ISPRS Journal of Photogrammetry and Remote Sensing 130 (2017) 63–80

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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

A higher order conditional random field model for simultaneous classification of land cover and land use

Lena Albert, Franz Rottensteiner *, Christian Heipke

Institute of Photogrammetry and GeoInformation, Leibniz Universität Hannover, Nienburger Str. 1, D-30167 Hannover, Germany

ARTICLE INFO

Article history:
Received 10 August 2016
Received in revised form 2 April 2017
Accepted 11 April 2017
Available online 31 May 2017

Keywords:
Contextual classification
Land cover
Land use
Conditional random fields
Higher order potential
Aerial imagery

ABSTRACT

We propose a new approach for the simultaneous classification of land cover and land use considering spatial as well as semantic context. We apply a Conditional Random Fields (CRF) consisting of a land cover and a land use layer. In the land cover layer of the CRF, the nodes represent super-pixels; in the land use layer, the nodes correspond to objects from a geospatial database. Intra-layer edges of the CRF model spatial dependencies between neighbouring image sites. All spatially overlapping sites in both layers are connected by inter-layer edges, which leads to higher order cliques modelling the semantic relation between all land cover and land use sites in the clique. A generic formulation of the higher order potential is proposed. In order to enable efficient inference in the two-layer higher order CRF, we propose an iterative inference procedure in which the two classification tasks mutually influence each other. We integrate contextual relations between land cover and land use in the classification process by using contextual features describing the complex dependencies of all nodes in a higher order clique. These features are incorporated in a discriminative classifier, which approximates the higher order potentials during the inference procedure. The approach is designed for input data based on aerial images. Experiments are carried out on two test sites to evaluate the performance of the proposed method. The experiments show that the classification results are improved compared to the results of a non-contextual classifier. For land cover classification, the result is much more homogeneous and the delineation of land cover segments is improved. For the land use classification, an improvement is mainly achieved for land use objects showing non-typical characteristics or similarities to other land use classes. Furthermore, we have shown that the size of the super-pixels has an influence on the level of detail of the classification result, but also on the degree of smoothing induced by the segmentation method, which is especially beneficial for land cover classes covering large, homogeneous areas.

1. Introduction

1.1. Motivation and goals

Geospatial land use databases contain important information with high benefit for several users, especially in the field of urban management and planning. The number of possible applications of such data increases with a higher level of detail, both in terms of the size of the geometrical entities as well as the diversity of land use classes. Because of fast changes of the land use due to urban growth and land use conversion, such geospatial databases become outdated quickly. The update process requires enormous efforts (Champion, 2007), so that the automation of this task on the basis of remote sensing data is desirable.

In contrast to *land cover*, which describes the physical material of the earth's surface (e.g. *grass*, *asphalt*), *land use* reveals the socio-economic function of a piece of land (e.g. *residential*, *agricultural*). Land use objects are typically composed of different land cover elements, which may be arranged in complex structures. On the other hand, land cover of a specific type can be a part of different land use objects. Thus, land cover and land use classification based on remote sensing data are tasks that pursue different objectives (Barnsley and Barr, 2000). Whereas land cover classification focuses on the assignment of class labels to (frequently small) image sites, the goal of land use classification is to assign a land use label to larger spatial entities which form a functional unit and which are typically represented by polygonal objects in a land use database. Land cover can be derived directly from spectral

* Corresponding author.
E-mail addresses: albert@ipi.uni-hannover.de (L. Albert), rottensteiner@ipi.uni-hannover.de (F. Rottensteiner), heipke@ipi.uni-hannover.de (C. Heipke).

http://dx.doi.org/10.1016/j.isprsjprs.2017.04.006
0924-2716/

characteristics of remote sensing data, but this is more difficult for land use; the classification of land use additionally requires information about the composition and arrangement of different land cover elements within a land use object. As a consequence, a procedure consisting of a sequence of two classification steps is frequently applied to obtain land use information from remote sensing imagery, e.g. (Albert et al., 2014a; Hermosilla et al., 2012). In a first step, land cover information is derived by a classification of remote sensing data. The second step consists of a land use classification, often based on segments or objects from a geospatial database, in which some of the features are derived from the results of the first step. It is the main drawback of this approach that wrong decisions taken during land cover classification cannot be reversed at later stages, which can easily lead to misclassifications of land use.

An important source of information to be tapped for land cover and land use classification is *context*. First, there are *spatial* dependencies between neigbouring sites. Both for land cover and for land use, certain classes are more likely to occur next to each other than others. For instance, a *residential* land use object is usually located next to a land use object of the class *street*. On the other hand, neighbouring land cover sites are typically small and, thus, likely to belong to the same class; to give an example, in a forest, a land cover site belonging to the class *tree* is usually surrounded by other trees rather than buildings or sealed area. Second, there is a mutual *semantic* dependency between land cover and land use. The classification of land use requires information about land use, but the knowledge about the current land use facilitates the determination of land cover, because, for instance, the land cover *tree* is more likely to occur inside a land use object of class *forest* than inside a *traffic* object. This additional hint becomes particularly beneficial for high-resolution aerial image data, where the appearance of many land cover classes is heterogeneous, making land cover classification more challenging.

In this paper, we present an approach for the simultaneous classification of land cover and land use that considers semantic as well as spatial context. Land cover classification is carried out at the level of super-pixels (Achanta et al., 2012), whereas land use is determined at the level of objects from a geospatial database, represented by polygons that are assumed to be correct. The rationale for this assumption, which also corresponds to the application scenario in our experiments, is that in regions such as Central Europe, the object boundaries in geospatial data bases are usually related to property boundaries which are kept up-to-date by governmental surveying authorities, whereas property owners are not obliged to report changes in land use, so that the update of the semantic information contained in such a database lags behind. Thus, our approach can be seen as a contribution to the automatic derivation of this semantic information from aerial imagery. If higher semantic accuracy is required, it can be the first step of a semi-automatic scheme for database updating, in which land use objects for which the classification result is incompatible with the information contained in the database are highlighted to a human operator for visual inspection, e.g. (Helmholz et al., 2014).

The method presented in this paper is supervised, i.e. it requires training data for both land cover and land use objects. The statistical framework for contextual classification is given by Conditional Random Fields (CRF), which in the standard formulation require unary potentials, typically the output of a local classifier, and pairwise potentials for the context model (Kumar and Hebert, 2006). Our method requires aerial images and derived products such as digital surface models (DSM), as well as the polygons from a geospatial database as input. Whereas the standard CRF framework is applied for modelling spatial context, the dependencies between the land cover and the land use layers of the CRF require the formulation of a high-order potential and a specific inference method for obtaining the optimal configuration of class labels in both layers. The scientific contribution of our method can be summarized as follows:

- *Simultaneous contextual classification of land cover and land use:* Unlike existing two-step approaches (Hermosilla et al., 2012; Walde et al., 2014; Albert et al., 2014a), where the semantic dependency between land cover and land use is only considered in one direction (i.e. from land cover to land use), our method integrates land cover and land use classification in a unified approach, allowing both tasks to interact in the classification procedure. This avoids early decisions and is expected to improve the classification accuracy, in particular for the land cover layer.
- *CRF with higher order potentials:* Standard CRF-based models only consider interactions between pairs of objects to be classified (Kumar and Hebert, 2006). However, such a model is not appropriate in our case, where one (large) land use object may interact with many (small) land cover objects. Thus, we propose a new type of higher order potential in our CRF-based approach which we expect to provide a better model of the complex statistical dependencies between land cover and land use and which, unlike other models for higher order potentials (Kohli et al., 2009), can also deal with different class structures.
- *Inference for CRF with higher order potentials:* Unlike existing techniques placing restrictions on the formulation of these potentials (Kohli et al., 2009), our method can cope with a generic model for the complex interactions between land cover and land use. Inference is carried out in an iterative way. In each iteration, we determine a partial solution for one of the two layers (land cover, land use), keeping current optimal state in the other layer fixed. As a consequence, the higher order potential is simplified to a simple unary potential based on the output of a local classifier, allowing the application of standard inference techniques for CRF. This procedure is inspired by Roig et al. (2011), but we use a different and more complex model for the higher order potential.
- *Semantic context features:* The local classifier used to represent the higher order potential in the approximate inference technique is based on a comprehensive set of contextual features that expresses the local arrangement of land cover primitives in a land use object and vice versa. These features are computed on the basis of the partial solution that is an intermediate result in the inference procedure, considering the current beliefs for all classes as weights.

The remainder of this paper is structured as follows. We start with a review of related work in Section 1.2. Our approach for land cover and land use classification is presented in Section 2, whereas Section 3 describes the experimental evaluation of our approach based on two test sites in Germany. Conclusions and an outlook are given in Section 4.

1.2. Related work

We start this review of related work by a discussion of papers that deal with different strategies for land use classification from high-resolution remote sensing data, focussing on the overall strategy and the way in which land cover is integrated into the process. In the second part of this review, we will discuss methodological aspects related to context-based classification, the use of multi-layer CRF, and CRF with higher order potentials.

High-resolution satellite or aerial images are the data source most frequently used for deriving land use information (Walde et al., 2014; Hermosilla et al., 2012; Albert et al., 2014a). The definition of the entities to be classified depends on the application.

For the reasons given in Section 1.1, we are interested in methods classifying objects whose boundaries are derived from a geospatial database. These objects can correspond to cartographic units, such as urban blocks from a digital landscape model (Walde et al., 2014), or to administrative units, such as cadastral plots (Hermosilla et al., 2012). In such a context, the class structures to be discerned depend on the object catalogue of the geospatial database. These object catalogues can have a considerable depth, but not all of these classes can be distinguished given remote sensing data. For the type of database also considered in this paper, Albert et al. (2016) investigated which classes can be discerned using high-resolution aerial imagery. That land use database forms a part of the German Authoritative Real Estate Cadastre Information System (ALKIS®). ALKIS® contains about 190 different land use classes that are hierarchically structured in several semantic levels; the object catalogue is described in AdV, 2008). The experiments showed that only 9 classes corresponding to object types at the two coarsest semantic level could be differentiated, the main problems being related to a lack of training data for some classes and to the similar appearance of some classes in the data. The land use classes discerned in this paper are selected on the basis of (Albert et al., 2016).

Most of the land use classification approaches apply a two-step processing strategy, e.g. (Hermosilla et al., 2012; Helmholz et al., 2014; Albert et al., 2014a). First, a pixel- or segment-based land cover classification is performed. In a second step, the classification results are transferred to the land use sites. For this purpose, the land cover elements within a land use site are analyzed with regard to their spatial composition and arrangement in order to determine land use. These characteristics are described by contextual features, which, together with spectral and other features are applied to determine land use in a classification process. Such contextual features can generally be divided into two groups: *spatial metrics* and *graph-based measures*. Spatial metrics quantify the spatial configuration and composition of the land cover elements within a land use unit. These features describe the size and the shape of the land cover segments, e.g. the proportion of building pixels (Hermosilla et al., 2012), and their spatial configuration, e.g. the position of building segments in relation to the boundaries of the land use object or other building segments (Novack and Stilla, 2015). The second group of features are adopted from graph theory (Barnsley and Barr, 1996). These features describe the frequency of the local spatial arrangement of land cover elements within a land use object. Barnsley and Barr (1996) calculate an *adjacency-event-matrix* for each land use object based on pixel-based land cover information. From this matrix, descriptive features are derived, e.g. the normalized number of edges between certain land cover classes. Walde et al. (2014) adapt the adjacency-event-matrix so that it can be determined based on land cover segments rather than pixels. These features are used in the context of two-stage procedures, considering the land cover classes to be known and, thus, constant. In contrast, we consider these class labels to be uncertain and, thus, use a slightly different definition of some of these features, taking into account the current beliefs as weights for the currently optimal class labels in the iterative optimization process.

Land use classification can be based on heuristic rules that require the definition of parameters such as thresholds (Banzhaf and Höfer, 2008), or it can be carried out in a supervised way, using classifiers such as Random Forests (Walde et al., 2014; Albert et al., 2014a) or Support Vector Machines (SVM) (Montanges et al., 2015). Some approaches also consider spatial context features in order to model spatial dependencies between neighbouring land use objects in the classification process. An example is given by Hermosilla et al. (2012), who adapt graph-based measures for the classification of land use objects, e.g. to encode the number of class correspondences with neighbouring entities. In contrast to such an implicit method for considering context, CRF allow to model context explicitly in a statistical framework. Albert et al. (2014a) presented a two-step land use classification approach using CRF, which was extended to include an iterative inference procedure in (Albert et al., 2015). Two separate CRFs are applied for land cover and land use classification. Both CRFs model spatial dependencies between neighbouring sites, i.e. pixels (Albert et al., 2014a) or super-pixels (Albert et al., 2015) in the case of land cover and segments in the case of land use classification. Spatial dependencies between neighbouring land use objects have also been modelled explicitly in graphical models by Novack and Stilla (2015) and Montanges et al. (2015). Novack and Stilla (2015) use high-resolution space borne synthetic aperture radar (SAR) data for the classification of urban land use. In order to get initial land use labels, they apply three different classifiers (RF, Logistic Regression, Nearest Neighbors) and aggregate class membership values per land use object. CRF are applied in post-processing in order to obtain a smoothed result. Montanges et al. (2015) propose an approach for the classification of urban land use in cells of a 200 m grid based on satellite images, digital surface models and additional thematic information. In their approach, two probabilistic graphical models (Markov Random Fields (MRF), CRF) are applied separately to achieve a spatial smoothing of the results. Both graphical models combine the output of a SVM based on features derived from land cover and structural information with a model of the spatial relations of neighbouring land use cells. In both approaches, the model of the spatial dependencies is based on a smoothness assumption, which is not justified for land use classification. The larger the size of the land use objects, the more likely the classes of neighbouring land use objects will be different. To the best of our knowledge, no land use classification approach exists which, on the one hand, considers a supervised classifier for modelling the spatial dependencies, and, on the other hand, models the complex semantic relationship between land cover and land use explicitly in a graphical model.

