AUTOMATIC CLASSIFICATION OF AERIAL IMAGERY FOR URBAN HYDROLOGICAL APPLICATIONS

A. Paul [a, *], C. Yang [a], U. Breitkopf [a], Y. Liu [a, b], Z. Wang [a, c], F. Rottensteiner [a], M. Wallner [d], A. Verworn [d], C. Heipke [a]

[a] Institute of Photogrammetry and GeoInformation, Leibniz Universität Hannover, Germany
(paul, yang, breitkopf, rottensteiner, heipke)@ipi.uni-hannover.de; [b] xkyqyzm@gmail.com, [c] wzy553504799@gmail.com
[d] BPI Consulting Engineers, Hannover, Germany, (wallner, verworn)@bpi-hannover.de

Commission ICWG II/III

KEY WORDS: Classification, hydrologic application, coefficient of imperviousness, Random Forests, Conditional Random Fields

ABSTRACT:

In this paper we investigate the potential of automatic supervised classification for urban hydrological applications. In particular, we contribute to runoff simulations using hydrodynamic urban drainage models. In order to assess whether the capacity of the sewers is sufficient to avoid surcharge within certain return periods, precipitation is transformed into runoff. The transformation of precipitation into runoff requires knowledge about the proportion of drainage-effective areas and their spatial distribution in the catchment area. Common simulation methods use the *coefficient of imperviousness* as an important parameter to estimate the overland flow, which subsequently contributes to the pipe flow. The coefficient of imperviousness is the percentage of area covered by impervious surfaces such as roofs or road surfaces. It is still common practice to assign the coefficient of imperviousness for each particular land parcel manually by visual interpretation of aerial images. Based on classification results of these imagery we contribute to an objective automatic determination of the coefficient of imperviousness. In this context we compare two classification techniques: Random Forests (RF) and Conditional Random Fields (CRF). Experimental results performed on an urban test area show good results and confirm that the automated derivation of the coefficient of imperviousness, apart from being more objective and, thus, reproducible, delivers more accurate results than the interactive estimation. We achieve an overall accuracy of about 85% for both classifiers. The root mean square error of the differences of the coefficient of imperviousness compared to the reference is 4.4% for the CRF-based classification, and 3.8% for the RF-based classification.

1. INTRODUCTION

A noticeable increase of flash flooding events in frequency and intensity caused by heavy rainfall can be observed for parts of Germany over the last 30 years. One of the reasons for the increase of flooding events, also noticeable globally, lies in climate change. Meanwhile, the ongoing urbanisation leads to an increase in impervious areas, which drives existing sewer systems to their limits. Consequently, there is a general interest in adaptation and planning of efficient sewer systems to minimize the damage due to flooding events. For that reason, urban drainage models are needed which transform precipitation into runoff. An important input parameter for the simulation of this transformation is the *coefficient of imperviousness,* which describes the proportion of area covered by impervious surfaces such as roofs or road surfaces relative to the catchment area. Typically, it is determined independently for each piece of land stored in a geographical information system, e.g. for each parcel of a cadastral database. It is common practice to determine the coefficient of imperviousness interactively based on the visual interpretation of aerial images: the cadastral boundaries are superimposed to an orthophoto and the operator estimates the coefficient of imperviousness for each piece of land. Apart from being time-consuming, this procedure leads to very subjective decisions, and thus to a poor reproducibility of the results.

In this paper, we address the problem of automatic supervised classification of aerial imagery with a focus on the determination of coefficient of imperviousness. We want to assess how the automatic classification of such data can contribute to making the determination of the coefficient of imperviousness more objective. We use input data derived from aerial imagery: multispectral orthophotos, digital surface models (DSM) and a digital terrain model (DTM). In addition, we use the cadastral database of the study area, which provides the boundaries of the cadastral parcels as well as with the information about its land use. The data must cover the entire catchment area for which the hydrological analysis is to be performed. We compare two supervised classification techniques, Random Forests (RF; Breiman, 2001) and Conditional Random Fields (CRF; Kumar and Hebert, 2006). These methods are used to determine the land cover for each pixel of the area of interest, which forms the basis for the determination of the coefficient of imperviousness.

Our processing pipeline starts with the extraction of different spectral, textural, structural and three-dimensional features using the available data. A ranking of the extracted features based on RF feature importance is undertaken to find out which features are the most relevant ones. In the second step, five land cover classes (*asphalt, building, tree, grass* and *bare soil)* are determined by supervised classification, using both CRF and RF. After training the classifiers, a land cover label is predicted for each pixel in the test area based on its features, resulting in a land cover map. Finally, this land cover map is post-processed in order to determine the required coefficient of imperviousness. For that purpose, we combine the classes supposed to be impervious (*asphalt, building*) and the classes that are permeable (*grass, bare soil*). Class *tree* requires a special treatment, because the material below the tree is relevant for the surface runoff. As the surface property is not directly visible, we assume that in areas corresponding to roads, trees will overhang

https://doi.org/10.5194/isprs-archives-XLII-3-1355-2018 | © Authors 2018. CC BY 4.0 License.

the roads, and thus *tree* pixels are considered to be impervious, otherwise they belong to the class permeable (see section 3.4 for details). Having thus defined impervious and permeable surfaces, the coefficient of imperviousness is simply the percentage of impervious pixels inside a certain area.

Our scientific contributions can be summarized as follows:

- We propose an efficient pipeline of determining coefficient of imperviousness, consisting of feature extraction, automatic classification and post processing, based on high resolution orthophoto and digital height models.
- For the task of automatic classification, we compare different classification methods (Random Forests and Conditional Random Fields), and highlight their benefits.
- We compare the coefficient of imperviousness determined by automatic classification to those estimated by a human operator in order to assess whether the automated processing chain leads to more accurate predictions.

The remainder of this paper is organized as follows. Section 2 gives an overview on related work in classification, with a focus on hydrological applications. In Section 3 we present automatic classification methods used in this work. Section 4 describes the experimental evaluation of our approach for real urban watershed. We conclude the article with an outlook and a discussion of future works in Section 5.

