Active Learning in the Spatial Domain for Remote Sensing Image Classification

André Stumpf, Nicolas Lachiche, Jean-Philippe Malet, Norman Kerle, and Anne Puissant

Abstract—Active learning (AL) algorithms have been proven useful in reducing the number of required training samples for remote sensing applications; however, most methods query samples pointwise without considering spatial constraints on their distribution. This may often lead to a spatially dispersed distribution of training points unfavorable for visual image interpretation or field surveys. The aim of this study is to develop region-based AL heuristics to guide user attention toward a limited number of compact spatial batches rather than distributed points. The proposed query functions are based on a tree ensemble classifier and combine criteria of sample uncertainty and diversity to select regions of interest. Class imbalance, which is inherent to many remote sensing applications, is addressed through stratified bootstrap sampling. Empirical tests of the proposed methods are performed with multitemporal and multisensor satellite images capturing, in particular, sites recently affected by large-scale landslide events. The assessment includes an experimental evaluation of the labeling time required by the user and the computational runtime, and a sensitivity analysis of the main algorithm parameters. Region-based heuristics that consider sample uncertainty and diversity are found to outperform pointwise sampling and region-based methods that consider only uncertainty. Reference landslide inventories from five different experts enable a detailed assessment of the spatial distribution of remaining errors and the uncertainty of the reference data.

I. INTRODUCTION

MAchine learning algorithms have become important tools for the extraction of environmental information from remote sensing images. State-of-the-art supervised algorithms, such as support vector machines (SVMs), artificial neural networks, and ensemble-based learning methods [1], have already been developed, among others, for land cover analysis [2], [3], biophysical parameter estimation [4], [5], change and anomaly detection [6], [7], and geomorphological mapping [8], [9].
Supervised algorithms are adaptable to a broad range of problems but depend on the availability of training data, which, in remote sensing, are typically obtained through cost- and labor-intensive field work or time-consuming visual image interpretation. Active learning (AL) has evolved as a key concept to reduce annotation costs and generally refers to systems where the learning algorithm receives some control over the selection of additional training data during several iterations [10], [11]. Common strategies are to query the labels for samples with high uncertainty, which can be measured by the ambiguity of the posterior probabilities [12], the distance to the decision boundary [13], or the disagreement of a classifier committee [14].

Recently, several AL heuristics have been introduced for remote sensing applications and demonstrated promising results in the classification of multi- and hyperspectral images [15]–[17]. Iterative retraining of the learning algorithm is typically a computational bottleneck and can be addressed with batch-mode query functions [18], [19] that select more than one sample per iteration. In addition to uncertainty criteria, these models also incorporate sample diversity in order to reject candidates that are highly uncertain but are largely redundant among each other [15].

Most AL approaches assume that annotation costs linearly depend on the size of the required training set (e.g., pixels in most cases) and, therefore, aim to reduce the overall number of queried samples. However, recent studies on practical applications of AL have demonstrated that real annotation costs can vary considerably depending on how the labeling is performed, among instances and annotators, and may change with the number of queried instances [20], [21]. Leaving real annotation costs unconsidered in many cases still hinders a successful transfer of AL algorithms to real-world applications [20].

The acquisition of training data for remote sensing applications depends to a large degree on the spatial distribution of the queried instances; however, very few studies have integrated spatial constraints into the AL algorithm [22], [23]. Most proposed AL algorithms select samples only according to their position in feature space, which typically yields a pointwise dispersed distribution of the training data in geographic space and incurs the risk of revisiting (during image interpretation or field work) the approximately same spatial location several times. Human scene interpretation generally involves the assessment of high-level contextual features [24], whereas pointwise queries do not exploit full interpreter knowledge of the spatial context around a particular point. This suggests that queries, which allow focusing on certain spatial subsets of an area of interest, are a strategy that is more aligned with human perception.

To address such issues, we propose region-based query strategies that select compact spatial batches with high sample uncertainty and diversity. Some key ideas of this approach have already been discussed in [25], whereas here, a more complete formulation, including new query criteria, considerations of class imbalance and expert uncertainties, and a thorough experimental evaluation, is provided. The proposed region-based AL is employed for a two-class problem and tested on very-high-resolution (VHR) optical remote sensing images depicting geographic sites affected by large-scale landslide events. The visual interpretation of VHR images is a common approach for the collection of training data for applications such as land cover classification, and it is still the prevailing standard for landslide inventory mapping typically performed by trained experts who annotate affected areas with marker tools on the image. Landslide inventory mapping is, therefore, a challenging problem for the development of semiautomatic image analysis techniques and a suitable example to investigate the effects of different labeling strategies.

This paper is organized in seven sections. Section II introduces closely related works on AL, and Section III details the developed methodology and baseline methods used for comparison. Section IV describes the data sets used for a series of experiments. Section V details the experimental design to evaluate the performance of the AL heuristics and the labeling time required by users. Section VI discusses corresponding results, and Section VII concludes on the outcomes of this work.

II. RELATED WORKS ON AL

A review of advances in AL methods has been recently given in [11], and a comprehensive overview of AL algorithms developed for the analysis of remote sensing images was provided in [15]. As in the present study, most approaches assume that all target classes are known a priori to initialize the AL algorithm, whereas probabilistic approaches that also allow for the discovery of new classes are more suitable if this assumption is no longer fulfilled [26], [27]. The general underlying idea of most AL approaches is to initialize a machine learning model using a small training set and to exploit the model state and/or the data structure to iteratively select the most valuable samples that should be labeled by the user and added to the training set. With relatively few queries and labeled samples, an AL strategy should ideally yield at least the same accuracy as an equivalent classifier trained with many randomly selected samples.

Large-margin heuristics based on SVMs are frequently adopted for the design of AL algorithms in remote sensing applications [15] and have already been used as a base learner for a conceptual open-source software implementation to combine the benefits of object-oriented image analysis and AL [28]. Committee-based heuristics that select samples according to the maximum disagreement of an ensemble of classifiers have achieved promising results in a recent AL benchmark challenge [29] but are less frequently adopted for remote sensing [17]. For both approaches, iterative retraining of the classifier is typically a computational bottleneck for large data sets, and it has been demonstrated that batch-mode query functions, which consider the uncertainty and diversity of the samples [18], [19], are able to reduce the number of iterations significantly.

A very different approach to AL has been taken in recent studies for semantic image segmentation by iteratively exploring hierarchical data representations [30], [31]. They highlighted the benefits of integrating additional spatio-contextual features into the feature vector but did not explicitly consider the distribution of the queried samples in geographic space. Pascoli et al. [23] considered spatial distances among points to enhance margin-based sampling for pointwise queries.
Increasing the distances among the queried points beyond the range of spatial autocorrelation generally encourages querying less correlated and more representative samples. Liu et al. [22] formulated an AL heuristic as a traveling salesman problem to minimize the travel distances to the most uncertain points and, thereby, attempt to reduce the overall traveling costs. They showed that incorporating the uncertainty at each location as a form of reward for the traveled distances performed better than baseline methods considering only the distance. More recently, Demir et al. [32] have addressed the issue that traveling costs in field surveys are typically not directly related to the Euclidean distance but depend on more complex variables such as the terrain and route network. They demonstrated that considering auxiliary information on the terrain and the route network helps reduce the traveling time compared with batch-mode AL [18] without considering spatial costs.

