ADAPTIVE SPATIAL SAMPLING WITH ACTIVE RANDOM FOREST FOR OBJECT-ORIENTED LANDSLIDE MAPPING

A. Stumpf [a, b, c]***, N. Kerle*** [c]***, A. Puissant*** [a]***, N. Lachiche*** [d]***, Jean-Philippe Malet*** [b]

[a] Laboratoire Image, Ville, Environnement, CNRS ERL 7230, Université de Strasbourg, 3 rue de l'Argonne, F- 67083 Strasbourg, France

[b] Institut de Physique du Globe de Strasbourg, CNRS UMR 7516, Université de Strasbourg / EOST, 5 rue René Descartes, F-67084 Strasbourg, France

[c] ITC-Faculty of Geo-Information Science and Earth Observation, University of Twente, Department of Earth Systems Analysis, Hengelosestraat 99, P.O. Box 6, Enschede, 7500 AA, The Netherlands

[d] Image Sciences, Computer Sciences and Remote Sensing Laboratory, CNRS UMR 7005, Université de Strasbourg, Bd Sébastien Brant - BP 10413, F-67412 Illkirch, France

1. INTRODUCTION

Landslide inventory mapping is indispensable for landslide hazard and risk- assessment [1], the quantification of erosion rates [2], seismic hazard assessment [3], and after large events triggered by earthquakes or heavy rainfalls crucial to provide information on affected areas in time for support disaster response. VHR remote sensing imagery to perform such tasks is now commonly available, whereas the development of robust operational techniques to accelerate the mapping process is still lacking behind. A number of recent studies addressed this problem developing object-oriented rule-based classifiers [4-7] that function without training data but may often require adjustments of the thresholds. Recently proposed supervised approaches fall into pixel-based studies using parametric classifiers [8] and approaches using object-oriented features and non-parametric learning algorithms [9]. While only object-oriented approaches can fully exploit the rich textural and spatial information content of VHR resolution images, both techniques achieved high mapping accuracies and but still require an extensive amount of training data. The acquisition of training data is typically associated with significant costs and an optimal training set should therefore be as small, while still representative for the target classes. In the domain of machine learning active learning has evolved as key concept to address such issues [10] and recently becomes more commonly used to reduce the labelling costs for remote sensing image classification of [11]. Active learning techniques have been adopted for the supervised segmentation of remote sensing images [12] but to the best of our knowledge yet no studies have been dedicated to the exploration of active learning heuristics for object-oriented image analysis. Furthermore, except a study presented by Liu et al. [13] the significance of spatial domain is often neglected in the sampling design leading in the worst case to queries with disperse spatial distribution and therefore inappropriate for an interactive labelling of a training set through field work or image interpretation. This study extends upon previous work that investigated the suitability of the Random Forest (RF) framework [14] for object-oriented mapping of landslides from VHR remote sensing images. The aim is to incorporate an active learning heuristic to focus the labelling process on most uncertain image section and test if the resulting training set provides accuracy enhancements when compared to standard stratified random sampling.

2. DATA AND METHODS

2.1. Test site and data

The study area is located in the Brazilian Serrana mountains north of Nova Friburgo in the Rio de Janeiro state. On 11-12 January 2011 the area was affected by heavy rainfalls (>250 mm) which triggered thousands of landslides and claimed more than 1500 victims [15]. Geoeye-1 images of the region were recorded on the 20 January and pre-event imagery from the same sensor was available for at the 26 May 2010. A reference inventory polygons (Fig.1 b) was digitized through visual interpretation of both images and in combination with topographic data. For all experiments a subset comprising an area of 9 km^2 was selected (Fig.1 a).

