Automatic Generation of Training Data for Land Use and Land Cover Classification by Fusing Heterogeneous Data Sets

SYLVIA SCHMITZ[1,2], MARTIN WEINMANN[1], UWE WEIDNER[1], HORST HAMMER[2] & ANTJE THIELE[1,2]

Abstract: Nowadays, automatic classification of remote sensing data can efficiently produce maps of land use and land cover, which provide an essential source of information in the field of environmental sciences. Most state-of-the-art algorithms use supervised learning methods that require a large amount of annotated training data. In order to avoid time-consuming manual labelling, we propose a method for the automatic annotation of remote sensing data that relies on available land use and land cover information. Using the example of automatic labelling of SAR data, we show how the Dempster-Shafer evidence theory can be used to fuse information from different land use and land cover products into one training data set. Our results confirm that the combination of information from OpenStreetMap, CORINE Land Cover 2018, Global Surface Water and the SAR data itself leads to reliable class assignments, and that this combination outperforms each considered single land use and land cover product.

1 Introduction

A wide range of research in the field of environmental science relies on information on land use and land cover (LULC). For instance, such knowledge is used to develop climate models, to plan cities and agricultural land use efficiently, or to support forest management (PIELKE et al. 2011; HANSEN et al. 2013). In order to respond to the demand for up-to-date LULC products, a variety of methods has been developed to derive appropriate class assignments from remote sensing data. The use of remote sensing data, in particular satellite data, is justified by the fact that large areas can be covered, in most cases even with short revisit time. In recent years, the accessibility of remote sensing data has increased significantly due to the use of Earth observation satellites such as Sentinel, Landsat or TerraSAR-X. Because of the large amount of data available, machine learning methods offer high potential for the automatic analysis and interpretation of these data. For the task of LULC classification, particularly methods related to supervised learning are used. These methods include, among others, Random Forest (VAN BEIJMA 2014), Support Vector Machine (SVM) (HUANG et al. 2002) and Convolutional Neural Networks (CNNs) (MOHAMMADIMANESH 2019). In particular, algorithms with a high number of trainable parameters, such as those found in the area of deep learning, require many annotated training data for successful learning processes. The high volume and diversity of the training data significantly contributes to the convergence of the training process, as well as the ability of the classifier to generalize to unseen samples. Manual

[1] Karlsruhe Institute of Technology, Institute of Photogrammetry and Remote Sensing, Englerstraße 7, D-76131 Karlsruhe, E-Mail: [Sylvia.Schmitz, Antje.Thiele, Martin.Weinmann, Uwe.Weidner]@kit.edu

[2] Fraunhofer Institute of Optronics, System Technologies and Image Exploitation IOSB, Gutleuthausstraße 1, D-76275 Ettlingen, E-Mail: [Sylvia.Schmitz, Antje.Thiele, Horst.Hammer]@iosb.fraunhofer.de

annotation of data usually requires a considerable amount of expert knowledge and can only be realized with an enormous amount of time.
One apparent approach to bypass manual annotation is to use freely available LULC products e.g. crowdsourced data such as OpenStreetMap (OSM) (HAKLAY & WEBER 2008) or products developed and provided as part of Earth observation programs e.g. Copernicus. Such products contain a high amount of valuable information suitable for an initial labelling of remote sensing data. Especially the tagged data layers from OSM often serve as class labels within the training of classification and segmentation methods. For example, BRIAN & KOTARO (2016) extract OSM polygons of the categories *natural* and *landuse* to train a Random Forest in order to realise LULC mapping using time-series of satellite imagery. Further, to generate training data for forest/non-forest classification, PEKKARINEN et al. (2009) use class labels extracted from a map of the CORINE Land Cover (CLC) programme.
However, available LULC products also comprises some disadvantages that interfere with the training of robust and reliable classifiers. For example, one LULC product may have an insufficient minimum mapping unit (MMU), while others do not cover the entire training area, do not represent all target classes, or contain misclassifications. The validity of the training labels derived from a LULC product is therefore not ensured. However, the reliability of training data in supervised learning algorithms is a fundamental factor for the performance of the resulting classifier. Another issue concerning this strategy is that LULC products build on the analysis of data from the past. Thus, they may not represent the state at the acquisition time of the data to be annotated.
In this paper, we present an approach to generate reliable annotations of remote sensing data with minimal manual labelling effort by fusing various sources of information. Instead of using only one LULC product as annotation, several complementary products are combined. The fusion of information eliminates the weaknesses of individual products and enhances their combined advantages at the same time. Besides various LULC products, information about the LULC state at the time of data acquisition are included. This information is derived from characteristic features of the data to be annotated. Using the Dempster-Shafer evidence theory, all information are combined to create a reliable training data set. We demonstrate the performance of the approach applying the automatic annotation on SAR data recorded over the German Wadden Sea.
This paper is organized as follows: in Section 2, we introduce the study area and the data collected in that region. In Section 3, all sources of information are described that are included in the automatic annotation and the approach to fuse these information is presented. Subsequently, in Section 4, the results that can be achieved with the approach are shown. Finally, in Section 5, we summarize the results, draw conclusions accordingly and give including suggestions for future work.