In contrast to the aforementioned works, this study is focused on labeling time to identify samples during visual image interpretation rather than on traveling time needed in field surveys.

### III. Proposed Methods

The developed AL algorithm follows the query-by-committee (QBC) strategy, where the next sample is chosen according to the maximum disagreement of an ensemble [14]. The random forest (RF) algorithm [33] is used to generate tree ensembles with 500 fully grown trees. The disagreement of the ensemble is quantified with vote entropy $H$, which is computed as

$$H = - \sum_{i=0(1)}^p p_i \log(p_i)$$

where $p_i$ denotes the fractions of the trees voting for the respective classes $i \in \{0, 1\}$. Here, only a binary classification problem (0—nonlandslide; 1—landslide) is considered, but the vote entropy can be easily extended to multiple probabilities in a multiclass setting, and it has been suggested to normalize the measure with respect to the number of classes [34]. It should also be noted that other uncertainty measures can be used when the learner is a probabilistic or large-margin-based classifier [11]. Based on the notion of uncertainty, a simple strategy would be to iteratively select the sample with the highest vote entropy to be labeled by the user. However, since classifier retraining is computationally expensive, it is generally desirable to query samples in batches, and for the aforementioned reasons, it appears advantageous to focus labeling efforts on compact spatial subsets of the area of interest.

#### A. Region-Based QBC ($QBC_R$)

The problem of finding the most uncertain region $W^x$ on the map is addressed by sliding window $W$ with a desired window size $w$ ($w_x = w_y$) on entropy map $M_H$ to compute the mean local vote entropy $\mu_H$ at each position $(x, y)$, as expressed in

$$\mu_H(x, y) = \frac{1}{w^2} \sum_{i=-w_x}^{w_x} \sum_{j=-w_y}^{w_y} M_H(x + i, y + j).$$

In (2), $i$ and $j$ denote indexes on a regular raster grid. The sliding window could be directly employed on the regular grid of an input image if the classification process is performed at the pixel level; however, with large images ($>1000 \times 1000$ pixels), computation will be rather slow and will hinder a near-real-time user interaction. If learning and classification are performed on the level of segments resulting from a presegmentation of the image (see Section IV-B), image segments can be represented by their centers of gravity, marking the spatial point coordinates to which the entropy values of the segments are assigned. For an efficient search of the local maximum, the centers of gravity are first projected onto a regular grid $G$ (search grid) with a certain grid cell size $g$ ($g_x = g_y$) to compute the mean entropy of all points per grid cell (see Fig. 1). The sliding window is subsequently applied to the search grid to locate the global maximum. Fig. 1(b) shows the effect of changing the resolution of the search grid at a constant search window size ($w = 100$ m) and illustrates the uncertainty in the position of the maximum, which typically is $<=$ $g/2$. As a tradeoff between good location accuracy and fast computation, we set $g = 20$ m except for search window sizes $w < 60$ m, where $g = w/3$ was used.

Tests with finer search grids did not reveal any enhanced learning performance. Based on this sliding-window method, a region-based query function can be formulated to query, in

\[H = - \sum_{i=0(1)}^p p_i \log(p_i)\]
each iteration, the labels of all samples contained in region $W^x$ maximizing $\mu_H$. The corresponding query function is formulated as 

$$W^x = \arg \max_{W \in M_H} \mu_H(W). \quad (3)$$

Subsequently, all samples with a center of gravity that is spatially contained within $W^x$ are queried. For a binary classification, it is then convenient to ask the user to identify only the positive examples and automatically assign all nonlabeled samples as negative examples. This method is further referred to as QBC$_R$ and described in the pseudocode under Algorithm 1. Note that step 4 creates a buffer around all $W$ queried in previous iterations to avoid querying previously labeled parts of the image. The algorithm yields queries with high average uncertainty of the samples within the batch but does not guarantee that the queried samples provide complementary information. In an extreme case, all samples in the batch can be highly uncertain but carry largely redundant information. One of the key requirements for batch-mode AL is, therefore, to enforce diversity among the samples queried in each iteration [15], [18], [29].

**Algorithm 1: QBC$_R$**

**Inputs:**

- $X^i$: Initial training set
- $n$: Number of iterations
- $w$: Size of the region/spatial-batch
- $U$: Pool of unlabeled samples
- $g$: Grid cell size of search grid $G$

**Output:**

- $S^i$: Set of unlabeled samples to be included in the training set in each iteration

FOR $i \leq n$ DO
1. Train an RF with the current training set $X^i$
2. Use the RF to cast votes and compute the vote entropy for each unlabeled sample $s \in U$
3. Project the location of all $s \in U$ on regular search grid $G$ with a desired resolution $g$ to compute $\mu_H$ with a sliding window of size $w$
4. IF $i > 1$ THEN Mask out all parts of $G$ within a distance $< w/2$ from any queried $W$ in earlier iterations
5. Find region $W^x$ according to (3)
6. Query the labels for all samples $S^i$ contained within $W^x$
7. Add all samples $s \in S^i$ to training set $X^{i+1}$ and remove them from the unlabeled pool $U$
ENDFOR

**B. Region-Based QBC Considering Sample Diversity (QBC$_{RD}$)**

The combination of criteria for sample uncertainty and diversity has already been addressed in SVM-based studies, using clustering techniques to partition candidate sets of uncertain samples and avoid sampling of redundant instances that fall into the same clusters [18], [19], [35]. The spread of queried samples in the input space and along the separating hyperplane is thereby increased, which leads to a better representation of the data space and the decision boundary. For the region-based query strategy developed in this work, the composition of the batches is preconstrained by the coverage of the spatial window, and the proposed clustering techniques are, therefore, not directly applicable. However, distance metrics used for clustering can also be employed to directly measure the dispersion of samples in feature space and thereby quantify the diversity of the batch. To this end, two distinct measures are proposed. The first one, which is formulated in (4), is $\sigma_d$, which represents the standard deviation of the Euclidean distances $\rho_k(X, c)$ between each of the samples in the candidate batch ($c \in W^m$) and their respective nearest training point ($s \in X$). Thus

$$\sigma_d = \sqrt{\frac{1}{|W^m| - 1} \sum_{c \in W^m} \left( \rho_k(X, c) - \rho_k(X, c) \right)^2}. \quad (4)$$