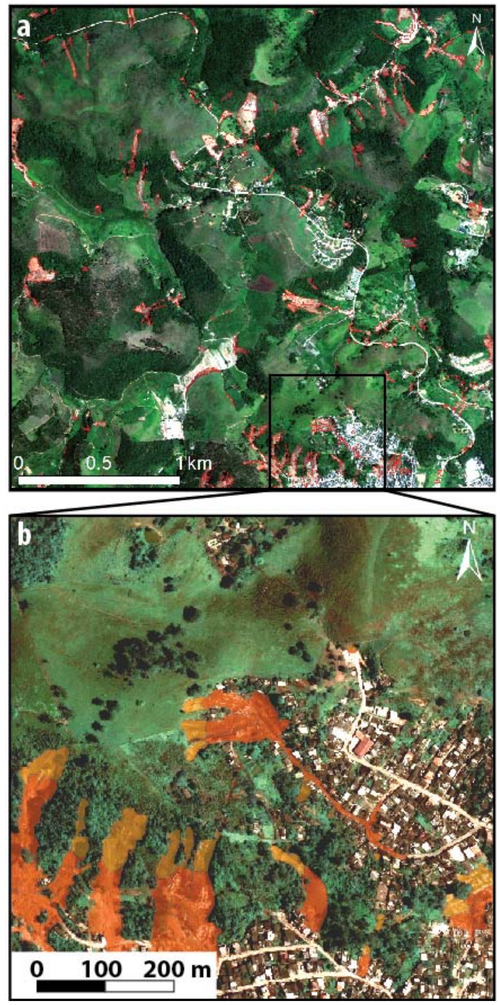


Fig. 1 (a) Post-disaster imagery of the study site. (b) Detailed view of the reference inventory (landslide source: light orange, track/deposit: dark orange) showing landslides that affected the suburbs of Nova Friburgo.

2.2. Image segmentation and feature extraction

Image segmentation was performed on the post-landslide image with equal weights of the panchromatic and multispectral bands using the eCognitions multi-resolution segmentation [16] with a scale factor of 10. For the resulting approximately 400.000 objects a set of 106 features describing intensity values, band ratios, texture, shape, neighborhood relationships and topographic location of each objects was calculated and the a reference class was assigned considering the overlap (majority vote) with the reference inventory.

2.3. Active Random Forest

The adopted active learning approach follows the query-by-committee strategy [17] where samples, which the committee is the most uncertain about, are repeatedly queried for labeling and added to the training sample. A RF with 500 trees was used as a base classifier. One possible measure for classification uncertainty of a RF is the vote entropy which is calculated as in Eq.1.

$$e_i = - \sum_{i \in (0,1)} p_i * \log(p_i) \qquad \text{Eq.1}$$

p_0 is the thereby the fraction of the trees that label the sample as non-landslide and $p_1$the fraction that label the same sample as landslide. To avoid spatially disperse queries which are impractical for interactive annotation the algorithm also considers the spatial domain through region-based queries. For this purpose the mean e_i of all objects is calculated within a circular sliding window, all objects from the region with the lowest e_i on the current map (Fig. 2 b) are added to the training set , and the classifier is retrained. For the presented experiments we considered a circular sampling area with radius of 100 m. This extent can be adjusted according to the accessibility of the terrain and the requirements of the user, and controls the size of the batch added to the training sample in each iteration. This adaptive strategy was tested over ten iteration and compared to random queries on a spatial coverage stratification [18] (Fig. 2 a). The first sampling area was forced to contain at least one landslide sample since both techniques require examples from both classes to initiate. In each iteration (n=10) the classification accuracy was assessed on the remaining datasets not used for training. To also assess the stability of the learning strategies both routines were repeated ten times.