Multi-layer CRF have been applied to several tasks in image analysis, e.g. hierarchical classification of building façades (Yang and Förstner, 2011), classification of scenes with occlusions (Kosov et al., 2013) and multi-temporal classification of remote sensing data (Hoberg et al., 2015). These methods are based on pairwise CRF models, i.e., models in which the layers are also linked only by edges between pairs of nodes from different layers. Furthermore, most of these multi-layer CRF approaches aim to distinguish an identical class structure in each layer, with the exception of the multi-scale approach in (Hoberg et al., 2015) where the class structure complies with a part-based object model, combining object parts at finer scale and compound objects at coarser scale. However, in our scenario there is no such hierarchical relation between the class structures at different layers. Albert et al. (2014b) proposed a two-layer CRF for the classification of land cover and land use, where the statistical dependencies between land cover and land use are modelled explicitly by inter-layer edges. This approach also makes use of pair-wise potentials to encode co-occurrence or smoothness constraints. However, complex dependencies between more than two variables, like the configuration of several land cover segments within a land use object, cannot be modelled appropriately in this way. Whereas the number of land cover objects of a given type inside a land use object will have an impact on the result in such a scenario, this is not the case for the spatial arrangement of these objects which also may contain very useful information.

By defining a higher order potential, it is possible to model complex dependencies between more than two random variables explicitly. For instance, Kohli et al. (2009) have presented a class of higher order potentials referred to as P^N-Potts model. This model

has been applied for the extraction of road networks (Wegner et al., 2013) and buildings (Montoya-Zegarra et al., 2015) from aerial images. Both approaches apply higher order CRF to encode prior knowledge about objects (roads, buildings). The generation and selection of relevant cliques as object hypotheses is realized in a data-driven sampling strategy following certain shape constraints.

In general, inference on higher order potentials is challenging. For the P^N-Potts model, an efficient solution based on graph cuts can be obtained, but it places restrictions on the models that may be used for the spatial interaction potential. Furthermore, no class labels are derived for the segments that form a higher-order clique. Standard inference algorithms can effectively approximate a solution for potentials involving only a limited number of variables if generic models are used for these potentials. This is problematic in our case, where the specific structure of the graphical model leads to a generic formulation of higher order potentials involving a large number of variables.

Roig et al. (2011) propose an iterative method for efficient inference in higher order CRF. In each iteration, they determine a partial solution by minimizing an approximated energy function based on the original formulation of the joint posterior for all class labels. The approximation is achieved by simplifying the higher order potentials to unary potentials. As a consequence, the partial solution can be determined using standard inference algorithms. In each iteration, the higher order terms are updated based on the previous partial solution. However, their higher order potentials model quite simple dependencies. Designed for the simultaneous classification of objects in different views of a scene, they are used to consider occlusions of objects within one view as well as the consistency of the classification result amongst different views.

Multi-stage inference procedures have also been proposed by Munoz et al. (2010) for 2D scene analysis based on image data and by Xiong et al. (2011) for 3D scene analysis based on terrestrial point clouds. Both approaches rely on hierarchical segmentations of image or point cloud data, where each segmentation result forms one level in the hierarchy. In contrast to Roig et al. (2011), their methods do not rely on graphical models to capture contextual relations. Instead, they model contextual dependencies between and within the hierarchies by using contextual features in a sequence of classifiers. Whereas Munoz et al. (2010) proceed only down the hierarchy and stop at the bottom level, the inference procedure by Xiong et al. (2011) is also designed for reversed and iterative processing. By using contextual features, the authors circumvent the difficulties associated with modelling complex dependencies by higher order potentials. Compared to a standard inference algorithm in CRF, all steps in the inference procedure, i.e. all classifiers, have to be trained beforehand. Furthermore, the selection of adequate context features requires a certain degree of knowledge about the characteristics of the contextual relations.

2. Methodology

2.1. Prerequisites

The classification approach presented in this paper is designed for high-resolution, multispectral image data and height information. A geospatial land use database is needed for defining the geometric outlines of the land use objects in the form of polygons. Furthermore, training data are required for both land cover and land use. For land use classification, the training data consist of database objects with known land use labels. The objects to be used for training need to be inspected by a human operator before processing the data in order to make sure that their class labels are correct. The training data for land cover consist of fully labelled image subsets, usually generated by manual annotation. The key component of our methodology is the CRF-framework, which can be adapted to various sensor data by extracting different kinds of features. However, the use of height information is recommended because it has been found to contribute important information to both classification tasks.

2.2. Conditional random fields

Conditional Random Fields were introduced for image classification by Kumar and Hebert (2006). CRF are undirected graphical models, consisting of nodes n and edges e. The nodes represent the image sites, e.g. pixels or segments. Nodes are connected by edges, which model statistical dependencies between the class labels and data at the associated nodes. The class labels of all image sites are combined in a label vector $\mathbf{y} = [y_1, \ldots, y_i, \ldots, y_n]$, where $i \in S$ is the index of an image site and S is the set of all image sites. The goal is to assign the most probable class label configuration $\mathbf{y}$ from a set of classes $L = [l_1, \ldots, l_m]$ to all image sites simultaneously considering the data $\mathbf{x}$. CRF are discriminative classifiers, thus, directly modelling the posterior probability $P(\mathbf{y}|\mathbf{x})$ of the label vector $\mathbf{y}$ given the data $\mathbf{x}$:

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})}\left(\prod_{i\in S}\phi(y_i,\mathbf{x}) \cdot \prod_{h\in H}\psi_h(\mathbf{y}_h,\mathbf{x})^{\omega^h}\right). \quad (1)$$

In Eq. (1), $\phi(y_i,\mathbf{x})$ is called the *association potential* and $\psi_h(\mathbf{y}_h,\mathbf{x})$ represents the *clique potential*. The association potential $\phi(y_i,\mathbf{x})$ models the relation between the class label y_i at site i and the observations $\mathbf{x}$. The clique potential $\psi_h(\mathbf{y}_h,\mathbf{x})$ models the relations between the labels y_j of all nodes n_j belonging to a clique $h \in H$ (i.e. the set of class labels $\mathbf{y}_h$) considering the observations $\mathbf{x}$, where h refers to the index of a clique and H is the set of all cliques. Possible cliques can be composed of two (pairwise cliques) or more nodes (higher order cliques). The partition function $Z(\mathbf{x})$ acts as a normalization constant. The parameter ω^h determines the weight of the clique potential relative to the association potential, and, thus, defines the influence of the clique potential in the classification process. CRF represent a general framework, which allows to introduce various functional models for both types of potentials (Kumar and Hebert, 2006). One can choose arbitrary discriminative classifiers with a probabilistic output $P(y_i|\mathbf{x})$ for the association potential. For the clique potential, the definition of a functional model depends on the order of the clique, which imposes restrictions on efficient inference. For pairwise potentials, one can also apply any discriminative classifier with a probabilistic output. In higher order CRF, the corresponding higher order potentials may be limited to specific formulations of functional models (Kohli et al., 2009) in order to allow efficient inference.

2.3. Two-layer graphical model

2.3.1. Graph structure

In order to realize a simultaneous classification of land cover and land use, where both classification tasks mutually support each other, we design a graphical model consisting of two layers. The layers correspond to hierarchical levels and are arranged one above the other, connected by inter-layer edges. We distinguish a *land cover layer* and a *land use layer*. Each layer consists of nodes and intra-layer edges. Fig. 1 illustrates the design of the two-layer graphical model. We want to determine the class labels for land cover y_i^c and land use y_k^u in the corresponding layers. The superscript indicates whether the variable belongs to the land cover (c) or land use (u) layer.

Both layers differ with respect to the image entities represented by the nodes and the employed features. Moreover, the class structures to be distinguished are different in both layers, i.e.

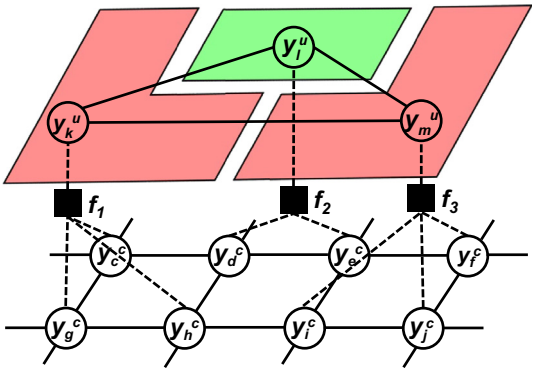

Fig. 1. Graphical model consisting of two layers: land cover (c) and land use (u) layer. Nodes are depicted as circles, intra-layer edges as solid lines. Inter-layer edges connect all spatially overlapping image sites in both layers, thus, forming higher order cliques indicated by the factor nodes f_1, f_2 and f_3 (black squares).

$y_i^c \in L^c = [l_1^c, \ldots, l_m^c]$ for the land cover layer and $y_k^u \in L^u = [l_1^u, \ldots, l_o^u]$ for the land use layer, where m and o denote the number of classes in the land cover and land use layers, respectively. In the land cover layer, the nodes correspond to super-pixels extracted from the image data, whereas in the land use layer the nodes correspond to GIS-objects from a geospatial database. The geometry of the image entities remains unchanged during the inference procedure. Examples for the shape of both types of sites are shown in Fig. 2. We use SLIC (Simple Linear Iterative Clustering) super-pixels (Achanta et al., 2012), whose size and compactness can be controlled by parameters in order to enable a certain adaptation to image boundaries in heterogeneous areas. In homogeneous areas, SLIC super-pixels tend to have a compact shape. Fig. 3 shows the shape of the extracted super-pixels in a heterogeneous urban environment and in a more homogeneous rural scene.

The intra-layer edges model the spatial neighbourhood of each node in the respective layer. The neighbourhood of a node n_i is composed of its first-order spatial neighbours, i.e. all sites that share a common boundary with the site represented by node n_i. Inter-level context, i.e. the statistical dependencies of land cover and land use, are modelled via inter-layer edges. These edges connect all spatially overlapping image sites in both layers, thus, forming a higher order clique indicated by the factor nodes in Fig. 1.

We apply CRF according to Eq. (2) for land cover and land use classification arranged in a two-layer graphical model:

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})}\Big(\prod_{i\in S^c}\phi^c\left(y_i^c,\mathbf{x}\right)^{\omega^1}\cdot\prod_{i\in S^c}\prod_{j\in N_i^c}\psi^c\left(y_i^c,y_j^c,\mathbf{x}\right)^{\omega^2} \cdot\prod_{k\in S^u}\phi^u\left(y_k^u,\mathbf{x}\right)^{\omega^3}\cdot\prod_{k\in S^u}\prod_{l\in N_k^u}\psi^u\left(y_k^u,y_l^u,\mathbf{x}\right)^{\omega^4} \cdot\prod_{h\in H}\psi^{cu}\left(\mathbf{y}_h^c,\mathbf{y}_h^u\right)^{\omega^5}\Big). \quad (2)$$

A superscript is added to the variables to indicate whether the variable refers to the land cover (c) or land use (u) classification. The label configurations in both layers are combined in the label vector $\mathbf{y} = \{\mathbf{y}^c, \mathbf{y}^u\}$. The *association potentials* $\phi^c(y_i^c,\mathbf{x})$ and $\phi^u(y_k^u,\mathbf{x})$ model the relations between class labels y_i^c, y_k^u and the data $\mathbf{x}$, where $i \in S^c$ and $k \in S^u$ are the indices of the image sites and S^c and S^u are the sets of all image sites in the land cover and land use layer, respectively. $\psi^c(y_i^c,y_j^c,\mathbf{x})$ and $\psi^u(y_k^u,y_l^u,\mathbf{x})$ represent the *pairwise intra-layer interaction potentials*, which model the spatial dependencies between neighbouring sites within each layer in consideration of the data $\mathbf{x}$. N_i^c and N_k^u define the neighbourhood of the image sites in both layers. $\psi^{cu}(\mathbf{y}_h^c,\mathbf{y}_h^u)$ describes the *inter-layer higher-order potential*, which models the relations between the labels y_j of all nodes n_j belonging to a clique $h \in H$ (i.e. the set of class labels $\mathbf{y}_h^c$ in the land cover layer and $\mathbf{y}_h^u$ in the land use layer), where h refers to the index of a clique and H is the set of all cliques. The parameters $\boldsymbol{\Omega} = (\omega^1,\omega^2,\omega^3,\omega^4,\omega^5)$ determine the impact of each potential relative to the first potential term (i.e. $\omega^1 \equiv 1$).