2. RELATED WORK

We start the review by a discussion of strategies for hydrological modelling, focusing on runoff simulation, based on remotely sensed data. We then give a brief review of hand-crafted feature selection and automatic classification based on these features for deriving land cover information from remote sensing data.

Many works proposing methods to simulate the rainfall-runoff have relied on remotely sensed data. Kite et al. (1992) apply semi-distributed watershed modelling, which requires prior knowledge about land cover classes (*bare ground*, *forests* and *grass*). Deguchi and Sugio (1994) evaluate the applicability of satellite imagery for estimating the percentage of impervious area in a scene with the goal of using it for runoff simulation in urban areas. In their work, land cover classification is achieved by clustering algorithms. Schmugge et al. (2002) propose methods for estimating snowmelt runoff by using the hydrological state variables such as snow cover and water equivalent. These variables are derived relying on physical models, using the characteristics of the land surface such as emissivity or temperature as input. The necessary physical parameters are determined from the visible and near-infrared bands of remotely sensed imagery. Kite and Pietroniro (1996) claim that rainfall-runoff modelling requires the separation of pervious and impervious surfaces and, thus, the information about land cover and the corresponding surface permeability. Lee et al. (2004) confirm this statement and show that imperviousness is one of the most critical parameters for runoff simulation. In state-of-the-art techniques for modelling runoff, imperviousness is typically represented by the coefficient of imperviousness.

Methods for the supervised classification of satellite or aerial imagery have been used in hydrological applications for many years. Frankhauser (1999) apply the maximum likelihood method to classify land cover using a digital orthophoto and use the classification results to determine the surface imperviousness in urban areas. In (Tyrna and Hochschild, 2010), a support vector machine (SVM; Cortes and Vapnik, 1995) is applied to the classification of satellite imagery and digital elevation data to determine *impervious* and *pervious* surfaces, aiming at modelling surface runoff. However, the authors do not assess the performance of classification with respect to the runoff modelling. Tokarczyk et al. (2015) propose a workflow for runoff simulation by using the imagery captured by an unmanned aerial vehicle (UAV). The classification results of images into two (*impervious, pervious*) or three classes (*rooftop, streets, vegetation*), achieved by maximum likelihood classification or boosting, serve as input to the next processing stage for the prediction of surface runoff. They state that their boosting classifier had achieved the largest accuracy for the three classes, and that more advanced classification methods would bring more accurate results, which are essential for the imperviousness determination. Hartcher and Chowdhury (2017) use high resolution aerial images as input for determining imperviousness based on a non-parametric parallelepiped classifier. The major constraint of their work is that the classification method cannot deal with the intra-class variability of the appearance of impervious surfaces, e.g. caused, by shadows or different surface properties.

The deterministic factor of runoff simulation is the quality of the imperviousness of the land cover surface, which is heavily affected by the performance of classification methods. Supervised classification using hand-crafted features have been applied in land cover application for many years. These features can be pixel-based, e.g. raw intensities of colour bands, or based on segments such as super-pixels, e.g. mean intensities inside a segment. There are many options for classifiers for deriving land cover information, e.g. the classical maximum likelihood classifier (Lillesand et al., 2003), Logistic Regression (Maas et al., 2016), or Random Forest (Gislason et al., 2006). Incorporating context as an additional information source was found to improve the classification results (Schindler, 2012). Hermosilla et al. (2012) consider context by defining specific context features that were fed to a standard classifier. Statistical models of context lead to models assuming that neighbouring pixels are more likely to belong to the same class than to different classes. This leads to a smoothing of the classification results. An important example for a statistical method for contextual classification are Conditional Random Fields (CRF; Kumar and Hebert, 2006). The advantages of smooth labelling methods, in particular CRF, were confirmed in Schindler (2012) and, with a focus on land cover and land use information, in (Albert et al., 2017). In this paper, we adopt Random Forest and Conditional Random Fields for the determination of land cover classifiers because the former one has shown good performance in remote sensing, while CRF additionally incorporate local context information (Schindler, 2012).

3. AUTOMATIC CLASSIFICATION AND DETERMINATION OF COEFFICIENT OF IMPERVIOUSNESS

For the classification of land cover, we used a multispectral digital orthophoto (DOP), a Digital Surface Model (DSM) and a Digital Terrain Model (DTM). We aim at a pixel-based classification using features derived from the input data. Sections 3.1 and 3.2 give an outline of the classification methods used in this study, whereas Section 3.3 describes the features used for classification. The land cover information is used to determine coefficient of imperviousness for the parcels

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLII-3-1355-2018 | © Authors 2018. CC BY 4.0 License.

of a cadastral database, which requires some post-processing as described in Section 3.4.

3.1 Random Forests (RF)

RF was introduced by Breiman (2001) and has been shown to be a powerful classifier in remote sensing applications (e.g. Schindler, 2012). RF consists of many randomized decision trees, each of which learned from a bootstrap data set, i.e. from a subset of the training samples that is drawn independently from the available training data. In the classification procedure, a feature vector is presented to each tree, and each tree casts a vote for the most likely class. The output of the RF is the class receiving the highest number of votes. If a posterior probability P(y_i | $\mathbf{f}_i(\mathbf{x})$) for the class label y_i of pixel i given the feature vector $\mathbf{f}_i(\mathbf{x})$ is required for further processing, it can be derived by dividing the number of votes n_c for class c by the total number of trees:

$$P\big(y_i = c|\mathbf{f_i}(\mathbf{x})\big) = \frac{n_c}{n_T}, \qquad (1)$$

where the notation $\mathbf{f}_i(\mathbf{x})$ implies that the feature vector $\mathbf{f}_i$ determined for pixel i may be a function of the data in a neighbourhood that may comprise the entire image $\mathbf{x}$.