Here, $|W^m|$ denotes the cardinality of the candidate set. In general, a larger $\sigma_d$ indicates a higher feature space spread of the contained samples in relation to the already acquired training data (see Fig. 2). The corresponding query function formulated in (5) can be used to select the region with the highest standard deviation out of $m$ candidate batches. Thus

$$W^x = \arg \max_{W^m \in M_H} \sigma_d(W^m). \quad (5)$$

A second measure is the cross-entropy $H^x$ between the training and the candidate set. The calculation of $H^x$ depends mainly on the mean logarithmic distances of the samples in the candidate batch ($c \in W^m$) to their $k$-nearest neighbors (kNNs) in the training set ($s \in X$) and is defined as

$$H^x(W^m, X) = \log(v_D |X|) - \psi(k) + \frac{D}{|W^m|} \sum_{c \in W^m} \log \rho_k(X, c) \quad (6)$$

where $v_D$ is the volume of the unit ball in $\mathbb{R}^D$, $D$ is the number of features, $\psi$ is the digamma function, and $|X|$ denotes the cardinality of the training set [36]. The measure $H^x$ is proposed here as a utility criterion to quantify the distinctiveness of the samples in the candidate batch with respect to the already available training data. The corresponding query function is formulated as

$$W^x = \arg \max_{W^m \in M_H} H^x(W^m). \quad (7)$$

**Fig. 2. Example calculations of diversity measures $\sigma_d$ and $H^x$ for two candidate batches in a simplified 2-D feature space.** Candidate batch B is preferable since contained samples have a higher spread away from the already known training points and provides a better representation of the boundary between the two classes.
The implementation in [37] was used to compute $H^x$ for each candidate batch setting $k = 1$. Fig. 2 illustrates that candidate regions with a higher $\sigma_d$ or $H^x$ are generally preferable since the contained samples yield a better exploration of previously undersampled parts of feature space.

To avoid the exploration of feature dimensions, which are less relevant or even useless for the separation of the considered classes, it is possible to constrain the distance computation to only those features that are beneficial according to the current model state. The RF algorithm offers an internal measure for variable importance, which is based on the instances that are left unconsidered (or out-of-bag (OOB)) in the bootstrap samples used for the construction of the individual classification trees [33]. The properties of different variants of this measure have been analyzed in recent studies with artificial and real-world data sets [38]–[40]. Here, the mean decrease in accuracy when a variable is randomly permuted is adopted, since it is more robust in the presence of many potentially correlated variables [41]. Threshold $t_i$ is introduced to define the minimum variable importance for which a feature will still be included in the calculation of the kNN distances.

Algorithm 2: QBC\textsubscript{RD}

**Inputs:**
- $X^i$: Initial training set
- $n$: Number of iterations
- $w$: Size of the region/spatial-batch
- $m$: Number of preselected candidate regions
- $t_i$: Minimum variable importance threshold for the inclusion of a feature in the distance calculation
- $fun$: Diversification function (5) or (7)
- $U$: Pool of unlabeled samples
- $G$: Grid cell size of search grid $G$

**Output:**
- $S^i$: Set of unlabeled samples to be included in the training set in each iteration

FOR $i \leq n$ DO
1. Train an RF with the current training set $X^i$
2. Use the RF to cast votes and compute the vote entropy for each unlabeled sample $s \in U$
3. Project the location of all $s \in U$ on regular search grid $G$ with a desired resolution $g$ compute $\mu_H$ with a sliding window of size $w$
4. IF $i > 1$ THEN Mask out all parts of $G$ within in a distance $< w/2$ from any queried $W^i$ in earlier iterations
5. Find the $m$ nonoverlapping candidate regions $W^m$ according to (3)
6. Select out of $m$ candidate regions the final region $W^\times$ that maximizes the selected diversity function $fun$ computed on all features with a variable importance $> t_i$
7. Query the labels for all samples $S^i$ contained within $W^\times$
8. Add all samples $s \in S^i$ to training set $X^{i+1}$ and remove them from the unlabeled pool $U$
ENDFOR

Fig. 3. Learning curves of an RF classifier (average and standard deviations over ten runs) in dependency of the size of the training set sampled (a) with SPCOSA and simple random sampling and (b) with the pointwise AL scheme (QBC\textsubscript{P}) at different batch sizes.

In [18], [19], and [35], the criteria for uncertainty and diversity are combined by preselecting a number of $m$ candidates with high uncertainty and by choosing among the candidates the instances with the highest diversity. A similar method is adopted in this work, whereas out of $m$ nonoverlapping candidate regions ($W^m$), the one with the highest diversity ($\sigma_d$ or $H^x$) is selected for the final query. This algorithm integrates criteria for uncertainty and diversity and from now on will be referred to as QBC\textsubscript{RD} (Algorithm 2), where QBC\textsubscript{RD} and QBC\textsubscript{RDH} are the two different versions depending on the adopted diversity criteria.

C. Baseline Methods for Comparison

The ability of AL heuristics to reduce the annotation costs is commonly assessed through comparison with simple random sampling. Nevertheless, systematic or stratified sampling designs are more frequently applied in remote sensing studies since they usually yield more accurate estimates of environmental variables than simple random sampling [42]. A stratification of the study area according to previously existing maps or preliminary image analysis is often adopted to encourage a representative sampling of all target classes. In cases where it is difficult to obtain a priori information about the suitable strata, sampling schemes that target a homogenous spatial distribution of the sampling points can be applied without prior information. An implementation of such a spatial coverage sampling (SPCOSA) scheme has been recently proposed [43]. It makes use of the k-means algorithm to divide a given area into spatially homogenous clusters and provides additional points at the most undersampled localitions. Fig. 3(a) indicates the amount of performance increase of an RF classifier trained with samples randomly derived and with SPCOSA and demonstrates that SPCOSA is preferable to assess the effectiveness of AL heuristics in the spatial domain.

The second baseline method used in this study is a simple pointwise AL heuristic (QBC\textsubscript{P}), which, initialized with just one sample per class, queries in each iteration the unlabeled instance with the highest vote entropy of the RF at each iteration. This method is very similar to the query-by-bagging algorithm proposed in [44], whereas the training of an RF involves not only bagging but also random feature selection during the construction of each tree. Fig. 3(b) illustrates that QBC\textsubscript{P} run in batch-mode yields only suboptimal results and that QBC\textsubscript{P} run with single queries per iteration is the more interesting benchmark.
IV. STUDY SITES

The first study site (Brazil) used for the experimental evaluation of the methods is located in the Serrana Mountains around Nova Friburgo (Brazil) and extends over 10.5 km$^2$. On January 11–12, 2011, the region was affected by heavy thunderstorms, triggering thousands of shallow landslides [45]. The second study site (China) is located at the eastern margin of the Tibetan Plateau around Wenchuan County and extends over 36 km$^2$. On May 12, 2008, a Mw 7.9 earthquake occurred in the region, triggering approximately 60,000 landslides [46].

A. Geospatial Data Sets

Geoeye-1 satellite imagery of the Brazil site was acquired after the event on January 20, 2011 [see Fig. 4(a)] and before the event on May 26, 2010. The ground spatial resolution of the data set is 2 m for the multispectral and 0.5 m for the panchromatic bands. A further pre-event image with a coarser resolution (30 m) was acquired by the Landsat 5 satellite on September 4, 2010. A pre-event image that resembles optical sensors with intermediate resolution (e.g., such as the current SPOT-5 or the forthcoming Sentinel-2 [47]) was simulated by resampling the Geoeye-1 image of May 26, 2010 to a resolution of 10 m using bicubic interpolation. All images were orthorectified using the ASTER Global Digital Elevation Model [48, ASTER GDEM] with a spatial resolution of 30 m. Between the pre- and post-event Geoeye-1 images, 22 tie points were manually selected to update the rational polynomial function (RPC) and thereby enhance the coregistration. The residual root mean square error (RMSE) of the updated RPC model was 14.7 m.