2.3.2. Association potential

The association potential predicts how likely node n_i is to belong to a class y_i given the data $\mathbf{x}$. The data are taken into account in the form of site-wise feature vectors $\mathbf{f}_i^c(\mathbf{x})$ and $\mathbf{f}_k^u(\mathbf{x})$ for the nodes in the land cover and land use layer, respectively. The site-wise feature vectors contain image-based and geometrical features. Both association potentials take values proportional to the probability of y_i^c and y_k^u given the site-wise feature vectors $\mathbf{f}_i^c(\mathbf{x})$ and $\mathbf{f}_k^u(\mathbf{x})$, i.e. $\phi^c(y_i^c,\mathbf{x}) \propto P(y_i^c|\mathbf{f}_i^c(\mathbf{x}))$ for the land cover layer and $\phi^u(y_k^u,\mathbf{x}) \propto P(y_k^u|\mathbf{f}_k^u(\mathbf{x}))$ for the land use layer, respectively. We choose the Random Forest (RF) classifier (Breiman, 2001) for determining the association potentials of both layers. However, each classification is based on a different set of features (cf. Section 2.5). RF has proven to be an efficient classifier, also in remote sensing applications, e.g. (Schindler, 2012). In training, an ensemble of randomized decision trees is generated. During classification, an unknown sample is classified by each tree based on the corresponding feature values, the tree thus casting a vote for the class it considers to be the most likely one. The sum of the votes for a class divided by the total number of trees defines the value of the association potential for that class. Some parameters of the RF classifier have to be set beforehand. The most important ones (Breiman, 2001) are the maximum number of samples per class to be used for training, the minimum number of samples in a node required to split the node if they belong to different classes, the maximum depth of a tree and the number of trees in the forest. Due to considerable differences in the structure of both classification tasks, these parameters have to be selected individually. The relevance of each feature in the classification process can be analyzed based on the permutation importance measure (Chehata et al., 2011), which can be easily obtained from the RF in the training procedure (Breiman, 2001). The overall importance values are determined per feature. For this purpose, the accuracy achieved using the correct feature values is compared to the accuracy obtained after permuting the values of the feature to be analyzed.

2.3.3. Pairwise intra-layer potential

This potential models the dependencies of the labels of nodes n_i and n_j being adjacent within one layer, considering the data $\mathbf{x}$. The data are taken into account in the form of an interaction feature vector $\boldsymbol{\mu}_{ij}(\mathbf{x})$ for each edge. We apply the RF classifier for determining the intra-layer interaction potentials of both layers. In contrast to our previous work (Albert et al., 2014a), we apply a statistical classifier also for land cover classification instead of using just data-dependent smoothing. Super-pixel segmentation merges pixels with similar characteristics, anyway, which leads to a smoothing effect. A statistical classifier favours more probable class configurations given the data. The probability of a certain class relation is learned from real-world occurrences in representative training data. Thus, the interaction potential is modelled as the joint posterior probability of both labels y_i^c and y_j^c given $\boldsymbol{\mu}_{ij}^c(\mathbf{x})$, i.e. $\psi^c(y_i^c,y_j^c,\mathbf{x}) \propto P(y_i^c,y_j^c|\boldsymbol{\mu}_{ij}^c(\mathbf{x}))$ for the land cover layer, and of both labels y_k^u and y_l^u given $\boldsymbol{\mu}_{kl}^u(\mathbf{x})$, i.e. $\psi^u(y_k^u,y_l^u,\mathbf{x}) \propto P(y_k^u,y_l^u|\boldsymbol{\mu}_{kl}^u(\mathbf{x}))$ for the land use layer. This corresponds to a standard classification task. Thus, it is also possible to model the interaction potential by applying RF. The difference to the RF used for the association potential is that any combination of classes at neighbouring nodes

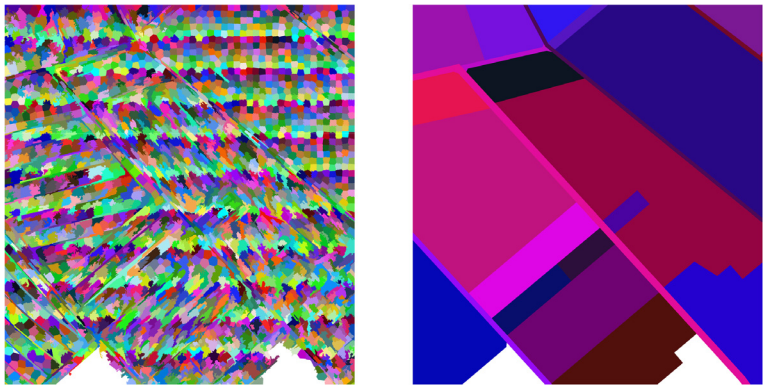

Fig. 2. Region images representing super-pixels (left) and land use objects (right). The colours are assigned randomly. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Fig. 3. SLIC super-pixels, each containing 100 pixels, superimposed to an orthophoto of an urban scene (left) and a rural scene (right). The compactness parameter (Achanta et al., 2012) is set to 20 in a range of [1;100]. In heterogeneous areas, SLIC super-pixels adapt to image boundaries. In homogeneous areas, SLIC super-pixels tend to have a compact shape.

is considered as a single class, which leads to a very high number of classes to be distinguished. In our case, the interaction feature vectors $\boldsymbol{\mu}^c_{ij}(\mathbf{x})$ in the land cover layer correspond to the element-wise differences of the site-wise feature vectors of two adjacent nodes, whereas the interaction feature vectors $\boldsymbol{\mu}^u_{kl}(\mathbf{x})$ in the land use layer are established by a concatenation of both feature vectors. In previous tests, this has been shown to be an appropriate way of incorporating the data in the estimation of the intra-layer interaction potential in the two layers.

2.3.4. *Inter-layer higher order potential*

This potential models the semantic dependencies between land cover labels y^c_i and land use labels y^u_k of all nodes n^c_i and n^u_k belonging to a higher order clique h, which is formed by all image sites in both layers having a spatial overlap. The inter-layer potential takes a value proportional to the joint probability of the set of class labels $\mathbf{y}^c_h$ in the land cover and $\mathbf{y}^u_h$ in the land use layer of all nodes n_h being connected in a higher order clique h, i.e. $\psi^{cu}(\mathbf{y}^c_h, \mathbf{y}^u_h) \propto P(\mathbf{y}^c_h, \mathbf{y}^u_h)$. The data are not taken into account for the estimation of this potential. Thus, the inter-layer potential only depends on class labels and their respective beliefs. The modelling of this higher order potential is not straightforward, as it is subject to several restrictions. A simple smoothing constraint, as considered by the P^N-Potts model, is not appropriate in our case, because the potential should not favour all land cover sites to take the same label. Furthermore, the class labels to be assigned to the compound segments differ from those at the land cover sites, anyway; such labels would not be determined in the P^N-Potts model at all. In order to model the complex relationship more adequately, a generic formulation is required. The higher order clique may contain a large number of random variables, where the number of its constituent nodes can differ. Hence, learning all possible class configurations by a statistical classifier becomes computationally intractable. Furthermore, by involving a large number of variables in a generic formulation of a higher order potential, inference becomes intractable as well. In order to allow the model to be learned based on representative training data as well as to enable efficient inference, we propose an iterative inference algorithm presented in Section 2.4.

2.4. Iterative inference procedure

In the inference step, the most probable label configuration $\mathbf{y}$ is determined for all nodes in a CRF simultaneously. This is based on maximizing the posterior probability $P(\mathbf{y}|\mathbf{x})$ of the labels given the data. Exact inference is computationally intractable for multi-class problems (Kumar and Hebert, 2006). Therefore, only approximate methods can be used. An approximate solution can be obtained by an iterative optimization method based on message passing techniques, e.g. Loopy Belief Propagation (LBP) (Frey and MacKay, 1998). We apply max-sum LBP for inference in the two-layer CRF. However, due to the limitations imposed by the higher order potential, we propose a joint iterative inference procedure inspired by Roig et al. (2011). In each iteration, we determine the most probable label configuration at each layer separately. As a result, we obtain the partial solutions $\tilde{\mathbf{y}}^c_t$ for the land cover layer and $\tilde{\mathbf{y}}^u_t$ for the land use layer. During this optimization process, the higher order terms are simplified to unary terms. There is one such term for each layer, i.e. $\phi^{u\to c}(y^c_i, \mathbf{y}^u_h) \propto P(y^c_i|\mathbf{y}^u_h)$ for the land cover layer and $\phi^{c\to u}(y^u_k, \mathbf{y}^c_h) \propto P(y^u_k|\mathbf{y}^c_h)$ for the land use layer. This allows an

approximate solution by LBP. In order to define the influence of the higher order terms in the classification process individually, we apply different weights $\omega^{5,c\to u}$ and $\omega^{5,u\to c}$ per layer. In both layers, we obtain unary terms by applying two simplifying assumptions: First, the higher order potential in each layer only depends on the class labels of the other layer, and second, these class labels are set to constants in each iteration. The constants are derived from the partial solutions $\tilde{\mathbf{y}}^c_{t-1}$ and $\tilde{\mathbf{y}}^u_{t-1}$ obtained in the previous step. Whereas Roig et al. (2011) update their higher order potential exclusively based on the current labelling obtained in the partial solution, our approach also considers the beliefs for all labels obtained in the partial solutions. The inference procedure is repeated until the classification result does not change anymore.

We apply a statistical classifier for both potential terms. Instead of introducing the class labels including their beliefs to a standard classification method directly, we encode the complex relationship within a higher order clique by using descriptive contextual features. These features encode the inter-level context, i.e. the complex dependencies between land cover and land use. For each image site, they are derived from the classification result of the other layer. The original input data are not taken into account. The set of features differs for both classification tasks. The extracted features are combined in additional site-wise feature vectors $\mathbf{m}^c_i(\mathbf{y}^u_h)$ and $\mathbf{m}^u_k(\mathbf{y}^c_h)$ for the nodes in the land cover and land use layer, respectively, and are introduced to a standard classification method. Thus, both inter-layer potentials take values proportional to the probability of y^c_i and y^u_k given the site-wise feature vectors $\mathbf{m}^c_i(\mathbf{y}^u_h)$ and $\mathbf{m}^u_k(\mathbf{y}^c_h)$, i.e. $\phi^{u\to c}(y^c_i,\mathbf{y}^u_h)\propto P(y^c_i|\mathbf{m}^c_i(\mathbf{y}^u_h))$ for the land cover layer and $\phi^{c\to u}(y^u_k,\mathbf{y}^c_h)\propto P(y^u_k|\mathbf{m}^u_k(\mathbf{y}^c_h))$ for the land use layer, respectively. Again, we apply the RF classifier (Breiman, 2001) for modelling these potentials.

The individual steps of the inference procedure are described below. In the first step of the procedure, a certain number n_{LBP} of iterations of LBP are performed at each layer, separately. In LBP, the higher order terms do not change and affect the message passing process as additional unary terms. We obtain partial solutions for land cover and land use by calculating temporary beliefs and inferring a label. The standard LBP algorithms at each layer are then interrupted in order to refine the inter-layer potential terms based on the partial solutions. For this purpose, we update the inter-level context features based on the partial solutions obtained in the previous step of the inference procedure. The contextual features $\mathbf{m}^c_i, \mathbf{m}^u_k$ are calculated based on spatially overlapping image sites. Afterwards, we derive new values for the inter-layer higher order potential terms at each layer by applying the respective classifier based on the updated site-wise feature vectors $\mathbf{m}$. Then, at each layer LBP continues at the point where it was stopped before the update step. Again, the higher order potential is considered as an additional unary term in LBP, but using the updated values, which affects the further evolution of the messages being passed. The procedure is repeated until a maximum number n_{lt} of iterations is reached. In the last step of the procedure, the final beliefs are calculated for each node based on the current messages and the node and edge potentials. The label having the maximum belief is assigned to each node.

Initial values of the inter-level context features are derived from an initial classification of land cover and land use based on features derived from image and height data, before starting the inference procedure. The initial classification procedure is explained in detail in Section 2.6.

2.5. Feature extraction

In our approach, feature extraction is designed for input data derived from high-resolution aerial images, such as digital surface models (DSM), digital terrain models (DTM) and orthophotos. The features for land cover and land use classification refer to different image entities. In the land cover layer, features are extracted for super-pixels. In the land use layer, features are extracted for land use objects, which are defined by the polygonal representation of the GIS-objects of a geospatial land use database. In the following description of the extracted features, the term 'segments' refers to super-pixels as well as land use objects. We distinguish three different sets of features: *image-based* and *geometrical features*, which remain unchanged during the inference procedure, and *contextual features*, which are updated at each iteration in the inference procedure. The contextual features are derived from the partial solutions obtained in each step of the inference procedure. For each image site, the partial solutions provide an output label, i.e. the label with the maximum belief, as well as the beliefs for all classes.

We extract a similar set of image-based and geometrical features for the nodes of each layer. The set of image-based features consists of spectral, textural and three-dimensional features. The extracted image-based and geometrical features are listed in Table 1. Here, NDVI is the normalized difference vegetation index and nDSM refers to the normalized DSM, which is defined as the difference between DSM and DTM. Structural features are derived from a histogram of the gradient orientations (HOG) weighted by their magnitude per segment. The textural features are derived from the 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 certain spatial configurations within each segment. The geometrical features are determined from the polygonal representation of each segment.