In the training procedure for a tree, a subset of the bootstrap dataset is used for learning the tests in the tree nodes and, thus, the structure of the tree. For that purpose, a set of tests involving n_F features is chosen randomly. These tests area applied to split the training samples arriving at that node into two subsets, and a splitting criterion such as the Gini Index is used to select the optimal split of that node (Criminisi et al., 2013). This procedure is repeated recursively until a stopping criterion is fulfilled. One criterion is the number of samples arriving at a node; recursion is terminated if this number is smaller than a threshold $minS_{split}$, in which case the node is declared to be a leaf. Another criterion is the maximum depth d_{max} of a tree. After the structure of the tree has been determined, the remaining training samples are passed through the tree and each leaf is assigned to the most frequent class labels among the training samples arriving at that leaf.

In our experiments, we used the RF classifier from the Python machine-learning library Scikit-learn (Pedregosa et al., 2011). We tuned the hyper-parameters using the tool Parfit (2017) to achieve better results. More information about the parameters is presented in section 4.

3.2 Conditional Random Fields (CRF)

CRF (Kumar and Hebert, 2006) provide a flexible framework for contextual classification. CRF are undirected graphical models. The underlying graph consists of nodes n and edges e, where the nodes represent the image sites (in our case: pixels) and the edges link adjacent nodes. Given a label vector $\mathbf{y} = [y_1, \dots, y_i, \dots, y_n]$, where $i \in S$ is the index of a pixel and S is the set of all image sites, the goal is to assign the most probable class labels $\mathbf{y}$ for all image sites simultaneously considering all the observed data $\mathbf{x}$, which can be achieved by maximizing the posterior probability $P(\mathbf{y}|\mathbf{x})$ (Kumar and Hebert, 2006):

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z}\prod_{i\in S}\varphi(y_i,\mathbf{x}) \cdot \prod_{j\in N_i}\psi(y_i,y_j,\mathbf{x}). \quad (2)$$

In (2), $\varphi(y_i,\mathbf{x})$ are the *association potentials* and $\psi_{ij}(y_i,y_j,\mathbf{x})$ are the *interaction potentials*. The partition function Z acts as normalization constant, which transforms the potentials into probabilities, and N_i is the neighborhood of image site i.

3.2.1 Association potential: the association potential indicates how likely a node i belongs to class y_i given the observations $\mathbf{x}$. The observations $\mathbf{x}$ are represented by site-wise feature vectors $\mathbf{f}_i(\mathbf{x})$, and the association potential for node i is modelled to correspond to the posterior of a local discriminative classifier based on $\mathbf{f}_i(\mathbf{x})$, i.e. $\varphi(y_i,\mathbf{x}) = P\big(y_i|\mathbf{f_i}(\mathbf{x})\big)$. We apply the RF classifier for the association potential due to its powerful classification ability, using eq. (1) to deliver the posterior probability. In this case, we use the OpenCV implementation of RF (OpenCV, 2016).

3.2.2 Interaction potential: the interaction potentials models the dependencies of the labels at adjacent nodes n_i and n_j, by considering all the observations $\mathbf{x}$, the neighborhood consists of four direct neighbors of each pixel in an image grid. We use the contrast-sensitive Potts model as our interaction term, since it produces good results (Schindler, 2012; Albert et al., 2017) and represents a good trade-off between accuracy and computation time:

$$\ln\psi(y_i,y_j,\mathbf{x}) = \delta(y_i = y_j)\cdot\left[\beta_0 + \beta_1\cdot\exp\left(-\frac{\left\|\mathrm{f_i}(\mathbf{x})-\mathrm{f_j}(\mathbf{x})\right\|^2}{2\cdot\sigma^2}\right)\right]. \quad (3)$$

This model yields a data-dependent smoothing of the resultant label image. The function $\delta(\cdot)$ returns 1 if the argument is true and 0 otherwise. The hyper-parameter β_0 specifies the degree of smoothing that is applied independently from the data, whereas β_1 is the weight of the data-dependent smoothing term. The parameter σ^2 is the mean value of the squared distances of the two feature vectors and is determined during training.

3.2.3 Training and Inference: The association and interaction potentials are trained separately. For interaction potential, the parameter σ^2 is computed from the training samples, whereas the hyper-parameters β_0 and β_1 are defined by the user. Given the site-wise feature vectors of an image, the most probable label configuration of the graphical model is determined for all nodes simultaneously by maximizing the posterior in eq. (2). As exact inference is computationally intractable (Kumar and Hebert, 2006), an approximate solution for the optimal label configuration is determined by Loopy Belief Propagation (Frey and MaxKay, 1998).

3.3 Feature Extraction

An efficient classification requires an appropriate set of discriminative features, which are extracted from the input data. Following Albert et al. (2017), we divide the features into three categories: spectral, textual and structural, and three-dimensional geometrical. For each pixel i, we extract features based on the three categories and concatenate them to define a site-specific feature vector $\mathbf{f}_i(\mathbf{x})$.

The original spectral features are the original grey values of the DOP. The normalized difference vegetation index (NDVI) derived from the near infrared and red band of the orthophotos is included in our feature vector, because it is suitable for the discrimination of vegetation from non-vegetation. Furthermore, we apply a colour space transformation and obtain the hue, saturation and intensity for each pixel. In addition to these basic features, we determine their means and variances from a local neighbourhood whose size depends on the image resolution (we used 13 x 13 pixels in our experiments). More features are determined by applying Gaussian filters (σ = 2 and σ = 5, referred to as *Gauss*$_2$ and *Gauss*$_5$, respectively). In addition, we use the orientation and magnitude of the intensity gradients; the latter were computed based on derivative of Gaussian filters of width σ = 2. Thus, we obtain a total of 42 features in the spectral group.