For the China site, Ikonos-2 imagery was acquired after the event on July 1, 2008 [see Fig. 4(d)]. The ground spatial resolution of Ikonos-2 is 4 m for the multispectral bands and 1 m for the panchromatic band. A pansharpened (1-m pixel resolution) and orthorectified product with an expected geolocation RMSE of 5 m is used. At both sites, a number of terrain variables (slope, hillshade at the times of image acquisition, hydrological flow direction, and the average distances to the mountain ridge [47]) were calculated.

Using a semivariogram analysis with an exponential model fit [49], it is possible to measure the spatial autocorrelation of the gray values within the images and gain quantitative insights into the spatial structure of the observed areas. Fig. 5 shows that, depending on the spectral bands, the effective range of the spatial autocorrelation varies between 93 and 245 m for the Brazil data set and between 207 and 468 m for the China data set. The significantly increased range of the autocorrelation in the near-infrared (NIR) band is related to large vegetated areas with relatively homogenous NIR reflectance, whereas the reflectance in the visible bands captures changes in the surface color with higher spatial frequencies. It provides an indication for the minimum size of the sampling windows, which should be sufficiently large to capture spatial variability beyond locally autocorrelated characteristics. Note that only random subsamples of the images (5%) were used for the semivariograms (see Fig. 5); however, no significant changes were observed when using different random subsamples.

In order to obtain reliable estimates of the performance and variance of AL routines, it is typically necessary to repeat
the routine several times with variable starting conditions. Therefore, it is mostly not feasible to integrate user interaction directly in the evaluation of AL algorithms, and most studies rely on readily prepared ground truth data sets. It has, however, been documented that landslide mapping by multiple experts is subjected to considerable uncertainties [50], [51], and it is generally desirable to assess uncertainties of ground truth data used for algorithm evaluation [52], [53]. To address this issue for the Brazil data set, five alternative landslide inventories (E1–E5) are independently prepared by five expert geomorphologists through joint interpretation of the two available VHR resolution images and the DEM.

Even in the absence of an ultimate ground truth, the availability of multiple-expert maps allows for the estimation of a lower bound for the mapping error. Optimistically assuming that the majority vote is correct, expert judgements that contradict the majority vote can be regarded erroneous, and the lower error bound can be estimated as a function of the expert disagreement [52]. For the five expert inventories, a lower error bound of 1.1% can be estimated for any of the five maps. This low error rate applies to the full map and results, to some degree, from the sparsely distributed landslide class. A higher error rate of 16.1% results if only pixels marked by at least one of the experts would be considered. To analyze further the interexpert variability, the F-measure \( F \), which is not prone to class imbalance [54], was computed for each pair of experts. The highest agreement between two experts was indicated by \( F = 0.86 \) (and the lowest by \( F = 0.71 \)). Considering that a learning algorithm that perfectly predicts the mapping of one expert would inevitably commit errors with respect to the mapping of a second expert, the expert agreement suggests an upper bound for the accuracy achievable with any automatic method on this particular problem. In this context, the uncertainty of each expert can be regarded as a type of inherent stochastic noise [55], which cannot be fitted without sacrificing the generalization to the maps of the other experts. A possible strategy to reduce this type of noise would be to combine the different ground truths (e.g., by majority vote). However, this would also lead to positively biased classification results in comparison to a real-world scenario where multiple ground truths are typically not available.

In order to select one of the Brazil inventories for the experiments, the number of clicks used by the experts is assessed and compared as a proxy for the degree of detail of the mapping. While it is generally difficult to assess the level of expertise, the amount of clicks can serve as an indicator for the level of detail and time investment by the expert. Fig. 4(b) illustrates that the most clicks were used for the elaboration of inventory E1, which was therefore chosen as a reference in the benchmarking of the different AL heuristics. For the significantly larger China data set, it was only feasible to obtain one reference landslide inventory, which is displayed in Fig. 4(d).

B. Segmentation and Feature Extraction

The multiresolution image segmentation algorithm [56, MRIS] was adopted, considering only spectral information of the post-event images and equal weights for all spectral bands. The region-growing algorithm comprises a scale parameter, which is a threshold for the maximum allowed increase in the segment’s variance during a merging operation. It has been demonstrated that statistical methods can help constrain the choice of the scale parameter for knowledge-based image classification [57], whereas oversegmentation with a small scale factor was found to yield higher classification accuracy in a supervised framework [9]. Consequently, the scale parameter was set to 20, corresponding to strong oversegmentation of landslides and most other landscape features. For all segments, a broad set of features describing the spectral characteristics, texture, shape, topographic variables and contrasts to neighbors are calculated. The object features are selected according to previous studies on the object-oriented mapping of landslides [9], [58]. Additional features, such as flow accumulation [59], average distance to the ridge [60], mean contrast to neighbors [61], and certain band ratios [62], are selected considering the known prevalence of landslides in certain topographic positions and typical spectral characteristics such as high contrast to surrounding vegetated areas.

Only certain combinations of all available features are used in the experiments to simulate typical scenarios of accessible data sets (see Section V-A). The class membership of the segments is assigned according to the maximum overlap with the reference landslide inventory maps (E1 for the BRAZIL data set). For the spatial queries described in the next section, segments are represented by their center of gravity. The segmentation results, the extracted features, and the implementation of the methods are available at http://eost.unistra.fr/recherche/ipgs/dgda/dgda-perso/andre-stumpf.

V. EXPERIMENTAL DESIGN

All experiments are carried out using the RF algorithm with 500 trees. To address class imbalance, stratified bootstrap sampling [63] is used for the implementation of the region-based methods so that each tree in the ensemble is built on a balanced training sample. The advantages of this strategy are discussed in Section VI-D. The number of maximum AL iterations is defined before the start of each run, and each run is repeated at least ten times with random seeds in order to estimate the mean prediction accuracy values and standard deviations on the unlabeled set. The initiation of each run is performed through stratified random sampling in order to ensure the presence of at least one example per class for the first iteration.

All segments of the data sets are labeled, and therefore, learning can be performed on the full map. This is different from the experimental designs of recent AL studies in remote sensing, where the data sets comprised reference labels only for subsets of the image [15], [17]–[19]. To avoid preconstraining the pool and to provide the query function with access to the full unlabeled pool, random prepartitioning of the data set into an unlabeled pool and test set was also avoided. Instead, for each iteration, the query functions move instances from the unlabeled pool into the training set, and the test set error is assessed on the full unlabeled pool of samples. For all experiments, the test set error, the OOB error, the system runtime, and the identifier of queried samples are recorded. In addition,
Fig. 6. Expert labeling time for the Brazil data set with (a) marker-based labeling of regions \(N = 30\) according to different window sizes and (b) with pointwise labeling \(N = 50\). (c) Marker-based labeling times per segments and pixels contained within each region.

the labeling time required by the user was experimentally evaluated, as described in the next section. Altogether, this permits the analysis of three important aspects of an AL system: 1) How often will the user be asked (number of iterations)? 2) How long will it take the user to answer the query (labeling time)? 3) How long will the user have to wait between the queries (computational time)?