Contextual features encode the inter-level context. We extract different sets of features for land cover and land use classification. These features are divided into *spatial metrics* and *graph-based measures*. The first group of features, belonging to the set of spatial metrics, is inspired by Munoz et al. (2010) and Xiong et al. (2011). These features are based on the pixel-based proportion of land cover labels within a land use object as well as land use labels within each land cover super-pixel, weighted by their respective beliefs. For this purpose, we map the classification results to the pixel level, where each pixel is assigned the beliefs per class of the segment-based classification result. Subsequently, we estimate the average of the pixel-wise beliefs per class within each segment:

$$x^l_{bel}=\frac{1}{\sum_{l\in L}\sum_{i\in K} bel(y_i)}\sum_{i\in K} bel(y_i) \tag{3}$$

In Eq. (3), the contextual features x^l_{bel} are calculated per class label $l\in L^t$ with $t\in\{c,u\}$ from its respective belief values $bel(y_i)$ for all pixels i within each segment K. By mapping the land use results to the pixel-level, we can consider the fact that some super-pixels may correspond to more than one land use object. All land use objects having a spatial overlap with the respective super-pixel contribute to the feature calculation according to their degree of overlap. This is also true for the land use layer, where a land use object is typically not totally congruent with the union of its spatially overlapping super-pixels. Furthermore, the minimum distance of each land cover super-pixel to the closest land use object boundary is used as a feature for land cover classification. Each distance value is normalized by the maximum extent of the land use object, i.e. the diagonal of the corresponding minimum enclosing rectangle.

For land use classification, we calculate the first and second order central image moments in order to describe the pixel-based distribution of each land cover class within a land use object. For that purpose, we derive a binary image per land cover class from the classification result beforehand. The image moment of order zero is also extracted and represents the area covered by each land cover class. In our case, the calculated image moments

Table 1
Pool of spectral, textural, three-dimensional (3D) and geometrical features extracted per segment. For the spectral and 3D features, the mean, standard deviation (std), minimum (min) and maximum (max) are estimated from the pixel-based feature values within each segment. For the features derived from the histogram of gradient orientations (HOG), the first (1st) and second (2nd) min. and max. and some ratio values are calculated. Each of the textural and geometrical features represents only one measure (no entries in the third column). MER is the min. enclosing rectangle.

Group	Basis/band	Features
Spectral	Grey value 'Red'	Mean, std, min, max
	Grey value 'Green'	Mean, std, min, max
	Grey value 'Blue'	Mean, std, min, max
	Grey value 'Near infrared'	Mean, std, min, max
	NDVI	Mean, std, min, max
	Hue	Mean, std, min, max
	Saturation	Mean, std, min, max
	Intensity	Mean, std, min, max
Textural	GLCM	Haralick energy, contrast, Correlation, homogeneity
Structural	HOG	Mean, std, min and max (1^{st}, 2^{nd}), Ratio (1^{st}min/1^{st}max), ratio(2^{nd}min/2^{nd}max), Ratio (1^{st}min/2^{nd}max), ratio(2^{nd}min/1^{st}max)
3D	nDSM	Mean, std, min, max
Geometrical	Segments	Area, perimeter, convexity, compactness, Side length ratio MER, elongated shape, Polar distance, shape index, fractal dimension

per land cover class label are normalized by the total number of pixels within the land use object. In order to ensure meaningful feature values, we first estimate the centre of gravity and the principal direction of the land use polygon. The information about its orientation in space is used to define a local coordinate system into which the pixel-based land cover results are transformed. By doing this, the features describing the land cover distribution are always related to the principal direction of the polygon, which is particularly important for a better comparability.

Graph-based measures are derived from an adjacency-event-matrix (Barnsley and Barr, 1996), which is computed from the co-occurrences of the land cover labels in a 4-neighbourhood at each pixel within the land use object. The matrix entries are normalized by the total number of entries. We use the normalized number of co-occurrences between each possible configuration of land cover classes in order to model the spatial arrangement of land cover segments within a land use object.

In total, the pool of available features for estimating the association and intra-layer potential consist of 63 image-based and geometrical features in the land cover and land use layer, respectively. The numbers of contextual features used in both layers are different, because some of these features are extracted for only one layer, i.e. minimum distances to land use boundaries for the land cover sites and central image moments and graph-based measures for the land use sites. The belief-weighted area features are calculated for the image sites in both layers, but the number differs due to a different number of class labels such a measure is calculated for. In the land cover layer, 21 contextual features are available for each node n_i^c to determine the inter-layer potential. For the land use layer, the set of contextual features consists of 84 features. From the pool of image-based, geometrical and contextual features, we select the most relevant features for classification based on the permutation importance measure obtained by the RF classifier (Breiman, 2001) and combine them in a feature vector per potential term. This subset contains the best trade-off between achieving a high classification accuracy and computational complexity. This means that any additional feature does not lead to a significant improvement of the classification accuracy (cf. Section 3.2.1).

2.6. Training

CRF being a supervised classification technique, the parameters of the potentials are learned. In our approach, the association, the intra-layer interaction and inter-layer potentials are trained separately using representative training data, which implies the training of the RF classifiers. Besides, the user has to define the weights $\omega^2, \omega^4, \omega^{5,c\to u}$ and $\omega^{5,u\to c}$ as well as the number n_{It} of iterations of the inference procedure and the number n_{LBP} of iterations in each LBP step. These parameters are determined empirically. During the training of the intra-layer interaction potentials, the relations between adjacent nodes are learned. Thus, the labels of adjacent image sites have to be known during training, which requires fully-labelled training data for the corresponding layer.

As mentioned before, the classification tasks in both layers are based on contextual features. In order to train the classifier appropriately, these features have to be available for the training step. However, the input required for the extraction of contextual features, i.e. the classification results for each layer, is typically not available during training. One could use training labels, but this would require the land cover labels for all pixels inside the training land use objects to be known. As the land use objects may be very large, this might lead to a prohibitively large effort for generating the training data. Therefore, an initial classification is carried out solely based on the association potentials without considering context, and the classification results thus obtained serve as input for the initial estimation of the contextual features.

3. Evaluation

3.1. Test data and test setup

The experiments are carried out to evaluate the performance of the presented approach. First, we analyze the relevance of the extracted features for the classification and select the most relevant features for further processing. Furthermore, we investigate the influence of the size of the super-pixels on the classification result in order to determine the level of detail that represents a good trade-off between accuracy and computation time. This investigation is based on the basic assumption that a smaller size of the super-pixels leads to an increase in computation time, which is not necessarily reflected in an improvement in accuracy. Besides, we compare the results we obtain by applying an iterative inference procedure to the results of the two-step processing strategy presented in Albert et al. (2014a) as well as to the results of a non-contextual classification based only on the association potentials of the proposed graphical model.

We perform our experiments on two test sites having different characteristics. The first test site is located in Hameln, Germany, and shows various urban, but also some rural characteristics, e.g. residential areas with detached houses, densely built-up areas, industrial areas, a river, forest, cropland and grassland. The test area has a size of 2 km × 6 km. The second test site covers the town of Schleswig, Germany, and its surroundings and has a size of 6 km × 6 km. This test area has rural characteristics, but also contains several villages and a small town. For each test site, an orthophoto, a DTM and a DSM are available. The orthophoto has a ground sampling distance (GSD) of 0.2 m and consists of four channels (near-infrared, RGB channels). The GSD of the DSM and the

DTM in Hameln are 0.5 m and 5 m, respectively; the corresponding GSDs in Schleswig are 0.28 m (DSM) and 1 m (DTM). The Hameln data were acquired in spring when deciduous trees had no foliage yet. The Schleswig data were acquired in summer and show a dense foliage of deciduous trees. In both cases, the orthophotos and the DTM were the standard products offered by the governmental surveying authorities of the federal states in which the two cities are situated.[1] The DSMs were generated from the images forming the basis of orthophoto generation by image matching. Furthermore, GIS-objects of the German geospatial land use database forming a part of the Authoritative Real Estate Cadastre Information System (ALKIS®) (AdV, 2008) are used to define the land use objects, which correspond to the nodes in the land use layer. These objects represent blocks, which may be composed of several parcels belonging to the same land use class. The nodes of the land cover layer correspond to SLIC super-pixels. The segmentation is performed on a three-channel image, where the channels correspond to the nDSM, i.e. the height above ground, the intensity and the NDVI extracted from the input data. The use of these three secondary channels instead of the original grey values enables a better adaptation to boundaries of certain land cover segments. We extract SLIC super-pixels of the size of 100, 400 and 2500 pixels in order to evaluate the influence of the size of the super-pixels on the classification result. The sizes of the super-pixels have been chosen exemplary to represent land cover information in three different levels of resolution. The SLIC compactness parameter is set to 20 in a range of $[1;100]$, which has been shown in previous tests to allow for a good adaptation to land cover boundaries in the image.

For training and evaluation, reference data are available for both layers. The reference data for the land cover layer consist of pixel-wise reference labels for 40 image tiles (Hameln) and 26 image tiles (Schleswig), each of size 200 m × 200 m, obtained by manual annotation. The reference data for the land use layer consist of the manually corrected geospatial land use database for the entire test areas, divided into 12 blocks (Hameln) and 36 blocks (Schleswig), respectively, each of size 1000 m × 1000 m. The reference for each super-pixel is assigned to the most frequent class label among its constituent pixels. However, the simple "winner-takes-all"-strat egy for the assignment of the reference label to each super-pixel leads to inaccuracies in the reference data. In the training process, we consider these uncertainties by eliminating uncertain samples, i.e., we only use super-pixels with at least 75% consistent pixels as training samples.

We distinguish eight land cover classes, i.e. $L^c = \{$*building (build.), sealed area (seal.), soil*, *grass*, *tree*, *water*, *car*, *others*$\}$. In land use classification, ten classes are discriminated, i.e. $L^u = \{$*residential (res.), non-residential (non-res.), urban green (green), traffic*, *square*, *cropland (cropl.), grassland (grassl.), forest (for.), water body (water)*, *others*$\}$. Some of these land use classes correspond to the coarsest hierarchical level of the object catalogue (*for.*, *water*, *others*), whereas others correspond to the second level (*res.*, *non-res.*, *green*, *traffic*, *square*, *cropl.*, *grassl.*) (AdV, 2008). The topic of the maximum semantic level of detail that can be discerned was investigated in (Albert et al., 2016).

The optimal parameters for the RF classifiers applied for each potential term as well as for the inference procedure were determined empirically. Table 2 shows the parameter settings used for the two test sites Hameln and Schleswig, respectively.

The quantitative evaluation is based on cross validation. For that purpose, the reference data are divided into groups, where each group contains a certain number of the 1 km^2 blocks of land use reference data mentioned above combined with spatially overlapping land cover reference data. In each test run, we use one group for the evaluation and all others for training. From all available training samples in a test run, we randomly select N_t samples per class (cf. Table 2) in order to avoid a bias for classes for which a large number of samples is available. In all test runs, each group contributes to the evaluation once. In the Hameln data set, the test data is divided in 12 groups, so that each group consists of only one 1 km^2 block. The reason for this lies in the overall low number of training samples for land use available in this test site. This is why we process each block in this test area individually to ensure that in each test run 11 blocks are available for training. Due to a higher number of training samples, the test data in Schleswig is only divided into two groups, where each group consists of 18 blocks. We obtain a confusion matrix by a site-wise comparison of the classification result to the reference for each layer separately. In addition, a pixel-based comparison is carried out for both classification tasks in order to show the improvement on an per-area basis. The quantitative evaluation is based on the overall accuracy, kappa index, correctness, completeness and quality values derived from the confusion matrix, e.g. (Rutzinger et al., 2009).

Table 2
Parameter settings of the RF classifiers and the inference procedure for processing the test data sets Hameln and Schleswig. For each RF classifier, the maximum number of samples (N_t), the maximum depth (max. depth), the minimum number of samples in a node belonging to different classes required to split the node further (min. samp. count) and the number of trees (No. trees) are set individually for the association (AP), intra-layer (IP) and inter-layer (HOP) potentials in the land cover (LC) and the land use (LU) layer. For the inference procedure, a weight ω^t per potential term with $t \in \{1,2,3,4,(5,u\rightarrow c),(5,c\rightarrow u)\}$, the number n_{lt} of iterations of the inference procedure and the number n_{LBP} of iterations in each LBP step are defined.

Parameter	Test site Hameln			Test site Schleswig		
LC-layer	AP^c	IP^c	$HOP^{u\rightarrow c}$	AP^c	IP^c	$HOP^{u\rightarrow c}$
N_t	15,000	15,000	1000	5000	15,000	15,000
max. depth	30	25	5	10	25	5
min. samp. count	5	5	5	20	5	25
No. trees	100	150	250	250	150	250
weight ω^t	1.0	0.5	0.25	1.0	2.0	0.25
LU-layer	AP^u	IP^u	$HOP^{c\rightarrow u}$	AP^u	IP^u	$HOP^{c\rightarrow u}$
N_t	1500	1000	1000	1500	1000	500
max. depth	30	20	25	25	30	20
min. samp. count	5	5	10	10	10	5
No. trees	150	150	200	200	250	250
weight ω^t	1.0	0.5	1.0	1.0	1.0	2.0
Inference						
n_{lt}	3			3		
n_{LBP}	5			10		

3.2. Results

3.2.1. Feature importance

We want to investigate the impact of the extracted features on the classification result in order to determine an appropriate set of features. Therefore, the relevance of each feature in the classification process is analyzed based on the permutation importance measure (Breiman, 2001). The overall importance values per feature are analyzed for each potential in the two-layer CRF separately. The ten features with the highest overall importance values (in percent), i.e. the most relevant features, are shown in Table 3 for the classification of association and intra-layer interaction potentials and in Table 4 for the inter-layer potentials in land cover and land use layers, respectively. The values shown in Tables 3 and 4 refer to the test site Hameln. These rankings are essentially confirmed in the Schleswig test site, and, therefore, this evaluation is omitted.