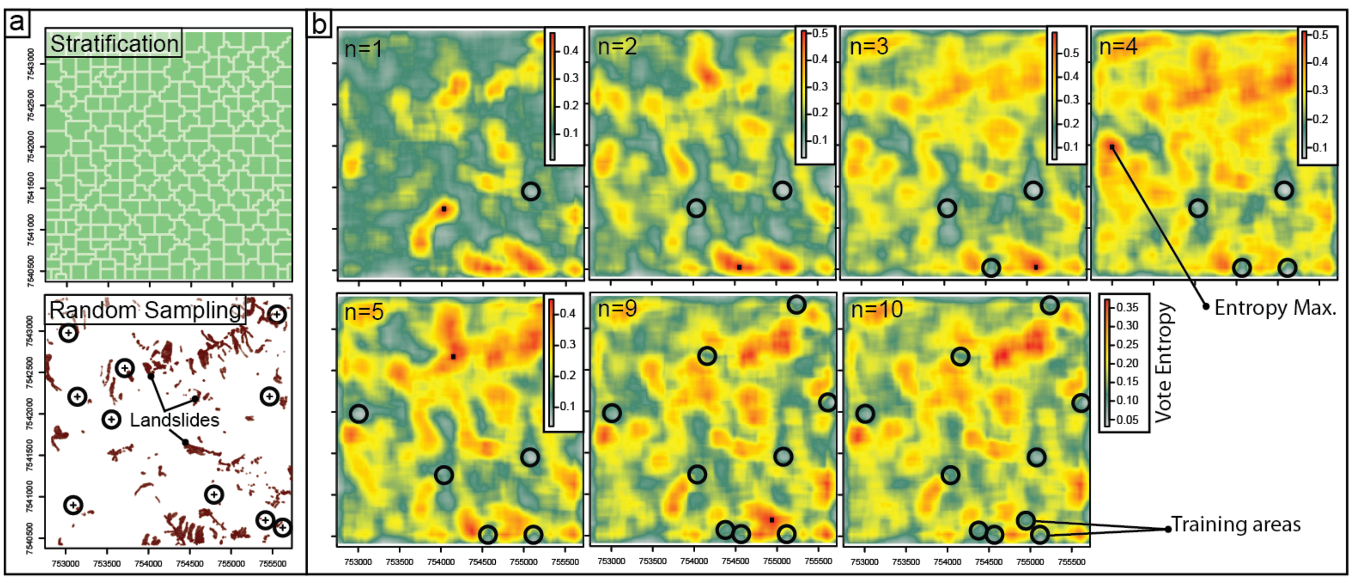
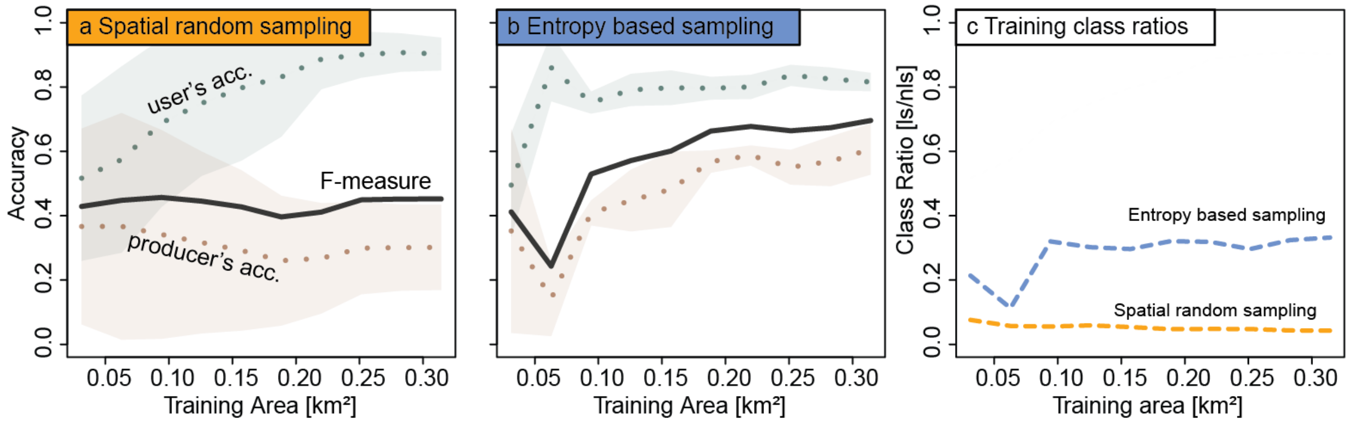


Fig. 2 (a) Spatial stratification and random selection of 10 sampling areas. (b) Iterative entropy-based selection of 10 sampling areas.

3. RESULTS

Fig, 3 demonstrates that the proposed active sampling strategy yields steeper learning curves and higher classification accuracies with lower variance than spatial random sampling. As landslide affected areas and similar image objects feature generally high classifier uncertainties (Fig. 2 b) the entropy-based heuristics also alleviates the class-imbalance problem because it focuses the labeling efforts on affected areas and thereby favors a more balanced class ratio within the training sample (Fig, 3 c). The proposed algorithm is simple and efficient to enhance mapping accuracies at reduced labeling cost while furthers enhancements may be possible considering the density and clustering of the data and an additional step for backward feature selection.



Fig, 3 Comparison of learning curves for (a) spatial coverage random sampling (b) entropy-based queries based on an active Random Forest and (c) development of the class ratios in the training samples of both routines.