2 Study area – German Wadden Sea

The presented approach is to be used on the automatic annotation of SAR data captured from tide-influenced environments. With the superior goal of generating high-resolution geodata for coastal monitoring, airborne SAR data were acquired over the German Wadden Sea. As representative test areas, data were recorded over Otzumer Balje, as well as from the Elbe estuary Medem channel

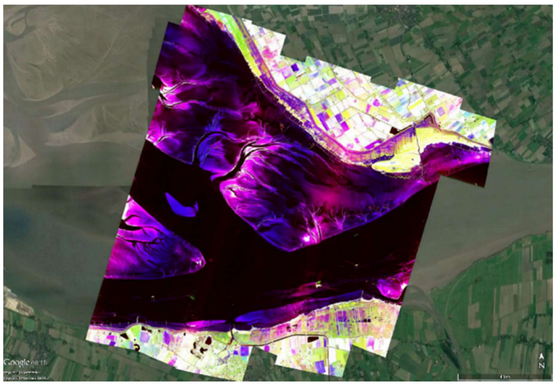 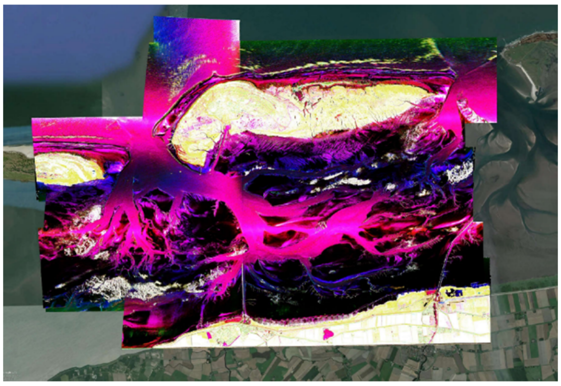

Fig. 1: Polarimetric S-band SAR image mosaic, captured over the Medem channel (left) and Otzumer Balje (right). In the background, an image of the corresponding region from GoogleEarth is displayed

(see Fig. 1). The data acquisition was performed by the F-SAR system, developed at the German Aerospace Center (Deutsches Zentrum für Luft- und Raumfahrt; DLR) (HORN 2009). F-SAR is an airborne SAR system that is capable of simultaneously capturing fully polarimetric SAR data at different wavelengths. In the measurement campaigns, the X- and S-band antennas were used, which are able to realize single-pass polarimetric interferometry. In addition to the single-pass interferometry, repeat-pass measurement mode was used to assure baseline flexibility. The flights took place in February and July 2019 on days with low tide. We have chosen times with low tides to observe large areas of dry fallen mudflats, as the structures found there are of particular interest for coastal protection. By observing variations in the morphology of tide ways, tidal flats, beaches and dunes, it is possible, for instance, to predict the formation of sandbanks and the erosion of dunes.
The presented approach automatically generates reference data for a part of the collected data in order to enable the training of various classifiers by means of supervised learning processes. These classifiers shall serve to distinguish between water, mudflats, different vegetation classes and man-made objects.

3 Automatic data annotation

With the aim of generating training data, a representative section of available airborne data is annotated automatically with respect to five classes: *water, mudflats*, *high vegetation* (e.g. trees, high shrubs), *ground* (including fields, meadow and arable land) and *man-made objects*. An overview of the approach for the generation of training data is illustrated in Fig. 2. The input data comprises classified polygons from OSM data, the CORINE Land Cover (CLC2018) map, information from the Global Surface Water (GSW) data set as well as interpreted SAR data (coherence mask and entropy mask). In order to enable a pixel-wise fusion, all information are first mapped to a uniform grid. To simplify the fusion we introduce a homogenisation of the class nomenclature. Consequently, there are up to five class assignments for each pixel to be annotated. These are regarded as evidences and fused using the Dempster-Shafer rule of combination. The resulting combined basic probability assignment function is used to determine a final class label whose certainty is quantified with a confidence value. Finally, the resulting training data set is composed of all class labels whose corresponding confidence value exceeds a fixed threshold value.