The textural features are derived from a Grey Level Co-Occurrence Matrix (GLCM) (Haralick et al., 1973), which is computed from the co-occurrences of the intensity values at pixel pairs in a certain spatial configuration within a window of 5 x 5 pixels. We computed a GLCM for four directions (0°, 45°, 90°, 135°), in all cases using a distance of 1. From the resultant four GLCMs, the Haralick features *Energy*, *Contrast*, C*orrelation* and *Homogeneity* were derived, so that in total we used 16 textual features.

Structural features are derived from the histogram of orientations of gradients (HOG), weighted by their magnitude, which is computed from a window of 32 x 32 pixels centred at each pixel. From the HOG, we derive the mean and variance of the histogram entries, the number of histogram entries that are larger than the mean (*NoM*), and the angle between the first and second maximum (*Angle*). In addition, we determine a distance transform feature, which delivers the distance of a pixel to the nearest edge, the latter derived by a Canny edge extractor based on gradients determined by derivative of Gaussian filters of width σ = 2. In total, we had 5 structural features.

The 3D features are derived from the DSM and the DTM. The first one is the normalised DSM (nDSM), which is a model of the height differences between DSM and DTM. Furthermore, the gradient orientations and magnitudes of the DSM and the nDSM, all computed based on derivative of Gaussian filters of width σ = 2, and the mean and Gaussian curvatures of the DSM are determined. From the magnitudes of the nDSM gradient we also computed their means and variances in a neighbourhood of size 13 x 13 pixels. Thus, in total we extracted 9 3D features for each pixel.

Altogether, the site-specific feature vectors $\mathbf{f_i}(\mathbf{x})$ consist of 72 features. Each feature is scaled linearly to the interval [0;1], discarding the 2% smallest and largest feature values, respectively, for robustness to outliers. An overview of our feature pool is presented in Table 1.

We rank our features based on the feature importance of the Scikit-learn RF implementation, which is based on the expected fraction of samples for which a specific feature is selected (Scikit-learn, 2018). In our experiments based on RF, we compared the classification results achieved using all features to those that can be obtained when using only the 20 most important features.

3.4 Post-processing for determining the coefficient of imperviousness

In the classification procedure, we differentiate five land cover classes (*asphalt* / *as*, *building* / *bld*, *tree* / *tr*, *low vegetation* / *lv* and *bare soil* / *bs*). This information is used to derive coefficient of imperviousness for each *subbasin*. A subbasin is defined as an area which is connected to a specific pipe of the sewer system in which surface water will drain. The subbasins are delineated on the basis of individual plots corresponding to a residential unit (typically consisting of a building and the attached garden and parking space) or a street. In order to compute the coefficient of imperviousness, we first use the five land cover classes to derive the information about impervious and permeable surfaces in study area. The classes *building* and *asphalt* are merged to define the class *impervious*, whereas *bare soil* and *low vegetation* are considered to belong to class *permeable*. The assignment of pixels classified as *tree* depends on the material of the ground underneath the tree and not on the fact that the image shows a tree. If the ground is covered by asphalt, the pixel should be considered to be *impervious*, otherwise it should belong to class *permeable*. However, this information cannot be derived from the image data, because trees occlude the ground. Consequently, we use the GIS data for that purpose. We extract all road objects from the GIS and use them to generate a binary road mask. Pixels marked to correspond to roads in that mask that are labelled as *tree* are assigned to class *impervious*; all other *tree* pixels are treated as *permeable*.

Group	Basic Feature	Derived Feature
spectral	Red	*Mean, Variance, Gauss*$_2$*, Gauss*$_5$
	Green	*Mean, Variance, Gauss*$_2$*, Gauss*$_5$
	Blue	*Mean, Variance, Gauss*$_2$*, Gauss*$_5$
	NIR	*Mean, Variance, Gauss*$_2$*, Gauss*$_5$
	NDVI	*Mean, Variance, Gauss*$_2$*, Gauss*$_5$
	Hue	*Mean, Variance, Gauss*$_2$*, Gauss*$_5$
	Saturation	*Mean, Variance, Gauss*$_2$*, Gauss*$_5$
	Intensity	*Mean, Variance, Gauss*$_2$*, Gauss*$_5$
	Intensity Gradient	*Orientation, Magnitude*
textural	GLCM	*Haralick Energy* *Haralick Contrast* *Haralick Entropy* *Haralick Homogeneity*
structural	HOG	*Mean, Variance, NoM, Angle*
	Distance Transform	*Distance to nearest edge*
3D	nDSM	
	DSM	*Mean and Gaussian Curvatures*
	DSM Gradient	*Orientation, Magnitude*
	nDSM Gradient	*Orientation, Magnitude, Mean, Variance*

Table 1. Pool of spectral, textural, structural and 3D features extracted per pixel. NIR: Near infrared band of the multispectral DOP. See the main text for the definition of the features.

After this heuristic correction step (which may introduce errors due to GIS road objects that contain some areas covered by low vegetation, e.g. in the centre between two lanes, or in non-road objects where trees overhang areas reserved for parking and, thus, are covered by asphalt), we compute the coefficient of imperviousness for all subbasins. The coefficient of imperviousness is defined as the percentage of *impervious* pixels inside the specific area.

4. EXPERIMENTS

The experiments are carried out to evaluate the effectiveness and the limitations of the proposed automatic classification methods for urban hydrological applications. We also compare different classifiers and different variants of these classifiers.