A. Labeling Time

The costs of field surveys are highly dependent on the particular application and environmental conditions. It is consequently difficult to estimate related costs \textit{a priori}, and repeated field experiments that would be required to support the development of an algorithm are not practical. However, for many applications, sampling is based on image interpretation, and the time requirement for image interpretation can be easily assessed. Two alternative labeling strategies for image interpretation are considered, one being the pointwise annotation of individual samples queried by a pointwise AL method, and the second being the annotation with a brush-like tool within compact regions queried by a region-based AL method. To evaluate the required labeling time for both strategies, sequences of post-event image subsets around points \(N = 50\) queried by a pointwise AL strategy \(QBC_p\), and around regions \(N = 30\) queried by a region-based AL method \(QBC_R\) are presented to a domain expert. For the labeling of regions, the expert is provided with a brush-like marker tool with adjustable size and asked to mark all the landslide area, assuming that the unmarked area will be automatically assigned as background. For the Brazil data set, the expert was asked to mark five nested subsets \(20, 60, 100, 150,\) and \(200\ m\) of each displayed region successively from the smallest to the largest nested subset [see Fig. 6(a)]. The required annotation time for each region and each of the respective subsets is recorded in order to assess not only the overall time per region but also the relationship between the window size and required time. For the China data set, the experiment is performed at only one window size of \(200\ m\). The choice of the window size is guided by the mean effective range of the spatial autocorrelation as determined through the semivariogram analysis in Section IV-B. For the pointwise queries, labeling is performed with a single click, signaling a positive or a negative example. For both labeling strategies, the expert is given the possibility to zoom in and out as needed, and the annotation time is recorded from the moment the subset was displayed until the expert completed the identification.

Fig. 6 provides a comparison of the labeling time for regions and individual points in the Brazil data set. In both cases, the probability density distributions of the labeling times show a bimodal distribution [see Fig. 6(a) and (b)] that can be linked to two categories of queries, namely, simple regions (unique homogeneous land cover, no landslides) and more complex regions (abundance of landslides, complex landslide boundaries).

B. Sensitivity Analysis for Different Sampling Window Sizes

Experiments on the Brazil data set are dedicated to analyzing the effects of different window sizes on the performance of the region-based AL heuristics and to compare their performance with the pointwise query function and SPCOSA. Considering the range of spatial autocorrelation, it can be assumed that a sufficiently large window size would encourage the inclusion of uncorrelated samples in one batch and, at the same time, guarantee a distance between sampling regions beyond the range of autocorrelation. In order to test this assumption, the region-based AL heuristics (see Section III-A and B) are tested at five different window sizes with edge lengths \(w = \{20, 60, 100, 150, 200\ \text{m}\}\). The diagonals of these squared windows range from 28.2 to 282.4 m and covered a distance range in which relevant semivariance changes are observed within the Brazil study site (see Fig. 5). The parameters for the \(QBC_{RD}\) heuristics that determine the number of candidate regions and the minimum variable importance are kept constant at \(m = 3\) and \(t_i = 0.01\), respectively. For these experiments, a stacked feature vector including object features from the pre- and post-event Geoeye-1 imagery (see Table I) is used.

C. Sensitivity Analysis of Parameter Settings in \(QBC_{RD}\)

Further experiments are carried out to evaluate the sensitivity of the parameters that control the number of candidate regions and the minimum variable importance for the calculation of \(\sigma_D\) and \(H^*\) within \(QBC_{RD}\). Four respective experiments are dedicated to test the influence of the preselection of more or fewer candidate regions with \(m = \{2, 3, 4, 5\}\), keeping the window size constant at an intermediate value of \(w = 100\ \text{m}\) and \(t_i = 0.01\). In a second setup, five different values of \(t_i = \{0, 0.01, 0.02, 0.03\}\) are tested, keeping the number of candidate regions and the window size constant at \(m = 3\) and \(w = 100\ \text{m}\), respectively. Learning is performed with all features from pre- and post-event Geoeye-1 images (see Table I) for the Brazil data set, and each AL routine is repeated 30 times.
methods, and hence, only data set already evaluated the performance of those different tests were performed to compare availability (see Table II) are evaluated. On the China data set, monitoring. In this context, four different scenarios of data often only available in historical archives and from sensors with hours after a given event. Pre-event images, contrarily, are aster images with submeter resolution can be acquired within operational and forthcoming steerable VHR satellites, postdis-

starting from randomly seeded runs to assess mean prediction accuracy values and standard deviations.

D. Performance on Different Data Sets

A final set of experiments is carried out in order to assess the robustness of the proposed methods on different data sets. These included tests on the China data set at a window size of \( w = 200 \, \text{m} \) and a simulation of data scenarios typically encountered in disaster response mapping using the Brazil data set at a window size of \( w = 100 \, \text{m} \). With the wide range of operational and forthcoming steerable VHR satellites, postdisaster images with submeter resolution can be acquired within hours after a given event. Pre-event images, contrarily, are often only available in historical archives and from sensors with coarser spatial resolution dedicated to long-term environmental monitoring. In this context, four different scenarios of data availability (see Table II) are evaluated. On the China data set, tests were performed to compare \( \text{SPCOSA}, \text{QBC}_P, \text{QBC}_{R_P}, \) and \( \text{QBC}_{R_{DH}} \). The abovementioned experiments on the Brazil data set already evaluated the performance of those different methods, and hence, only \( \text{QBC}_{R_{DH}} \) and \( \text{QBC}_{R_{S}} \) are tested to evaluate their robustness in the different scenarios. The learning curves are averaged over 30 random seeded runs, and the parameters for the diversity criteria are kept constant at \( t_i = 0.01 \) and \( m = 3 \).

VI. RESULTS AND DISCUSSION

A. Sampling Window Size

Fig. 7 depicts the outcome of the experimental comparisons of the baseline methods (\( \text{SPCOSA}, \text{QBC}_P \)) and the proposed region-based AL heuristics on the Brazil data set. For all tested window sizes \( (w = \{60, 150\}) \) are not displayed in Fig. 7), the region-based queries generally yield steeper learning curves, outperforming \( \text{SPCOSA} \) and pointwise query method \( \text{QBC}_P \) in terms of runtime, labeling time, and number of iterations.

The region-based algorithms query a significantly greater overall number of samples and are computationally more complex than pointwise queries, leading to longer runtimes per iteration. However, since the overall number of iterations is lower when querying a limited number of spatial batches, significantly less overall computational runtime is required for the same accuracy level [see Fig. 7(a)].

