class-specific) image features, which are descriptive enough for the classification task at hand, and use them as input to typical classifiers such as Support Vector Machines (SVMs, e.g. in Mantero et al., 2005), Multi-Layer Perceptrons (MLPs, e.g. in Bruzzone et al., 2004), or Random Forests (RFs, e.g. in Hänsch and Hellwich, 2010b). However, most of such classifiers are designed for real-valued input only and cannot easily be extended to the complex domain. This problem is solved by extracting real-valued features which are then used by the classifier. This feature extraction step states a couple of problems: Firstly, the extracted features are mostly hand-crafted and preselected for a specific classification task. The design and preselection of hand-crafted features usually involves manual trial and error. The creation of a descriptive feature set for typical tasks such as land use classification continues to be an active field of research. Despite an abundance of available features capturing polarimetric (such as polarimetric decompositions see e.g. Cloude and Pottier, 1996) or textural (e.g. as in He et al., 2013) information, it is unclear which combination yields the best results within the applied classification framework. Secondly, the computation of a diverse and descriptive set of features increases the computational load of the whole processing chain. On the one hand does the computation time increase. The features need to be computed from the PolSAR image and the computation time of many classifiers increases with a higher dimensional input space. On the other hand does the memory footprint of the method increase as well, since all the computed features need to be kept in memory to be accessible by the classifier. Thirdly, the more-or-less arbitrary projection of complex-valued data into the real-valued domain causes at best a dependency of the classification result on the used projection, i.e. feature extraction method. In the worst case it means a loss of valuable information and thus sub-optimal results.

In principle, there are two approaches to cope with the aforementioned problems: Either avoiding the extraction of real-valued features and aiming to model the relation between the PolSAR data itself and the class label directly, or using an exhaustive feature set which at least potentially contains all information necessary to solve the classification task.

Besides generative statistical models as discussed above, only very few approaches belong to the first group and aim to work directly on the complex-valued PolSAR data instead of extracting real-valued features. One example are complex-valued MLPs which have been applied to the task of land use classification e.g. in Hänsch (2010). Within these networks, input, weights, as well as output are modelled as complex-valued numbers. Thus, they can directly be applied to complex-valued data such as PolSAR images at the cost of a slightly more complicated training procedure as MLPs usually use bounded and analytical functions which do not exists for the complex domain. Another possibility are SVMs with complex-valued kernels tailored to the analysis of PolSAR images as in Moser and Serpico (2014). However, setting the corresponding hyperparameters is a non-trivial task and often results in time consuming grid search within the parameter space.

Approaches of the second group that are based on quasi-exhaustive feature sets usually aim to reduce this set to a meaningful subset which is then used by a standard classifier. Typical examples are principle component analysis (Licciardi et al., 2014), independent component analysis (Tao et al., 2015), linear discriminant analysis (He et al., 2014), or genetic algorithms (Haddadi et al., 2011). Other works rely on classifiers such as Random Forests (RFs) which are able to handle high-dimensional and partially undescriptive feature spaces due to an inbuilt feature selection. Random Forests have been extensively applied to the classification of remotely sensed data in general and PolSAR images in particular (see e.g. Belgiu and Dragut, 2016 for a recent review). The work in Hänsch (2014) extracts hundreds of real-valued features from a given PolSAR image and uses a RF to focus on the most descriptive ones. The work in Tokarczyk et al. (2015) pushes this idea even further by extracting thousands of (easy to compute) features and relying on boosted decision stumps to select relevant features for land cover classification from optical images. As exhaustive feature sets are less likely to be biased towards specific classification tasks and more likely to contain useful information for the task at hand, those approaches are very generic in the sense that they achieve remarkable results for various classification problems (with retraining, but without redesigning the processing chain i.e. implementing a different feature extraction). However, the large feature sets represent a considerable burden regarding memory footprint and computation time.

The solution to this problem is to parametrize feature extraction and include it into the optimization problem during the training of the classifier. This feature learning works on the data itself (or very basic features) which allows to compute task-optimal features on the fly and only when and where needed. Convolutional Networks (ConvNets) are the most well known examples of modern feature learning approaches and have been applied to close range images (e.g. Ranzato et al., 2007) as well as to remote sensing images (e.g. in Mnih and Hinton, 2012). As ConvNets are - similar to MLPs - originally designed for real-valued data, most methods (e.g. Zhou et al., 2016) extract real-valued features in order to classify PolSAR images, while only few works use complex-valued networks which are tailored to the complex values of PolSAR images (e.g. in Hänsch and Hellwich, 2010a; Zhang et al., 2017). Despite the potential power and large success of ConvNets in other computer vision domains, their sensitivity to training parameters as well as their need for large amount of (labelled) data often lead to results that are barely competitive to conventional approaches to classification of remote sensing images (Tokarczyk et al., 2012).

Another example of modern feature learning approaches are RFs where the internal node tests of the individual trees are especially designed for image data: Instead of treating the input samples as n-dimensional (feature) vectors in $\mathbb{R}^n$, image or feature patches (i.e. elements of $\mathbb{R}^{w \times w \times c}$ with patch size w and c channels) are used. The corresponding node tests are then defined over those patches, e.g. by performing comparisons of random pixel pairs within the patch. These RFs are heavily used in the computer vision community to analyse close-range optical images (see e.g. Criminisi and Shotton, 2013; Fröhlich et al., 2012), but also have found their way into remote sensing (e.g. Fröhlich et al., 2013). The work of Hänsch (2014) applies these kind of Random Forests in a sophisticated multi-stage framework for generic object classification from image data with a focus on PolSAR images: A first step extracts several low-level features from the provided image data. As this feature set might contain redundant as well as non-descriptive features for a specific classification task, Random Forests are used which on the one hand reject meaningless features due to their built-in feature selection and on the other hand are able to perform a spatial analysis of the provided feature maps. This first classification is then further processed by a semantic segmentation and a second classification step based on a similar Random Forests which exploits segment-based features.

This paper proposes to adapt these Random Forests further to be applied to complex-valued data directly and thus combines feature learning and a classifier that can directly process PolSAR images. Thus, the proposed Random Forest circumvents all above discussed problems by completely skipping the explicit computation of features, which saves computation time and memory. Instead, the internal structure of the Random Forest is adapted to the statistics of PolSAR images by following the ideas of e.g. (Fröhlich et al., 2012; Hänsch, 2014), but without the extraction of initial feature maps. Instead, the family of node test functions is extended and defined over the space of image patches containing Hermitian matrices (see Section 3.1). This allows to process the corresponding PolSAR images faster, while maintaining high accuracy. Furthermore, the saved memory can be spent on an increased model complexity of the RF (i.e. more and higher trees) which potentially leads to an even increased classification performance.

The resulting classification framework consists only of the RF itself, i.e. no data preprocessing (besides covariance/coherency matrix computation), no explicit feature extraction, and no post-processing are performed. Consequently, the classifier is very generic in the sense that it can be applied to different classification tasks or data from different sensors by simply retraining the model. This is illustrated in Section 4 by classifying airborne fully-polarimetric data from the E-SAR sensor as well as space-borne dual-polarimetric data from TerraSAR-X without any changes to the overall processing chain or adaptation of hyperparameters.

