Available online at www.sciencedirect.com

ScienceDirect

Procedia Environmental Sciences 3 (2011) 123–129

Procedia Environmental Sciences

Combining Random Forests and object-oriented analysis for landslide mapping from very high resolution imagery

André Stumpf [a,b], Norman Kerle [a]

[a] *Faculty of Geo-Information Science and Earth Observation-ITC, University of Twente, Hengelose Straat 99, P.O.Box 6, 7500 AA, The Netherlands*

[b] *Laboratoire Image, Ville, Environnement - ERL 7230 CNRS, Université de Strasbourg, 3 rue de l'Argonne 67000 Srasbourg, France*

Abstract

The increasing availability of very high resolution (VHR) remote sensing images has been leading to new opportunities for the cartography of landslides in risk management and disaster response. Object-oriented image analysis has become one of the key-concepts to better exploit additional spatial, spectral and contextual information. The multitude of additional object attributes calls for the use of advanced data mining and machine learning tools to identify the most suitable features and handle the non-linear classification task. In this study we used the Random Forest algorithm for the selection of useful features and object classification in the context of landslide mapping. A workflow for image segmentation, feature extraction, feature selection and classification was developed and tested with multi-sensor optical imagery from four different test sites. Due to class imbalance and class overlap between landslide and non-landslide areas the classifier can be heavily biased towards over- and under-prediction of the affected areas. This is a common issue for many real-world applications and a procedure to estimate a well-adjusted class ratio from the training samples was designed and tested. A number of potentially useful object metrics was evaluated and it was demonstrated that topographically guided texture measures provide significant enhancements. Employing 20 % of the image objects for training accuracies between 73.3 % and 87.1 % were achieved at four different test sites.

Selection and/or peer-review under responsibility of [name organizer]

Keywords: landslide mapping; VHR satellite images; image segmentation; object-oriented image analysis; Random Forest

1. Introduction

Comprehensive landslide inventories are the most commonly used source for quantitative landslide hazard and risk assessment at regional scales (van Westen et al. 2006). Manual interpretation of aerial photographs and field work to date remains the most frequently followed approach for the elaboration of

doi:10.1016/j.proenv.2011.02.022

inventory maps in scientific studies and by administrative bodies. Despite its time-consuming and "labor-intensive" nature, however, results still include a large degree of subjectivity and may vary considerably among different experts (e.g. Galli et al. 2008).

A number of recent events in China (2008), Italy (2009), Haiti (2010), and Brazil (2011) illustrate the short-term availability of comprehensive VHR satellite images and strongly contrast the lack of reliable machine-aided mapping workflows. Previously proposed workflows for the analysis of optical data largely focused on the signals of individual pixel (e.g. Borghuis et al. 2007; Nichol and Wong 2005). More advanced object-oriented methods can make use of a manifold of additional image features such as texture, shape, topography and spatial context but proposed approaches rely on a the manual selection of suitable features and hard coded thresholds for object classification (Barlow et al. 2006; Lu et al. in press; Martha et al. 2010). Based on samples, image segmentation and the Random Forest (RF) (Breiman 2001) algorithm this study targeted the elaboration and testing of a workflow for feature selection and classifier training for landslide mapping on images from variety of state-of-the-art optical sensors.

2. Data and Methods

The analysed VHR imagery comprised an aerial photograph, IKONOS, Quickbird and Geoeye-1 of recently affected sites at Haiti, China, Italy and and France, respectively (Figure 1). Topographic datasets were available from various sources and resampled to a resolution of 10m to ensure consistency. Reference data were available through landslide inventories obtained from a careful interpretation the VHR remote sensing datasets by experts and all except the Haiti inventory were validated through in the field.

The multi-resolution segmentation algorithm (Baatz and Schäpe 2000) in eCognition 8.64 was used for image segmentation. An increase in the algorithm's scale factor corresponds to coarser image segmentation with larger, more heterogeneous objects and we tested 15 scale factors between 10 and 100. At each scale 96 object attributes, including color, texture, shape and topographic derivatives, were calculated. An initial screening for generally useful object-metrics was carried out taking into account all landslide objects (O_{LS}) and an equal amount of randomly sampled non-landslide objects (O_{NLS}). For each test site and scale the respective populations were analysed in a RF-based feature selection approach described in Diaz-Uriarte and Alvarez de Andres (2006).