Multiplying the mean time requirements for pointwise labeling and marker-based labeling of regions [see Fig. 6(a) and (b)] with the number of iterations, the time expenditure of a user of the proposed methods is estimated. Fig. 7(b) shows that all three region-based queries \( (\text{QBC}_{R_P}, \text{QBC}_{R_{DH}}, \) and \( \text{QBC}_{R_{S}} \)) yield steeper learning curves at all window sizes and result in a reduction in labeling time compared with \( \text{QBC}_P \). Depending on the window size, the average labeling time was reduced from 37.9\% \((w = 200 \, \text{m})\) to 72.2\% \((w = 60 \, \text{m})\) when comparing the learning curves of \( \text{QBC}_{R_{DH}} \) to the pointwise heuristic \( \text{QBC}_P \) over all accuracy levels. At intermediate window sizes \((20 \, \text{m} < w < 200 \, \text{m})\), these gains tended to be more important (from 57.9\% to 72.2\%).

It is shown in Fig. 7(a) and (b) that at \( w = 200 \, \text{m} \), the region-based heuristics quickly reach a plateau, whereas \( \text{QBC}_P \) approximates and slightly exceeds the same accuracy level in later iterations. Consequently, the proposed methods still yield important overall time savings in early iterations but provide slightly lower final accuracy values when selecting a large window size. This is probably because a large window size \((w = 200 \, \text{m} \), diagonal \( = 283 \, \text{m} \)) significantly beyond the range of spatial autocorrelation \((143 \, \text{m} \); see Fig. 5) does not help increase the variance captured per query region. The fact that the batch size (window size) cannot be arbitrarily increased without sacrificing the performance of the AL model is consistent with other studies using batch-mode AL with uncertainty and diversity criteria on remote sensing data sets [18], [19] and points toward a lower sample variance per batch when the ratio between \( |X| \) and the number of model updates (iterations) decreases.

Since larger sampling windows result in fewer iterations and, consequently, lower computational runtimes, changing the window size constitutes a tradeoff parameter that enables the

| Table I: Overview of the Data Sets Resulting From Image Segmentation and Feature Extraction |
|---|---|---|---|
| Brazil | China |
| landslide | non-landslide | landslide | non-landslide |
| spectral features (post-event Geoye-I/IKonos-2) | 10 | 9 |
| texture features (post-event Geoye-1/IKonos-2) | 65 | 56 |
| shape features (post-event Geoye-1/IKonos-2) | 7 | 7 |
| contrast to neighbours (post-event Geoye-1/IKonos-2) | 6 | 12 |
| spectral features (pre-event Geoye-1) | 7 | - |
| spectral features (pre-event SPOT like) | 7 | 7 |
| spectral features (pre-event Landsat 5) | 11 | 5 |
| topographic features | 6 | - |

<table>
<thead>
<tr>
<th>Table II: Overview of the Different Tested Postdisaster Scenarios at the Brazil Study Site</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description</td>
</tr>
<tr>
<td>Scenario 1 Post-event VHR image + topographic data</td>
</tr>
<tr>
<td>Scenario 2 Post-event VHR image + pre-event Landsat type image + topographic data</td>
</tr>
<tr>
<td>Scenario 3 Post-event VHR image + pre-event Sentinel/SPOT type image + topographic data</td>
</tr>
<tr>
<td>Scenario 4 Post-event VHR image + pre-event VHR image + topographic data</td>
</tr>
</tbody>
</table>
Fig. 7. Learning curves for comparison between the baseline sampling methods (SPCOSA and \(QBC_P\)) and the proposed region-based queries (\(QBC_{R\mu}\), \(QBC_{RD\sigma}\), and \(QBC_{RDH}\)) on the Brazil data set for three of the five tested window sizes. Accuracy gains are plotted with respect to (a) the computational runtime, (b) the labeling time derived by multiplying mean labeling times as detailed in Section V-A with the number of iterations, and (c) the number of iterations (early iterations, including 20 additional runs for the estimation of mean and variance). The mean time reduction rate was computed over all accuracy levels as the mean difference between the \(QBC_{RDH}\)-curve and the nearest corresponding points on the \(QBC_P\)-curve. For better visibility, the learning curves for SPCOSA are only presented in the first column.

user to choose between more frequent labeling on small batches (smaller window size) or accepting longer waiting times between the labeling of fewer larger and more time consuming but fewer spatial batches (larger window size). In the experiments, a good tradeoff was provided by \(w = 100\) m (diagonal = 141 m), which approximately coincides with the mean effective range of the autocorrelation measured on the gray values of the image bands (143 m; see Fig. 5).

Regarding performance differences between \(QBC_{R\mu}\), \(QBC_{RD\sigma}\), and \(QBC_{RDH}\), Fig. 7(c) shows the learning curves for the early iterations from 30 random seeded runs. At the smallest window size \((w = 20)\) \(QBC_{RDH}\) and \(QBC_{R\mu}\) yield very similar learning curves, whereas \(QBC_{RD\sigma}\) performs the worst over the first 20 iterations. This suggests that \(\sigma_D\) becomes a less robust diversity measure when computed with fewer samples in the training and candidate set. In addition, \(QBC_{RDH}\) appears to be more robust to changes in the window size and outperformed \(QBC_{R\mu}\) and \(QBC_{RD\sigma}\), particularly at \(w = \{150, 200\}\) m.

On the one hand, the integration of an explicit diversity criteria enables further reduction in the labeling costs, whereas on the other hand, the higher computational complexity of \(QBC_{RD\sigma}\) and \(QBC_{RDH}\) results in longer runtimes when compared with the simpler \(QBC_{R\mu}\) heuristic [see Fig. 7(a)].
Fig. 8(a) and (c). For the choice of queries sent to the user, which is an important issue for the different labeling and computational costs, it illustrates the considerably lower number of queries sent to the user, which is an important issue for the design of an AL system.

B. Influence of Parameters m and t_i and m in QBC_{RD}

Fig. 8 illustrates the performance of QBC_{RDσ} and QBC_{RDH} on the Brazil data set according to changes in the main parameters m and t_i. The AL heuristic, which does not consider the sample diversity (QBC_{Ru}), marks the lower bound of the obtained learning curves. This confirms that the proposed diversity criteria are useful in further reducing the labeling time in the early iterations of the training process.

Furthermore, it demonstrates that, within the evaluated parameter range, the performances of QBC_{RDσ} and QBC_{RDH} are at least competitive with QBC_{Ru}, no matter which values are set for m and t_i.

For the tested parameter values, both heuristics tend to perform better with a higher number of m candidate regions [see Fig. 8(a) and (c)]. For the choice of t_i, the results give no clear indication that discarding features with low variable importance yields any advantage. For QBC_{RDH}, the best performance is achieved using all available features in each iteration for the computation of the cross-entropy (see Fig. 8(d); t_i = 0.00). This suggests the use of a simpler version of QBC_{RDH} without the computation of the variable importance and parameter t_i. However, the remaining experiments are carried out with the default settings introduced above (m = 3; t_i = 0.01) to not positively bias the final results.

C. Performance on Different Data Sets

Outcomes of the evaluation of SPCOSA, QBC_{P}, QBC_{Ru}, and QBC_{RDH} on the China data set are presented in Fig. 9 and show a significantly better learning performance in terms of runtime and labeling time of the proposed region-based methods than SPCOSA and QBC_{P}. In terms of labeling time, QBC_{RDH} reduces the required time (mean time reduction over all accuracy levels) by 56% when compared with the pointwise queries [see Fig. 9(b)] and outperforms the purely uncertainty-based QBC_{Ru} method. QBC_{RDH} also yields the highest overall accuracy resulting in F = 0.76. Due to its lower complexity, QBC_{Ru} provides slight advantages in terms of computational runtime [see Fig. 9(a)]. The performance increase with the developed region-based methods is consistent with the experimental results on the Brazil data set.