4.1 Test dataset and test setup

Experiments are based on a case study in the city of Osnabrück (Germany). The ground sampling distance (GSD) of the DOP is 20 cm. It has four bands (RGB-IR) and was acquired in 2014. The DSM and the DTM were based on airborne laser scanner data acquired in 2011. Both the DSM and the DTM were provided at a coarser resolution (GSD = 50 cm) and were resampled to the GSD of the orthophoto by applying bilinear interpolation. Moreover, we had building and plot outlines corresponding to cadastral parcels of the German Authoritative Real Estate Cadastre Information System (ALKIS) of 2013. A reference for land cover consisting of the five classes defined in Section 3.4 was generated in four test areas (area 0 – area 3) by manual annotation. These annotated areas cover a total of 1.6 km^2 (Figure 1). Due to the temporal acquisition dates of the DOP, the DSM/DTM and the ALKIS data, we observed data inconsistences: some pixels in the DSM did not represent the same object as the corresponding pixels in the orthophoto or in ALKIS. For instance, some building outlines in ALKIS were outdated compared to the DOP, and the DOP also showed some buildings that had not been available at the time of the acquisition of the DSM. Pixels affected by such inconsistences are marked as *invalid* and are not used in any training and evaluation procedure. Figure 1 presents the test data and the four areas with reference data. The reference consists of almost 40 million pixels (excluding *invalid* pixels). In our experiments, 17% of the labelled pixels (about 6.8 million) were used for training. The training areas are marked by black rectangles in Figure 1. The distribution of classes is rather unbalanced in both the training and test sets. For instance, only about 3% of the training pixels belong to *bare soil*, whereas about 30% of the pixels are labelled as *low vegetation*.

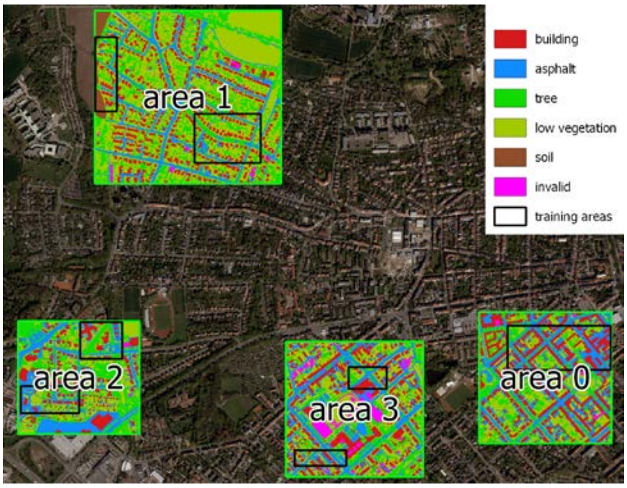

Figure 1. The test data and the labelled areas (areas 0-3). Pixels inside the black rectangles are training samples

To test the runoff simulation based on automatic classification, a hydrological expert defined the subbasins of the sewer system in test area 1. On the basis of ALKIS outlines, individual plots of land that are drained by the same part of the sewer were merged to subbasins. This resulted in a total of 23 subbasins in area 1 with an average size of about 5000 m^2. The study catchment with its subbasins is shown in Figure 2.

In our experiments, we used the samples inside the black rectangles in Figure 1 for training our classifiers (RF and CRF) and the remaining labelled samples for evaluation. For the RF, we used three variants based on different hyper-parameter settings. In all cases, unless noted otherwise, we used the default parameters of Scikit-learn (Pedregosa et al., 2011). In variant RF-STD, we used all the 72 features defined in Section 3.3 for classification. The RF consisted of n_T = 30 decision trees, and we used the default number $n_F = \sqrt{72} = 8$ features for learning the node tests. The minimum number of samples for splitting a node was set to $minS_{split}$ = 2. For the maximum tree depth, we used the default setting of Scikit-learn, d_{max}. = ∞, which implies that this criterion is not used in training. No measures were applied to obtain a balanced class distribution of the training samples. In variant RF-FS20 we only used the 20 most important features according to the feature importance (Section 3.3.1) and the same hyper-parameter setting as in RF-STD, which implies $n_F = \sqrt{20} = 4$. Finally, in variant RF-TND, we tuned the hyper-parameters in a procedure where we only used half of the samples in the black rectangles in Figure 1 for training and the rest for validation. For the RF that delivered the association potentials of the CRF classifier, we used all 72 features and n_T = 200, $n_F = \sqrt{72} = 8$, $minS_{split}$ = 5 and d_{max}. = 25. In this case, 100000 samples per class drawn randomly from all training samples were used for training. The parameters of the contrast-sensitive Potts model were set to $\beta_0 = \beta_1 = 0.5$.

After land cover classification, we determined the coefficient of imperviousness for each subbasin. This resulted in four different variants of the coefficient of imperviousness for the four classification results (COI_RF-STD, COI_RF-FS20, COI_RF-TND, COI_CRF). A reference for the coefficient of imperviousness (COI_REF) was generated by applying the post-processing procedure described in Section 3.4 to the reference labels. In order to assess the reliability of visual inspection by humans, we also asked six (non-expert) individuals to estimate the coefficient of imperviousness based on a visual inspection under professional guidance; this variant is referred to as COI_VIS. For each of these variants, a mean coefficient of imperviousness $\overline{SF}$ was computed by averaging the coefficient of imperviousness over all 23 subbasins.

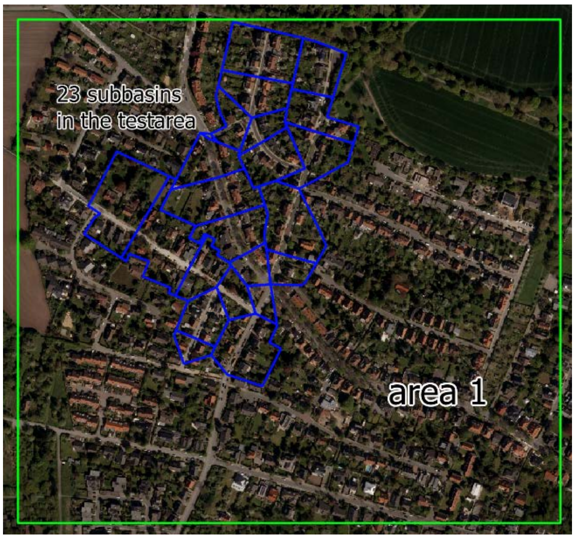

Figure 2. The study catchment with 23 subbasins (blue) in area 1 (green rectangle)

In order to evaluate our results for land cover classification, we compared them to the reference and determined a confusion matrix as well as metrics derived from the confusion matrix, only considering samples outside the training areas. We focus on the *overall accuracy (OA)* i.e. the percentage of pixels that are assigned the correct class label by the classification process, and the *F1* score, i.e. the harmonic means of the completeness and correctness per class; we also report the average F1 score ($\overline{F1}$). In addition, we evaluate the coefficient of imperviousness. For that purpose, we compute the difference between the coefficient of imperviousness obtained by the variants and the reference (COI_REF) for each subbasin. From these differences, we determine the root mean square error (*RMSE*) and the mean difference Δ_{avg} of the 23 subbasins.