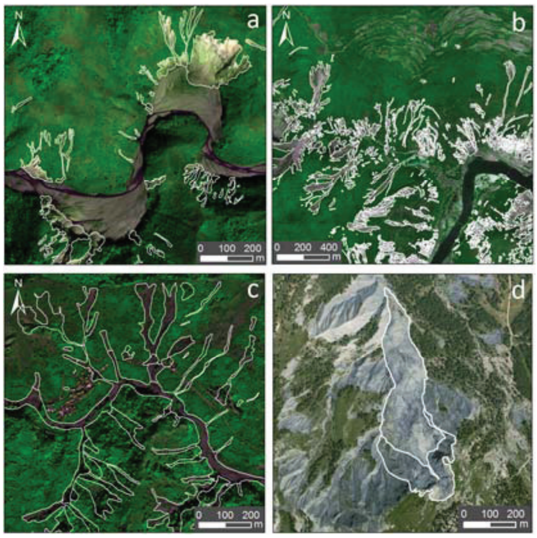

Figure 1: Analysed subsets at the different test sites. a) Momance River (Haiti) b) Wenchuan (China), c) Messina area, (Italy) d) Barcelonnette basin (France). White outlines indicate the landslide areas.

In a subsequent step a scenario with 20% of the data available for training was simulated in order to investigate the achievable classification accuracies. Due to the combined effects of class-overlap and class-imbalance classifiers can be heavily biased toward one of the classes (e.g. Denil and Trappenberg 2010). In the present study this would have caused an undesirable under- or over-estimation of the affected areas, if the original class-balance or an equally resampled class- balance would have been used, respectively. To account for such effects a procedure for the iterative estimation of class errors from subsets of the training sample was implemented and tested. Initially 20% of the data were sampled for training from the entire populations. This training sample (*Tr*) was then again split repeatedly into training (Tr_{sub}) and testing subsets (Te_{sub}) with iteratively increasing class-imbalances. In each cycle 20% of the O_{LS} and a stepwise increasing portion of O_{NLS} is sampled randomly and with replacement from *Tr* into Tr_{sub}. Starting from a Tr_{sub} with a balanced ratio of O_{LS} and O_{NLS} (β_i =1) a RF is trained and tested on the remainder subset Te_{sub}. In each following the ratio β_i was increased (β_i +0.1) by sampling more O_{NLS} into $Tr_{sub.}$ The underlying assumption was that a class ratio that provides a balanced error rate among subsets of the training data (further termed β_n) would also enhance the error balance for the entire population. To test the efficiency of this approach O_{NLS} in the original training sample *Tr* was undersampled according to β_n, the adjusted sample was used to train a RF and the classification error was assessed on the previously unseen 80% of the data. In an additional step entire training sets *Tr* were repeatedly sampled with altering class-balances to investigate if the estimated classification accuracies and learning curves were representative were representative for different subsets of the data.

3. Results and discussion

3.1. Feature selection

Color, topographic variables and topographically guided versions of Haralick's original texture measures (Haralick 1973) were ranked with a high variable importance at all test sites and segmentation scales. On average only about one third of the pre-selected metrics were detected as useful. The ranking of less important features, and especially the overall selected number, varied considerably among the different test sites and segmentation scales (Figure 2). The inclusion of shape metrics yielded very little accuracy enhancements. At larger segmentation scales they added some discriminant power but could not compensate a loss of fidelity that was observable for other features such as topographic variables. A RF classifier trained with 20% of all O_{LS} and an equal number of O_{NLS} yielded higher accuracies in the classification of unseen image objects if a reduced feature space was adopted (results not shown here). This was observed for all test sites and a representative subset of tested segmentation scales.