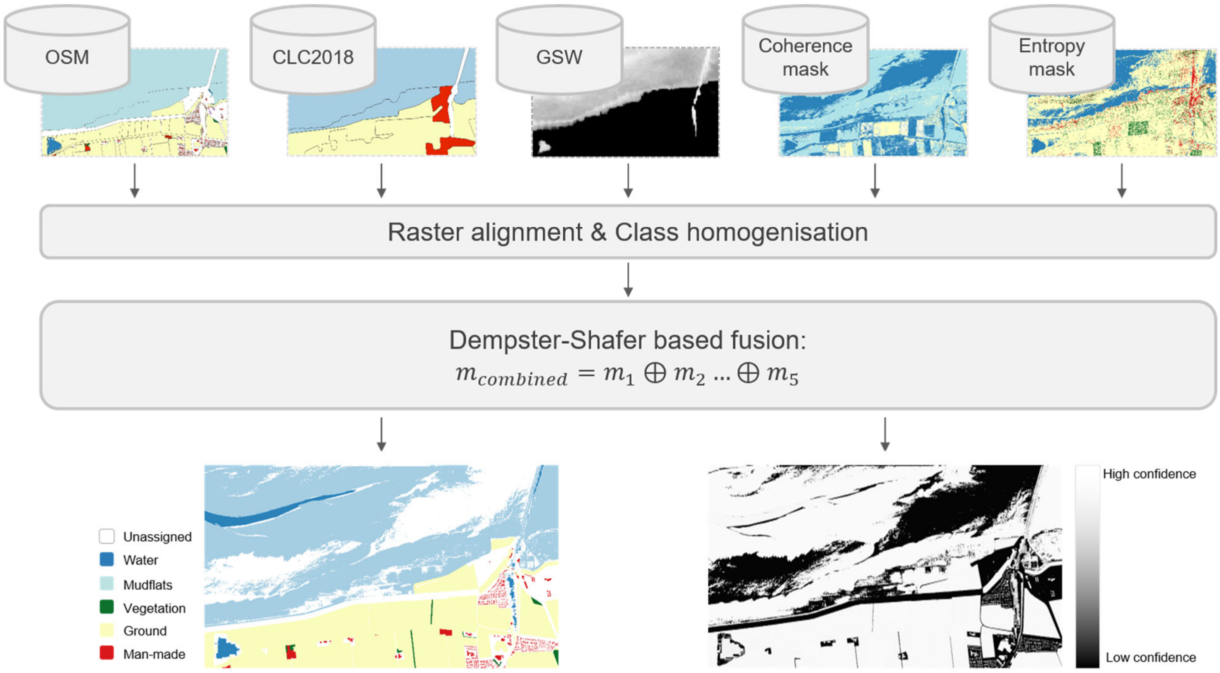

Fig. 2: Overview of the approach to generate annotated training data by fusing several LULC products (OSM, CLC2018) and interpreted SAR data (Coherence mask, Entropy mask)

3.1 Information base

The following three LULC products serve as the information basis for automatic annotation:

1. OpenStreetMap (OSM) data (HAKLAY & WEBER 2008), a crowdsourced data set to which everyone can voluntarily make his own contribution.
2. CORINE Land Cover 2018 (CLC2018) (BÜTTNER et al. 2004), a data product developed within the European Earth observation programme Copernicus by the analysis and interpretation of satellite images and in-situ data.
3. Global Surface Water (GSW) (PEKEL et al. 2016), a data set, which was also developed as part of Copernicus to reflect the temporal distribution of water surfaces over the last 3.5 decades.

The main characteristics of the data sets are listed in the first three columns of Tab. 1 and described in detail in the following sections.

3.1.1 OpenStreetMap data

The collaborative project OpenStreetMap aims at creating and maintaining freely available map material, which captures mainly information related to land use, transport infrastructure and buildings as well as the localisation of national and coastal borders. Since 2004 information has been collected on a voluntary basis by more than two million OSM users. The generation of map sections is a two-step process. First, raw data is recorded, for example by GPS measurements collected while following roads, paths or rivers. In a second step, users transfer the data in form of polygons, lines and points to the map and assign them with attributes, so-called *tags*. These tags consist of key – value pairs. The key indicates the general topic or category of an object (e.g. *landuse*) and the value assigns the explicit class (e.g. *crop*).

Tab. 1: Properties of each included source of information.

	OSM	CLC2018	GSW	COHERENCE MASK	ENTROPY MASK
# CLASSES	61	44	-	3	5
SPATIAL RESOLUTION	-	MMU: $25\,\mathrm{ha}$	$30\,\mathrm{m} \times 30\,\mathrm{m}$	$1\,\mathrm{m} \times 1\,\mathrm{m}$	$1\,\mathrm{m} \times 1\,\mathrm{m}$
TEMPORAL COVERAGE	2004 to today	2017 to 2018	1984 to 2018	July 2019	July 2019
DATA BASIS	Voluntarily contributed data	Sentinel-2 Landsat-8	Landsat-5 Landsat-7 Landsat-8	F-SAR data	F-SAR data

Concerning the goal of annotating acquired SAR data with LULC classes, the publicly available OSM data can make a significant informational contribution. The information we require within our approach predominantly lies in polygons that are tagged with the category (key) *natural*, *waterways*, *buildings*, *man-made* and *landuse*. Thus, all polygons located within the test area and annotated with one of those keys are extracted from the OSM data along with their attributes. The information contained therein forms an important building block for the automatic data annotation approach.

3.1.2 CORINE Land Cover 2018

The EU's CORINE Land Cover (CLC) programme was initiated in 1985 with the aim of generating and providing of standardised localised geographical information on the land cover of all EU member states. The latest data set produced within the project refers to the status of LULC in 2018 (CLC2018). Research groups from 39 member countries are involved in the development of the LULC map. On a national basis, satellite data are evaluated by visual interpretation and semi-automated processes. In addition, information from in-situ data and GIS data are partially integrated. The resulting product maps LULC with a minimum mapping unit (MMU) of 25 hectares to 44 classes, following the hierarchical 3-level CORINE nomenclature. The coarsest level (Level 1) consisting of *artificial surfaces*, *forest and semi natural areas*, *wetlands* and *water bodies* corresponds directly to the five class assignments that we want to map in our training data.
As shown in Fig. 2, the CLC2018 data set provides less detailed spatial information compared to polygons extracted from OSM data. However, it offers complete coverage of the test area in contrast to the sparse coverage of OSM polygons. Thus, it complements the OSM data on the one hand and serves for the validation of the OSM polygons on the other hand.

3.1.3 Global Surface Water

The GSW-Explorer provides water data sets developed by the European Commission's Joint Research Centre in order to support scientists and decision-makers, especially in the field of water management. The data sets were developed based on the analysis of more than 3 million satellite images acquired by Landsat 5, 7, and 8 between March 1984 and October 2015. The resulting maps represent information for the entire globe on the local and temporal occurrence of water surfaces, as well as their changes. Natural and artificial water surfaces are mapped with a spatial resolution of 30 m × 30 m.

Since the test area consists mainly of water and mudflats, which need to be distinguished, information regarding water occurrences are of particular importance. For this reason, we include the Surface Water Occurrence map of the GSW-Explorer in the annotation process. This map displays the spatial and temporal variations of surface water in a single product, giving the frequencies with which water surfaces occur. With the background knowledge that the data to be annotated were captured at low tide, this information supports the distinction between water, mudflats and land.