Fig. 10 displays the results of the two region-based AL heuristics that integrate diversity criteria (QBC_{RDσ} and QBC_{RDH}) when tested in the four different scenarios for the Brazil data set. Unsurprisingly, both algorithms reach the best performance when pre- and post-event images are VHR (scenario 4). Degrading the pre-event images to a 10-m resolution (scenario 3) yielded a rather small loss in accuracy (F-measure), which was, on average, less than 1% beyond the tenth iteration. The use of medium resolution (scenario 2) or the complete absence of pre-event imagery (scenario 1), however, led to a significant loss in accuracy, which was, on average, 5.5% and 5.7% beyond the tenth iteration.

Within the first ten iterations, QBC_{RDH} yielded greater improvement (steeper learning curve) from scenario 4 to scenario 3 compared to QBC_{RDσ}. This demonstrates that QBC_{RDH} heuristics exploit the additionally available data more...
The displayed results were obtained applying with (b) regular bootstrap sampling and (c) stratified bootstrap sampling where the majority class is downsampled to the size of the minority class in each iteration. Heuristics could be noted. In general, the results suggest that both AL significantly performance difference between the two heuristics efficiently in the first iterations. Beyond the tenth iteration, no sampling observed in all experiments.

Fig. 11. (a) Ratio of landslide and nonlandslide samples ($\beta$) in the training set queried with regular and stratified bootstrap sampling. Effects of class imbalance with (b) regular bootstrap sampling and (c) stratified bootstrap sampling where the majority class is downsampling to the size of the minority class in each iteration. The displayed results were obtained applying $QBC_{RDH}^\infty$ at $w = 100$ m on the Brazil data set but are representative for positive effects of stratified bootstrap sampling observed in all experiments.

D. Effects of Class Imbalance and Stratified Bootstrap Sampling

Class imbalance is a frequent challenge for the application and design of machine learning algorithms [54] and intrinsic to landslide mapping where affected areas usually cover only minor fractions of the landscape.

In certain cases, iterative resampling can be effective to approximate a class balance in the training set that will lead to a balance of user’s and producer’s accuracy on the test set [9]. The underlying assumption is that the training set was sampled at random and, therefore, represents well the underlying class distribution. However, this does not typically apply to a training set obtained with an AL algorithm, since the queries are intentionally biased toward the regions in feature space where class overlap and class imbalance become less pronounced. It has been previously reported that this bias typically yields a more balanced training sample and can alleviate the class imbalance problem [54].

Fig. 11(a) illustrates that this also applies in our experiments, where the class imbalance in the queried training sets is considerably lower than in the full data sets. It is observed that the queries are particularly directed toward regions with abundant occurrences of landslides, which must be regarded as a positive attribute of the AL heuristics. Fig. 11(a) also shows that class imbalance in the training set becomes slightly higher when using stratified bootstrap sampling [63] for the tree construction (all described experiments).

Fig. 11(b) displays the results of $QBC_{RDH}^\infty$ when using regular bootstrap sampling, yielding a considerable bias toward the nonlandslide class, an underdetection of the affected area, and a relatively low producer’s accuracy. Integrating stratified bootstrap sampling in the tree construction during the AL iterations leads to convergence during the learning process and provides a significantly higher F-measure than training with the regular bootstrap sampling [see Fig. 11(c)]. The inversion of the balance between user’s and producer’s accuracy when stratified bootstrap sampling is applied can be explained by the combined effect of queries and downsampling, which yields an overrepresentation of the landslide class in the overlap region of the two classes in feature space [64].

E. Spatial Distribution of Errors and Uncertainties

To investigate the causes of the remaining classification errors, 40 iterations of $QBC_{RDH}^\infty$ are performed with the initially chosen standard parameter settings ($m = 3; t_i = 0.01; w = 100$ m) on the Brazil data set. The resulting landslide map and related errors are displayed in Fig. 12(a). The spatial batches queried during 40 iterations [see Fig. 12(c)] cover 0.4 km$^2$ and correspond to 3.8% of the entire study area. The achieved accuracy values [see Fig. 12(b)] are consistent with the mean accuracy values estimated through cross-validation in previous experiments, and $F = 0.73$ is above the lowest ($F = 0.71$) and below the highest ($F = 0.86$) interexpert agreement rate. Assuming that all queried samples were correctly labeled, the accuracy of the full map would be $F = 0.81$. In this context, it is important to note that a pointwise labeled training set contributes significantly less to the overall accuracy of the entire map.

The obtained landslide map is used to analyze the distribution of the classification errors in dependence of a few factors that are suspected to favor misclassification, namely, proximity to the landslide boundary, occurrence of shadows, and certain terrain positions. Fig. 12(a) illustrates that errors are clustered in certain regions and particularly along landslide boundaries. This is illustrated through quantitative comparison of the frequency density distributions of the distances to...
Fig. 12. (a) Final map results for the Brazil study site after 40 iterations with $QBC_{RDH} (m = 3, \tau = 0.01, \text{and } w = 100 \text{ m}).$ (b) Accuracy measure for the classification of the test set and the full map, assuming that all queried samples have been correctly labeled. (c) All 40 regions queried for training.

landslide boundaries of false positives (FPs), false negatives (FNs), and all landslide samples [see Fig. 13(a)]. FNs concentrate along the boundaries of larger clearly visible landslides and, while introducing errors in the overall area, do not typically hinder the detection of the respective landslides. On the contrary, FPs are rather located some distant from actual landslides.

The presence of shadows [see Fig. 13(b)] could not be clearly related to an increased misclassification rate, whereas FPs had a slightly higher object brightness related to infrastructure and other anthropogenic features with high reflectance. Many of such FPs are located in the upper parts of the slopes, whereas FN errors display slight prevalence at the lower slopes [see Fig. 13(c)].

Analogous to the ensemble vote entropy (1), the five available expert mappings [see Fig. 4(a)] enable the computation of a map of the vote entropy of expert geomorphologists. The vote entropy of the experts can be considered as a measure for the uncertainty of the reference data, and it is interesting to compare this inherent uncertainty with the uncertainty of the RF classifier and the distribution of errors. Fig. 13(d) clearly shows that image objects with high expert vote entropy are also much more frequently evaluated as uncertain by the classifier. Fig. 13(e) complements this picture, showing that most of the correctly classified objects coincide with areas of low expert uncertainty, whereas FNs are particularly concentrated in areas with high expert vote entropy. It can be inferred that the true class of the latter is highly uncertain, and efforts to increase classification accuracy for those samples will be prone to failure because an improvement with respect to one expert map will result in a performance decrease with respect to the opposing opinion of another expert. On the contrary, more than half of the FPs overlap with areas where all experts agreed [see Fig. 13(e)]. The fact that all experts agreed on those samples suggests that they exploit relatively unambiguous image features in their judgement, making those samples interesting candidates for the design of enhanced object features. This group of samples might also be more reliable to evaluate algorithmic improvements since class membership does not depend on the respective ground truth.