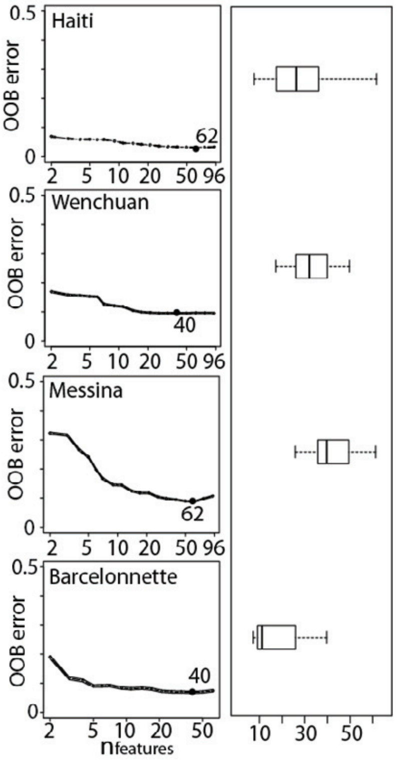

Figure 2: Feature selection history at the four test sites with the smallest segmentation scale. The out-of-bag (OOB) error is an error estimate intrinsic to the Random Forest approach (see Breiman 2001). The black dot indicates the variable combination with the smallest OOB error. Boxplots indicate the variability of the number of selected features among all 15 segmentation scales.

3.2. Estimates of optimal class-balance β_n from the training samples Tr

For all cases we observed a strong over-prediction of landslide areas if a class-balanced training sample (β_i=1) was employed, and a strong under-prediction if the natural class-balance was used for the training. The iterative procedure described in section 2 was useful to monitor the effects of the changing β_i on the user's and producer's accuracies that the RF achieved with subsets ($Tr_{sub,}$ Te_{sub}) of the training data (Figure 3). Subsequently, *Tr* (20% of all objects) was adjusted according to the estimated β_n by under-sampling the O_{NLS}. Although, β_n estimates did not solve the problem entirely they provided a significantly better balance between user's and producer's accuracies (Table 1) than could have been achieved with the natural class distribution or equally balanced training samples. The overall accuracy in terms of area and the balance of error rates was generally better with smaller scale factors leading to a finer segmentation.

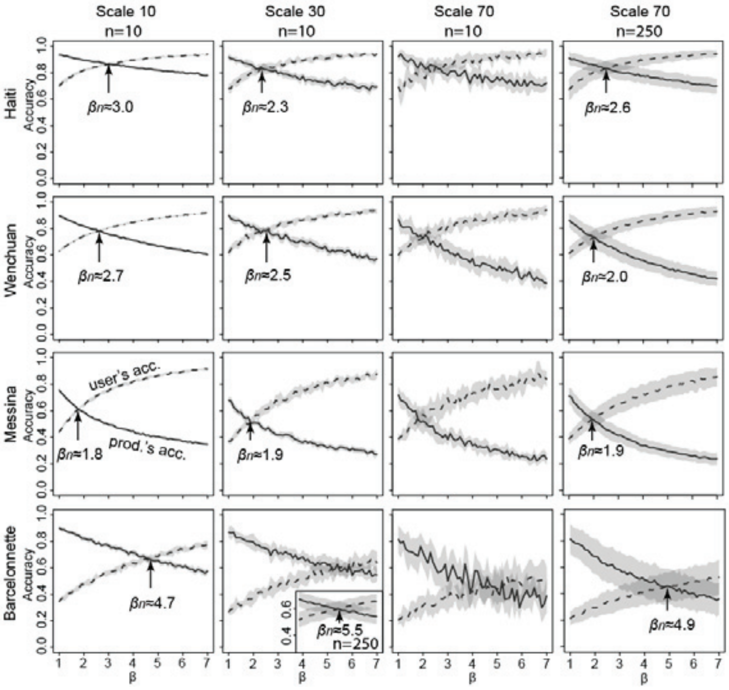

Figure 3: Estimates of the class balance (β_n) that lead to a balance of the mean user's (dashed black line) and mean producer's accuracies (solid black line) from iterative resampling of *Tr* into subsets. For each ratio value β the resampling of *Tr* into Tr_{sub} and Te_{sub} was repeated ten times. Each time a RF classifier was trained and tested and the mean accuracies for each β were calculate as the average over all replication runs (n=10). The grey margins show the corresponding standard deviations. For learning curves with high variance (Barcelonnette at a scale 30, all at a scale of 70) additional figures from 250 replicate runs (n=250) are presented.

