Remote Sensing of Environment 115 (2011) 2564–2577

Contents lists available at ScienceDirect

Remote Sensing of Environment

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

Object-oriented mapping of landslides using Random Forests

André Stumpf [a,b,*], Norman Kerle [a]

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

[b] École et Observatoire des Sciences de la Terre, Institut de Physique du Globe de Strasbourg, UMR 7516 CNRS, Université de Strasbourg, 5, rue René Descartes, 67084 Strasbourg Cedex, France

ARTICLE INFO

Article history:
Received 19 November 2010
Received in revised form 18 May 2011
Accepted 19 May 2011
Available online 24 June 2011

Keywords:
Landslide mapping
VHR satellite images
Image segmentation
Object-oriented
Random Forest

ABSTRACT

Landslide inventory mapping is an indispensable prerequisite for reliable hazard and risk analysis, and with the increasing availability of very high resolution (VHR) remote sensing imagery the creation and updating of such inventories on regular bases and directly after major events is becoming possible. The diversity of landslide processes and spectral similarities of affected areas with other landscape elements pose major challenges for automated image processing, and time-consuming manual image interpretation and field surveys are still the most commonly applied mapping techniques. Taking advantage of recent advances in object-oriented image analysis (OOA) and machine learning algorithms, a supervised workflow is proposed in this study to reduce manual labor and objectify the choice of significant object features and classification thresholds. A sequence of image segmentation, feature selection, object classification and error balancing was developed and tested on a variety of sample datasets (Quickbird, IKONOS, Geoeye-1, aerial photographs) of four sites in the northern hemisphere recently affected by landslides (Haiti, Italy, China, France). Besides object metrics, such as band ratios and slope, newly introduced topographically-guided texture measures were found to enhance significantly the classification, and also feature selection revealed positive influence on the overall performance. With an iterative procedure to examine the class-imbalance within the training sample it was furthermore possible to compensate spurious effects of class-imbalance and class-overlap on the balance of the error rates. Employing approximately 20% of the data for training, the proposed workflow resulted in accuracies between 73% and 87% for the affected areas, and approximately balanced commission and omission errors.

1. Introduction

During the last century (1903–2004) approximately 16,000 people were killed by landslides in Europe (Nadim et al., 2006), while in other parts of the world even single events can have comparable dimensions (20,000 in Peru, 1970, 29,000 in China, 2008)(Kjekstad & Highland, 2009; Petley, 2009). The mean annual costs of landslides in Italy, Austria, Switzerland and France are estimated between USD 1–5 billion for each of the countries (Kjekstad & Highland, 2009). The assessment of associated risks, a prerequisite for disaster mitigation, is still a difficult task, with comprehensive landslide inventories being the most commonly used source for quantitative landslide hazard and risk assessment at regional scales (van Westen et al., 2006).

Landslide inventories have traditionally been prepared combining the visual interpretation of aerial photographs and field work, which to date remains the most frequently followed approach for the elaboration of inventory maps in scientific studies and by administrative bodies (Hervás & Bobrowsky, 2009). Despite its time-consuming and labor intensive nature, however, results still include a large degree of subjectivity (Galli et al., 2008), and incur the risk of omissions due to limited site access or aerial survey campaigns only being mounted with some delay, when landslide traces are starting to disappear.

Notable advances are being made in the detection of surface-displacements from active (e.g. Cascini et al., 2010) and passive (e.g. Debella-Gilo & Kääb, 2011) spaceborne sensors, allowing for detailed monitoring of ground-deformations. Those techniques depend on a coherent signal over time and are applicable for the mapping of slow to extremely slow moving landslides (<13 m/month after Cruden & Varnes, 1996) with a sparse vegetation cover. For the automated mapping of dormant landslides under forest high-resolution surface models from airborne laser scans provide new opportunities (e.g. Booth et al., 2009). However, most hazardous landslides reach considerable velocities and can typically only be mapped in a post-failure stage, for which optical airborne and satellite images are the commonly chosen data sources. Large events with thousands of individual landslides such as recently in China (earthquake, 2008),

* Corresponding author at: ITC-Faculty of Geo-Information Science and Earth Observation of the University of Twente, Department of Earth Systems Analysis, Hengelosestraat 99, P.O. Box 6, Enschede, 7500 AA, The Netherlands.
E-mail addresses: stumpf24883@itc.nl (A. Stumpf), kerle@itc.nl (N. Kerle).

0034-4257/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.rse.2011.05.013

Haiti (earthquake, 2010) and Brazil (rainfall, 2011) illustrate the immense challenges posed for any non-automated mapping approach.

The large fleet of existing and planned very high resolution (VHR) satellites allows to record inexpensive imagery within days or even hours after a given landslide event, and a number of studies have already addressed the development of more automatic techniques for landslide mapping with VHR images (Barlow et al., 2006; Borghuis et al., 2007; Hervás & Rosin, 1996; Joyce et al., 2008; Lu et al., 2011; Martha et al., 2010; Nichol & Wong, 2005; Rau et al., 2007; Whitworth et al., 2005). Most of them targeted the mapping of fresh features after rapid slope failures, but a few works also demonstrated the potential of optical data for the identification of slow-moving and dormant landslides (Hervás & Rosin, 1996; Whitworth et al., 2005).