11. REFERENCES

[1] C. J. van Westen, E. Castellanos, and S. L. Kuriakose, "Spatial data for landslide susceptibility, hazard, and vulnerability assessment: An overview," *Engineering Geology,* vol. 102, pp. 112-131, 2008.

[2] F. Guzzetti, F. Ardizzone, M. Cardinali, M. Rossi, and D. Valigi, "Landslide volumes and landslide mobilization rates in Umbria, central Italy," *Earth and Planetary Science Letters,* vol. 279, pp. 222-229, 2009.

[3] E. L. Harp, D. K. Keefer, H. P. Sato, and H. Yagi, "Landslide inventories: The essential part of seismic landslide hazard analyses," *Engineering Geology,* 2010.

[4] T. Martha, N. Kerle, C. J. van Westen, and K. Kumar, "Characterising spectral, spatial and morphometric properties of landslides for semi-automatic detection using object-oriented methods," *Geomorphology,* vol. 116, pp. 24-36 2010.

[5] T. R. Martha, N. Kerle, C. J. van Westen, V. Jetten, and K. V. Kumar, "Segment Optimization and Data-Driven Thresholding for Knowledge-Based Landslide Detection by Object-Based Image Analysis," *Geoscience and Remote Sensing, IEEE Transactions on,* vol. 49, pp. 4928-4943, 2011.

[6] P. Lu, A. Stumpf, N. Kerle, and N. Casagli, "Object-Oriented Change Detection for Landslide Rapid Mapping," *Geoscience and Remote Sensing Letters, IEEE,* vol. PP, pp. 701-705, 2011.

[7] T. Lahousse, K. T. Chang, and Y. H. Lin, "Landslide mapping with multi-scale object-based image analysis – a case study in the Baichi watershed, Taiwan," *Nat. Hazards Earth Syst. Sci.,* vol. 11, pp. 2715-2726, 2011.

[8] A. C. Mondini, F. Guzzetti, P. Reichenbach, M. Rossi, M. Cardinali, and F. Ardizzone, "Semi-automatic recognition and mapping of rainfall induced shallow landslides using optical satellite images," *Remote Sensing of Environment,* vol. 115, pp. 1743-1757, 2011.

[9] A. Stumpf and N. Kerle, "Object-oriented mapping of landslides using Random Forests," *Remote Sensing of Environment,* vol. 115, pp. 2564-2577, 2011.

[10] B. Settles, "Active Learning Literature Survey," University of Wisconsin–Madison2010.

[11] D. Tuia, M. Volpi, L. Copa, M. Kanevski, and J. Munoz-Mari, "A survey of active learning algorithms for supervised remote sensing image classification," *Selected Topics in Signal Processing, IEEE Journal of,* vol. PP, pp. 1-1, 2011.

[12] P. Mitra, B. Uma Shankar, and S. K. Pal, "Segmentation of multispectral remote sensing images using active support vector machines," *Pattern Recognition Letters,* vol. 25, pp. 1067-1074, 2004.

[13] A. Liu, G. Jun, and J. Gho, "Spatially Cost-sensitive Active Learning," presented at the Ninth SIAM International Conference on Data Mining Sparks, Nevada, 2009.

[14] L. Breiman, "Random Forests," *Machine Learning,* vol. 45, pp. 5-32, 2001.

[15] A. L. C. Netto, A. M. Sato, A. d. S. Avelar, L. G. G. Vianna, I. S. Araújo, D. L. C. Ferreira, P. H. Lima, A. P. A. Silva, and R. P. Silva, "January 2011: the extreme landslide disaster in Brazil," presented at the Second World Landslide Forum, Rome, Italy, 2011.

[16] M. Baatz and A. Schäpe, "Multiresolution Segmentation – an optimization approach for high quality multi-scale image segmentation," presented at the Angewandte Geographische Informationsverarbeitung XII, Salzburg, 2000.

[17] Y. Freund, H. S. Seung, E. Shamir, and N. Tishby, "Selective Sampling Using the Query by Committee Algorithm," *Machine Learning,* vol. 28, pp. 133-168, 1997.

[18] D. J. J. Walvoort, D. J. Brus, and J. J. de Gruijter, "An R package for spatial coverage sampling and random sampling from compact geographical strata by k-means," *Computers & Geosciences,* vol. 36, pp. 1261-1267, 2010.