2. PolSAR data and distance measures

Synthetic Aperture Radar (SAR) measures the amplitude (and phase) of the backscattered echo of an emitted microwave pulse. Polarimetric SAR transmits and receives the microwave in two orthogonal polarizations which results in the measurement of a scattering matrix $\mathbf{S}$ (Lee and Pottier, 2009):

$$\mathbf{S} = \begin{bmatrix} S_{HH} & S_{HV} \\ S_{VH} & S_{VV} \end{bmatrix} \tag{1}$$

where H, V denote horizontal and vertical polarization, respectively. Assuming reciprocity in the mono-static case and representing the scattering information in vector form leads to target vectors $\mathbf{s}$ and $\mathbf{k}$ in lexicographic and Pauli basis, respectively:

$$\mathbf{s} = [S_{HH}, \sqrt{2}S_{HV}, S_{VV}]^T \tag{2}$$

$$\mathbf{k} = \frac{1}{\sqrt{2}}[S_{HH} + S_{VV}, S_{HH} - S_{VV}, 2S_{HV}]^T \tag{3}$$

Distributed targets are characterized by multiple scatterers within a resolution cell, which leads to interference of the individual echoes causing the so called speckle effect. If the speckle is fully developed (i.e. large amount of scatterers in a resolution cell) the target vectors are distributed according to a complex-variate zero-mean Gaussian distribution (Goodman, 1963), which is fully described by its covariance matrix $\mathbf{\Sigma}$. If target vectors are given in lexicographic basis, $\mathbf{\Sigma}$ is called polarimetric covariance ($\mathbf{\Sigma} = E[\mathbf{s}\mathbf{s}^\dagger] = \mathbf{C}$),[1] if target vectors are given in Pauli basis, $\mathbf{\Sigma}$ is called polarimetric coherency ($\mathbf{\Sigma} = E[\mathbf{k}\mathbf{k}^\dagger] = \mathbf{T}$). The expectation $E[\cdot]$ is usually computed as spatial average within a small local window. If the target vectors are complex-Normal distributed, the resulting sample covariance/coherency matrices are Wishart-distributed - an assumption that plays an important role in many statistical models and in the derivation of many distance measures for PolSAR data (see Section 2.2).

[1] $(\cdot)^\dagger$ denotes conjugate transpose.

Since distributed targets are fully described by their covariance/coherency matrix, PolSAR images are often given in this format, i.e. each pixel contains a Hermitian matrix. The two standard approaches for the analysis of PolSAR images are either to apply statistical models to these Hermitian matrices directly or to extract real-valued features to be used in a (usually) descriptive framework. This paper follows a different and less common approach to apply a descriptive model (i.e. Random Forests as discussed in the following Section 3) directly to the complex-valued data relying on neither predefined statistical models and assumptions, nor predefined hand-crafted features. This approach merely requires a measure for pixelwise comparisons, i.e. a distance measure $d(\mathbf{A}, \mathbf{B})$, which in case of PolSAR data has to be defined over Hermitian matrices $\mathbf{A} = (a_{ij})$ and $\mathbf{B} = (b_{ij})$ (with $1 \leqslant i,j \leqslant k$ where k is the number of polarimetric channels). The literature provides several such measures, which are briefly introduced and discussed in the following subsections.

2.1. Norm-based distances

The span of a PolSAR image represents the total power of the backscattered signal. A very simple comparison of two covariance/coherency matrices is the absolute difference d_A of their span:

$$d_A(\mathbf{A}, \mathbf{B}) = \left| \sum_{i=1}^{k} a_{ii} - \sum_{i=1}^{k} b_{ii} \right| \tag{4}$$

Eq. (4) uses the fact that for a Hermitian matrix $\mathbf{M} = (m_{ij}), m_{ij} \in \mathbb{C}$ $m_{ij} = m_{ji}^*$, where $(\cdot)^*$ denotes complex conjugate and therefore $m_{ii} \in \mathbb{R}$. This distance measure is usually only used in single-channel SAR images, since it ignores all polarimetric information.

An only slightly more sophisticated measure is the Euclidean distance d_E of the (real-valued) main diagonal elements of both matrices that contain the variance of the individual components of the target vectors:

$$d_E(\mathbf{A}, \mathbf{B}) = \sqrt{\sum_{i=1}^{k} (a_{ii} - b_{ii})^2} \tag{5}$$

A distance measure that takes all matrix elements into account is based on the Frobenius norm in Eq. (6) where $|z|$ measures the amplitude of $z \in \mathbb{C}$.

$$d_F(\mathbf{A}, \mathbf{B}) = ||\mathbf{A} - \mathbf{B}||_F = \sqrt{\sum_{i=1}^{k} \sum_{j=1}^{k} |a_{ij} - b_{ij}|^2} \tag{6}$$

None of these distance measures is related to the Wishart distribution or other PolSAR characteristics and thus the derived distance is not explicitly related to the statistical information of the data.

2.2. Wishart-based distances

A standard distance measure for PolSAR data is directly derived from the Wishart distribution as the (normalized) logarithm of its probability density (Lee et al., 1994). This Wishart distance d_W as defined in Eq. (7) (where $|\cdot|, Tr(\cdot)$, and $(\cdot)^{-1}$ denote matrix determinant, trace, and inverse, respectively) is not a distance metric in a strict mathematical sense, since it barely fulfills any of the necessary requirements: It is not symmetric, not subadditive, and the minimum value $d_W(\mathbf{A}, \mathbf{A})$ is not constant but depends on $\mathbf{A}$.

$$d_W(\mathbf{A}, \mathbf{B}) = \log(|\mathbf{B}|) + \mathrm{Tr}(\mathbf{B}^{-1}\mathbf{A}) \tag{7}$$

A simple modification (Anfinsen et al., 2007) leads to the symmetric Wishart distance d_{WS} in Eq. (8) which is symmetric but $d_{WS}(\mathbf{A}, \mathbf{A})$ is still not constant:

$$d_{WS}(\mathbf{A},\mathbf{B}) = \frac{1}{2}(d_W(\mathbf{A},\mathbf{B}) + d_W(\mathbf{B},\mathbf{A})) = \frac{1}{2}\left(\log(|\mathbf{AB}|) + \mathrm{Tr}(\mathbf{AB}^{-1} + \mathbf{BA}^{-1})\right) \quad (8)$$

More sophisticated distance measures are derived from likelihood ratio hypothesis tests. One example is the Bartlet distance d_B in Eq. (9) which tests whether two Wishart densities are equal or not (Conradsen et al., 2003; Kersten et al., 2005):

$$d_B(\mathbf{A},\mathbf{B}) = \ln\frac{|\mathbf{A}+\mathbf{B}|^2}{|\mathbf{A}||\mathbf{B}|} \quad (9)$$

The Bartlet distance is a semimetric, since all requirements of a metric besides the subadditivity are fulfilled.

The revised Wishart distance d_{RW} in Eq. (10) is derived from a very similar likelihood ratio test (Kersten et al., 2005):

$$d_{RW}(\mathbf{A},\mathbf{B}) = \log\left(\frac{|\mathbf{B}|}{|\mathbf{A}|}\right) + \mathrm{Tr}(\mathbf{B}^{-1}\mathbf{A}) \quad (10)$$

Like the Wishart distance it is not symmetric, but a symmetric version can be obtained in the same way resulting in the symmetric revised Wishart distance d_{RWS}:

$$d_{RWS}(\mathbf{A},\mathbf{B}) = \frac{1}{2}(d_{RW}(\mathbf{A},\mathbf{B}) + d_{RW}(\mathbf{B},\mathbf{A})) = \frac{1}{2}\mathrm{Tr}\left(\mathbf{AB}^{-1} + \mathbf{BA}^{-1}\right) \quad (11)$$

It is a semimetric as well, since it fulfills all conditions besides the triangle inequality.

2.3. Geodesic distances

Since covariance matrices are Hermitian positive definite matrices which form a Riemannian manifold (Pennec et al., 2006), affine invariant metrics have been proposed that lead to proper distance measures in this space (Barbaresco, 2009). The geodesic distance d_G in Eq. (12) corresponds to the shortest path between two points on the manifold (where $||\cdot||_F$ denotes the Frobenius norm defined in Eq. (6)):

$$d_G(\mathbf{A},\mathbf{B}) = ||\log(\mathbf{A}^{-\frac{1}{2}}\mathbf{B}\mathbf{A}^{-\frac{1}{2}})||_F \quad (12)$$

This distance is computationally very expensive. The log-Euclidean distance d_{LE} in Eq. (13) has been proposed as a simpler geodesic distance (Arsigny et al., 2006) which is still invariant with respect to similarity transformations.

$$d_{LE}(\mathbf{A},\mathbf{B}) = ||\log(\mathbf{A}) - \log(\mathbf{B})||_F \quad (13)$$

3. Random Forests

A Random Forest (RF), first introduced in Ho (1998) and Breiman (2001), consists of a set of multiple decision trees and leverages the advantages of single decision trees (e.g. being applicable to different kinds of data, interpretability, simplicity) while avoiding their limitations (e.g. high variance, prone to overfitting). The core idea of RFs is to introduce a random process during tree creation to generate multiple, equally accurate but still slightly different decision trees. If successful, many trees will agree on the correct estimate of the target variable (e.g. class label) for most samples. Although the remaining trees give wrong answers, their answers will not be consistent if the ensemble is sufficiently diverse. Therefore, the correct answer obtains the majority of all votes on average. An in-depth discussion of Decision Trees, Random Forests, and Ensemble Learning is beyond the scope of this paper (but can be found e.g. in Criminisi and Shotton, 2013; Hänsch, 2014). Instead, the following subsections give a brief introduction in tree creation, training, as well as application, and focus on the specific parts that have been adapted to images in general and PolSAR data in particular.