3.2 Information from SAR data

The three selected products do not necessarily reflect the state of LULC at the desired time, but represent results that may be outdated or show a mixture of past occurrences. However, for the annotation of currently available remote sensing data, it is of great importance to map LULC at the time of acquisition. This is particularly important in the tidal areas we are investigating, as water and mudflats are constantly changing. In order to be able to map the corresponding water levels correctly, it is necessary to include information derived directly from present SAR data. In order to obtain relevant information, meaningful SAR features are extracted. To differentiate between water, mudflats and land areas, we make use of the multi-path constellation that was applied during data acquisition. Based on two co-registered S-band SAR images with VV polarisation, a coherence image is calculated. Due to the reflective properties of water, low coherence values result in the corresponding areas. With the addition of X-band amplitude images, a threshold value method is used to assign the label *water*, *mudflat* or *land* to each pixel. For example, pixels with low coherence and low amplitude belongs to class *water*. The resulting coherence mask can reliably distinguish between these three classes. However, the differentiation of different land cover classes (*ground*, *vegetation*, *man-made*) is not is not performed. Therefore, another classification mask based on polarimetric features is generated. Following the approach proposed by CLOUDE & POTTIER (1997), the complex-valued polarimetric scattering matrix is decomposed to extract entropy (H), alpha (α) and anisotropy values. Based on the resulting H/α-plane, each pixel receives one of the class labels: *water*, *mudflat*, *ground*, *vegetation* or *man-made*. The corresponding interpretation of the H/α-plane for class assignments is shown in Fig. 3.

The maps resulting from the two methods (hereinafter referred to as entropy mask and coherence mask) are used in the annotation process to incorporate knowledge about the LULC at the appropriate time.

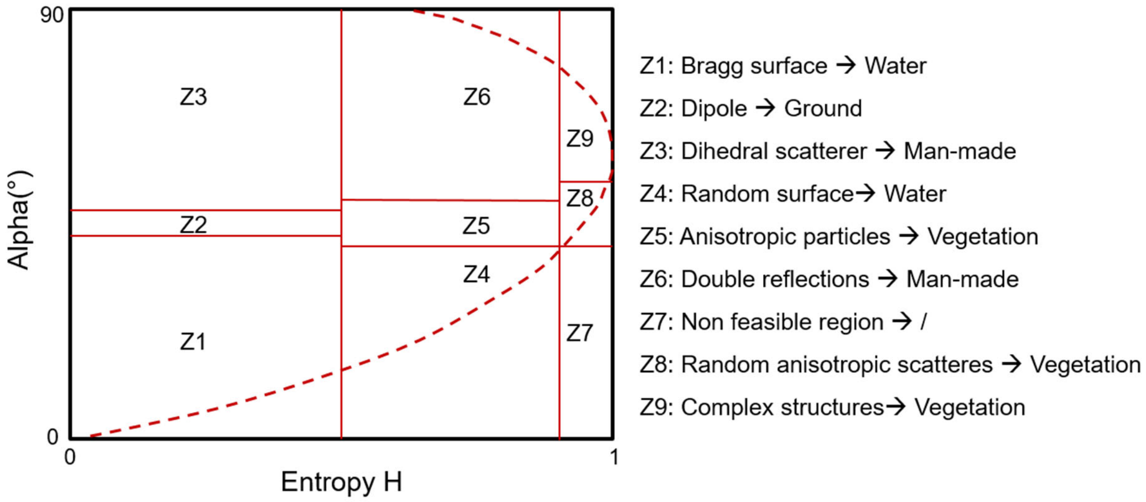

Fig. 3: The H-alpha plot (Cloude & Pottier, 1997) shows how Entropy and alpha values are used to distinguish between five LULC classes

3.3 Homogenisation of class nomenclatures

To simplify the final fusion, information stored in the presented LULC products are mapped to up to five output classes: *water*, *mudflats*, *high vegetation*, *ground* and *man-made objects*. To this

Tab. 2: Mapping of OSM *tags*, CLC nomenclature and water occurrences to five output classes.

	OSM		CLC2018	GSW
WATER	Natural:	*water*	Water bodies	Occurrence > 0.95
	Waterway:	*all values*		
MUDFLATS	Natural:	*wetland, mud*	Wetlands	0.95 > Occurrence > 0.5
	Landuse:	*salt_pond*		
VEGETATION	Landuse:	*vineyard, orchard, greenfield*	Forest and semi-natural areas	Occurrence < 0.5
	Natural:	*grassland, greenfield, scrub, heath, forest*		
GROUND	Landuse:	*greenhouse_horticulture, farmland, meadow*	Agricultural areas	
	Natural:	*fell, bare_rock, sand, rock, cliff*		
MAN-MADE	Landuse:	*industrial, commercial, retail, quarry, construction, allotments, farmyard, garages*	Artificial surfaces	
	Man_made:	*all values*		
	Office:	*all values*		
	Buildings:	*all values*		
	Shop:	*all values*		

end, a custom assignment scheme, presented in Tab. 2 is applied. For the mapping of information from OSM data, several key – value pairs are grouped and one of the output classes per group is assigned. In the case of CLC2018 data, the class mapping is straightforward adopting the Level-1 labels. The GSW data provide values between 0% and 100%. Two thresholds were chosen to separate *water*, *mudflat* and *land* (see Tab. 2). In this case we do not differentiate between *vegetation*, *ground* and *man-made*, because the GSW data do not contain any information about different land types.