Proposed approaches may be generally classed into pixel-based and object-based techniques, both including methods for the analysis of monotemporal and multitemporal imagery, and often making use of ancillary datasets such as digital elevation models (DEMs). Pixel-based approaches include unsupervised (Borghuis et al., 2007) and supervised classification (Joyce et al., 2008), as well as change detection techniques (Hervás et al., 2003; Nichol & Wong, 2005; Rau et al., 2007). Although those techniques consider to some extent additional geometric constrains, such as minimum size, minimum slope or non-rectangular shapes, they rely mainly on the spectral signal of individual pixels. To exploit better the information content of local pixel neighborhoods, Hervás and Rosin (1996) conducted a systematic statistical evaluation of texture measures for landslide mapping and found texture features after Haralick et al. (1973) especially useful to highlight hummocky surfaces often associated with landslides. Similarly, more recent studies concluded that the integration of texture improves the image classification and may yield more accurate maps (Carr & Rathje, 2008; Whitworth et al., 2005).

In general there is an emerging agreement in the remote sensing community that unsatisfactory results of per-pixel analysis can often be attributed to the fact that geometric and contextual information contained in the image is largely neglected (e.g. Blaschke, 2010). This is especially true at higher resolutions, with a higher spectral variance leading to increased intra-class variability and typically lower classification accuracies (Woodcock & Strahler, 1987). Further challenges arise due to the typically lower number of spectral bands of modern VHR sensors and a higher sensitivity to co-registration errors at higher resolutions. To address such issues object-oriented analysis (OOA), also often referred to as object-based image analysis (OBIA), became a widely spread concept for many geoscientific studies to exploit geometric and contextual image information of multi-source data (Blaschke, 2010).

Image segmentation and classification resemble human cognition to some degree and have inspired a number of researchers to transfer existing knowledge in machine executable rule sets. Such rule sets have already been used for landslide mapping as a self-contained classification scheme (Barlow et al., 2003), prior to supervised classification (Barlow et al., 2006), for the post-processing of pixel-based classification (Danneels et al., 2007), and for change detection with multi-temporal images (Lu et al., 2011). Martha et al. (2010) emphasized the importance of exploiting a range of features as widely as possible, and developed a complex set of decision rules, including 36 particular thresholds, to detect and classify landslides of 5 different types in the High Himalayas.

Expert rule sets are a very transparent solution for the exploitation of domain knowledge but comprise two main limitations: (i) the difficulty to decide which descriptive features are actually significant, and (ii) their restricted generic applicability for different input data types and under variable environmental conditions. Professional OOA software solutions readily provide hundreds of potentially useful object metrics, and further customized features enrich this great variety. They allow the user high flexibility in setting up efficient automated processes, but the selection of significant features remains a challenging and time-consuming task.

Feature selection in high-dimensional datasets is an important task in many fields such as bioinformatics (Saeys et al., 2007) or hyperspectral remote sensing (e.g. Guo et al., 2008), and typically targets a better performance of the algorithm classifying the data and/or the investigation of causal relationships. A few object-oriented studies already addressed statistical feature selection for land cover mapping from VHR imagery (e.g. Laliberte & Rango, 2009; Van Coillie et al., 2007), but no such efforts have been in the context of landslide mapping. Little is known about the robustness, efficiency, scale-dependency and generic applicability of the object-features and thresholds proposed in individual studies. Considering the great variety of landslide types, environmental conditions and available imagery this largely prevents the transferability of proposed methods and the development of operational workflows.

Machine learning algorithms, such as Random Forests (RF, Breiman, 2001), have demonstrated excellent performance for the analyses of many complex remote sensing datasets (Gislason et al., 2006; Lawrence et al., 2006; Watts et al., 2009). RF is based on ensembles of classification trees and exhibits many desirable properties, such as high accuracy, robustness against over-fitting the training data, and integrated measures of variable importance (Diaz-Uriarte & Alvarez de Andres, 2006). However, like many other statistical learning techniques RF is bias-prone in situations where the number of instances is distributed unequally among the classes of interest. Under class-imbalance in fact most classifiers tend to be biased in favor of the majority class, and vice versa may underestimate the number of cases belonging to the minority class (He & Garcia, 2009). Experiments on synthetic datasets suggest that such biases are combined effects of class imbalance and an overlap of the classes in feature space (e.g. Denil & Trappenberg, 2010). As landslides typically cover only minor fractions of a given area, class-imbalance is an inherent issue that affects the probabilistic assessments of slope susceptibility (Van Den Eeckhaut et al., 2006), and may complicate the application of machine learning algorithms for image-based inventory mapping.

The objective of this study was to investigate the applicability and performance of the RF learning algorithm in combination with OOA to reduce the manual labor in landslide inventory mapping with VHR images. Assuming that a sample-based framework combining both techniques could be a flexible and efficient solution for many real-world scenarios, VHR imagery recorded by state-of-the-art systems (Geoeye-1, IKONOS, Quickbird, and airborne) at four different sites was analyzed. To achieve an accurate and robust image classification it was of particular interest to determine which image object metrics efficiently distinguish landslide and non-landslide areas. Training and testing samples were derived from existing landslide inventories, and a RF-based feature selection method (Diaz-Uriarte & Alvarez de Andres, 2006) was adopted to evaluate the capability of a broad set of object metrics (color, texture, shape, topography) and their sensitivity to changing scales of the image segmentation. Class-imbalance and -overlap were expected to be critical points for the application of the RF, and we further investigated if an iterative resampling scheme could be used to design training sets that lead to a balance between commission and omission errors. The efficiency of this approach was evaluated at each test site with different segmentation scales and in scenarios where 20% of the image objects would be available for training.