[1] http://www.geodaten.niedersachsen.de/datenangebot/geobasisdaten/ for Lower Saxony (Hameln) and http://www.schleswig-holstein.de/DE/Landesregierung/LVERMGEOSH/Service/serviceGeobasisdaten/geodatenService_Geobasisdaten_mehrLesen.html for Schleswig-Holstein (Schleswig). Accessed 16 March 2017.

Table 3
Overview of ten most important features for the classification of association and intra-layer interaction potentials in land cover and land use layers, respectively, ordered by their overall importance value (imp.) obtained by RF for the test site Hameln. SP corresponds to the startpoint and EP to the endpoint of an edge.

Rank	Association (LC-layer)		Intra-layer (LC-layer)	
	Imp. (%)	Feature f^c	Imp. (%)	Feature μ^c
1	4.03	Mean NDVI	4.05	Mean NDVI
2	4.01	Min. NDVI	3.62	Min. NDVI
3	3.40	Mean nDSM	3.09	Mean nDSM
4	3.07	Min. NIR	2.53	Mean Hue
5	2.92	Max. nDSM	2.45	Mean Blue
6	2.73	Mean NIR	2.37	Mean Intensity
7	2.62	Min. Hue	2.25	Min. nDSM
8	2.60	Min. nDSM	2.21	Mean NIR
9	2.57	Haralick homogenity	2.20	Haralick homogenity
10	2.37	Min. Saturation	2.13	Max. NDVI
		Feature f^u		Feature μ^u
1	3.32	Min. NDVI	1.73	Fractal dimension (SP)
2	2.96	Fractal dimension	1.70	Min. NDVI (SP)
3	2.68	Mean nDSM	1.67	Min. NDVI (EP)
4	2.48	Mean grad. histogram	1.54	Fractal dimension (EP)
5	2.46	Mean NDVI	1.26	Mean nDSM (SP)
6	2.35	Area	1.20	Mean nDSM (EP)
7	2.33	Mean NIR	1.18	Std. intensity (SP)
8	2.23	Shape index	1.13	Compactness (EP)
9	2.14	Std. intensity	1.13	2. min. grad. histogram (SP)
10	2.07	Std. Blue	1.11	Shape index (EP)

Table 4
Overview of ten most important features for the classification of inter-layer potentials in land cover and land use layer, respectively, ordered by their overall importance value (imp.) obtained by RF for the test site Hameln.

Rank	Inter-layer (LC-layer)		Inter-layer (LU-layer)	
	Imp. (%)	Feature m^c	Imp. (%)	Feature m^u
1	7.68	Bel.-weight. area (res.)	3.09	Area (sealed)
2	7.27	Bel.-weight. area (grassl.)	2.83	Bel.-weight. area (build.)
3	7.14	Bel.-weight. area (non-res.)	2.68	Bel.-weight. area (water)
4	6.99	Bel.-weight. area (traffic)	2.65	Bel.-weight. area (sealed)
5	6.85	Bel.-weight. area (water)	2.57	Moment m_{00} (sealed)
6	6.77	Bel.-weight. area (green)	2.51	Bel.-weight. area (tree)
7	6.73	Bel.-weight. area (cropl.)	2.32	Area (tree)
8	6.12	Bel.-weight. area (square)	2.27	Bel.-weight. area (grass)
9	6.06	Bel.-weight. area (others)	2.22	Moment m_{20} (sealed)
10	5.75	Bel.-weight. area (for.)	2.19	Moment m_{02} (sealed)

For land cover and land use classification, the set of the most relevant features differs. For land cover classification, the most important features comprise exclusively image-based and three-dimensional features. The set of image-based features involves statistical values (mean, minimum and maximum values) of the NDVI, intensity, hue, saturation and the blue and near infrared values as well as the textural feature Haralick homogenity. The NDVI-derived features are the most important ones, both for the determination of association and intra-layer interaction potential. Thus, all land cover classes and their relations can be distinguished most clearly based on NDVI information. The three-dimensional features describing the mean, minimum and maximum values of the nDSM are also among the ten most important features. Thus, height information also contributes significantly to the discrimination of land cover classes. The most relevant features are nearly the same for the nodes and edges. In the contrast, the geometrical features do not contribute much information (rank 26 for the most important geometrical feature). This is to be expected due to the similar geometrical appearance of super-pixels (e.g. nearly the same area for all super-pixels), which is a consequence of the applied segmentation method. However, geometrical features are much more important for the classification of land use due to the characteristic geometrical shape of certain land use objects. For instance, roads often have an elongated shape, whereas residential parcels are much more compact. The most important geometrical features are fractal dimension, shape index, compactness and area, where the fractal dimension is in first place in the ranking of the edge features and in second place in the ranking of the node features. Besides geometrical features, the set of most relevant features for land use classification also contains features derived from the histogram of gradient orientations, such as the mean and the second minimum value. Nevertheless, image-based features also contribute important information for land use classification. Actually, the most important features for nodes contain the mean and minimum value of the NDVI and the mean near infrared value, which can also be found among the six most important features for land cover classification. These are complemented by the mean of the nDSM and the standard deviation of the intensity and the blue value. The most relevant features for the inter-layer interaction potential are the belief-weighted area features derived from the per-class beliefs of the partial solutions for both, land cover as well as land use classification. For the land use classification, most of the graph-based measures are of less importance. Nevertheless, the feature quantifying the co-occurrence of neighbouring pixels

both of land cover class *sealed* within a land use object is on rank 14. Moreover, the second order central moments describing the spatial distribution of *sealed* within a land use object and the relative area of *sealed* and *tree* are among the most relevant features.

We select the 20 (land cover layer) and 30 (land use layer) most relevant features for the association potential based on their overall importance value for estimating the association and intra-layer potential per layer. Moreover, the 20 most important contextual features are used for estimating the inter-layer potential in both layers, respectively. For determining the number of features required, we investigated the predicted overall accuracy achieved by progressively including features in the classification according to their importance values, starting with the most relevant one. The chosen number of features corresponds to the saturation point of the overall accuracy; including additional features does not increase the overall accuracy further.

3.2.2. Land cover classification

A quantitative evaluation of the results obtained by applying the iterative inference procedure for a super-pixel size of 400 ($\approx 4 \times 4$ m^2) in both test sites is presented in Table 5. The evaluation is carried out on a per-pixel basis. For comparison reasons, the results of an independent non-contextual RF classifier for the same super-pixel size are also listed. The results are not compared to the results of the two-step processing strategy, because the land cover classification is not improved during the two-step procedure (used as input for the land use classification).

The result of the iterative inference procedure based on super-pixels of size 400 yields a mean overall accuracy of about 83.7% in Hameln and of 82.5% in Schleswig, which is an improvement of 1.8% and 5.1%, respectively, compared to the result of an independent RF classification. The mean kappa index is improved by applying the iterative inference procedure by 2.2% and 6.4%, respectively. Compared to the result of an independent RF classification, there is no land cover class for which both completeness and correctness decrease at the same time. The correctness decreases for the class *building*, but for all other classes, the iterative inference procedure leads to an improvement of the correctness. The completeness decreases for the classes *car* and *others* (in both test sites) as well as *water* (only in Hameln). Due to a larger decrease in completeness than increase in correctness, the quality value decreases for these classes. For all other classes, an improvement of the quality value is achieved. The completeness and correctness for the class *soil* are improved in Hameln by 4.3% and 13.8%, respectively, and in Schleswig by 22.6% and 4.6%, respectively. This land cover class benefits most from the context information about the present land use class. Moreover, in both test sites the class *tree* benefits from the consideration of contextual knowledge. For the class *sealed*, the correctness and completeness are also improved in both test sites. However, the improvement in Hameln (approx. 1.5% for both accuracy measures) is much lower compared to Schleswig, where the correctness and completeness increase by 11.3% and 3.4%, respectively. In Schleswig, the correctness and completeness increase also for the classes *grass* and *water*, even though the improvement (max. 8.2%) is smaller than for the class *soil*. In Hameln, these classes show only an improvement in correctness, whereas the completeness decreases. The decrease in completeness for the classes *car* and *others* goes along with an increase in correctness, which exceeds 16% for the class *car* and 29% for the class *others* in both test sites. For the class *building*, the completeness increases in Hameln by 2.4% and in Schleswig by 4.7% and the correctness decreases in Hameln by 0.6% and in Schleswig by 1.9%.

Fig. 4 shows four examples of land cover classification results in the test site Hameln, which are obtained for super-pixels of size 400 by a RF classifier and by applying the iterative inference procedure. Besides, the pixel-wise reference data are depicted in the figure. The images in the first row in Fig. 4 show a crossroads in an agricultural environment. The bordering croplands are in different states of the vegetation cycle: the cropland area on the left is already covered by low vegetation, whereas the cropland area on the right still consists of soil. Both approaches classify the low vegetation correctly as *grass*. However, some individual super-pixels of *soil* are misclassified as *sealed* by the RF classifier, which leads to a noisy appearance of the classification result in these areas (cf. Fig. 4 (b)). As the land cover class *sealed* is unlikely to appear in the land use *cropland*, the classification errors are eliminated during the iterative inference procedure. All super-pixels within the *cropland* objects are correctly assigned to the land cover classes *soil* (in the case of no vegetation) or *grass* (in the case of low vegetation), thus, leading to a more homogeneous solution compared to the result of the non-contextual RF classifier (cf. Fig. 4(c)). As a second example, a scene consisting of a forest crossed by a road is depicted in the images in the second row in Fig. 4. Again, the result of the RF classifier is very noisy, i.e. many super-pixels are erroneously classified as *building*, *grass* and *car*, both in the forest and road area (cf. Fig. 4(e)). The correct assignment of the super-pixels to the class *tree* is difficult due to the fact that the trees appear without leaves, which leads to confusion with other classes. By considering the land use during the iterative inference procedure, unlikely land cover classes can be eliminated and are substituted by land cover classes which are more likely to appear in such land use objects,

Table 5
Overall accuracy (OA) [%], kappa index (Kap.) [%], completeness (Cp.), correctness (Cr.) and quality (Q.) values [%] based on a pixel-based evaluation for the land cover classes *build.*, *seal.*, *soil*, *grass*, *tree*, *water*, *car* and *others* obtained for the test sites Hameln and Schleswig by applying a non-contextual RF classifier (RF_{400}) and the iterative inference procedure ($CRF_{iter,400}$) based on super-pixels of size 400.

	Test site Hameln						Test site Schleswig					
	RF_{400}			$CRF_{iter,400}$			RF_{400}			$CRF_{iter,400}$		
	Cr. [%]	Cp. [%]	Q. [%]	Cr. [%]	Cp. [%]	Q. [%]	Cr. [%]	Cp. [%]	Q. [%]	Cr. [%]	Cp. [%]	Q. [%]
build.	87.0	90.0	79.3	86.4	92.4	80.7	85.7	83.7	73.4	83.8	88.4	75.5
seal.	73.0	82.4	63.1	74.5	84.1	65.3	61.2	77.1	51.8	72.5	80.5	61.6
soil	80.7	70.5	60.4	94.5	74.8	71.7	81.4	53.1	47.3	86.0	75.7	67.5
grass	86.8	80.7	71.9	89.3	80.5	73.4	79.2	77.2	64.2	81.2	79.7	67.3
tree	78.8	83.8	68.4	79.1	88.0	71.4	82.9	87.7	74.3	84.7	92.2	79.1
water	94.6	82.5	78.8	96.4	80.2	77.9	90.1	65.4	61.0	98.3	69.1	68.3
car	45.4	21.1	16.8	62.0	15.4	14.1	25.6	34.4	17.2	81.5	6.7	6.6
others	18.7	4.0	3.4	47.8	0.1	0.1	13.9	25.3	9.9	70.2	6.1	5.9
OA [%]		81.9			83.7			77.4			82.5	
Kap. [%]		76.7			78.9			71.2			77.6	