3.4 Fusion methodology

As a result of the processing steps described above, for each pixel, we obtain up to five statements regarding the class label. In order to come to an overall decision about the class label, we fuse the partially conflicting statements using the Dempster-Shafer theory (DST). The DST, which has been developed by DEMPSTER and SHAFER (1968, 1992), provides a tool for knowledge representation and reasoning with uncertainty in expert systems. It can be applied to combine information (evidences) from different sources into an overall statement. In contrast to traditional probability theory principles, the Dempster Shafer approach allows for the representation of ignorance.

Formally, the theory is described by the following definitions:

(1) *Frame of discernment*: In DST, all mutually exclusive hypotheses form a finite non-empty set θ, commonly known as frame of discernment. $\Omega(\theta)$ denotes the power set comprising all 2^{θ}subsets of θ, including itself and the null hypotheses $\emptyset$. For the case we investigate, the distinction between five LULC classes, this results in the frame of discernment:

$$\theta = \{'water', 'mudflats', 'vegetation', 'ground', 'man\text{-}made'\}$$

The resulting power set $\Omega(\theta)$ contains all 32 combinations of these hypotheses, but only some subsets are of interest. One relevant example is the subset

$$A = \{'vegetation', 'ground', 'man\text{-}made'\},$$

which implies that the LULC class is one of the land classes *vegetation*, *ground* or *man-made.*

(2) *Basic probability assignment function*: To quantify the certainty with which each subset of $\Omega(\theta)$ is supported by a distinct evidence, the basic probability assignment (bpa) function *m* is introduced. It is defined as a mapping that assigns a value between 0 (no belief) and 1 (total belief) to every element of the power set. The formal representation is given by:

$$m :\ \Omega(\theta) \rightarrow [0, 1]. \qquad \text{Eq. 1}$$

with

$$m(\emptyset) = 0, \qquad \sum_{A \subset \Omega(\theta)} m(A) = 1. \qquad \text{Eq. 2}$$

The bpa of an individual hypothesis $m(A)$ quantifies the portion of the total belief committed exactly to the subset A. However, it does not provide any information about the amount of support across the different subsets of A. This shall be demonstrated by an example related to our application. Based on the information extracted from GSW, which indicate that the frequency of water occurrences was 0, we can assign a high portion of the total belief to the subset $A = \{'vegetation', 'ground', 'man\text{-}made'\}$. Note that by this the amount of support of the individual classes, e.g. *vegetation*, is not given.

To represent ignorance, the total amount of belief 1 is given to the full set θ and consequently no belief (0) to any other subset of θ. Applied to our example, this representation is reasonable if a dataset does not provide any information for one considered pixel. This occurs especially in connection with the sparse OSM data.

(3) *Dempster's rule of combination*: In order to combine several bpa functions, derived from evidence from independent sources, Dempster's rule of combination can be applied. Given two bpa functions m_1 and m_2 in the same frame of discernment θ, a joint bpa is calculated using the following equation:

$$m_{1,2}(A) = (m_1 \oplus m_2)(A) = \frac{1}{K-1} \sum_{B \cap C = A} m_1(B)\, m_2(C), \qquad \text{Eq. 6}$$

$$\text{with } K = \sum_{B \cap C = \emptyset} m_1(B)\, m_2(C) \; > 0. \qquad \text{Eq. 7}$$

Thus, K indicates the amount of conflict among the evidences, obtained from two different sources of information. The larger the conflict, the less informative is the resulting combination of sources. The following generalisation results for the combination of more than two sources of evidence to an aggregated hypothesis:

$$m(A) = (m_1 \oplus m_2 \oplus \dots \oplus m_n)(A), \qquad \text{Eq. 8}$$

$$\frac{1}{1-K} \sum_{\cap_{i-1}^{n} E_i = A} m_1(E_1) \cdot m_2(E_2) \cdot \dots \cdot m_n(E_n). \qquad \text{Eq. 9}$$

We use the previously explained evidence theory to fuse the information contained in OSM, CLC2018, GSW, entropy mask and coherence mask for each pixel. As already indicated, the frame of discernment is determined by:

$$\theta = \{'water', 'mudflats', 'vegetation', 'ground', 'man\text{-}made'\}.$$

The first substantial step of our approach includes the assignment of bpa functions to the respective evidence, derived from the different information sources. In other words, we need to quantify the reliability of the evidence that we extract from the selected LULC products. This is accomplished by including prior knowledge about the class significance of the different data sets. This prior knowledge is obtained by comparing class predictions of the data sets to a manually labelled training area. Following the approach proposed by DENG et al. (2006), we construct bpa functions $m_i^{\varphi_k}$ for each class c_i $(i = 1 \dots N)$ in every data set φ_k $(k = OSM, CORINE, GSW, entropy\ mask, coherence\ mask)$ based on recall and precision rates. Using the labelled training area, that contains the true classes c_i, we compute the confusion matrices C_{φ_k} to quantify the performance of a data set φ_k:

$$C_{\varphi_k} = \begin{bmatrix} n_{11} & n_{12} & \dots & n_{1N} \\ n_{21} & n_{22} & \dots & n_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ n_{N1} & n_{N2} & \dots & n_{NN} \end{bmatrix}. \qquad \text{Eq. 10}$$