3.1. Tree creation and training

Random Forests are a set of (usually) binary trees that can be used for regression and classification, as well as for a combination thereof. Each tree is a graph with a single root node (i.e. no ingoing connections from other nodes), multiple internal or split nodes (i.e. one in- and two outgoing connections), and multiple terminal nodes or leafs (i.e. no outgoing connection). Tree creation and training are based on a training set $D = \{(\mathbf{x},\mathbf{y})_i\}_{i=1,\dots,N}$ of N samples $\mathbf{x}$ for which the corresponding value of the target variable $\mathbf{y}$ is known. For classification, $\mathbf{y}$ is usually a class label (i.e. $\mathbf{y} = y \in \mathbb{N}$) or the class posterior. The samples $\mathbf{x}$ are usually assumed to be a real-valued feature vector, i.e. $\mathbf{x} \in \mathbb{R}^n$. If the RF consists of T trees, the training data is often resampled (with replacement) into T bags $D_t \subset D$ ($1 \leqslant t \leqslant T$), each containing N samples (Bagging, Breiman, 1996). Starting at the root node, the corresponding bag D_t is propagated through each tree t. Each non-terminal node asks a simple yes-or-no question to the data points, which are then shifted to the left or right child node depending on their answer. This recursive splitting ends, when certain stopping criteria are fulfilled, e.g. when the maximum tree height is reached, the current node contains too few samples or samples of only one class, etc. In this case a leaf node is created, which stores information about the target variable, e.g. the posterior class distribution estimated from the samples that reached this particular leaf.

One of the most important points that determines the success of a RF is the nature of its node tests. Firstly, besides bagging, they are the second random process which is supposed to ensure a high diversity among the trees. Consequently, the set of different node tests should be as large and diverse as possible. Secondly, they should be rather efficient in terms of computation and memory since they are stored and applied thousands to millions of times within a single tree. Thirdly, in order to serve as a proper node test, they should establish a robust (not necessarily meaningful, though) order of the data. If $\mathbf{x} \in \mathbb{R}^n$, a standard node test takes the form $x_i < \theta?$ where x_i is the i-th component of $\mathbf{x}$. The split point θ can be defined in many different ways ranging from random sampling to fully-optimized selection. The work of Hänsch and Hellwich (2015) evaluates several split point selection methods and shows, that the median of the projected values is sufficiently stable, leads to balanced and accurate trees, while being easy to compute. Furthermore, to obtain splits that are better tailored towards a specific task, each node randomly creates multiple tests. From these split candidates, the best test is selected based on a quality criterion which is usually based on the drop of impurity ΔI (Eq. (14)):

$$\Delta I = I(P(y|D_n)) - p_L I(P(y|D_{n_L})) - p_R I(P(y|D_{n_R})) \quad (14)$$

$$I(P(y)) = 1 - \sum_{i=1}^{C} P(y_i)^2 \quad (15)$$

where n_L, n_R are the left and right child nodes of node n with the respective data subsets D_{n_L}, D_{n_R} (with $D_{n_L} \cup D_{n_R} = D_n$ and $D_{n_L} \cap D_{n_R} = \varnothing$) and the corresponding prior probabilities $P_{L/R} = |D_{n_{L/R}}|/|D_n|$. The node impurity is measured by the Gini impurity (Eq. (15)) of the corresponding local class posteriors $P(y|D_n)$ of node n estimated from the local subset $D_n \subset D_t \subset D$ of the training set.

Node tests of the form $x_i < \theta?$ produce linear axis-aligned splits through the feature space. Different test functions producing higher-order decision boundaries have been proposed (see e.g. Criminisi and Shotton, 2013), but are seldom used since they are

computational more expensive and can be approximated by subsequent linear splits. For structured data such as images, however, more sophisticated node tests have been proven to be more efficient (Lepetit and Fua, 2006; Fröhlich et al., 2012) since they implicitly analyse the local spatial structure e.g. by performing comparisons between random pixel pairs within a patch. This work extends these ideas to the characteristics of PolSAR images and models the samples as patches of a PolSAR image, i.e. $\mathbf{x} \in \mathbb{C}^{w \times w \times k \times k}$, where w is the spatial patch size and k is the number of channels of the PolSAR data (i.e. $k = 2$ for dual-, $k = 3$ for fully-polarized data). An operator $\phi : \mathbb{C}^{\tilde{w} \times \tilde{w} \times k \times k} \rightarrow \mathbb{C}^{k \times k}$ is applied to one, two, or four regions $R_r \subset \mathbf{x}$ $(r = 1, \ldots, 4)$ of size $\tilde{w}_r \times \tilde{w}_r$ inside a patch $\mathbf{x}$ (where $\tilde{w}_r < w$). Possible operators are the center/average value of the region or the region element with minimal/maximal span within the region (see Eq. (16)).

$$\mathbf{C}_R = \phi(R) = \begin{cases} R(\tilde{w}/2, \tilde{w}/2) \\ \frac{1}{\tilde{w}_R^2} \sum_{i=1}^{\tilde{w}} \sum_{j=1}^{\tilde{w}} R(i,j) \\ R(i^*, j^*) \qquad \text{with } (i^*, j^*) = \underset{0<i,j<\tilde{w}}{\text{argmin}} \text{ span } R(i,j) \\ R(i^*, j^*) \qquad \text{with } (i^*, j^*) = \underset{0<i,j<\tilde{w}}{\text{argmax}} \text{ span } R(i,j) \end{cases} \tag{16}$$

The operator as well as region position and size are randomly selected. The outputs of the operator for each region are compared to each other by Eqs. (17)–(19), where $\tilde{\mathbf{C}}$ is a covariance matrix randomly selected from the whole image:

$$\text{1-point projection}: \quad d(\mathbf{C}_{R_1}, \tilde{\mathbf{C}}) \quad < \theta \tag{17}$$

$$\text{2-point projection}: \quad d(\mathbf{C}_{R_1}, \mathbf{C}_{R_2}) \quad < \theta \tag{18}$$

$$\text{4-point projection}: \quad d(\mathbf{C}_{R_1}, \mathbf{C}_{R_2}) - d(\mathbf{C}_{R_3}, \mathbf{C}_{R_4}) \quad < \theta \tag{19}$$

These projections (illustrated in Fig. 1) analyse local spectral and textural properties and are based on a proper distance measure $d(\mathbf{A}, \mathbf{B})$.

In the case of real-valued color images or feature maps (i.e. $\mathbf{x} \in \mathbb{R}^{w \times w \times c}$ where c is the number of channels), ϕ often acts as channel selection (as e.g. in Lepetit and Fua, 2006) and local averaging (as e.g. in Fröhlich et al., 2012). In this case are $\mathbf{A} = a, \mathbf{B} = b$ with $a, b \in \mathbb{R}$ and usually $d(a,b) = a - b$ or $d(a,b) = |a - b|$.

In the case of PolSAR data, each pixel contains a Hermitian matrix, which leads to the need of a distance over the space of complex-valued matrices. Section 2 states and discusses some common choices for such a distance measure which are comparatively evaluated in Section 4. It should be noted that a single RF can apply any subset of available distance measures and does not need to be limited to using only one: Which distance is used is simply another random variable that is fixed during tree creation.