2. Study sites and data

VHR images collected in the immediate aftermath of two recent major earthquakes, as well as from two sites affected by non-seismic landslides, were used in this study (Table 1). The areas are characterized by a great diversity of environmental settings, landslide processes and image acquisition conditions, and in this manner

Table 1
Overview of analyzed images and topographic data.

Test site	Haiti	Wenchuan	Messina	Barcelonnette
Sensor	Geoeye-1	IKONOS	Quickbird	Aerial photograph
Spectral bands	4-band multispectral	4-band multispectral	4-band multispectral	3-band natural Color
Pixel size (multispectral/panchromatic) [m]	2/0.5	4/1	2.4/0.61	0.5/–
Sensor Tilt [°]	2.7	15.7	3.1	n.a.
Nominal collection azimuth [°]	343.8	62.7	343.3	n.a
Solar zenith angle [°]	45.6	19.2	45.6	–
Sun angle azimuth [°]	150.2	119.3	161.7	–
Date (days after the event)	13/01/2010 (1)	23/05/2008 (11)	10/8/2009 (8)	07/2004 (n.a.)
Test area [km²]	1	4	1	1
landslide affected areas [%]	9.6	15.1	19.6	8.7
DEM resolution (Source resolution)	10 m (1 m LiDAR DSM)	10 m (20 m contour lines)	10 m (1 m LiDAR DSM)	10 m (1 m IFSAR DSM)

simulate realistic test cases with imagery that is typically available shortly after major events.

2.1. Test site 1: Momance River — Haiti

On January 12, 2010 an earthquake with a moment magnitude of 7.0 struck Haiti. It was caused by the rupture of a previously unknown system of a blind thrust fault (Hayes et al., 2010) and claimed approximately 230,000 victims. Landslides caused extensive yet unquantified damage at several locations (Eberhard et al., 2010), and an increased frequency of slope failures and debris flows can be expected during future rainfall events. The study site is located at the Enriquillo fault line, which forms a tectogenetic valley followed by the Momance River. The slopes are between 20 to 50° steep and show a large number of shallow debris and rock slides. Most of the gentler terrain sections are under agricultural use by hundreds of scattered family farms. Due to erosion bare soils are exposed at several locations, and the valley bottom is covered by fluvial gravel bars and fresh landslide deposits. Geoeye-1 imagery was recorded one day after the event.

2.2. Test site 2: Wenchuan town — China

The rupture of the Longmenshan fault system on May 12, 2008 ($M_L = 8.0$) triggered more than 60,000 individual slope failures (Gorum et al., in press), and approximately 30,000 of the 80,000 casualties can be attributed to the impact of landslides (Tang et al., 2010). The county capital, Wenchuan town, is located on both sides of the Min River at 1330 m.a.s.l. and is surrounded by steep terrain with average slopes of approximately 30°. The town and its surroundings were seriously affected by a large number of mainly shallow translational landslides, which are concentrated on the steepest slopes in proximity to the drainage lines. Already before the event those terrain units were rather sparsely vegetated and showed bedrock outcrops at several locations. The main land cover types are degraded mountain forest and terraced field crops, which extend to slopes of up to 35°. Because the harvest was underway at the time of the initial rupture, many fields were barren and showed similar spectral characteristics as newly triggered landslides. IKONOS imagery was acquired 11 days after the main shock.

2.3. Test site 3: Messina — Italy

On the 1st of October 2009 a series of debris flows struck several catchments a few kilometers south of the city of Messina/Sicily. The debris flows were triggered by extraordinarily intense rainfall in the afternoon of that day, which had been preceded by prolonged intense rainfall at the end of September. Thirty-one people were killed during the event and the direct economic loss was estimated as almost US$ 825 million (Civil-Protection-Sicily, 2010). The affected area comprises ten small and medium size catchments that rise from sea level to about 700 m in the Peloritani Mountains. The present land cover types comprise bare ground, crop-, shrub- and grassland, deciduous forest and rural built-up areas. Most of the landslides were initiated as shallow debris flows or slides at the upper slopes, and evolved into rapid hyper-concentrated flows along their way through the drainage network. The Quickbird imagery was recorded 7 days after the event.

2.4. Test site 4: Barcelonnette Basin — France

The Barcelonnette Basin is located in the South French Alps and characterized by a mountain climate with Mediterranean influence. The area is known for the large number of slow-moving active landslides, and in the present study a small subset comprising the Super Sauze active slow-moving mudslide (Malet, 2003) was examined. The task here was mainly to distinguish the landslide body from the surrounding badlands, and since the affected area is one compact object this rather corresponds to an image segmentation task. The available imagery is a natural color aerial photograph recorded in summer 2004.

2.5. Landslide inventories

The reference inventories for Wenchuan, Messina and Barcelonnette are based on field work and visual interpretation of aerial photographs as well as VHR satellite imagery. As detailed field investigations of the earthquake-induced landslides in Haiti have not yet been completed, the corresponding inventory is based on the interpretation of remote sensing products only. To minimize the risk of miss-mapping we considered pre-and post-event VHR satellite imagery from multiple sensors (IKONOS, Geoeye-1, WorldView-2) and a post-event LiDAR DEM for the manual delineation of affected areas.