Here, n_{ij} represents the number of pixels belonging to the true class c_i and classified by the data set φ_k as c_j. Note that we exclude all unclassified pixels. With the precision rate r_{ij}^r and recall rate r_{ij}^p given by:

$$r_{ij}^r = \frac{n_{ij}}{\sum_{j=1}^{N+1} n_{ij}} \quad \text{and} \quad r_{ij}^p = \frac{n_{ij}}{\sum_{i=1}^{N+1} n_{ij}}, \qquad \text{Eq. 11}$$

we derive the corresponding matrices:

$$C_{\varphi_k}^r = \begin{bmatrix} r_{11}^r & r_{12}^r & \dots & r_{1N}^r \\ r_{21}^r & r_{22}^r & \dots & r_{2N}^r \\ \vdots & \vdots & \ddots & \vdots \\ r_{N1}^r & r_{N2}^r & \dots & r_{NN}^r \end{bmatrix} \quad \text{and} \quad C_{\varphi_k}^p = \begin{bmatrix} r_{11}^p & r_{12}^p & \dots & r_{1N}^p \\ r_{21}^p & r_{22}^p & \dots & r_{2N}^p \\ \vdots & \vdots & \ddots & \vdots \\ r_{N1}^p & r_{N2}^p & \dots & r_{NN}^p \end{bmatrix}. \qquad \text{Eq. 12}$$

Based on $C_{\varphi_k}^r$ and $C_{\varphi_k}^p$, two bpa functions $m_{\varphi_k,i}^r$ and $m_{\varphi_k,i}^p$ are derived for every class c_i with:

$$m_{\varphi_k,i}^r(\{c_i\}) = \frac{r_{ii}^r}{\sum_{j=1}^{N} r_{ji}^r} \quad \text{and} \quad m_{\varphi_k,i}^p(\{c_i\}) = \frac{r_{ii}^p}{\sum_{j=1}^{N} r_{ji}^p},\ c_i \in 2^{\theta}, \qquad \text{Eq. 13}$$

$$m_{\varphi_k,i}^r(\{\theta\}) = 1 - m_{\varphi_k,i}^r(\{c_i\}) \quad \text{and} \quad m_{\varphi_k,i}^p(\{\theta\}) = 1 - m_{\varphi_k,i}^p(\{c_i\}),\ c_i \in 2^{\theta}, \qquad \text{Eq. 14}$$

$$m_{\varphi_k,i}^r(\{A\}) = 0 \qquad \text{and} \quad m_{\varphi_k,i}^r(\{A\}) = 0,\ \forall A \in 2^{\theta} \setminus \{\{c_i\}, \theta\}. \qquad \text{Eq. 15}$$

Note that the considered classes c_i can represent both singletons as well as subsets such as $\{'vegetation',' ground','man\text{-}made'\}$ of θ.
To generate the final bpa functions $m_i^{\varphi_k}$, that reflect the ability of a dataset φ_k to recognise the class c_i the two bpa functions $m_{\varphi_k,i}^r$ and $m_{\varphi_k,i}^p$ are combined:

$$m_i^{\varphi_k} = m_{\varphi_k,i}^r \oplus m_{\varphi_k,i}^r. \qquad \text{Eq. 16}$$

After the generation of bpa functions based on a training area, we can produce a fused map for an independent test area. For each pixel p within the test area, every data set assigns a predicted class $c_p^{\varphi_k}$. According to the predictions, the corresponding bpa functions $m_i^{\varphi_k}$ with $i = c_p^{\varphi_k}$ are selected. For the case of unclassified samples, $m^{\varphi_k}(\Theta) = 1$ is used. For the fusion of the five bpa functions, we apply Dempster's rule of combination and obtain the combined bpa function:

$$m_p^{combined} = f(\varphi_1, c_p^{\varphi_1}) \oplus f(\varphi_2, c_p^{\varphi_2}) \oplus \dots \oplus f(\varphi_5, c_p^{\varphi_5}). \qquad \text{Eq. 17}$$

Here $f(\varphi_k, c_p^{\varphi_k})$ denotes the associated bpa $m_i^{\varphi_k}$, belonging to the predicted class $c_p^{\varphi_k}$ for pixel p. The last step contains the decision-making regarding the final class, depending on $m^{combined}$. For this, we apply the pignistic transformation (SMETS et al., 1994):

$$c_p^{combined} = \arg\max\left\{\sum_{A\subseteq\theta, x\in A} \frac{1}{|A|}\frac{m_p^{combined}(A)}{1-m_p^{combined}(\emptyset)}\right\},\ x\in\theta. \qquad \text{Eq. 18}$$

As a result, we obtain the final class label $c_p^{combined}$ for each pixel p. Furthermore, information about the reliability of the class assignments is stored for each label. To quantify this information, we use the maximum value calculated in the argument of Eq. 16, denoted as confidence value. For our goal of generating annotated training data, the complete annotation of each individual pixel is not necessary. More important is the correctness of the annotation of pixels that build up the training data set. Therefore, the intended training data set consists only of pixels whose class assignments are marked with a high confidence value.

4 Results

We use the presented approach in order to annotate a part of SAR images, captured over the German Wadden Sea. For the evaluation of the proposed fusion method, a test area was selected that covers a coastal strip south of Spiekeroog. Fig. 4 shows the SAR image of the test area overlaid with manually annotated polygons. The annotated area was divided into training and validation areas, whereby in both areas all output classes are included. The training area comprises 2,098,003 annotated pixels with a pixel size of 1 m × 1 m. The annotated pixels of this area are used to determine the bpa functions, which define the reliability of the data to be fused. For this purpose, we calculate the confusion matrices and derive the corresponding recall and precision matrices for each included LULC product.