Let the minimal and maximal region size be denoted by $\tilde{w}_{min}$ and $\tilde{w}_{max}$, respectively, and the maximal region distance to the patch center by a_{max}. Then the analyzed patch size is $w = 2 \cdot a_{max} + \tilde{w}_{max}$ and there are w^2 possible positions to place a region of size $\tilde{w}$. Furthermore, there are $b = \tilde{w}_{max} - \tilde{w}_{min} + 1$ possible region sizes, which results in $w^2 \cdot b$ possible ways to define a region within a given patch. 1-, 2-, and 4-point projections need 1, 2, and 4 regions, respectively, to which four different operators can be applied where the output is compared based on 10 different distances. This leads to $w^2 \cdot b \cdot (1 + 2 + 4) \cdot 4 \cdot 10$ possible node tests, i.e. to more than $2 \cdot 10^6$ possible tests assuming $\tilde{w}_{min} = 3$ and $\tilde{w}_{max} = a_{max} = 10$. This rough calculation neither takes into account the selection of the reference value for 1-point projections or the selection of the split point θ (which is defined as the median of the projected data but could also be randomly selected) - both would increase the number of possible tests - nor the correlation between different tests - which is decreasing the number of truly different node tests. Nevertheless, the set of possible node tests - and therefore implicit features - is tremendous. From this (virtual) set of features each node computes only a random subset and selects the best candidate from it (see discussion above, i.e. Eq. (14)). These simple features are then combined in a hierarchical manner while a sample is propagated through each tree. The final number of (combined) features equals the number of leaf nodes within the forest which is easily in the thousands or millions. It would be infeasible to store or even to compute all those features for all samples. Instead, only a small fraction is computed for each sample, i.e. exactly one combined feature per tree (corresponding to the path the sample takes through the tree) which consists of as many simple features (i.e. node tests) as there are split nodes from the root to the leaf that is reached by this sample (typically between 10 and 100). This is one of the reasons for the high descriptive power of Random Forests while they maintain a reasonable computational load in terms of speed and memory footprint.

3.2. Prediction

Once the Random Forest is created and trained, it can be used for prediction. A query sample is propagated through all trees, starting at the root node. It will end in exactly one leaf node $n_t(\mathbf{x})$ per tree t. The information assigned to these leafs, i.e. the class posterior $P(y|n_t(\mathbf{x}))$, is averaged to obtain the final class posterior $P(y|\mathbf{x})$:

$$P(y|\mathbf{x}) = \frac{1}{T} \sum_{t=1}^{T} P(y|n_t(\mathbf{x})) \tag{20}$$

4. Experiments

4.1. Data

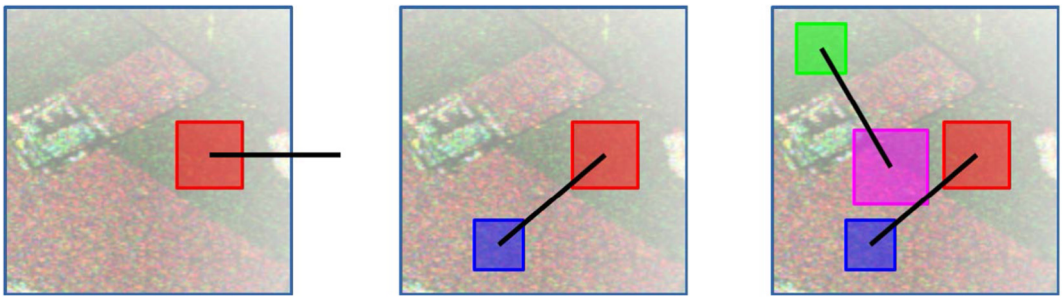
The proposed method is evaluated on two different datasets:

First, a fully polarimetric image acquired in 1999 by the E-SAR sensor (DLR) in L-band (1.25 GHz) over Oberpfaffenhofen, Germany, shown in Fig. 3a. The image size is 1390×6640 px with a resolution of approximately 1.5 m. This dataset contains manmade (such as urban areas and roads) as well as natural structures (such as forests and fields) and is manually labelled with five different classes, namely City, Road, Forest, Shrubland, and Field (see Fig. 3b).

Second, a dual-polarimetric image (providing only S_{HH} and S_{VV}) acquired in 2008[2] by TerraSAR-X (DLR) in X-band and spotlight mode over central Berlin, Germany, shown in Fig. 6a. The image size is 6240×3953 px with a resolution of approximately 1 m. It shows noise on a rather high level which was not further reduced. The Berlin dataset is taken over a dense urban area and contains, besides roads and buildings, the river Spree and one of the major parks of Berlin. It is manually labelled into six different categories which are illustrated in Fig. 6b, namely Building, Road, Railway, Forest, Lawn, and Water.

The image data is divided into five different folds, training data are randomly sampled from four stripes while the fifth stripe is used for testing. For each run f the balanced accuracy ba_f is measured as the average detection rate per class. The final balanced accuracy estimate ba is the average over all folds, i.e. $ba = \frac{1}{5} \sum_{f=1}^{5} ba_f$.

[2] 2008-07-4T05:25:11, descending.



(a) 1-point projection (b) 2-point projection (c) 4-point projection

Fig. 1. Different spatial projections within a node test function: An operator is applied to different regions within a patch (here denoted as colored boxes). A distance measure (illustrated as black line) is applied to the corresponding operator outputs. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4.2. Results

A RF with 30 trees of maximal height 50 is created with median-based split point selection. Different versions of node tests are evaluated which either apply a single one of the distance measures discussed in Section 2 or leave the RF the freedom to select among all of them within a node.

The obtained average accuracy for the Oberpfaffenhofen dataset is shown in Fig. 2a. The first column repeats the result from Hänsch (2014) (denoted as "R") which extracts a large set of real-valued features and use it as input to a RF. Besides this different input space the RFs are identical (i.e. identical parameters as well as training and evaluation procedures). The subsequent columns represent the results obtained by the proposed approach, i.e. skipping the explicit computation of real-valued features and working directly on the complex-valued PolSAR data.

While the real-valued features lead to a balanced accuracy of 89.4%, the different versions of the complex-valued RF obtain results around 87%. The best result is obtained by the log-Euclidean distance with 87.5%, which is a decrease of only 2%. The worst result of 83.2% is achieved by the distance only considering the span (i.e. the total intensity) disregarding polarimetry altogether. Despite being the weakest classification result, it is still somewhat surprising that this simple distance measure performs that well. The Euclidean distance of the real-valued elements of the main diagonal of the covariance matrix shows nearly the same performance (83.6%). The Frobenius norm is slightly better (85.3%), but is outperformed by distance measures optimized for PolSAR data. However, interestingly the increase in performance is (although statistically significant) only marginal. There is only a 4% difference between the best result of the log-Euclidean distance and the worst result of the span-based distance analysing only the total intensity.

Fig. 3c shows the semantic map of the test data using the log-Euclidean distance d_{LE}. The results are further summarized in the confusion matrix shown in Table 1.

If compared to the results obtained by extracting real-valued features as input to the RF, approximately the same accuracy ($\pm 1\%$) is maintained for all categories except the City class for which accuracy dropped from 94% to 87%. The drop of roughly 7% is nearly entirely caused by confusion with the Shrubland class (6%). Fig. 4a shows the spatial distribution of the differences of both methods. Both methods are able to lead to accurate semantic maps, while the majority of the errors are consistent (denoted in green) and are concentrated in areas known for being notoriously difficult to classify. On the one hand, those are roads which are very thin lines in this image. As both RFs use image patches, most of the available information stems from surrounding classes which explains why road pixels are mostly confused for city and field but seldom for shrubland and forest (which contain less labelled roads within the reference data). On the other hand, the consistently misclassified areas (marked by black rectangles) are mostly fields which are either classified as shrubland (due to high intensity values) or road (due to very low intensity values). The large building on the campus of the DLR (denoted by rectangle A in Fig. 4a) is usually misclassified as forest since material properties as well as the 45° orientation to the sensor results in volumetric backscattering. Interestingly, the proposed method improves the classification slightly in this region, most likely as it is able to perform a finer analysis of polarimetric properties.

Interestingly, the largest part of the changes in accuracy - the confusion between Shrubland and City (denoted by a black ellipse in Fig. 4a, Fig. 5 shows this part in larger detail) - appears to be partially caused by ambiguous labels in the reference data: Parts of the City class that are labelled as Shrubland (and thus are counted as misclassified) are actually larger areas of vegetation (parks and backyards) within the city. While the reference method declares those areas as belonging to the city, the proposed method recognizes them as vegetation.