3. Methods

At each test site we selected subsets (Fig. 1) that include landslides and spectrally similar objects, such as river plains, urban areas, roads, badlands and barren fields. A scalable segmentation algorithm (Section 3.1) was applied on the images from each area, and a comprehensive set of object metrics was calculated (Section 3.2). These processing steps (Fig. 2 a) were performed with eCognition® software, which implements nearest neighbor interpolation to resample coarser image layers to the resolution of the finer panchromatic layers. Subsequent to segmentation and metric calculation the landslide inventories compiled from field work and visual image interpretation (Section 2.5) were used to create a sample database with all objects assigned either as landslide objects (O_{LS}) or non-landslide objects (O_{NLS}, Fig. 2 b). Each image object containing at least 50% of landslide-affected area was labeled as O_{LS}, and all others

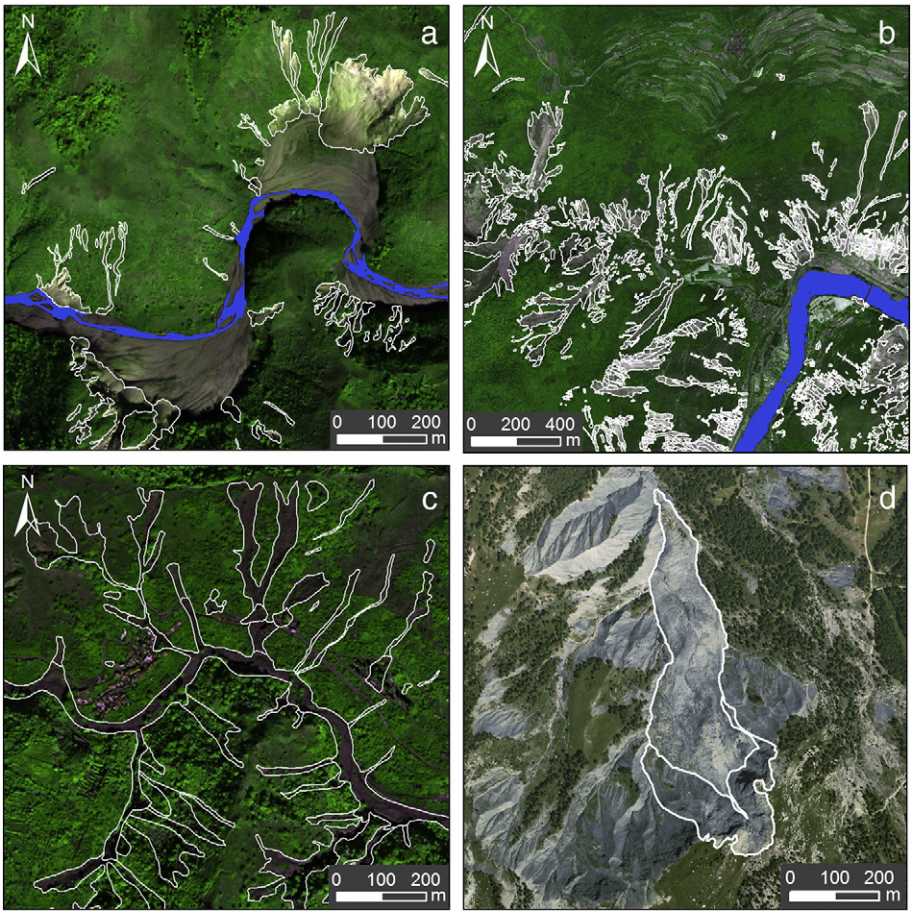


Fig. 1. Analyzed areas at the different test sites. a) Momance River, Haiti (Momance river in blue), b) Wenchuan, China (Min River in blue), c) Messina, Italy, d) Barcelonnette basin, France. White outlines indicate the landslide areas.

as O_{NLS}. Such a majority criterion was considered as the most logical choice because it minimizes the overall amount of miss-labeled areas, while retaining also marginal cases that may provide useful information for the classifier training. To evaluate a comprehensive set of object metrics (Table 2) for the discrimination of landslides and unaffected areas, all O_{LS} and an equally sized random sample of O_{NLS}, were used at all test sites and scales, respectively. They were introduced in the RF-based approach for feature evaluation and reduction (Fig. 2 c) proposed by Diaz-Uriarte and Alvarez de Andres (2006), and described in greater detail in Section 3.3.1.

Non-relevant features were subsequently removed and the data were split into training and testing sets (Fig. 2 d1). To account for spurious effects of class-imbalance and class-overlap, an iterative scheme for the adjustment of the training set was developed and tested (Fig. 2 d2, Section 3.3.2.). The classification accuracy of the approach was finally assessed on a test set comprising 80% of all image objects (Fig. 2 d3).

3.1. Image segmentation

Image segmentation generates the building blocks of OOA, and the delineation quality of the target objects has a direct influence on the accuracy of the subsequent image classification. Numerous image segmentation algorithms have been developed in the last decades and applied in remote sensing image analysis (Dey et al., 2010), all of them aiming at the delineation of relatively homogeneous and meaningful segments.