4.2 Land cover classification

4.2.1 Feature selection: The 72 features were ranked by feature importance delivered by the RF. Table 2 shows the 20 most important features according to that measure. It is not surprising that the nDSM and NDVI-based features are among the most relevant ones: the nDSM is crucial for differentiating elevated objects from objects on the terrain, whereas the NDVI helps to differentiate vegetation from other objects (e.g. Rottensteiner et al., 2007). We used these 20 features in variant RF-FS20.

R	Feature	*R*	Feature
1	*nDSM*	11	*NDVI: var.*
2	*∇DSM: mean mag.*	12	*Distance transform*
3	*NDVI: mean*	13	*Hue: Mean*
4	*NDVI: Gauss$_2$*	14	*Hue: Gauss$_5$*
5	*∇nDSM: mag.*	15	*Hue*
6	*NDVI*	16	*Blue: Gauss$_2$*
7	*NDVI: Gauss$_5$*	17	*Red: Gauss$_2$*
8	*∇DSM: mag.*	18	*Saturation: Gauss$_5$*
9	*∇nDSM: var. of mag.*	19	*Intensity : Mean*
10	*DSM: mean curvature*	20	*Hue: Gauss$_2$*

Table 2. The 20 most important features. *R*: Rank. ∇: Gradient; *mag.*: magnitude; *var.*: variance. For a definition of the features, cf. Section 3.3.

4.2.2 Tuning the RF hyper-parameters: To find the best values of the hyper-parameters for variant RF-TND, we split the training areas (black rectangles in Figure 1) horizontally into two subset of equal size and used the upper half for training and the lower half for validation. We used different hyper-parameter settings, trained the RF and classified the validation set. Then we compared the results with the reference labels of the validation set and determined the average F1 score ($\overline{F1}$) as our main criterion for evaluating the quality of the results.

We carried out tests using two different values for the number of features to be tested in the RF nodes (all features, i.e. n_F = 72, and the default setting $n_F = \sqrt{72}$ = 8). For n_F = 72, we did not compensate for the unbalanced class distributions in the training data (*bal.* = *no*), whereas for n_F = 8, we additionally investigated the effects of balancing the training samples to compensate for this effect, carrying out two groups of tests (*bal.* = *no* and *bal.* = yes, respectively; Pedregosa et al., 2011). In each of the resultant three groups of tests, we varied the parameters n_T and $minS_{split}$ within certain ranges. Table 3 shows the resultant average F1 scores ($\overline{F1}$) on the validation set.

Comparing the results for the first two groups of experiments in Table 3, both conducted without balancing the training samples, it is obvious that using the default setting for n_F ($n_F = \sqrt{72}$ = 8) outperforms the variant using all features (n_F = 72) for all combinations of n_T and $minS_{split}$. A comparison of the two groups of experiments conducted using n_F = 8 (*bal.* = *no* vs. *bal.* = *yes*; second and third group in Table 3) shows that balancing the training samples has a positive but almost negligible effect on the average F1 scores. In general, increasing the number of trees has a positive effect, independent of the other parameters; using a very large value for $minS_{split}$ (1000) leads to a deterioration of the results. The table shows that in general, the RF is relatively robust against varying the hyper-parameters, but careful tuning can lead to an improvement in the average F1 scores ($\overline{F1}$) of about 2.5%. The best average F1 score of 82.8% was achieved when using n_T = 100 decision trees, balancing the training data (*bal.* = *yes*), $n_F = \sqrt{72}$ = 8 features for learning the node tests and splitting a node if more than $minS_{split}$ = 2 samples reach that node in training. These values are used for variant RF-TND in the evaluation.

	$minS_{split}$					
n_T	2	10	100	1000	n_F	*bal.*
10	80.3	80.3	80.6	80.2	72	*no*
30	80.9	80.9	81.0	80.1		
60	81.0	81.0	81.0	80.1		
100	**81.1**	81.1	81.1	80.0		
10	81.5	81.7	82.0	81.5	8	*no*
30	82.4	82.5	82.4	81.7		
60	82.6	82.6	82.6	81.8		
100	82.7	**82.8**	82.6	81.8		
10	81.8	81.7	81.7	80.2	8	*yes*
30	82.5	82.5	82.3	80.6		
60	82.7	82.7	82.3	80.7		
100	**82.8**	82.7	82.4	80.8		

Table 3. Average F1 score ($\overline{F1}$) in [%] for different combinations of hyper-parameters of the RF classifier. n_T: number of trees; n_F: number of features used in a node test; $minS_{split}$: minimum number of samples in a tree node for splitting; *bal*: balancing the training set or not.

4.2.3 Evaluation: Table 4 presents the evaluation of the land cover classification results of all variants of RF and the CRF. In general, all RF variants and CRF delivers very similar results (all about 85% OA). Besides, all classifiers have difficulties in classifying the class *bare soil*, for which only very few (training and test) samples are available. Examples for some classification results are shown in Figure 3. In the subsequent paragraphs, we analyse the variants in more detail.

	F1 [%]					$\overline{F1}$	OA
Classifier	*bld.*	*as.*	*tr.*	*gr.*	*bs.*	[%]	[%]
RF-STD	91.5	82.3	85.8	84.5	60.3	80.9	85.4
RF-FS20	91.6	81.8	85.7	84.2	60.0	80.7	85.1
RF-TND	**91.7**	**82.6**	**86.2**	**85.0**	**60.4**	**81.2**	**85.8**
CRF	91.5	**82.6**	85.8	84.2	59.9	80.8	85.3

Table 4. Results of land cover classification for different variants of RF (cf. Section 4.1) and the CRF. Best scores are printed in bold font.