Table 1: Accuracy assessment for all test sites at three exemplary segmentation scales. RFs (n_{trees} = 500), trained with 20% of the landslide objects (O_{LS}) and β_n -fold amount of non-landslide objects ($O_{\overline{LS}}$) were tested. The mean accuracies from 50 replicated runs are provided. β_o indicates the natural distribution of the classes. The F-measure is commonly defined as the harmonic mean of user's and producer's accuracy and can be calculated in terms of correctly classified objects ($F_{objects}$) or area (F_{area}).

									ple *all* *jects*
									1.7
Haiti	30	2.3 (4.2)	0.828	0.871	0.849	0.883	564	1297	12.8
	70	2.6 (4.0)	0.885	0.724	0.796	0.885	149	387	14.3
	10	2.7 (3.4)	**0.813**	**0.811**	**0.812**	0.805	6535	17645	17.0
Wenchuan	30	2.5 (3.0)	0.812	0.771	0.791	0.803	570	1425	17.4
	70	2.0 (2.6)	0.777	0.753	0.765	0.799	125	250	16.5
	10	1.8 (4.2)	**0.729**	**0.746**	**0.737**	0.730	6135	11043	10.8
Messina	30	1.9 (4.1)	0.690	0.609	0.647	0.592	663	1260	11.3
	70	1.9 (3.7)	0.643	0.598	0.620	0.605	125	238	11.9
	10	4.7 (9.5)	**0.778**	**0.780**	**0.779**	0.765	1810	8507	10.8
Barcelonnette	30	5.5 (11.5)	0.747	0.759	0.752	0.674	237	1304	10.1
	70	4.9 (12.1)	0.633	0.886	0.733	0.653	46	226	8.9

3.3. Stability of the classification accuracy and impacts of scale, sample-balance and sample-size

To this point we already demonstrated that, with small segmentation scales and 20% of labeled data, the RF classifier can efficiently distinguish objects representing landslide affected areas. Furthermore, through an iterative resampling of the training data it is possible to approximate balanced errors of commission and omission. To further explore the limits of the proposed method a second experiment was conducted. Unlike previously not only one, but several sets training set *Tr* were repeatedly sampled form the entire population. For each *Tr* 20% of all O_{LS} and a stepwise increasing β-fold amount of O_{NLS} were sampled from the entire population. Because each of the resulting training sets included different subsets, class-balances and overall numbers of objects (depending on class balance and scale) they could be used to analyze the overall sensitivity of the RF classifier to combined effects of those changes.

The graphs in Figure 4 show the results of all simulations with increasing class-imbalances and hence overall object number in the training sample. For a small scale factor (10) the learning curves for the entire population are shown and reproduce the same β-ratios as previously estimated on the training data (Figure 3). However, at larger scales the β_n-estimates increasingly deviated from the ratios that would have actually led to balanced error rates. The increasing uncertainties must be attributed to the strongly reduced amount of the training objects at larger scales (resulting from an increased object size). For the Barcelonnette dataset for example a scale of 70 corresponds to an overall number of 3066 objects, consequently a training set of 613 objects, and hence training subsets of only about 18-75 for the estimation of β.

The overall accuracy of the classification can be measured in terms of correctly classified objects (Table 1, Figure 4, $F_{objects}$) and correctly classified area (Table 1, Figure 4, F_{area}). For the Haiti and Wenchuan test sites Figure 4 shows an efficient and relatively stable performance of the RF classifier, which is indicated by high $F_{objects}$ and relatively little dependencies on the segmentation scale and the amount and composition of training data. At both test sites F_{area} generally decreased for larger scales, which has to be attributed to an increasing mismatch of the image segments with the reference inventories. On the other hand, at the Messina and Barcelonnette test sites both the classifier performance and the areal accuracy decreased with larger scale factors. The decrease in the absolute object number at