The multi-resolution image segmentation (MRIS) implemented in eCognition® software is a frequently used algorithm in Earth science studies (Blaschke, 2010). MRIS is a region-growing segmentation algorithm which, starting from individual pixels, merges the most similar adjacent regions, as long as the internal heterogeneity of the resulting object does not exceed the user defined threshold scale factor (Benz et al., 2004). Proposed statistical optimization methods (e.g. Drăguţ et al., 2010) may allow an objectification of the choice of the scale parameter if the targeted objects or processes exhibit a single operational scale. However, slope failures and surrounding land cover elements feature several orders of magnitudes in volume and area, prompting other researchers to look for automatic optimisation at multiple scales (Martha et al., in press).

To evaluate the impact of changing segmentation scales on the feature space and class separability, image segmentation was performed at 15 different scales (10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 70, 80, 90, and 100). The segmentation results depend on data characteristics such as spatial resolution, the number of bands, image quantization and the scene characteristics. The same scale factor does not necessarily yield comparable objects in different scenes, but increasing the scale factor for the segmentation of the same dataset will generally lead to larger object sizes. Thus, it is possible to emulate increasingly coarser representations of the same scene and compare resulting trends among the tested sites.

The MRIS framework offers the possibility to assign different weights to spectral bands and shape of segments. All multi-spectral bands (blue, green, red, and near-infrared [NIR]) were equally weighted with a value of one, while the panchromatic channel of the satellite images was assigned a weight of four, allowing a balance of multispectral and finer panchromatic data in the segmentation. The shape criteria were weighted with zero and, consequently, not considered in the segmentation.

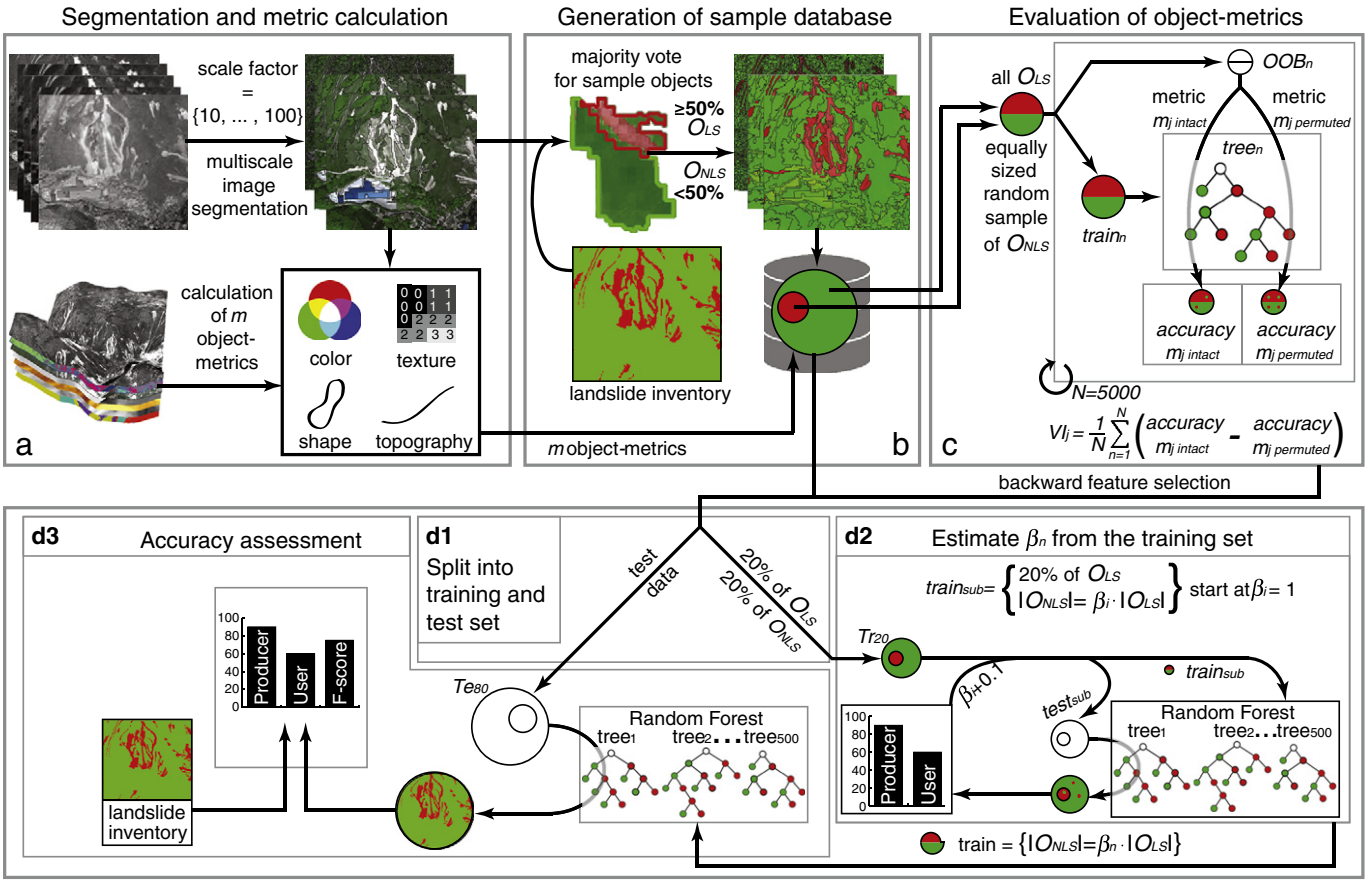


Fig. 2. Overview of the processing steps followed in this study. Explanations are given throughout the text in Section 3.