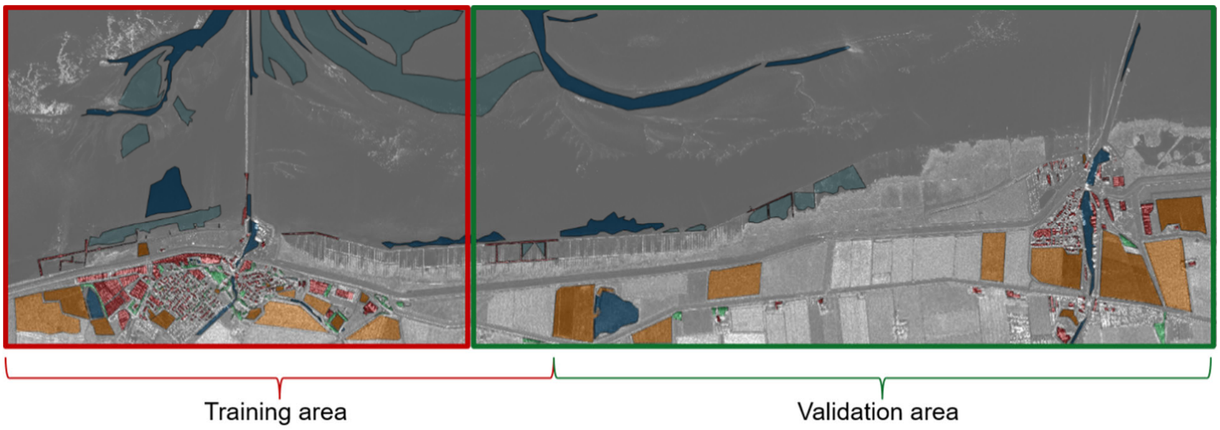

Fig. 4: Test area, divided into training and validation area. The S-band amplitude image is overlaid with manually annotated polygons

The validation area contains 1,851,813 annotated pixels. Within this area, class labels are determined by fusing five LULC products, using the presented approach. For the automatic selection of reliable annotations, the fused map is filtered using a threshold value of 0.9 on the provided confidence values. In this way, 69.10% of the pixels within the validation area are assigned a class label. 1,456,739 of these pixels have a manual annotation and are used for evaluation. In the following, we present the obtained results and compare them with the performance of the individual LULC products OSM and CLC2018. For the evaluation, the confusion matrices including precision and recall rates per class are shown. For the validation area, we obtain performance scores that are summarized in Tab. 3 to 5. With precision values between 0.87 and 0.99, the annotation resulting from the fusion provides highly reliable class information. The separation of the five classes is successful and only few misclassifications remain after the fusion and the pixel selection. The only poor performance value is the recall rate of 0.62 for *man-made objects*. This class is often mistakenly classified as *mudflats*, since tiny coastal buildings are not captured in any of the used LULC products. Compared to the performance of OSM and CLC2018, the strength of the fused result is particularly evident in the distinction between *water* and *mudflats*. Due to the constant change in the tidal influenced area, the static LULC products OSM and CLC2018 do not reflect the water level at the time of data acquisition. However, the fused result takes into account time developments and current information, so that significantly higher precision and recall rates can be achieved. As Tab. 4 shows, the performance of the OSM data is high regarding the distinction of the three land classes. This strength is also exploited in the fusion, so the strong performance is reflected in the fused result.

Tab. 3 Confusion matrix for the fused result (FUSED).

	Man-Made	Ground	Vegetation	Mudflats	Water	**Recall**
Man-Made	**46,433**	3,837	156	24,104	204	0.62
Ground	2,,319	**807,730**	0	0	0	0.997
Vegetation	864	1,567	**25,158**	0	0	0.91
Mudflats	0	0	0	**269,435**	0	1.0
Water	0	677	0	16,056	**282,903**	0.94
Precision	0.94	0.99	0.99	0.87	0.99	

Tab. 4 Confusion matrix for OSM data.

	Man-Made	Ground	Vegetation	Mudflats	Water	**Recall**
Man-Made	**46,697**	3,837	156	21,583	9,180	0.57
Ground	2,319	**807,730**	0	0	4,010	0.99
Vegetation	864	1,567	**25,158**	0		0.91
Mudflats	0	0	0	**269,473**	0	1.0
Water	0	732	0	107,611	**133,971**	0.55
Precision	0.94	0.99	0.99	0.67	0.91	

Tab. 5 Confusion matrix for CLC2018 data.

	Man-Made	Ground	Vegetation	Mudflats	Water	**Recall**
Man-Made	**63,582**	49,544	0	31,936	11,984	0.40
Ground	10,809	**904,358**	0	0	45	0.99
Vegetation	80	33,583	**0**	0	0	0
Mudflats	0	0	0	**269,473**	0	1.0
Water	4,290	87,624	0	168,787	**215,718**	0.45
Precision	0.81	0.84	0	0.57	0.95	

A qualitative assessment can be deduced from Fig. 5 that shows the annotation of the different LULC products (OSM, CLC2018; FUSED) over the validation area. It is noticeable that, in many cases, reliable class labels from the OSM data will prevail in the context of the fusion for the land area. In the tidal influenced area, the dominance of the OSM classes is significantly lower. Due to stronger conflicts between the fused LULC products in the tidal area, the occurrence of unclassified areas increases. By excluding pixels with high uncertainty, wrong class assignments are prevented. Thus, only those pixels remain in the tidal area for which a reliable distinction can be made between *water* and *mudflats*.