Random Forest: The standard RF variant (RF-STD) delivers promising results with an OA of 85.4% and an average F1 score of 80.9%. The RF with fine-tuned hyper-parameters (RF-TND) delivers the best results, not only in terms of the OA (85.8%) and average F1 score (81.2%), but also in all class-specific F1 scores. The biggest improvement in the class-specific F1 score occurs with *low vegetation* (0.5% compared to RF-STD). Using only the 20 most important features (variant RF-FS20) delivers a slightly worse result, with a decrease of 0.3% in OA and 0.2% in the average F1 score compared to the RF-STD. The biggest class-specific drop in F1 occurs with *asphalt* (0.5%). However, this small drop in performance is contrasted by a large reduction in computation time. For variant RF-FS20, training and testing only require 40% and 70% of the time needed by variant RF-STD, respectively.

Conditional Random Fields: in the CRF, we used all the 72 features for training and testing, and we did not apply any hyper-parameters tuning. Due to the properties of the model used for the interaction potential, the CRF delivers a smoother classification result, especially at the object boundaries (cf. Figure 3). Table 4 indicates that the result of CRF is very close to the one of RF-STD, but slightly worse than the one of RF-TND (0.5% in OA). This could be caused by the fact there was no tuning of the hyper-parameters of the RF used in the CRF. The CRF achieves the best F1 score for class *asphalt*. However, introducing the interaction potentials causes the training and inference time to be increased by factors of 3 and 20, respectively, compared to RF-STD.

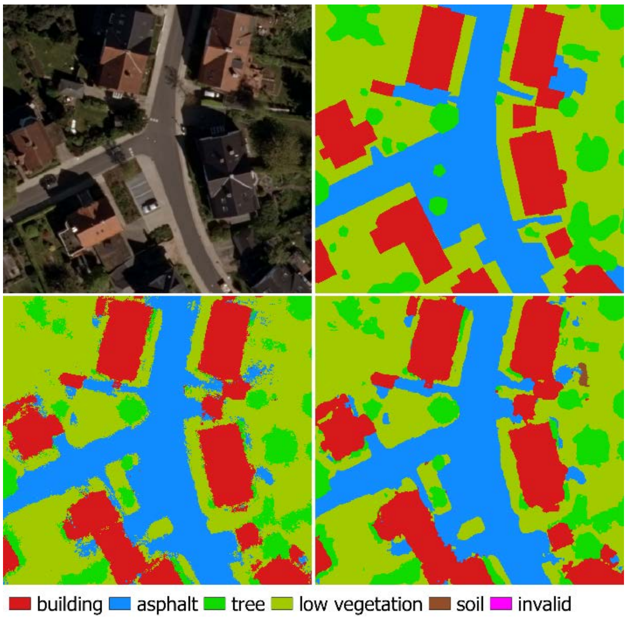

Figure 3: Classification results. Top left: input image; top right: reference labels; bottom left: results of RF-STD; bottom right: results of CRF.

4.3 Coefficient of imperviousness

Table 5 shows the average coefficient of imperviousness and the average difference to the reference as well as the RSME for all variants. Of the automatic methods, COI_RF-TND performs best, showing a mean difference Δ_{avg} of -0.6% and a RMSE of 3.8%. The coefficient of imperviousness computed from the results of the other RF variants are very promising as well (-0.9% and -1.1%, respectively). Interestingly (and somewhat surprisingly), COI_CRF delivers the worst results; compared to COI_RF-TND, the mean difference is larger by 1.4% and the RMSE by 0.6%. However, these numbers have to be interpreted with caution. As the reference was generated by using the same heuristics as the ones used to deal with class *tree*, it may not be perfect.

Variant	$\overline{SF}$ [%]	Δ_{avg} [%]	$RMSE$ [%]
COI_RF-STD	42.1	-0.9	**3.8**
COI_RF-FS20	42.3	-1.1	**3.8**
COI_RF-TND	41.8	**-0.6**	**3.8**
COI_CRF	43.2	-2.0	4.4
COI_VIS	51.5	-10.3	12.2
COI_REF	41.2	--.--	--.--

Table 5. Results of coefficient of imperviousness computed in different variants. All values are computed based on the 23 subbasins. Best results are print in bold font

It is interesting to see that the visual determination of the coefficient of imperviousness by humans (COI_VIS) shows both, the largest mean difference to the reference (-10.3%) and the largest RMSE (12.2%). It would seem that humans tend to over-estimate the coefficient of imperviousness by a considerable margin. It is also noteworthy that there exists a large difference in the coefficient of imperviousness determined by the six individuals (standard deviation of 9.3% with respect to the mean). It would seem that the visual interpretation of non-expert humans is subjective and may be not reproducible. Our results indicate that the coefficient of imperviousness derived by an automatic procedure based on classification can be more accurate and reproducible.

5. CONCLUSION

In this paper, we proposed and evaluated a methodology for determining the coefficient of imperviousness, which is a major parameter for urban drainage models, based on the supervised classification of aerial imagery and height data. We compared three variants of a RF classifier and a CRF. The results of land cover classification show no clear advantage of either classifier, both achieving an overall accuracy of about 85.5%. To determine the coefficient of imperviousness, the classification results had to be corrected to compensate for the occlusion of the ground surface by trees. This was achieved by a heuristic method taking into account information from a GIS. The best result for the coefficient of imperviousness is achieved on the basis of the classification results of the best RF classifier (RF-TND) with a mean difference of -0.6% and a root mean square error of 3.8% compared to the reference. This is considerably better than the results obtained by visual interpretation by six non-expert humans under a professional guidance. Visual interpretation did not only result in a high variance of the results between the individuals, but also in a considerably larger error compared to the reference than the automated methods. This indicates that the automated derivation of the coefficient of imperviousness, apart from being more objective and more reproducible, delivers more accurate results than the manual estimation.