3.2. Calculation of image object metrics

This section provides an overview of features adopted from previously studies (Table 2), and introduces a number of further object metrics that were calculated subsequent to the image segmentation (Fig. 2 a). Spectral features previously recommended in the literature (Table 2) comprise band intensities, band ratios, principal component (PC) transform and brightness, and respective mean values were calculated per image object. The mean brightness (B) was defined as the sum of the object means in the visible and panchromatic band ($\bar{c}_{i(vis)}$) divided by the number of corresponding bands (n_{vis}).

$$B = \frac{1}{n_{vis}} \sum_{i=1}^{n_{vis}} \bar{c}_{i(vis)}$$

The same bands were considered to calculate *MaxDiff* for each object, defined as the absolute value of the difference of the minimum

Table 2
Overview of features used to identify landslides in previous works and adopted for this study. Most of the studies combined several attributes and are listed only exemplarily. Number in brackets indicates the number of features used with the aerial photographs.

	Tested features	No.	Case study
Spectral information	Spectral bands	5 (3)	(e.g. Nichol & Wong, 2005)
	PC	4 (3)	(Forsythe & Wheate, 2003)
	Band ratios (blue/green, green/red, red/NIR)	3 (2)	(e.g. Rau et al., 2007)
	Brightness	1	(Martha et al., 2010)
	MaxDiff	1	This study
Texture	$GLCM_{all\ dir.}$ (Ent., Mean, Cor., Con., Stdv.)	25 (15)	(Carr & Rathje, 2008; Hervás & Rosin, 1996; Martha et al., 2010; Whitworth et al., 2005)
Geometric	Shape index, compactness, roundness	3	(Moine et al., 2009)
	Length–width ratio	1	(Martha et al., 2010; Martha et al., in press)
Auxiliary data	Hillshade	1	(Martha et al., 2010)
	Slope	1	(Borghuis et al., 2007; Danneels et al., 2007)
Combined metrics	Object direction/flow direction	1	(Martha et al., 2010; Martha et al., in press)
	$GLCM_{flow.dir.}$ (Ent., Mean, Cor., Con., Stdv.)	25 (15)	This study
	$GLCM$ (Ent., Mean, Cor., Con., Stdv.)	25 (15)	This study

object mean ($min(\bar{c}_{i(vis)})$) and the maximum object mean ($max(\bar{c}_{i(vis)})$), divided by the object brightness *B*.

$$MaxDiff = \frac{\left|min\left(\bar{c}_{i(vis)}\right) - max\left(\bar{c}_{i(vis)}\right)\right|}{B}$$

To quantify surface textures a variety of derivatives of the Grey Level Co-occurrence Matrix (*GLCM*) has been adopted in previous landslide studies (Table 2). Considering their large computational burden and frequent reports on strong correlations among several *GLCM* derivatives (Hall-Beyer, 2007; Laliberte & Rango, 2009), a subset of five texture measures was selected for this study. Those are *contrast* (*Con.*), *correlation* (*Cor.*), *entropy* (*Ent.*), *standard deviation* (*Stdv.*) and *Mean*. For a detailed formulation of the *GLCM* and derivatives we refer to Haralick et al. (1973) and here only recall that the co-occurrence frequencies of grey-levels are typically counted in symmetric matrices for pixels neighboring directly at 0° (N–S), 45° (NE–SW), 90° (E–W) or 135° (SW–NE), respectively. Rotation-invariance of a *GLCM* derivative can be achieved by calculating its mean or minimum value among all four directions (e.g. Pesaresi et al., 2008), or by summing up the four directional *GLCMs* ($GLCM_{all\ dir.}$) before the calculation of the derivative. The latter technique is implemented in eCognition (Trimble, 2011) and was used in this study to calculate five rotation-invariant texture measures per band directly for each image object.

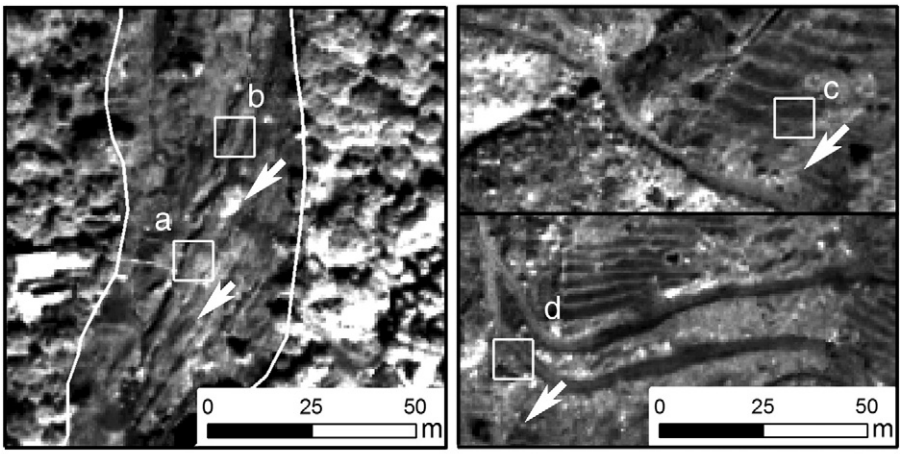
Rotation-invariance is desirable for many applications but fails to capture directional patterns in the grey-value distribution. Landslide-affected surfaces often show downslope-directed texture patterns that are potential diagnostic features to distinguish them from surfaces with texture patterns oriented at the strike of the slope (Fig. 3). In order to quantify such patterns better, additional directional texture measures were derived from two directional *GLCMs*; one computed along the hydrological flow direction ($GLCM_{flow\ dir.}$) and one perpendicular to it ($GLCM_{\perp flow\ dir.}$). For this purpose flow direction rasters (Jenson & Domingue, 1988) were derived from the respective DEMs (10 m resolution, Table 1) and their latticeswere superimposed on the images. For each resulting 10×10 m grid cell two directional *GLCMs* were calculated according to the direction (and the normal) indicated in the flow direction raster. Fig. 3 shows this exemplarily for *GLCM Correlation*, where the flow direction in each squared cell is aligned at 45°, and the two directional *GLCMs* consequently consider the grey-levels of pixels neighboring at 45° ($GLCM_{flow\ dir.}Cor.$) or 135° ($GLCM_{\perp flow\ dir.}Cor.$), respectively.