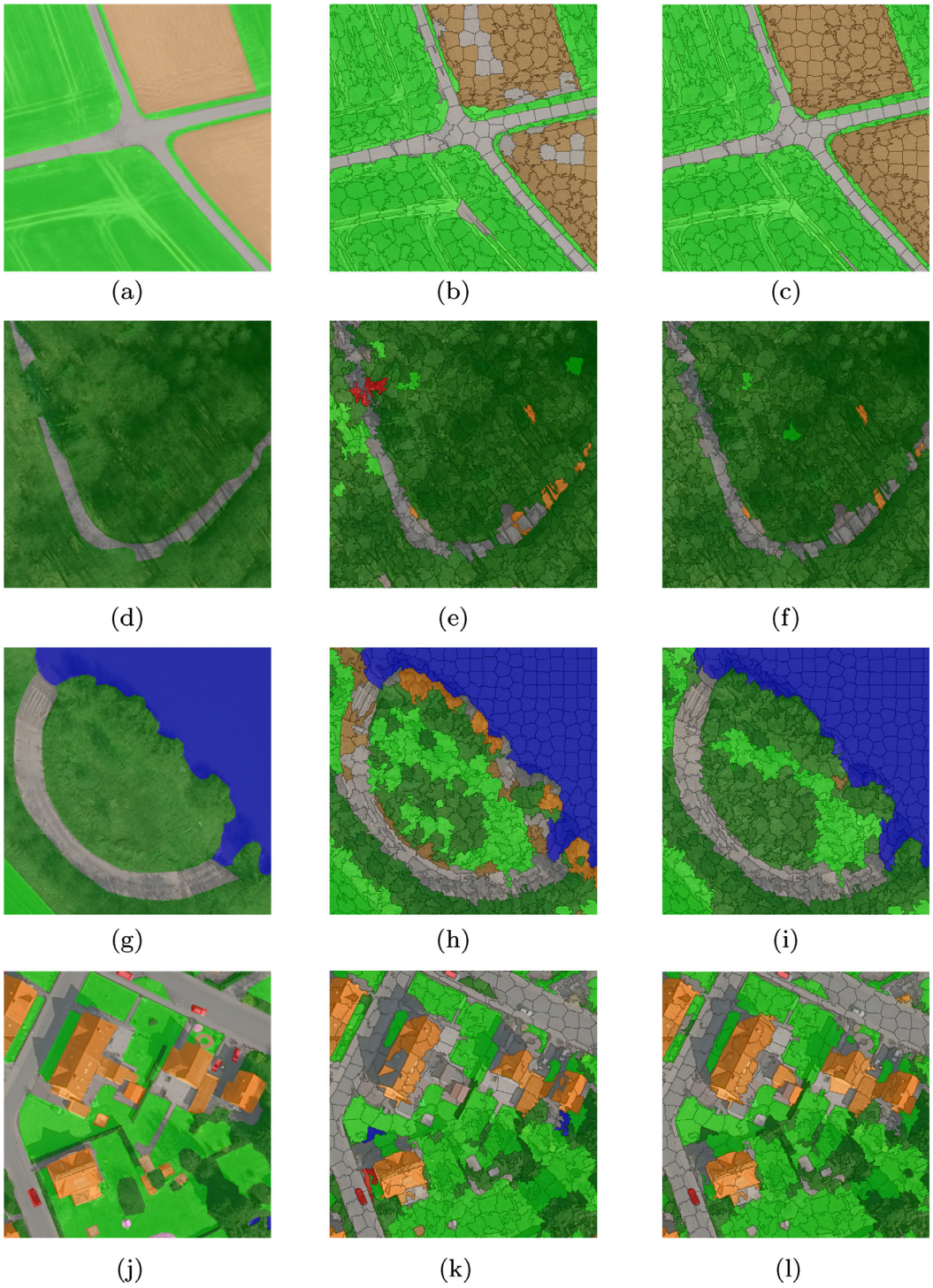

Fig. 4. Images of four different scenes (each row belongs to one scene) showing the pixel-based reference data (left column: a, d, g, j) and super-pixel based results of a non-contextual RF classifier (centre column: b, e, h, k) and the iterative inference procedure (right column: c, f, i, l) obtained for super-pixels of size 400, superimposed to an orthophoto. The colours indicate the land cover class label assigned to the image sites: *build.* (orange), *sealed* (grey), *soil* (brown), *grass* (light green), *tree* (dark green), *water* (blue), *car* (red), *others* (pink). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

i.e. the land cover class *tree* in forests and *sealed* in *traffic* objects (cf. Fig. 4(f)). As a consequence, the iterative inference procedure results in a more realistic representation of the real world scenario. The images in the third row in Fig. 4 show the shore zone of a river. Without considering context, the super-pixels at the border between river and land area are often misclassified as *building* and *sealed* (cf. Fig. 4(h)). In most of the cases, these super-pixels contain bushes which partly overlap the water surface. Thus, the characteristics of both land cover classes, especially with regard to the NDVI information, are mixed in the feature values. These ambiguities may cause the RF classifier to wrongly assign the land cover classes *building* and *sealed* to these super-pixels. During the iterative inference procedure, the spatial neighbourhood of land cover classes as well as the geometric delineation of the water body are considered. As a result, most of the super-pixels at the border are correctly assigned to the water or land surface (cf.

Fig. 4(i)). Besides, the knowledge about the existence and position of a road supports the super-pixels to take the land cover label *sealed* instead of *soil*, in spite of their quite similar spectral appearance. The last example (fourth row in Fig. 4) compares the results obtained by the non-contextual classifier and the iterative inference procedure for an urban scene. As mentioned before, the correctness of the class *building* decreases slightly by applying the iterative inference procedure, whereas the completeness increases. This effect mainly results from changes of the classification result at building boundaries, which becomes visible in Fig. 4(l). In general, the delineation of buildings is improved by merging or removing super-pixels (e.g. third building from the right in Fig. 4(l)). However, at the borders of some buildings, adjacent super-pixels are merged to the building area, although they predominantly contain another land cover (e.g. second building from the right in Fig. 4 (l)). This classification error occurs more frequently if the spectral appearance of the adjacent super-pixels is quite similar, which is especially true for the class *sealed*. If these super-pixels contain small parts of buildings, a wrong assignment leads to an improvement in completeness and a decrease in correctness for the class *building*. Fig. 4(l) shows some more positive effects of the iterative inference procedure. For instance, some classes being unlikely to occur in residential land use objects, such as *car* or *water*, are eliminated. However, this also concerns garden areas covered by *soil*, which occur rarely in the training data, and, thus, are expected to be unlikely to occur in residential areas.

Table 6 shows the land cover classification results obtained by the iterative inference procedure for super-pixels of the size of 100, 400 and 2500 pixels. By classifying super-pixels rather than pixels, a class is predicted for all pixels within a super-pixel, although these pixels may belong to different classes. Due to the "winner-takes-all"-strategy in the assignment of a ground truth label to each super-pixel, some classes typically covering only small sets of pixels, such as *car*, are often merged to other classes, and, thus, are not represented appropriately in the training data. On the other hand, the segmentation of super-pixels merges pixels with homogeneous characteristics, even though some individual pixels may show untypical characteristics. Thus, a smoothing effect of the land cover classification result is achieved, which complies with the naturally inherent characteristics of land cover in the real world. In order to determine the maximum level of accuracy which can be achieved by using super-pixels instead of pixels, we perform a pixel-wise comparison of the super-pixel-based reference data to the pixel-based ground truth for each of the extracted super-pixel sizes. The obtained accuracy measures are also reported in Table 6.

Table 6
Overall accuracy (OA) [%], kappa index (Kap.) [%] and quality (Q) values [%] based on a pixel-based evaluation for the land cover classes *build.*, *seal.*, *soil*, *grass*, *tree*, *water*, *car* and *others* obtained for the test site Hameln by applying the iterative inference procedure based on super-pixels of size 100 ($CRF_{iter,100}$), 400 ($CRF_{iter,400}$) and 2500 ($CRF_{iter,2500}$). Ref. refers to the quality values obtained by a pixel-wise comparison of the reference label per super-pixel to the pixel-wise ground truth.

	$CRF_{iter,100}$ 10 × 10 pixels (2 × 2 m²)		$CRF_{iter,400}$ 20 × 20 pixels (4 × 4 m²)		$CRF_{iter,2500}$ 50 × 50 pixels (10 × 10 m²)	
	Q [%]	$Q_{Ref.}$ [%]	Q [%]	$Q_{Ref.}$ [%]	Q [%]	$Q_{Ref.}$ [%]
build.	80.5	93.5	80.7	90.3	74.3	80.0
seal.	64.5	86.1	65.3	80.7	58.7	66.1
soil	66.3	95.7	71.7	93.9	74.3	87.5
grass	71.1	93.0	73.4	89.8	70.3	80.6
tree	69.3	92.6	71.4	88.5	67.3	78.7
water	80.9	94.0	77.9	91.1	78.6	84.8
car	30.2	58.2	14.1	37.2	–	3.7
others	3.5	70.7	0.1	54.5	–	24.6
OA [%]	82.6	95.6	83.7	93.3	80.8	87.0
Kap. [%]	77.6	94.3	78.9	91.4	75.1	83.2

The corresponding overall accuracy and mean kappa index for a super-pixel size of 2500 are 87% and 83.2%, respectively, which are improved by 6.3% and 8.2%, respectively, for super-pixels of size 400, and by 8.6% and 11.1%, respectively, for super-pixels of size 100. Thus, the maximum level of accuracy increases significantly with a smaller size of the super-pixels. However, this general improvement is not reflected in the classification results. The result of the iterative inference procedure based on super-pixels of size 2500 yields a mean overall accuracy of 80.8% and a mean kappa index of 75.1%, which are improved by 2.9% and 3.8%, respectively, for super-pixels of size 400 and by 1.8% and 2.5%, respectively, for super-pixels of size 100. Thus, the result obtained for super-pixels of size 400 is slightly better compared to the result obtained for super-pixels of size 100. Nevertheless, the results for some land cover classes are more accurate for a smaller size of the super-pixels, especially for the classes *water*, *car* and *others*. In general, smaller super-pixels represent the land cover segments more accurately in a geometric sense, and they can capture more details such as cars or small trees. Larger super-pixels partly cover different land cover classes, which leads to inaccuracies. Fig. 5 shows the results of the iterative inference procedure for three different super-pixel sizes in an urban scene. It is obvious that the level of detail increases with a smaller size of the super-pixels. For instance, cars are delineated in the result with the smallest size, but are merged to surrounding land cover classes by increasing the super-pixel size, thus, not being covered by an individual super-pixel. Depending on the proportion of different land cover classes within a super-pixel, it is most probable that the classifier yields the class label for the predominant land cover, i.e. the land cover covering the majority of its constituent pixels. As a consequence, the class *car* is not detected when using large super-pixels. This is also true for the class *others*, which mainly represents small structures in the test data sets. The quality of the class *building* obtained for a super-pixel size of 2500 decreases by more than 6.4% compared to the result obtained for super-pixels of size 400. This loss in accuracy is partly caused by some effects related to the inference procedure, which we have discussed before, but a large percentage also results from the restrictions imposed by the super-pixel segmentation. This becomes clear by comparing the quality values of the reference data, which also decreases by 10.3%. This large loss in accuracy may result from the fact that the boundaries of large super-pixels frequently do not match the building boundaries, which is particularly visible in Fig. 5(d). This is partly caused by inaccuracies in the DSM and similarities in spectral characteristics with neighbouring land cover classes. By reducing the size of the super-pixels from 400 to 100, the quality values of the classification results are quite similar, although the quality of the reference data further increases by 3.2%. By a visual comparison of the results obtained for a super-pixel size of 400 (Fig. 5(c)) and 100 (Fig. 5(b)), it becomes obvious that the classified building areas in both results are quite similar in spite of different classification entities. Small differences only occur at the building boundaries, where the super-pixels of size 100 provide a more detailed delineation of the building outline compared to super-pixels of size 400. On the other hand, small structures having similar characteristics to buildings are also captured by individual super-pixels, which leads to isolated misclassifications of buildings in residential land use objects as shown in Fig. 5(b). In contrast, the quality for land cover classes typically covering large areas, such as *soil*, increases with a larger size of the super-pixels. This increase in quality results from the smoothing effect of the super-pixel segmentation on the classification result, which is especially beneficial in homogeneous areas. For the classes *grass* and *tree*, which cover large, homogeneous areas as well as small-structured areas in urban environments, the best accuracy is achieved for a super-pixel size of 400. Thus, this super-pixel size represents a good

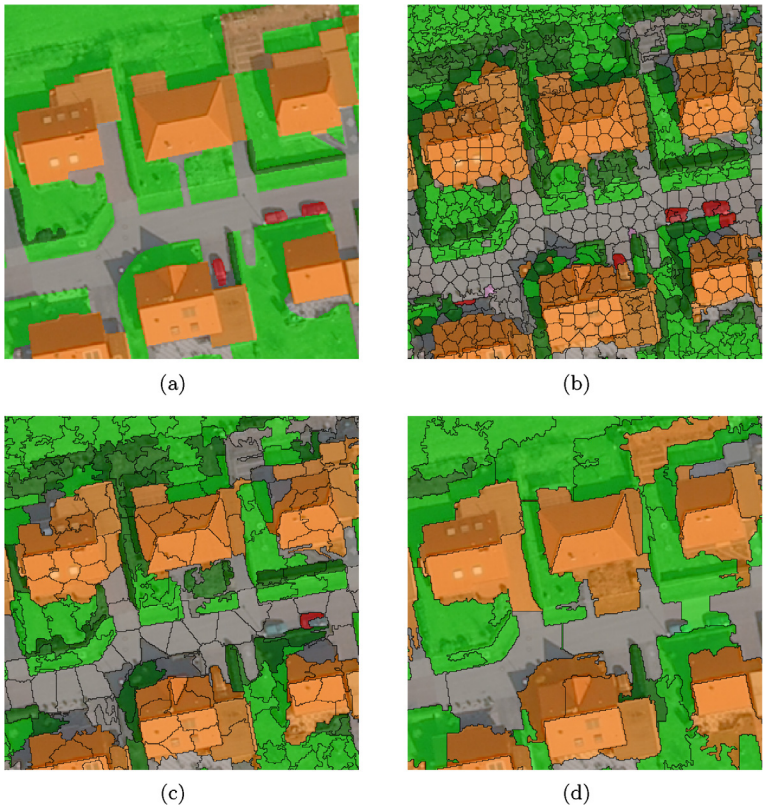

Fig. 5. Images of an urban scene showing the pixel-based reference data (a) and super-pixel based results of the iterative inference procedure obtained for super-pixels of size 100 (b), 400 (c) and 2500 (d), superimposed to an orthophoto. The colours indicate the land cover class label assigned to the image sites: *build.* (orange), *sealed* (grey), *soil* (brown), *grass* (light green), *tree* (dark green), *water* (blue), *car* (red), *others* (pink). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

trade-off between a high level of detail and a smoothing effect for these classes. It also has the advantage of speeding up the computations considerably compared to using a super-pixel size of 100.