Fig. 4b shows that the differences in the label maps of both methods do indeed go both ways: A pixel that was correctly labelled by the reference method can be wrongly labelled by the proposed method and a pixel that was wrongly labelled by the reference method can be correctly labelled by the proposed method. However, the first case happens slightly more often (2.7% of all labelled pixels) then the second (1.3% of all labelled pixels). This slight decrease of classification accuracy comes with the benefit that no features have to be extracted upfront (while the reference method extracts around 370 different features as input to the RF).

Similar findings are obtained based on the Berlin dataset with two important differences: On the one hand, the RF trained on real-valued features is strongly outperformed by the complex-valued RF (54.3% vs. 63.3%). On the other hand, the overall performance of all tested RF variations is considerably worse than for the Oberpfaffenhofen dataset: While on the latter the accuracy ranged between 83 and 89%, only 54–63% are achieved on the Berlin dataset. The reasons for this gap are threefold: (1) The class instances in Oberpfaffenhofen are large, often homogeneous segments within the image. In many parts they can be easily distinguished based on intensity (e.g. the forest is very bright) and texture (i.e. fields are very homogeneous). Within the Berlin dataset the class instances are smaller, much stronger spatially mixed, and more heterogeneous. (2) The Berlin image contains only dual-pol information, it has a higher resolution, and stronger noise. This leads to less stable features and a noisier computation of distance measures. Furthermore, common assumptions made by many PolSAR feature extractors as well as distance measures are more often violated (e.g. fully developed speckle and Wishart distributed sample

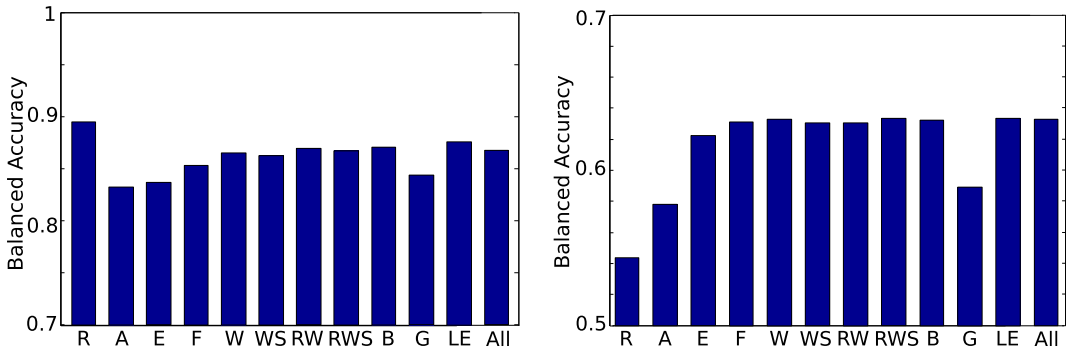


Fig. 2. Balanced accuracy with different distance measures: R: Real valued features (Hänsch, 2014); A: Span-based distance; E: Euclidean distance; F: Frobenius distance; W: Wishart distance; WS: Symmetric Wishart distance; RW: Revised Wishart distance; RWS: Symmetric revised Wishart distance; B: Bartlet distance; G: Geodesic Distance; LE: Log-Euclidean distance; All: All distance measures (from A to LE).

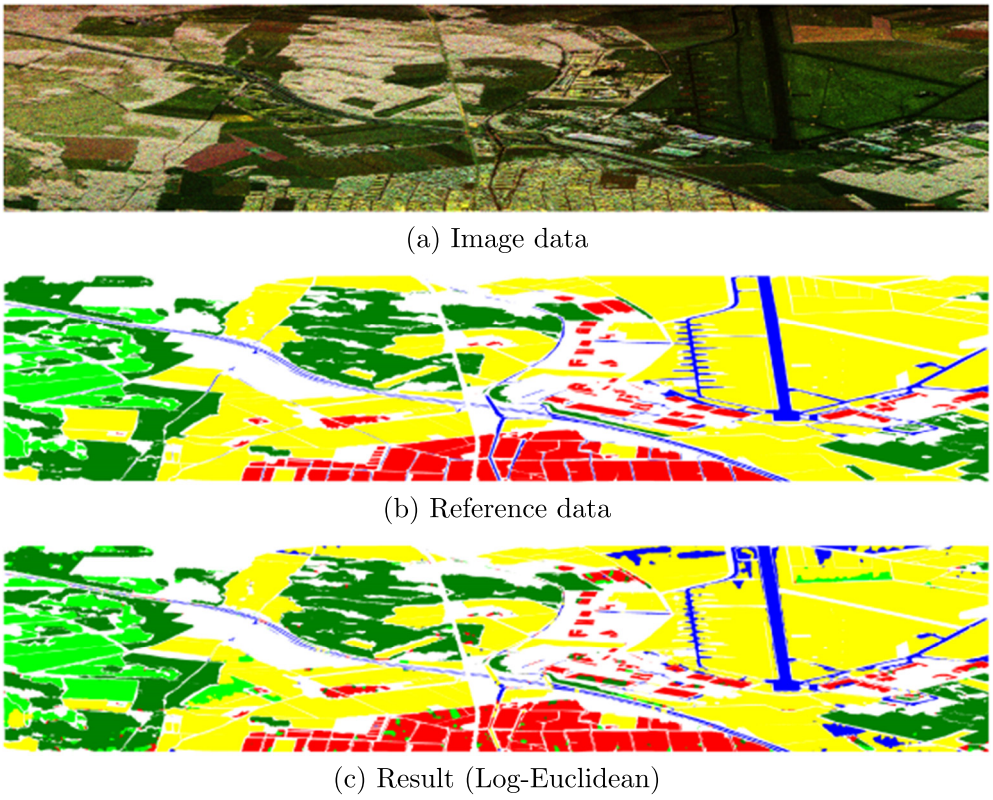


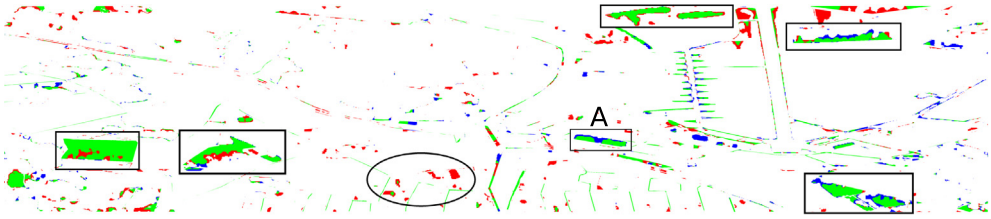
Fig. 3. Image and reference data as well as classification results for the Oberpfaffenhofen data set (PolSAR image acquired by E-SAR, DLR).

Table 1
Confusion matrix (Log-Euclidean) for the Oberpfaffenhofen data set. The numbers in brackets denote the change with respect to the reference method.

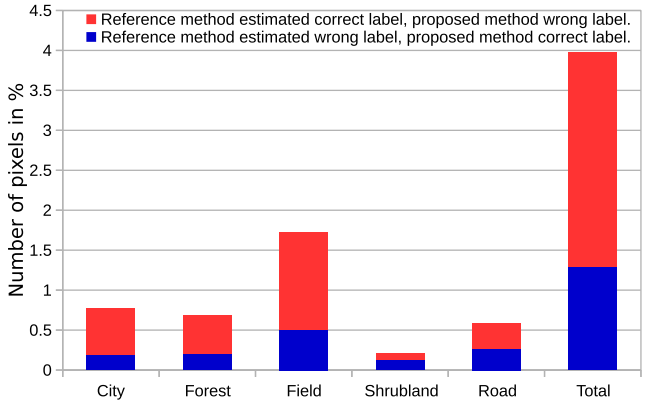
$ba = 87.5\%$	City	Forest	Field	Shrubl.	Road
City	**87**%(-7)	6%(+1)	0%(+0)	6%(+6)	1%(+0)
Forest	2%(+0)	**96**%(-1)	0%(+0)	2%(+1)	0%(+0)
Field	0%(+0)	0%(+0)	**93**%(-1)	4%(+0)	3%(+1)
Shrubl.	0%(-2)	2%(-1)	8%(+2)	**90**%(+1)	0%(+0)
Road	11%(+0)	1%(+0)	13%(-1)	2%(+1)	**73**%(+0)

covariance matrices). (3) The reference data of the Berlin dataset contains six instead of five classes, which were significantly harder to assign to the image due to strong layover artifacts and a heterogeneous class distribution.