Ratio features (*GLCM*) were subsequently calculated for each squared cell simply as the quotient of the texture measures computed in flow direction and their counterparts computed in the perpendicular direction. *Contrast*, *correlation*, *entropy*, *standard deviation* and *Mean* from $GLCM_{flow\ dir.}$ and their respective *GLCM* ratios are also referred to as topographically-guided texture measures. They were computed on all image bands in a 10×10 m grid and finally converted into raster layers with a pixel size of 10 m. This corresponds to 10 additional layers per band, where each image object (Section 3.1) obtains the mean layer value within its extent. Together with the texture measures from $GLCM_{all\ dir.}$ and a number of object metrics characterizing mean spectral values, shape and topographic metrics, a total of 96 and 62 features per image object were calculated for the satellite imagery and aerial photograph, respectively (Table 2).

3.3. Random Forests

Since the fundamental works on ensemble decision trees (e.g. Breiman, 2001), Random Forests (RF) have already provided promising results in fields such as genomics (Diaz-Uriarte & Alvarez de Andres, 2006), ecology (Cutler et al., 2007) and remote sensing (Lawrence et al., 2006; Watts et al., 2009).

Small changes in the training data induce a high variance in single classification trees and often lead to rather low classification accuracies (Breiman, 1996). The underlying idea of RFs is to grow multiple decision trees on random subsets of the training data and related variables. For the classification of previously unseen data, RFs take advantage of the high variance among individual trees, letting each tree vote for the class membership, and assigning the respective class according to the majority of the votes. Such ensembles demonstrate robust and accurate performance on complex datasets with little need for fine-tuning and in the presence of many noisy variables. Furthermore, integrated procedures for variable assessment and selection, and freely available high-



	Object	$GLCM_{all\ dir.}$ Cor.	$GLCM_{flow\ dir.}$ Cor.	$GLCM_{\perp flow\ dir.}$ Cor.	$GLCM_{\perp}$ Cor.
a	debris flow	0.877	0.879	0.812	1.083
b	debris flow	0.870	0.892	0.780	1.144
c	field	0.881	0.864	0.959	0.901
d	street	0.879	0.751	0.783	0.959

Fig. 3. Exemplary comparison between the rotation-invariant and topographically-guided *GLCM Cor.* at the Messina test site. The texture measures are calculated on the panchromatic channel. White arrows indicate the hydrological flow direction within the measured cells. For linear structures along the flow direction (debris flows) values of $GLCM_{flow\ dir.}$ *Cor.* tend to be lower and values of $GLCM_{\perp flow\ dir.}$ *Cor.* tend to be higher. Hence, their ratio (*GLCM Cor.*) is typically lower for linear structures aligned perpendicular to the flow direction (e.g. fields, streets).

quality software implementations, make RFs an interesting tool to be combined with OOA. In this work we extensively used the `randomForest` package (Liaw, 2010) and its extension for variable selection `varSelRF` (Diaz-Uriarte, 2010) implemented in the `R` statistical programming environment (R-Development-Core-Team, 2009).

3.3.1. Evaluation and selection of object metrics

As a starting point we were interested in object metrics that are actually helpful to distinguish landslides from other image objects, and in understanding how their performances depends on the scale of the image segmentation. For this purpose a RF–based variable importance measure was used to evaluate the object metrics at each test site with 15 different segmentation scales (10–100). RF offers a number of internal measures to estimate the importance of employed variables for the accuracy of a given classification. The properties of those measures have been intensively studied in recent years, and the so-called permutation importance is considered a computationally tractable choice for the screening of large datasets (Nicodemus et al., 2010). The permutation importance, subsequently termed variable importance (*VI*), is calculated as follows.

The original training data are resampled randomly (with replacement) to create a training set ($train_n$, Fig. 2 c) and build a classification tree. Considering a total number of m extracted object-features (Fig. 2 a) at each tree node a subset $\sqrt{m}$ features is randomly selected and tested for the best split. Approximately one third of the instances are left out of the training set and remain as *out-of-bag* sample (OOB_n, Fig. 2 c) that can be used to assess the classification accuracy of the tree. The importance of a feature m_j for the correct classification is estimated by permuting the feature values within the OOB_n sample and calculating the difference of prediction accuracies before and after the perturbation. The *VI* of the variable m_j (VI_j, Fig. 2 c) results from averaging the permutation importance of m_j over a large number of trees (N = 5000, Fig. 2 c). In the present study it provided a measure for the utility of the different object metrics to distinguish between O_{LS} and O_{NLS}. In order to give equal weight to both classes, at each test site all O_{LS} and an equal number of randomly sampled O_{NLS} were taken into account. *VIs* were calculated for all variables at the 15 different segmentation scales (Section 3.1), where the overall number of sample objects varied between a few hundred at the coarsest scale and more than 60,000 at the finest scale.