In the future, we want to investigate how the determination of the coefficient of imperviousness can benefit from using Convolutional Neural Networks (CNN, LeCun et al., 1998), which have been shown to outperform classification methods based on hand-crafted features (Zhu et al, 2017).

ACKNOWLEDGEMENTS

The authors thank SWO-Netz GmbH and Stadt Osnabrück, Fachdiesnt Geodaten for their support. Furthermore, we would like to thank Sara de Toffel for useful comments and support.

In addition, we would like to thank all participants of the project seminar 2017/2018 for their help in this project, and in particular Mareike Dorozynski, Najib Mahfuzh and Victoria Kröger.

Chun Yang is an associate member of the Research Training Group i.c.sens (GRK 2159), funded by the German Research Foundadtion (DFG).

REFERENCES

Albert, L., Rottensteiner, F., Heipke, C., 2017: A higher order conditional random field model for simultaneous classification of land cover and land use. ISPRS Journal of Photogrammetry and Remote Sensing 130, pp. 63-80.

Breiman, L., 2001: Random Forests. Journal of Machine Learning 45(1), pp. 5-32

Cortes, C., Vapnik, V., 1995: Support-vector networks. Machine Learning 20(3), pp. 273-297

Criminisi A., Shotton, J., 2013: Decision Forests for computer vision and medical image analysis. Advances in Computer Vision and Pattern Recognition, pp. 25-45

Deguchi, C., Sugio, S., 1994: Estimations for percentage of impervious area by the use of satellite remote sensing imagery. Water Science and Technology 29, pp. 135-144

Fankhauser, R., 1999: Automatic determination of imperviousness in urban areas from digital orthophotos. Water Science and Technology 39(9), pp. 81-86

Frey, B. and MacKay, D., 1998: A revolution: Belief propagation in graphs with cycles. Advances in Neural Information Processing Systems, pp. 479-485

Gislason, P. O., Benediktssoon, J. A., Sveinson, J. R., 2006: Random forests for land cover classification. Pattern Recognition Letter 27(4), pp. 294-300

Haralick, R. M., Shanmugan, K., Dinstein, I., 1973: Texture features for image classification. IEEE Transactions on Systems, Man and Cybernetics SMC-3(6), pp. 610-621

Hartcher, M. G., Chowdhury, R. K., 2017: An alternative method for estimating total impervious area in catchments using high-resolution color aerial photography. Water Practice and Technology 12, pp. 478-486

Hermosilla, T., Ruiz, L. A., Recio, J. A., Cambra-López, M., 2012. Assessing contextual descriptive features for plot-based classification of urban areas. Landscape and Urban Planning 106(1), pp. 124-137.

Kite, G. W., Kouwen, N., 1992: watershed modeling using land classifications. Journal of Water Resources Research 28(12), pp. 3193-3200.

Kite, G. W., Pietroniro, A., 1996: Remote sensing applications in hydrological modelling. Journal of Hydrological Sciences 41(4), pp. 563-591.

Lillesand, T. M., Kiefer, R. W., Chipman, J. W., 2003: Remote sensing and image interpretation. Hoboken, NJ, USA: Wiley.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE 86(11): 2278–2324.

Maas, A., Rottensteiner, F., Heipke, C., 2016: using label noise robust logistic regression for automated updating of topographic geospatial databases. In: ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences III-7, pp. 133-140.

Kumar, S., Herbert, M., 2006: Discriminative Random Fields. International Journal of Computer Vision 68(2), pp. 179-201.

Lee., J. G., Heaney, J. P., Asce, M., 2004: Estimation of urban imperviousness and its impacts on storm water systems. Journal of Water Resources Planning and Management 129(5), pp. 419-426.

OpenCV, 2016: Random Trees. OpenCV Reference Manual: https://docs.opencv.org/3.0-beta/modules/ml/doc/random_trees.html (accessed 22 March 2018).

Parfit, 2017: A package for parallelizing the fit and flexibly scoring of sklearn machine learning models, with optional plotting routines: https://github.com/jmcarpenter2/parfit (accessed 22 March 2018).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011: Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12, pp. 2825-2830.

Rottensteiner, F., Trinder, J., Clode, S., Kubik, K., 2007: Building detection by fusion of airborne laserscanner data and multi-spectral images: Performance evaluation and sensitivity analysis. ISPRS Journal for Photogrammetry and Remote Sensing 62(2), pp. 135-149.

Schindler, K., 2012: An overview and comparison of smooth labeling methods for land cover classification. IEEE Transactions on Geoscience and Remote Sensing 50(11), pp. 4534-4545

Scikit-learn, 2018: Forests of Random Trees. Scikit-learn Reference Manual, http://scikit-learn.org/stable/modules/ensemble.html#forest (accessed 22 March 2018).

Schmugge, T. J., Kustas, W. P., Ritchie, J. C., Jackson, J., Rango, A., 2002: Remote sensing in hydrology. Advances in Water Resources 25, pp. 1367-1385.

Tokarczyk, P., Leitao, J. P., Rieckermann, J., Schindler, K., Blumensaat, F., 2015: High-quality observation of surface imperviousness for urban runoff modeling using UAV imagery. Hydrology and Earth System Sciences 19, pp. 4215-4228.

Tyrna, B. G., Hochschild, V., 2010: Urban flash flood modelling based on soil sealing information derived from high resolution satellite data. HydroPredict Conference 2010, Prague, The Czech Republic.

Zhu, X.X., Tuia, D., Mou, L., Xia, G.S., Zhang, L., Xu, F., Fraundorfer, F., 2017. Deep learning in remote sensing: a comprehensive review and list of resources. IEEE Geoscience and Remote Sensing Magazine 5(4), pp. 8-36

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLII-3-1355-2018