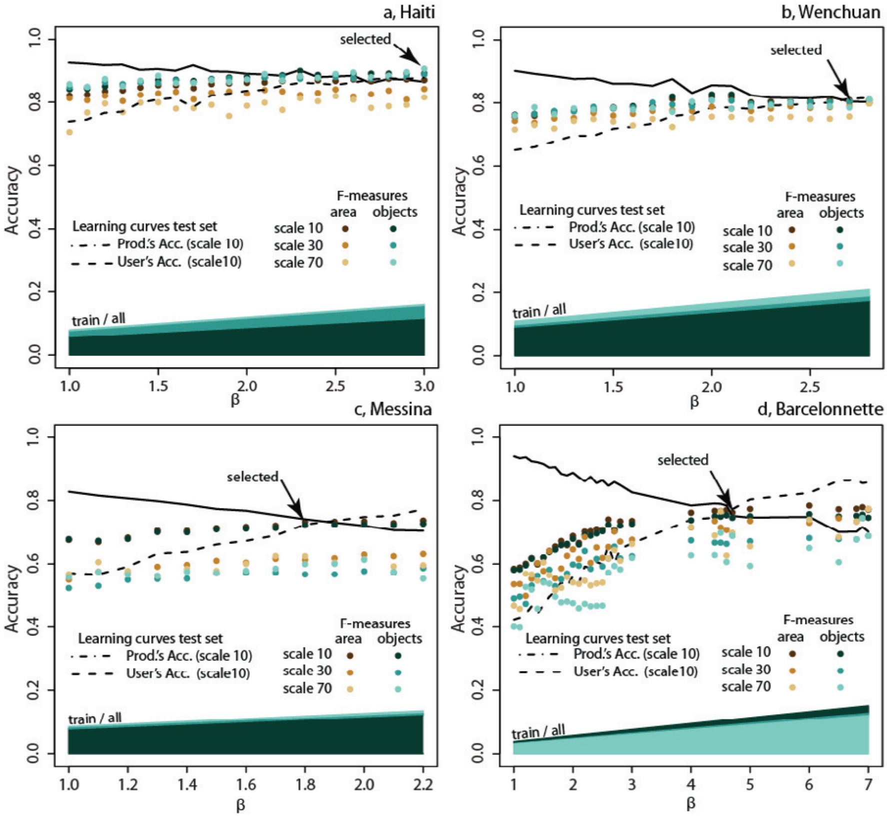

Figure 4: Stability of the classification accuracy depending on the class-balance, the percentage of training data, and the segmentation scale. The arrow indicates the β-value previously estimated from the training sample. The F-measure is commonly defined as the harmonic mean of user's and producer's accuracy. The vertical scattering of blue dots ($F_{objects}$) provides an indicator of the sensitivity of the classifier. The brown dots indicate the areal (F_{area}) accuracy and combine effects of segmentation and classification. Due to an increasing ratio of O_{NLS} (β) in the training sample the overall percentage of data used for training increases as well. The increase is indicated by the increasing bars at the bottom of the graphs for the three different scales, respectively.

with larger scales is similar as for the two other test sites (compare Table 1) and hence, the decreasing classifier performance must be rather attributed to a stronger class overlap in Messina and Barcelonnette datasets. This is also reflected by generally lower accuracies, and apparently increases with higher intra-segment heterogeneity at larger scales.

Unsurprisingly, adding further O_{NLS} to the training sample led in all cases to an increase in the overall accuracy. However, the trend saturated relatively quickly after balanced error rates were reached. In a real-world scenario the costs for the collection of further O_{NLS} may not be justified beyond that point. In summary, the RF classifier provided relatively high accuracies of up to 87% for the Haiti and Wenchuan test sites, while in the case of Messina the best model reached an accuracy of 73% (Figure 4, Table 1). Small segmentation scales and an availability of labeled training data in the range of 20% of the entire dataset led to a stable and efficient classification at all tested sites. Less training data might be sufficient as long as color, texture and morphology of the landslides are relatively unique within a given area, but for a more complex scene (Barcelonnette, Figure 4) we observed serious degradations of the results of if less than 10% of the data was employed for training.

4. Conclusion and future directions

Several factors such as the scene characteristics, selected object-metrics, class-imbalance and -overlap and the amount of available training data influence the correct recognition of landslide affected areas. We found that RF-based feature reduction enhanced the classifier performance in terms of accuracy and speed. Additional object-metrics, such as texture and auxiliary topographic data, are capable of considerably reducing the confusion among classes. Balanced error rates are desirable for landslide inventory mapping because over- or under-prediction of affected areas would lead to systematic over- and under-estimation of the hazard. A procedure to estimate the optimal class balance from the training sample was proposed in this study and demonstrated to enhance the balance of user's and producer's accuracy significantly.