Fig. 6c shows the semantic map of the test data using the log-Euclidean distance d_{LE}, while Table 2 shows the corresponding confusion matrix. Most classes are classified correctly with an accuracy of around 80%. The low average accuracy is mainly caused by the Road and Lawn classes with only 17% and 52%, respectively. The major confusion of the Lawn class is with the Forest class (35%), which is not surprising given the similar nature of both classes. In particular, there is a continuous spectrum of lawn, lawn with sparse trees, lawn densely populated with trees, forest with larger clearings, and dense forest. The discrete cut between those two classes is somewhat arbitrary and leads to partially ambiguous labels resulting in the observed confusion of both classes. It also means that a coarser class covering all vegetation (forest and lawn) would have achieved an accuracy of 87%. Most of the Road pixels have been classified as Building (43%). Partially, this is due to the



(a) Spatial distribution of differences in the obtained semantic maps.



(b) Fraction of labelled pixels for which either proposed or reference method (but not both) estimated the correct label.

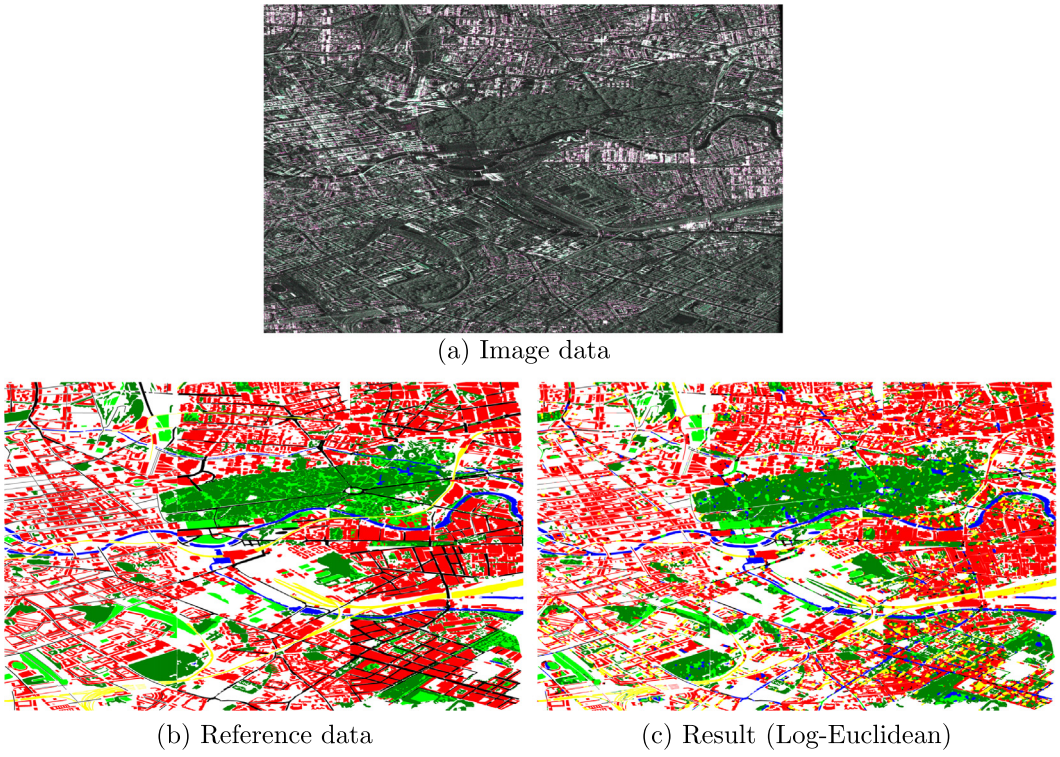
Fig. 4. Differences within the semantic maps of the Oberpfaffenhofen data set obtained by the reference method, i.e. a RF with real-valued features, and the proposed method. White: Both methods estimated correct label (or reference data does not provide a label); Green: Both methods estimated wrong label; Red: Reference method estimated correct label, proposed method wrong label; Blue: Reference method estimated wrong label, proposed method correct label. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

(a) PolSAR data (b) Optical data (c) Semantic map (d) Reference data

Fig. 5. A detail of the Oberpfaffenhofen dataset and the corresponding classification result. Examples of larger misclassified regions have been outlined in the images in the left, while the right side shows the corresponding semantic maps with City in red, Forest in dark green, Field in yellow, Shrubland in light green, and Road in blue. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

strong overlay of high buildings in direction of the sensor, which actually causes a fusion of buildings and streets in the image, when the streets are parallel to the flight direction. However, the road class lost 38% accuracy compared to the reference method (from 55% to 17%) which indicates again a loss of accuracy by using node tests which are purely based on covariance matrices while the reference method uses also features based on the original scattering vectors. Nevertheless, compared to the real-valued RF, all classes besides Road show an increased accuracy (Building by 20%, Railway by 11%, Forest 36%, Lawn by 13%, Water by 11%). Railway pixels share a major confusion with the building class (in both directions) which is probably caused by high backscattering intensities due to metal structures within both classes. Furthermore, similar to roads, railways are thin, elongated structures which are often surrounded by buildings leading to a certain overlap in the features within a patch.

In general, the observations based on these two very different datasets are consistent: Ignoring or mistreating polarimetric information by using distance measures not suited for PolSAR processing leads to suboptimal, but surprisingly good results. This



(a) Image data

(b) Reference data

(c) Result (Log-Euclidean)

Fig. 6. Image and reference data as well as classification results for the Berlin data set (PolSAR image acquired by TerraSAR-X, DLR).

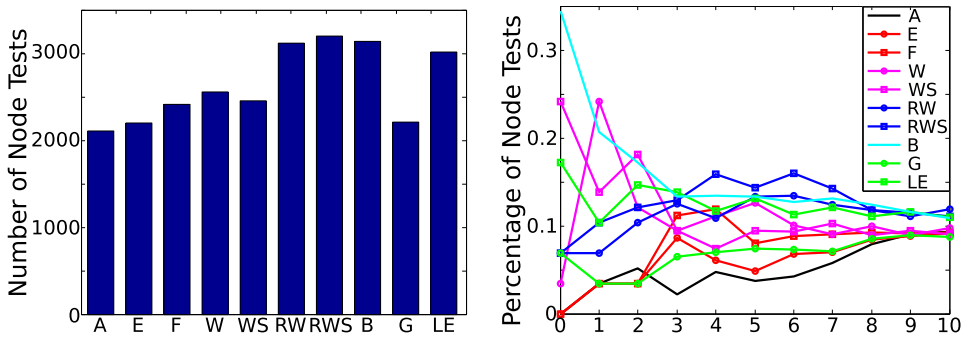
Table 2
Confusion matrix (Log-Euclidean) for the Berlin data set. The numbers in brackets denote the change with respect to the reference method.