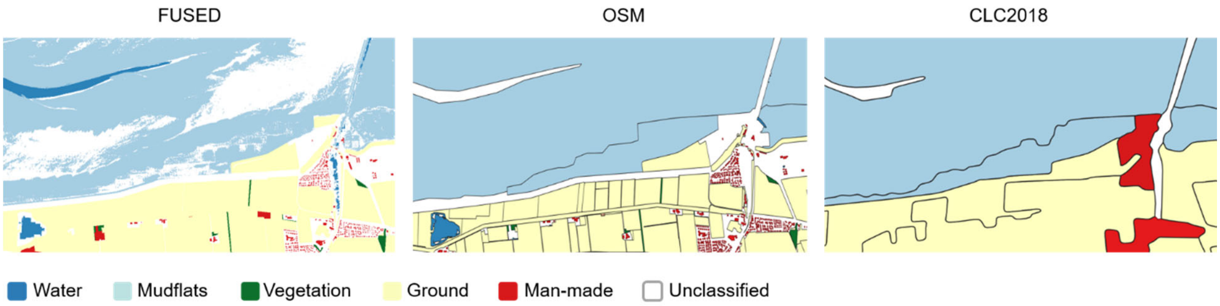

Fig. 5: Assigned classes by CLC2018, OSM and FUSED (our result)

5 Conclusion

In order to apply supervised learning to train LULC classifiers that rely solely on remote sensing data, a high amount of annotated data is necessary. In this paper, an approach is proposed to annotate remote sensing data automatically. Using the example of SAR data from the German Wadden Sea, we show how to use Dempster-Shafer theory to fuse information from LULC products in order to generate reliable training data. The presented results show that reliable class assignments are achieved by fusing data derived from OSM, CLC2018, GSW and from polarimetric and coherent features of the SAR data itself. The results testify that the fused map represents the LULC classes at the acquisition time of the SAR data more precisely than a single data set. It receives details in contrast to the roughly resolved CLC2018 map, and represents all required classes as opposed to the GSW map, coherence mask and entropy mask. By including information based on SAR features, the fusion result also maps the current water occurrence and distinguishes reliably between *water* and *mudflats*. This makes it more suitable as a training set than for example only OSM data. Additionally, the automatic selection of certain class assignments also greatly reduces misclassifications, which is especially relevant for the use as training data. In the future, we intend to use the proposed method for the annotation of spatially and temporally distributed SAR images to generate a diversified training data set. Such a data set will enable the robust training of CNN-based classifiers with a high number of parameters. In particular, a training from scratch is facilitated, which is essential for the analysis of SAR data to learn SAR-specific low-level and high-level features.

6 Acknowledgement

This study is part of the GeoWAM project that is funded by the German Federal Ministry of Transport and Digital Infrastructure within the framework of the Modernity Fund ("mFUND").

7 Literature

BÜTTNER, G., FERANEC, J., FAFFRAIN, G., MARI, L., MAUCHA, G. & SOUKUP, T., 2004: The CORINE Land Cover 2000 Project. EARSeL eProceedings **3**(3), 331-346.

DEMPSTER, A. P., 1968: A generalization of Bayesian inference. Journal of the Royal Statistical Society **30**, 205-247.

DENG, X., LIU, Q., DENG, Y. & MAHADEVAN, S., 2016: An improved method to construct basic probability assignment based on the confusion matrix for classification problem. Information Sciences **340**, 250-261.

HAKLAY, M. & WEBER, P., 2008: OpenStreetMap: User-generated street maps. IEEE Pervasive Computing **7**(4), 12-18.

HANSEN, M. C., POTAPOY, V. P., MOORE, R., HANCHER, M., TURUBANOVA, S. A., TYUKAVINA A., THAU, D., STEHMAN, S. V., GOETZ, S. J., LOVELAND, T. R., KOMMAREDDY, A., EGOROV, A., CHINI, L., JUSTICE, C. O. & TOWNSHEND, J. R. G., 2013: High-resolution global maps of 21st-century forest cover change. Science **342**(6160), 850-853.

HORN, R., JAEGER, M., KELLER, M., LIMBACH, M., NOTTENSTEINER, A., PARDINI, M., REIGBER, A. & SCHEIBER, R., 2017: F-SAR - Recent upgrades and campaign activities. 18th International Radar Symposium (IRS), 1-10.

HUANG, C., DAVIS, L. S. & TOWNSHEND, J. R. G., 2002: An assessment of support vector machines for land cover classification. International Journal of Remote Sensing **23**(4), 725-749.

MOHAMMADIMANESH, F., SALEHI, B., MAHDIANPARI, M., GILL, E. & MOLINIER, M., 2019: A new fully convolutional neural network for semantic segmentation of polarimetric SAR imagery in complex land cover ecosystem. ISPRS Journal of Photogrammetry and Remote Sensing **151**, 223-236.

PEKEL, J., COTTAM, A., GORELICK, N. & BELWARD, A. S., 2016: High-resolution mapping of global surface water and its long-term changes. Nature **540**(7633), 418-422.

PEKKARINEN, A., REITHMAIER, L. & STROBL, P., 2009: Pan-European forest/non-forest mapping with Landsat ETM+ and CORINE Land Cover 2000 data. ISPRS Journal of Photogrammetry and Remote Sensing **64**(2), 171-183.

PIELKE SR, R. A., PITMAN, A., NIYOGI, D., MAHMOOD, R., MCALPINE, C., FAISAL, H., GOLDEWIJK, K., NAIR, U., BETTS, R. & FALL, S., 2011: Land use/land cover changes and climate: modeling analysis and observational evidence. Wiley Interdisciplinary Reviews: Climate Change, **2**(6), 828-850.

SHAFER, D., 1992: Dempster-Shafer theory. Encyclopedia of Artificial Intelligence **1**, 330-331.

SMETS, P. & KENNES, R., 1994: The transferable belief model. Art. Intelligence **66**(2), 191-234.

VAN BEIJMA, S., COMBER, A. & LAMB, A., 2014: Random forest classification of salt marsh vegetation habitats using quad-polarimetric airborne SAR, elevation and optical RS data. Remote Sensing of Environment **149**, 118-129.