Achieved accuracies are in a similar range as the results of other recent studies on landslide mapping from optical imagery (Barlow et al. 2006; Lu et al. in press; Martha et al. 2010). Though the quantities of employed samples are not always explicitly mentioned (Barlow et al. 2006; Borghuis et al. 2007; Nichol and Wong 2005), most proposed solutions depend on the availability of some sort of training area to adjust the method. Once samples are provided, the workflow elaborated in this study has the potential to run fully automated with different image types, and liberates the user from the selection of appropriate features and thresholds.

Although the particular performance of the presented supervised framework will vary for different ground conditions and input datasets, the robust performance of the workflow in the tested cases raises confidence in its utility of landslide mappings on regional scale. One key point for future work in this direction is certainly the consideration of user interaction, which should involve techniques such as pre-clustering and active learning. Recently it has been demonstrated that active learning methods can help to reduce significantly the number of necessary training samples (e.g. Tuia et al. 2009), and there is still room for an improved consideration of the spatial context in such methods.

Further observations (not presented here) indicate that the predictive power of different features varies depending on the segmentation scale, and hence further improvements might possible by considering evidence hierarchically among different scales. The presented technique is especially suitable for high-dimensional data sets and would certainly benefit from further image features (e.g. steerable filters) and additional data (e.g. multi-temporal datasets).

Acknowledgements

The work described in this paper was (partially) supported by the project SafeLand “Living with landslide risk in Europe: Assessment, effects of global change, and risk management strategies” under Grant Agreement No. 226479 in the 7th Framework Programme of the European Commission. This support is gratefully acknowledged. In addition we thank the involved partners for the exchange of data.

References

[1]Baatz, M., & Schäpe, A. (2000). Multiresolution Segmentation – an optimization approach for high quality multi-scale image segmentation. In J. Strobl, T. Blaschke, & G. Griesebner (Eds.), *Angewandte Geographische Informationsverarbeitung XII* (pp. 12-23). Salzburg: Wichmann, Heidelberg

[2]Barlow, J., Franklin, S., & Martin, Y. (2006). High spatial resolution satellite imagery, DEM derivatives, and image segmentation for the detection of mass wasting processes. *Photogrammetric Engineering and Remote Sensing, 72*, 687-692

[3]Borghuis, A.M., Chang, K., & Lee, H.Y. (2007). Comparison between automated and manual mapping of typhoon-triggered landslides from SPOT-5 imagery. *International Journal of Remote Sensing, 28*, 1843–1856

[4]Breiman, L. (2001). Random Forests. *Machine Learning, 45*, 5-32

[5]Denil, M., & Trappenberg, T. (2010). Overlap versus Imbalance. In A. Farzindar, & V. Kešelj (Eds.), *Advances in Artificial Intelligence* (pp. 220-231): Springer Berlin / Heidelberg

[6]Diaz-Uriarte, R., & Alvarez de Andres, S. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics, 7*, 3

[7]Galli, M., Ardizzone, F., Cardinali, M., Guzzetti, F., & Reichenbach, P. (2008). Comparing landslide inventory maps. *Geomorphology, 94*, 268-289

[8]Haralick, R.M., K. Shanmugam and I. H. Dinstein (1973). Textural features forimage classification. *IEEE Transactions on Systems, Man and Cybernetics, 3*, 610-621

[9]Lu, P., Stumpf, A., Kerle, N., & Casagli, N. (in press). Object-oriented change detection for landslide rapid mapping. *IEEE Geoscience and Remote Sensing Letters*

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

[11]Nichol, J., & Wong, M.S. (2005). Satellite remote sensing for detailed landslide inventories using change detection and image fusion. *International Journal of Remote Sensing, 26*, 1913 - 1926

[12]Tuia, D., Ratle, F., Pacifici, F., Kanevski, M.F., & Emery, W.J. (2009). Active Learning Methods for Remote Sensing Image Classification. *Geoscience and Remote Sensing, IEEE Transactions on, 47*, 2218-2232

[13]van Westen, C.J., van Asch, T.W.J., & Soeters, R. (2006). Landslide hazard and risk zonation—why is it still so difficult? *Bulletin of Engineering Geology and the Environment, 65*, 167-184