$ba = 63.3\%$	Building	Road	Railway	Forest	Lawn	Water
Building	**84%**	1%	9%	5%	0%	1%
	(+20)	(-18)	(-3)	(+2)	(-1)	(+0)
Road	43%	**17%**	5%	10%	12%	13%
	(+28)	(-38)	(+0)	(+6)	(+7)	(-3)
Railway	19%	0%	**67%**	12%	2%	0%
	(+6)	(-8)	(+11)	(+2)	(-6)	(-5)
Forest	7%	1%	5%	**79%**	7%	1%
	(+3)	(-7)	(-12)	(+36)	(-11)	(-9)
Lawn	4%	1%	2%	35%	**52%**	6%
	(+0)	(-11)	(-8)	(+19)	(+13)	(-13)
Water	5%	6%	0%	4%	5%	**80%**
	(+1)	(-9)	(-3)	(+1)	(-1)	(+11)

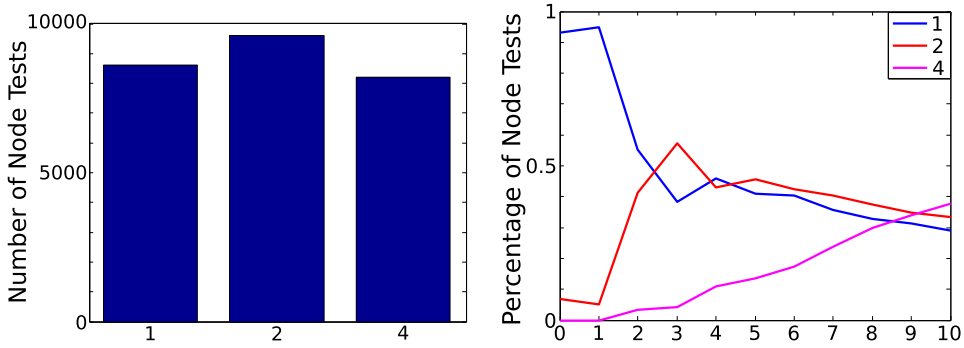
indicates that intensity as well as spatial information (i.e. textural features as implicitly learnt by the RF) are a very strong cue for the corresponding object class. Adding a proper exploitation of polarimetric information does increase the performance only slightly, but consistently. Which polarimetric distance measure is used, however, seems to play a minor role. An exception is the geodesic distance, which has shown very similar behaviour to the log-Euclidean distance in other works (D'Hondt et al., 2013), but falls far behind in this work. The reason might be the numerical instability of the three involved mathematical operators, namely matrix-logarithm, -root, and -inverse.

Unlike many other machine learning approaches, the RF framework is not a black box but allows deeper insights into how the given task has been solved. In particular it provides information about which node tests have been preferred. The left side of Fig. 7a shows the frequency with which the different distance measures have been selected within a RF that was allowed to select freely which measure to use at each node (applied to the Oberpfaffenhofen dataset). It should be noted that the results are highly consistent to the individual experiments above, i.e. Fig. 2a: The span-based measure is used, but less often than the Frobenius norm, which is less often applied than the PolSAR-tailored distance measures. The revised Wishart distance (and its symmetric version), as well as Bartlet and log-Euclidean distance have been (roughly equally) preferred by the RF, while the geodesic distance is selected less. The right side of Fig. 7a presents the same information per tree level as relative usage frequency of the different distance measures. It shows that the Bartlet, symmetric Wishart, and Log-Euclidean distance are preferred by nodes close to the root, while all norm-based measures (span, Euclidean, and Frobenius) are basically ignored. For medium tree levels the revised Wishart distances gain importance, while every measure is roughly equally selected at very high tree levels. The reason for this is that the trees tend to make simple decisions early, i.e. decisions that are able to correctly propagate large parts of different classes to different child nodes. Harder decisions that often involve more fine-grained feature extractions are made at medium tree levels. At very high tree levels the data has already been partitioned considerably, which leads to nodes with very few samples. Estimations based on few samples show a very high variance, i.e. are not very trustworthy and can only hardly be optimized. At some point the decision will be based on noise only which leads to a randomly selected test even if optimized test selection is performed (since the value of quality criterion will be random at that point).

A similar effect can be seen in Fig. 7b, which shows the selection frequency of the projection types (see Section 3.1): A simple comparison of a region property with a reference value (i.e. one-point projection) is preferred at nodes close to the root. This kind of projection is able to distinguish between patches that belong to different appearance clusters, e.g. bright, dark, certain backscatter types, etc. Two-point projections, that analyse local texture by approximating a gradient, gain fast in importance and are most often used at medium tree levels. The more complicated four-point projections are increasingly used at higher tree levels. Close to the leafs of the trees, all three possibilities are nearly equally often selected.



(a) Usage frequency of distance measures: A: Span-based distance; E: Euclidean distance; F: Frobenius distance; W: Wishart distance; WS: Symmetric Wishart distance; RW: Revised Wishart distance; RWS: Symmetric revised Wishart distance; B: Bartlet distance; G: Geodesic Distance; LE: Log-Euclidean distance; All: All distance measures (from A to LE).



(b) Usage frequency of different spatial projections, i.e. 1,2, and 4 denote 1-, 2-, and 4-point projections, respectively.

Fig. 7. Usage frequencies of different node tests. Left: Global, absolute frequency for the whole forest; Right: Relative frequency per tree level.

5. Conclusion

The majority of published works on PolSAR classification can be divided into two groups: On the one hand, sophisticated statistical models tailored to the special image characteristics of (Pol)SAR images and directly applied to the complex-valued scattering vectors or sample covariance/coherency matrices. These often lack generality to be applied to heterogeneous target classes within contemporary high-resolution data. On the other hand, discriminative approaches that start with the extraction of hand-crafted real-valued features which are subsequently used as input to a standard machine-learning classifier. Only little work has been done to define classifiers that can directly be applied to the PolSAR data itself, without the expensive (in terms of computation time and memory) computation of features. This work proposes a step in this direction and shows, that it is possible to define the node tests within a Random Forest directly over Hermitian matrices. This leads to savings with respect to time and memory (since there is no explicit feature extraction), while maintaining a high classification accuracy. Furthermore, it removes the dependency of the whole classification pipeline on the extracted features which might be descriptive only for a subset of possible classes while being suboptimal for other tasks. It should be noted that, if it is known that a certain image feature is highly descriptive for a particular application, it can still be included into the proposed RF as an additional node test.

One disadvantage of the proposed approach is that its current implementation can only be applied to PolSAR data, while the work of Hänsch (2014) can be applied to many different types of image data. However, it is easily possible to extend the current approach by including other distance measures that are tailored towards other types of imagery (including but not limited to InSAR or TomoSAR). Another disadvantage is a certain loss of resolution in the semantic map. Among the many features calculated in Hänsch (2014) from a PolSAR image, are several that are based on the scattering vectors, while others are based on covariance matrices. This allows the RF to select node tests that work on the original resolution of the data (which is lost to some extend if sample covariance matrices are computed by spatial averaging). However, both disadvantages hold for most of the state-of-the-art approaches for classification from PolSAR data and seem to be a small cost compared to the tremendous decrease in computation time and memory usage. One possible alternative to cope with this problem is to include node tests which work directly on the scattering vectors instead of using covariance/coherency matrices. The combination of both kinds of node tests would give the RF access to high-resolution details on the one hand (without the implicit smoothing during the computation of covariance/coherency matrices) and on the other hand still enables the usage of second order statistics as contained in these matrices.

The performance of any supervised machine learning method depends drastically on the amount (and quality) of available training data. While labelled PolSAR images are sparse (especially if the same sensor and acquisition mode has to be used), unlabelled images are readily available nowadays. Future work will therefore focus on the usage of unlabelled data by the RF in order to aid the classification task. To this aim three factors are necessary: (1) Online learning techniques have to be included into the framework as it will not be possible anymore to hold the whole dataset (i.e. multiple high-resolution PolSAR images) in memory. (2) The node

tests of the RF have to be able to exploit unlabelled data in order to find good splits. (3) Model capacity needs to be restricted to subspaces of the data manifold which show highest potential to be correctly classified. Otherwise, the decision trees will grow (along with increased memory load and computation time) while trying to distinguish indistinguishable subsets of the samples.

References

Anfinsen, S.N., Jenssen, R., Eltoft, T., 2007. Spectral clustering of polarimetric SAR data with Wishart-derived distance measures. In: Proceedings of 7th POLinSAR. Frascati, Italy.

Arsigny, V., Fillard, P., Pennec, X., Ayache, N., 2006. Log-euclidean metrics for fast and simple calculus on diffusion tensors. Magn. Reson. Med. 56 (2), 411–421.

Barbaresco, F., 2009. Interactions between symmetric cone and information geometries: Bruhat-tits and siegel spaces models for high resolution autoregressive doppler imagery. Emerg. Trends Vis. Comput., 124–163

Belgiu, M., Dragut, L., 2016. Random Forest in remote sensing: a review of applications and future directions. ISPRS J. Photogram. Rem. Sens. 114, 24–31.

Breiman, L., 1996. Bagging predictors. Mach. Learn. 24 (2), 123–140.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.

Bruzzone, L., Marconcini, M., Wegmuller, U., Wiesmann, A., 2004. An advanced system for the automatic classification of multitemporal SAR images. IEEE Trans. Geosci. Rem. Sens. 42 (6), 1321–1334.