3.2.3. Land use classification

A quantitative evaluation of the results obtained by the iterative inference procedure is presented in Table 7 for the test site Hameln and in Table 8 for the test site Schleswig. Besides, both tables contain the results obtained by the two-step processing strategy presented in our previous work (Albert et al., 2014a) as well as the results of a non-contextual RF classifier. The accuracy measures in both tables are derived by a site-wise comparison, i.e. number of objects, of the classification result to the reference.

For the test site Hameln, the result obtained by the iterative inference procedure based on super-pixels of size 400 achieves a mean overall accuracy of 78.3% and a mean kappa index of 72.2%. Compared to the results of the non-contextual RF classifier, the mean overall accuracy and mean kappa index are improved by 2.5% and 3.1%, respectively. However, the overall accuracy and mean kappa index do not change significantly compared to the results of the two-step processing strategy. Nevertheless, the correctness and completeness of individual classes differ. Compared to the non-contextual classification, the largest improvement of the quality by 13.1% is achieved for the class *forest*, which results from a large increase in completeness by 24.6% accompanied by a much lower decrease in correctness by 3%. Even compared to the two-step strategy, which already achieves an increase in completeness by 18%, the completeness is further improved by 6.6% by applying the iterative inference procedure. Thus, the class *forest* benefits the most from the consideration of context during classification in general (as shown by the large improvement for the contextual two-step classification approach), but also from the iterative processing strategy. For the classes *residential*, *non-residential*, *urban green*, *traffic* and *others*, the quality values are also improved compared to the result of a non-contextual classification. For the classes *residential* and *non-residential*, this improvement results from an increase in correctness and completeness by more than 2.6%. The improvement for the classes *urban green*, *traffic* and *others* results from an increase of one of the accuracy measures (completeness or correctness), where the other one remains nearly the same, i.e. a higher correctness in the case of *urban green* and *others* and a higher completeness in the case of *traffic*. Among all of these classes, the correctness for the class *others* shows the largest improvement by 16.9%. Compared to the two-step strategy, the differences are significantly smaller for these classes, thus, leading to maximal deviations between the quality values of about $\pm$1.8%. The quality is improved for the classes *residential* and *others* by 1.8% and 1.6%, respectively. In both cases, the completeness increases while the correctness remains almost constant (*residential*) or is even improved (*others*). The quality decreases slightly (max. 1.2%) for the classes *non-residential*, *urban green* and *traffic*, although the correctness for the class *urban green* increases by 2.2%. Among those classes for which lower quality values are achieved compared to a non-contextual classification (i.e. *square*, *cropland* and *water body*), the classes *cropland* and *water body* show the largest decrease by 12.7% and 11.6%, respectively. For the class *cropland*, this results from a large decrease in completeness by 14.3% and correctness by 5.2%. By comparing the results of the two-step approach with those obtained by the non-contextual RF classifier, it becomes obvious that the consideration of context information during classification leads to this large loss in completeness, which is not further degraded by applying the iterative

Table 7
Overall accuracy (OA) [%], kappa index (Kap.) [%], completeness (Cp.), correctness (Cr.) and quality (Q.) values [%] based on a site-wise evaluation for the land use classes *residential (res.)*, *non-residential (non-res.)*, *urban green (green)*, *traffic (traf.)*, *square*, *cropland (cropl.)*, *grassland (grassl.)*, *forest*, *water body (water)*, *others* in the test area Hameln obtained by applying a non-contextual RF classifier (RF_{400}), the two-step processing strategy ($CRF_{2-step,400}$) (Albert et al., 2014a) and the iterative inference procedure ($CRF_{iter,400}$) based on super-pixels of size 400. N_{smp}: Number of samples available for the respective class.

	Test site Hameln								
	N_{smp}	RF_{400}			$CRF_{2-step,400}$			$CRF_{iter,400}$		
		Cr. [%]	Cp. [%]	Q. [%]	Cr. [%]	Cp. [%]	Q. [%]	Cr. [%]	Cp. [%]	Q. [%]
res.	566	83.2	79.0	68.1	87.0	79.0	70.6	86.5	81.6	72.4
non-res.	450	70.5	73.3	56.1	73.8	80.9	62.9	74.7	78.0	61.7
green	366	64.2	79.8	55.2	64.4	83.1	56.9	66.6	79.5	56.8
traf.	1262	85.2	90.5	78.2	85.9	94.0	81.4	85.4	93.8	80.9
square	96	53.7	37.5	28.3	66.7	39.6	33.0	57.1	33.3	26.7
cropl.	77	80.6	70.1	60.0	79.6	55.8	48.9	75.4	55.8	47.3
grassl.	132	58.2	43.2	32.9	52.7	43.9	31.5	53.0	46.2	32.8
for.	106	72.6	57.5	47.3	74.8	75.5	60.2	69.6	82.1	60.4
water	66	80.0	45.1	40.5	100.0	23.9	23.9	81.5	31.0	28.9
others	232	47.8	41.4	28.5	61.8	40.5	32.4	64.7	41.8	34.0
OA [%]			75.8			78.4			78.3	
Kap. [%]			69.1			72.2			72.2	

Table 8
Overall accuracy (OA) [%], kappa index (Kap.) [%], completeness (Cp.), correctness (Cr.) and quality (Q.) values [%] based on a site-wise evaluation for the land use classes *residential (res.)*, *non-residential (non-res.)*, *urban green (green)*, *traffic (traf.)*, *square*, *cropland (cropl.)*, *grassland (grassl.)*, *forest*, *water body (water)*, *others* in the test area Schleswig obtained by applying a non-contextual RF classifier (RF_{400}), the two-step processing strategy ($CRF_{2-step,400}$) (Albert et al., 2014a) and the iterative inference procedure ($CRF_{iter,400}$) based on super-pixels of size 400. N_{smp}: Number of samples available for the respective class.

	Test site Schleswig								
	N_{smp}	RF_{400}			$CRF_{2-step,400}$			$CRF_{iter,400}$		
		Cr. [%]	Cp. [%]	Q. [%]	Cr. [%]	Cp. [%]	Q. [%]	Cr. [%]	Cp. [%]	Q. [%]
res.	970	70.9	86.5	63.9	76.2	90.9	70.8	76.1	88.8	69.4
non-res.	360	50.0	30.3	23.2	71.4	31.9	28.3	60.8	37.5	30.2
green	580	52.3	55.9	37.0	61.5	65.7	46.5	62.0	55.3	41.3
traf.	1209	78.5	88.3	71.1	79.2	88.7	71.9	80.4	88.5	72.8
square	141	48.4	22.0	17.8	53.6	36.9	28.0	54.3	27.0	22.0
cropl.	171	79.3	62.6	53.8	79.8	43.9	39.5	78.0	45.6	40.4
grassl.	435	69.7	72.0	54.8	61.7	77.2	52.2	57.7	82.3	51.4
forest	337	78.7	77.7	64.2	71.1	81.0	60.9	68.9	81.6	59.7
water	90	72.1	40.8	35.2	70.1	33.2	29.0	70.8	40.8	34.9
others	123	33.3	11.4	9.3	60.0	4.9	4.7	27.3	2.4	2.3
OA [%]			69.7			72.1			71.3	
Kap. [%]			63.2			66.2			65.2	

inference procedure. By analyzing the confusion matrix, the main reason for this loss of accuracy lies in the wrong assignment of cropland objects to the class *grassland*. As a result, the correctness of the class *grassland* decreases by 5.2% compared to the non-contextual classification. Nevertheless, the completeness is improved by 3%, which mainly results from a reduced number of wrong assignments of grassland objects to other classes, such as *urban green* and *traffic*. This increase in completeness is achieved by the iterative processing strategy, whereas the decrease in correctness can already be observed in the results of the two-step approach, thus, being caused by the consideration of context information during classification in general. As mentioned before, the second largest loss in quality compared to the results of the non-contextual RF classifier is achieved for the class *water body*, which mainly results from a large decrease in completeness by 14.1%, going along with a much smaller decrease in correctness by 1.5%. This is mainly caused by wrong assignments of water bodies to the class *traffic*. However, the completeness obtained by the two-step approach is even smaller, thus, many wrong assignments can be recovered during the iterative inference procedure. As a result, the quality is improved significantly by 5% compared to the two-step approach. In contrast, the class *square* suffers the most from the iterative processing, which is shown by a large decrease in quality of about 6.3% compared to the two-step approach. In comparison to the non-contextual classification, the quality decreases as well, even though the difference is much smaller (only 1.6%). Hence, the consideration of spatial and semantic dependencies is basically beneficial for this class, but the benefit is reduced by updating the context information based on refined land cover class predictions during the iterative inference procedure.

For the test site Schleswig, the accuracy obtained by the iterative inference procedure based on super-pixels of size 400 differs from the accuracy obtained for the test site Hameln, although the consideration of context again leads to a slight improvement in the overall accuracy and the mean kappa index by 1.6% and 2%, respectively. Compared to the non-contextual classification, the correctness is significantly improved for the classes *residential*, *non-residential*, *urban green*, *traffic* and *square*. For the classes *residential*, *non-residential* and *square*, the large increase in correctness goes along with a large increase in completeness. The completeness also increases for the classes *grassland* and *forest*, which goes along with a decrease in correctness. As the decrease in correctness is higher than the increase in completeness, the iterative inference

procedure achieves lower quality values for these classes compared to the non-contextual classification result. For the classes *cropland* and *others*, the completeness and correctness values decrease, which leads to lower quality values for these classes as well. Only for the class *water body*, the quality remains almost constant. Even though the overall improvement of the accuracy is smaller compared to the test site Hameln, the quantitative evaluation confirms again the benefit of incorporating contextual knowledge in the land use classification process. The impact of the iterative processing strategy on the results becomes obvious by comparing it to the two-step processing strategy. Although the overall accuracy and the mean kappa index are quite similar, the per-class accuracies differ. Basically, only a few classes benefit from an iterative processing, for instance the classes *non-residential*, *traffic*, *cropland*, *grassland* and *water body*, for which either the completeness or correctness value is improved. However, for the classes *non-residential* and *water body*, this improvement does not lead to a significant higher quality value. In contrast, the quality decreases for the classes *residential*, *urban green*, *square*, *forest* and *others*. Due to these quite heterogeneous

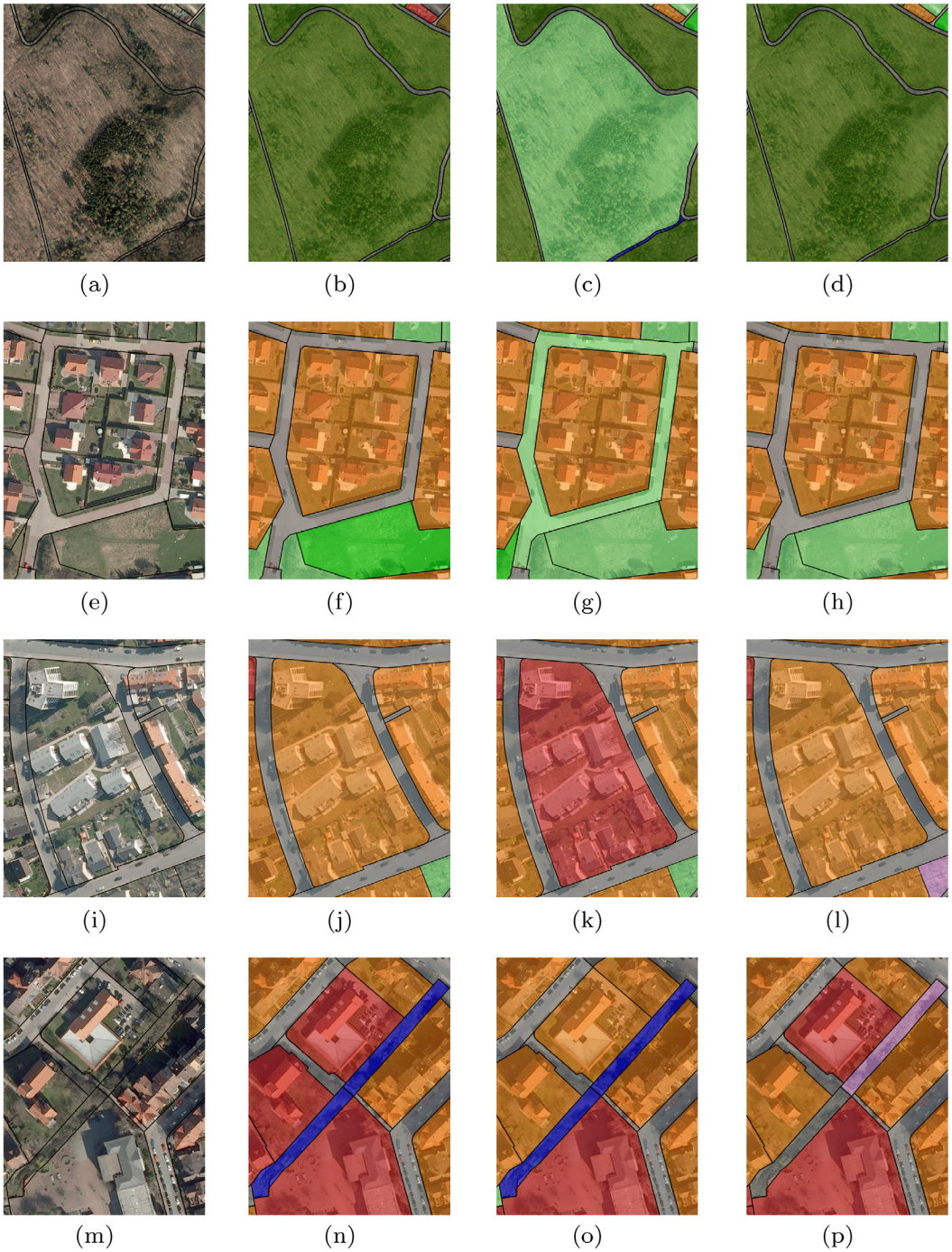

Fig. 6. Images of four different scenes (each row belongs to one scene) showing the boundaries of the land use objects (a, e, i, m), the reference data (b, f, j, n) and the results of a non-contextual RF classifier (c, g, k, o) and the iterative inference procedure (d, h, l, p) obtained for land use objects, superimposed to an orthophoto. The colours indicate the land use class label assigned to the image sites: *res.* (orange), *non-res.* (red), *green* (light green), *traffic* (grey), *square* (dark grey), *cropl.* (brown), *grassl.* (green), *for.* (dark green), *water body* (blue), *others* (pink). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

results for the test site Schleswig, it is difficult to infer a significant positive or negative impact of the iterative refinement process on the land use classification result.