VII. CONCLUSION

This work was dedicated to the integration of spatial constraints in the design of an AL algorithm to avoid a spatially dispersed distribution of queried samples and consequently reduce the sampling costs for remote sensing applications. Using the RF classifier and the QBC (QPC) framework, three different AL methods are proposed to select interesting regions rather than individual sampling points. The developed heuristics make use of criteria for sample uncertainty and diversity and are evaluated for landslide mapping for two reference data sets using mono- and multitemporal images from different satellite sensors. To address class imbalance, which is an inherent issue for landslide mapping and many other remote sensing applications, all region-based AL heuristics are implemented using stratified bootstrap sampling for the construction of individual trees in the ensemble. This strategy helps achieve significantly
better balance between the user’s and the producer’s accuracy and, furthermore, reduce the computational time for training and classification.

The overall performance is quantified by accuracy gains (F-measure) in dependency of the number of iterations, runtime, and labeling time and compared with spatial coverage sampling and a pointwise query function. The comparison demonstrated that querying compact spatial batches significantly reduced the required number of iterations and, therefore, the labeling time and the computational runtime. The size of the spatial search window can be regarded as a tradeoff parameter that controls the size of the queried batches. A semivariogram analysis is adopted to assess the range of the spatial autocorrelation as an a priori estimate of potentially suitable window sizes. A systematic assessment indicates that region-based heuristics are efficient over a range of different window sizes. More favorable results are achieved with window sizes close to the mean effective range, whereas significantly larger window sizes rather lead to performance decay.

Two of the proposed region-based AL heuristics combined \(QBC_{RD_{\sigma}}\) and \(QBC_{RDH}\) sample uncertainty with additional measures to select the most diverse batch among uncertain candidate regions. Particularly in early iterations, they frequently provide steeper learning curves than purely uncertainty-based heuristics \(QBC_{RU}\). \(QBC_{RD_{\sigma}}\) and \(QBC_{RDH}\) comprise two tuning parameters, namely, the number of candidate regions \(m\) and a variable importance threshold \(t_i\). The impact of parameter changes on the algorithm performance is tested through experimentation showing that both algorithms benefited from a higher number of candidate regions \((m > 2)\), whereas the omission of less important features \((t_i > 0.0)\) does not lead to an enhanced performance. In terms of required labeling time, \(QBC_{RDH}\) provides the best performance (particularly in early AL iterations), whereas the purely uncertainty-based \(QBC_{RU}\) requires less computational time and provides competitive results (particularly in later AL iterations). In comparison to pointwise queries, \(QBC_{RDH}\) helps reduce labeling time by 56.0% on the China data set and by up to 78.2% on the Brazil data set.

The results of 40 iterations of \(QBC_{RDH}\) on the Brazil data set are examined to gain a better understanding of the causes and distribution of remaining errors. Querying 3.8% of the study area for training provided a test set error of \(F = 0.73\). Remaining errors are found to be spatially clustered and particularly FNs concentrated at the boundaries of manually mapped landslides. Although it is well understood that reference data sets from experts and other sources typically bear considerable uncertainties, their impact on the evaluation of image classification and feature detection still remains unconsidered in most remote sensing studies [53], [65]. The availability of five expert inventories at the Brazil study site enables comparison of the uncertainty of the expert judgment with the uncertainty of the classification and the distribution of remaining errors.

A strong positive relationship between expert and classifier uncertainty is found as well as prevalence of FNs in areas where experts disagreed, showing that the corresponding samples are inherently difficult and should be considered as uncertain rather than as real errors. This issue and the fact that the classification accuracy already exceeded the agreement of the most disagreeing experts suggest that further studies must address uncertainties in the ground truth to achieve real enhancements for similar types of problems.

The region-based AL methods reduce labeling time when sampling through a visual interpretation of VHR images and are also a valuable tool to guide field sampling surveys that typically bear significantly higher costs for the analysis of low- and medium-resolution images. The employed criteria for sample uncertainty and diversity can be easily adapted for applications to multiclass problems, and the use of the proposed framework for applications such as land cover mapping could be an interesting direction for further research.

Acknowledgment

The authors would like to thank T. A. Lampert for proofreading and the fruitful discussions on ground truth uncertainty, and B. Koenig for his support in the elaboration of the landslide inventory maps.

References

André Stumpf received the Diploma in physical geography and geomatics from the Dresden University of Technology, Dresden, Germany, in 2009. He is currently working toward the Ph.D. degree in the Laboratoire Image, Ville, Environnement (LIVE), University of Strasbourg, Strasbourg, France.

His Ph.D. research is a joint research program between the LIVE and the Institut de Physique du Globe de Strasbourg, both at the University of Strasbourg, and the Faculty of Geo-Information Science and Earth Observation (ITC) at the University of Twente, Enschede, The Netherlands. The focus is on the development of image analysis techniques for the mapping and monitoring of landslide with optical remote sensing data. His principal research interests include image processing, machine learning, geostatistics, and geomorphological hazards.

Nicolas Lachiche received the Ph.D. degree in computer science from the University of Nancy, Nancy, France, in 1997. From 1997 to 1999, he was a Research Associate with the University of Bristol, Bristol, U.K. He is a Senior Lecturer in computer science with the University of Strasbourg, Strasbourg, U.K. His research interests include machine learning and data mining, particularly cost-sensitive learning and relational data mining, and their applications to other sciences, specifically bioinformatics, chemistry, and geography.

Jean-Philippe Malet received the Ph.D. degree in earth sciences and engineering geology (landslide numerical modeling) from the University of Strasbourg, Strasbourg, France, in 2003. From 2004 to 2006, he was a Research Associate with Utrecht University, Utrecht, The Netherlands, working on physical small-scale modeling of geomorphological processes. He is currently a Senior Researcher in geomorphology with the French National Research Council (Centre National de la Recherche Scientifique) and with the University of Strasbourg. His research interests include mapping, monitoring, and numerical modeling of geohazards; assessment of induced risks; and development of geophysical investigation tools.

Anne Puissant received the Ph.D. degree in geography from the University of Strasbourg, Strasbourg, France, in 2003. She is currently an Assistant Professor of spatial analysis, geographic information systems, and remote sensing with the Department of Geography, University of Strasbourg. Her research topics focus on the utility of geoinformatics for landscape analysis (characterization and dynamics of urban or natural processes), particularly questions of knowledge extraction, spatial analysis, and remote sensing image processing.

Norman Kerle received the M.A. degrees in geography from The Ohio State University, Columbus, OH, USA, in 1997 and the University of Hamburg, Hamburg, Germany, in 1998 and the Ph.D. degree in geography (volcano remote sensing) from the University of Cambridge, Cambridge, U.K., in 2002. He is currently an Assistant Professor of disaster geoinformation management with the Department of Earth Systems Analysis, Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, Enschede, The Netherlands. His principal research interests include the utility of geoinformatics for disaster risk management, particularly questions of assessment of vulnerability, risk and postdisaster structural damage, and object-oriented image analysis methods.