Cloude, S.R., Pottier, E., 1996. A review of target decomposition theorems in radar polarimetry. IEEE Trans. Geosci. Rem. Sens. 34 (2), 498–518.

Conradsen, K., Nielsen, A.A., Schou, J., Skriver, H., 2003. A test statistic in the complex Wishart distribution and its application to change detection in polarimetric SAR data. IEEE Trans. Geosci. Rem. Sens. 41 (1), 4–19.

Criminisi, A., Shotton, J., 2013. Decision Forests for Computer Vision and Medical Image Analysis. Springer Publishing Company, Incorporated.

D'Hondt, O., Guillaso, S., Hellwich, O., 2013. Iterative bilateral filtering of polarimetric SAR data. IEEE J. Select. Top. Appl. Earth Observ. Rem. Sens. 6 (3), 1628–1639.

Fröhlich, B., Bach, E., Walde, I., Hese, S., Schmullius, C., Denzler, J., 2013. Land cover classification of satellite images using contextual information. In: ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences II-3/W1, pp. 1–6.

Fröhlich, B., Rodner, E., Denzler, J., 2012. Semantic segmentation with millions of features: integrating multiple cues in a combined Random Forest approach. In: 11th Asian Conference on Computer Vision. Daejeon, Korea, pp. 218–231.

Goodman, N.R., 1963. Statistical analysis based on a certain multivariate complex Gaussian distribution (an introduction). Ann. Math. Stat. 34, 152–177.

Haddadi, A., Sahebi, M.R., Mansourian, A., 2011. Polarimetric SAR feature selection using a genetic algorithm. Can. J. Rem. Sens. 37 (1), 27–36.

Hänsch, R., 2010. Complex-valued multi-layer perceptrons - an application to polarimetric SAR data. Photogram. Eng. Rem. Sens. 9, 1081–1088.

Hänsch, R., 2014. Generic Object Categorization in PolSAR Images - and Beyond. Ph. D. thesis, TU Berlin, Germany.

Hänsch, R., Hellwich, O., 2010a. Complex-valued convolutional neural networks for object detection in polsar data. In: 8th European Conference on Synthetic Aperture Radar. Aachen, Germany, pp. 1–4.

Hänsch, R., Hellwich, O., 2010b. Random Forests for building detection in polarimetric SAR data. In: 2010 IEEE International Geoscience and Remote Sensing Symposium (IGARSS). Honolulu, USA, pp. 460–463.

Hänsch, R., Hellwich, O., 2015. Evaluation of tree creation methods within Random Forests for classification of PolSAR images. In: 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS). Milan, Italy, pp. 361–364.

He, C., Li, S., Liao, Z., Liao, M., 2013. Texture classification of PolSAR data based on sparse coding of wavelet polarization textons. IEEE Trans. Geosci. Rem. Sens. 51 (8), 4576–4590.

He, C., Zhuo, T., Ou, D., Liu, M., Liao, M., 2014. Nonlinear compressed sensing-based LDA topic model for polarimetric SAR image classification. IEEE J. Select. Top. Appl. Earth Observ. Rem. Sens. 7 (3), 972–982.

Ho, T.K., 1998. The random subspace method for constructing decision forests. IEEE Trans. Pattern Anal. Mach. Intell. 20 (8), 832–844.

Kersten, P.R., Lee, J.-S., Ainsworth, T.L., 2005. Unsupervised classification of polarimetric synthetic aperture radar images using fuzzy clustering and EM clustering. IEEE Trans. Geosci. Rem. Sens. 43 (3), 519–527.

Krylov, V.A., Moser, G., Serpico, S.B., Zerubia, J., 2011. Supervised high-resolution dual-polarization SAR image classification by finite mixtures and copulas. IEEE J. Select. Top. Signal Process. 5 (3), 554–566.

Kuruoglu, E.E., Zerubia, J., 2004. Modeling SAR images with a generalization of the Rayleigh distribution. IEEE Trans. Image Process. 13 (4), 527–533.

Lee, J.-S., Grunes, M.R., Kwok, R., 1994. Classification of multilook polarimetric SAR imagery based on complex Wishart distribution. Int. J. Rem. Sens. 15 (11), 229–231.

Lee, J.-S., Pottier, E., 2009. Polarimetric Radar Imaging: From Basics to Applications. Taylor & Francis, London, UK.

Lepetit, V., Fua, P., 2006. Keypoint recognition using randomized trees. IEEE Trans. Pattern Anal. Mach. Intell. 28 (9), 1465–1479.

Li, H.C., Hong, W., Wu, Y.R., Fan, P.Z., 2011. On the empirical-statistical modeling of SAR images with generalized gamma distribution. IEEE J. Select. Top. Signal Process. 5 (3), 386–397.

Licciardi, G., Avezzano, R.G., Frate, F.D., Schiavon, G., Chanussot, J., 2014. A novel approach to polarimetric SAR data processing based on nonlinear PCA. Pattern Recogn. 47 (5), 1953–1967.

Mantero, P., Moser, G., Serpico, S.B., 2005. Partially supervised classification of remote sensing images through SVM-based probability density estimation. IEEE Trans. Geosci. Rem. Sens. 43 (3), 559–570.

Mnih, V., Hinton, G.E., 2012. Learning to label aerial images from noisy data. In: Langford, J., Pineau, J. (Eds.), Proceedings of the 29th International Conference on Machine Learning (ICML-12). Edinburgh, Scotland, pp. 567–574.

Moser, G., Serpico, S.B., 2014. Kernel-based classification in complex-valued feature spaces for polarimetric SAR data. In: 2014 IEEE Geoscience and Remote Sensing Symposium (IGARSS). Quebec City, Canada, pp. 1257–1260.

Moser, G., Zerubia, J., Serpico, S.B., 2006. SAR amplitude probability density function estimation based on a generalized Gaussian model. IEEE Trans. Image Process. 15 (6), 1429–1442.

Nicolas, J.M., Tupin, F., 2016. Statistical models for SAR amplitude data: a unified vision through Mellin transform and Meijer functions. In: 2016 24th European Signal Processing Conference (EUSIPCO). Budapest, Hungary, pp. 518–522.

Pennec, X., Fillard, P., Ayache, N., 2006. A Riemannian framework for tensor computing. Int. J. Comput. Vis. 66 (1), 41–66.

Ranzato, M., Huang, F.J., Boureau, Y.L., LeCun, Y., 2007. Unsupervised learning of invariant feature hierarchies with applications to object recognition. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition. Minneapolis, USA, pp. 1–8.

Tao, M., Zhou, F., Liu, Y., Zhang, Z., 2015. Tensorial independent component analysis-based feature extraction for polarimetric SAR data classification. IEEE Trans. Geosci. Rem. Sens. 53 (5), 2481–2495.

Taravat, A., Latini, D., Frate, F.D., 2014. Fully automatic dark-spot detection from SAR imagery with the combination of nonadaptive Weibull multiplicative model and pulse-coupled neural networks. IEEE Trans. Geosci. Rem. Sens. 52 (5), 2427–2435.

Tison, C., Nicolas, J.M., Tupin, F., Maitre, H., 2004. A new statistical model for Markovian classification of urban areas in high-resolution SAR images. IEEE Trans. Geosci. Rem. Sens. 42 (10), 2046–2057.

Tokarczyk, P., Montoya, J., Schindler, K., 2012. An evaluation of feature learning methods for high resolution image classification. In: ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences I-3, pp. 389–394.

Tokarczyk, P., Wegner, J.D., Walk, S., Schindler, K., 2015. Features, color spaces, and boosting: new insights on semantic classification of remote sensing images. IEEE Trans. Geosci. Rem. Sens. 53 (1), 280–295.

Zhang, Z., Wang, H., Xu, F., Jin, Y.Q., 2017. Complex-valued convolutional neural network and its application in polarimetric SAR image classification. IEEE Trans. Geosci. Rem. Sens. (99), 1–12

Zhou, Y., Wang, H., Xu, F., Jin, Y.Q., 2016. Polarimetric SAR image classification using deep convolutional neural networks. IEEE Geosci. Rem. Sens. Lett. 13 (12), 1935–1939.