Fig. 6 shows four examples of land use classification results in the test site Hameln, which are obtained by a non-contextual RF classifier and by applying the iterative inference procedure. Besides, the reference data per land use object are depicted in the figure. The examples are chosen in order to highlight typical success and failure cases of the proposed approach. The first example (first row in Fig. 6) shows a *forest* land use object, which is wrongly classified as *urban green* by a non-contextual RF classifier. The consideration of contextual dependencies helps to assign the correct land use label. In this example, the land cover result within the forest area is significantly improved during the inference procedure, which, in turn, helps to improve the land use classification result. In the second example (second row in Fig. 6), a road having a non-typical shape is wrongly classified as *urban green* by the RF classifier. The consideration of the predominant land cover *sealed* within the land use object as well as the spatial neighbourhood of other roads and residential areas supports the assignment of the correct land use label *traffic* to this land use object. The images in the third row in Fig. 6 depict an example for the improved discrimination of the land use classes *residential* and *non-residential*. The land use object shown in this figure contains some buildings having a different roof structure than the other buildings in residential areas in the test area. As a consequence, the non-contextual classifier is quite uncertain about whether this land use object belongs to the class *non-residential* or *residential* (expressed by almost equal shares for the beliefs of the classes *non-residential* (0.6) and *residential* (0.4)), but erroneously favours the land use *non-residential*. By considering the context information in the classification process, the belief for the class *residential* increases to 0.6, so that the land use object is classified correctly. The last example (fourth row in Fig. 6) visualizes a failure case of the proposed approach. The images show a creek crossing an urban scene, which is represented by two land use objects of the class *water body*. Without considering context, these land use objects are correctly classified. However, the incorporation of context information in the classification process leads to wrong assignments to the classes *traffic* and *others*. By analyzing the beliefs obtained by the non-contextual RF classifier, it is shown that the decisions are taken on the basis of low beliefs in both cases, i.e. max. 0.35 for the class *water body*, and other land use classes are almost equally probable, such as *traffic* (max. 0.25) and *urban green* (max. 0.27). These uncertainties are also propagated during the inference procedure. This might have led to errors in the land cover classification, which, in turn, might have affected the land use classification in the next iteration.

3.3. Effectiveness of the approach

The analysis of the effectiveness of the proposed approach shows a significant improvement of the land cover classification result by applying the iterative inference procedure. This is achieved by eliminating isolated classification errors during inference, which are unlikely to appear in certain land use classes. The unlikely land cover classes are substituted by land cover classes which are more likely to occur in these land use objects. Consequently, the land cover classification result is much more homogeneous compared to the result of a non-contextual classifier. This effect is especially beneficial for land cover classes which are difficult to discriminate due to a spectral appearance similar to other classes, e.g. *sealed* and *soil*. Therefore, the land cover class *soil* benefits most from the context information about the land use class. Besides, the knowledge about the geometric delineation of the land use objects as well as the spatial neighbourhood of land cover classes helps to assign the super-pixels at land use borders to the correct land cover class, especially those where the characteristics of both land cover classes are mixed in the corresponding feature values due to partial overlap. Even for land cover boundaries within a land use object, e.g. building boundaries, the delineation is improved by considering the knowledge about the land use class. However, the accuracy for classes covering small areas, such as *cars* and *others*, decreases. These classes occur rarely in the training data, and, thus, are expected to be unlikely to occur in certain land use objects.

For land use classification, the consideration of context leads to an improvement of the classification accuracy. An improvement is especially achieved for land use objects for which the non-contextual classifier fails due to non-typical characteristics or similarities to other land use classes. In these cases, the context information about the composition and arrangement of certain land cover classes within the land use object as well as the spatial neighbourhood to other land use objects helps to assign the correct land use label. However, the overall accuracies do not change significantly compared to the results of the two-step approach (Albert et al., 2014a), although some classes benefit from the iterative processing strategy. In Hameln, the class *forest* benefits most from the consideration of context during classification as well as from the iterative processing strategy. For these land use objects, the land cover result is significantly improved during the inference procedure, which supports the assignment of the correct land use label. In contrast, for classes having a small number of training samples, the quality decreases compared to a non-contextual classification. The corresponding land use objects are assigned to classes which occur more frequently in the training data. Thus, these classes are not properly represented in the training data, which makes it more difficult to learn the contextual relationships appropriately.

4. Conclusions

We have proposed a two-layer CRF model for the simultaneous classification of land cover and land use, considering different kinds of context information. Spatial dependencies are modelled by pair-wise interaction potentials for land cover and land use, respectively. The complex statistical dependencies between land cover and land use are modelled explicitly by using higher order potentials. In order to allow inference in the two-layer higher order CRF, we propose an iterative inference procedure. The experiments show that the classification results are improved compared to the results of a non-contextual classification. The land cover classification result is much more homogeneous compared to the result of a non-contextual classifier. Furthermore, the delineation of land cover segments is improved. For the land use classification, an improvement is mainly achieved for land use objects showing non-typical characteristics or similarities to other land use classes. Furthermore, we have shown that the size of the super-pixels has an influence on the level of detail of the classification result, but also on the degree of smoothing induced by the segmentation method, which is especially beneficial for land cover classes covering large, homogeneous areas.

Nevertheless, further work is required in order to improve the classification result. Remaining problems may result from the fact that for some classes we currently have only a low number of training samples, thus, not all classes are properly represented in the training data. Therefore, we want to apply our approach on more test areas with different characteristics and more training data, especially for currently underrepresented classes.

In our current implementation, the training data are supposed to be correct. That is, the user is supposed to check the class labels of the land use objects to be used for training. In the future, we

might consider the use of label noise robust classifiers in order to be able to use the original class labels from the database for training (Maas et al., 2016).

Finally, the presented method is the first step of a scheme for updating the given geospatial database. Currently, the geometric delineation of the geospatial objects is assumed to be correct, which might not always be the case. Therefore, we aim to infer changes to the geometric outline of objects automatically, e.g. by splitting and merging objects.

Acknowledgements

We would like to thank the surveying authorities of Lower Saxony, the *Landesamt für Geoinformation und Landesvermessung Niedersachsen* (LGLN), and Schleswig-Holstein, the *Landesamt für Vermessung und Geoinformation Schleswig-Holstein* (LVermGeo), for providing data and for the support of this project.

References

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S., 2012. Slic superpixels compared to state-of-the-art superpixel methods. IEEE Transactions on Pattern Analysis and Machine Intelligence 34 (11), 2274–2282.

AdV, 2008. Alkis-objektartenkatalog 6.0. Tech. rep., Arbeitsgemeinschaft der Vermessungsverwaltungen der Länder der Bundesrepublik Deutschland. <http://www.adv-online.de/AAA-Modell/Dokumente-der-GeoInfoDok/>.

Albert, L., Rottensteiner, F., Heipke, C., 2014a. Land use classification using conditional random fields for the verification of geospatial databases. ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. II-4, pp. 1–7.

Albert, L., Rottensteiner, F., Heipke, C., 2014b. A two-layer conditional random field model for simultaneous classification of land cover and land use. ISPRS Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. XL-3, pp. 17–24.

Albert, L., Rottensteiner, F., Heipke, C., 2015. An iterative inference procedure applying conditional random fields for simultaneous classification of land cover and land use. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. II-3/W5, pp. 369–376.

Albert, L., Rottensteiner, F., Heipke, C., 2016. Investigation of maximum level of semantic resolution achieved by contextual land use classification. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. XLI-B4, pp. 11–18.

Banzhaf, E., Höfer, R., 2008. Monitoring urban structure types as spatial indicators with cir aerial photographs for a more effective urban environmental management. IEEE J. Select. Top. Appl. Earth Observ. Remote Sensing 1 (2), 129–138.

Barnsley, M.J., Barr, S.L., 1996. Inferring urban land use from satellite sensor images using kernel-based spatial reclassification. Photogr. Eng. Remote Sensing 62 (8), 949–958.

Barnsley, M.J., Barr, S.L., 2000. Monitoring urban land use by earth observation. Surveys Geophys. 21 (2–3), 269–289.

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

Champion, N., 2007. 2D building change detection from high resolution aerial images and correlation digital surface models. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. XXXVI-3/W49A, pp. 197–202.

Chehata, N., Mallet, C., Boukir, S., Guo, L., 2011. Relevance of airborne lidar and multispectral image data for urban scene classification using random forests. ISPRS J. Photogr. Remote Sensing 66 (1), 56–66.

Frey, B.J., MacKay, D.J., 1998. A revolution: belief propagation in graphs with cycles. Adv. Neural Inform. Process. Syst. 10, 479–485.

Haralick, R.M., Shanmugam, K., Dinstein, I.H., 1973. Textural features for image classification. IEEE Trans. Syst. Man Cybernet. SMC-3 (6), 610–621.

Helmholz, P., Rottensteiner, F., Heipke, C., 2014. Semi-automatic verification of cropland and grassland using very high resolution mono-temporal satellite images. ISPRS J. Photogr. Remote Sensing 97, 204–218.

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

Hoberg, T., Rottensteiner, F., Queiroz Feitosa, R., Heipke, C., 2015. Conditional random fields for multitemporal and multiscale classification of optical satellite imagery. IEEE Trans. Geosci. Remote Sensing 53 (2), 659–673.

Kohli, P., Ladicky, L., Torr, P.H., 2009. Robust higher order potentials for enforcing label consistency. Int. J. Comput. Vision 82 (3), 302–324.

Kosov, S., Rottensteiner, F., Heipke, C., 2013. Sequential gaussian mixture models for two-level conditional random fields. Proceedings of the 35th German Conference on Pattern Recognition (GCPR), vol. LNCS 8142. Springer, Heidelberg, pp. 153–163.

Kumar, S., Hebert, M., 2006. Discriminative random fields. Int. J. Comput. Vision 68 (2), 179–201.

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

Montanges, A.P., Moser, G., Taubenbock, H., Wurm, M., Tuia, D., 2015. Classification of urban structural types with multisource data and structured models. In: IEEE Joint Urban Remote Sensing Event (JURSE), pp. 1–4.

Montoya-Zegarra, J.A., Wegner, J.D., Ladický, L., Schindler, K., 2015. Semantic segmentation of aerial images in urban areas with class-specific higher-order cliques. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. II-3/W4, pp. 127–133.

Munoz, D., Bagnell, J.A., Hebert, M., 2010. Stacked hierarchical labeling. In: European Conference on Computer Vision (ECCV). Springer, Heidelberg, pp. 57–70.

Novack, T., Stilla, U., 2015. Discrimination of urban settlement types based on space-borne sar datasets and a conditional random fields model. ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences vol. II-3/W4, 143–148.

Roig, G., Boix, X., Ben Shitrit, H., Fua, P., 2011. Conditional random fields for multi-camera object detection. In: IEEE International Conference on Computer Vision (ICCV), pp. 563–570.

Rutzinger, M., Rottensteiner, F., Pfeifer, N., 2009. A comparison of evaluation techniques for building extraction from airborne laser scanning. IEEE J. Select. Top. Appl. Earth Observ. Remote Sensing 2 (1), 11–20.

Schindler, K., 2012. An overview and comparison of smooth labeling methods for land-cover classification. IEEE Trans. Geosci. Remote Sensing 50 (11), 4534–4545.

Walde, I., Hese, S., Berger, C., Schmullius, C., 2014. From land cover-graphs to urban structure types. Int. J. Geogr. Inform. Sci. 28 (3), 584–609.

Wegner, J.D., Montoya-Zegarra, J.A., Schindler, K., 2013. A higher-order crf model for road network extraction. In: IEEE Conference onComputer Vision and Pattern Recognition (CVPR), pp. 1698–1705.

Xiong, X., Munoz, D., Bagnell, J.A., Hebert, M., 2011. 3-d scene analysis via sequenced predictions over points and regions. In: IEEE International Conference on Robotics and Automation (ICRA), pp. 2609–2616.

Yang, M.Y., Förstner, W., 2011. A hierarchical conditional random field model for labeling and classifying images of man-made scenes. In: IEEE International Conference on Computer Vision Workshops, pp. 196–203.