Diaz-Uriarte and Alvarez de Andres (2006) proposed to compute the *VI* from a large RF (N = 5000) to obtain an initial variable ranking and then proceed with an iterative backward elimination of the least important variables. In each iteration the least important 20% of the features are dropped, a new RF (N = 2000) is trained with the remaining feature set, and the *OOB* sample is used to assess its miss-classification rate (*OOB error*). The final features set is selected according to the RF that produces the lowest *OOB error* (Fig. 4). In the present study this procedure was used to determine the set of object metrics that were used for the construction of the final RF classifiers (Fig. 2 d1–3).

3.3.2. Balancing of error rates and accuracy assessment

At all four test sites landslides covered only minor fractions of the scene (Table 1). This is a typical situation leading to an imbalance between O_{LS} and O_{NLS}, and potentially introduces a bias of the classification towards the over-represented non-affected area. Preliminary test runs adopting naturally imbalanced training sets indeed demonstrated serious underestimation of the landslide class, suggesting the presence of the class-imbalance problem. Such biases are undesirable in any manual or automated landslide mapping, because an over- or underestimation of the affected areas would generally lead to a respective over- or underestimation of the associated hazards and risks.

Numerous methods to account for such effects have been proposed in the context of different statistical learning techniques. For logistic regression they may involve prior corrections and weighting methods (King & Zeng, 2001) or asymptotical coefficient estimates for infinite

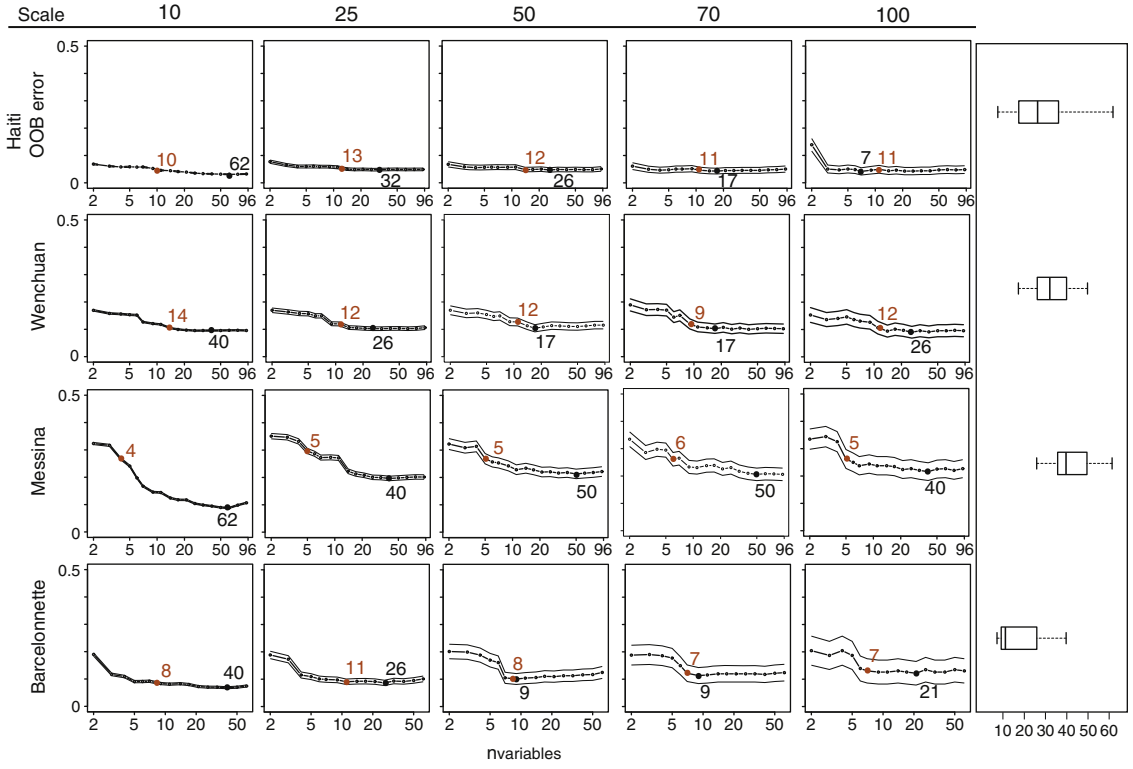


Fig. 4. Feature selection histories at the four test sites shown exemplarily for 5 of the 15 segmentation scales. The *OOB error* is drawn with a margin of one standard error against the number of selected variables. Black dot: Model with the smallest *OOB error* and number of selected metrics. Brown dot: Highest ranked texture measure with the respective rank. Boxplots indicate the variability of the number of selected features among all 15 segmentation scales.

class imbalance (Owen, 2007). Many approaches have also been developed for nonparametric learning algorithms such as RF, and may be grouped into resampling, cost-sensitive learning and kernel methods (He & Garcia, 2009). None of the methods proposed so far provides a universal advantage in all situations, but it has been demonstrated that an undersampling of the majority class is a beneficiary strategy in many different setups (Blagus & Lusa, 2010; Burez & Van den Poel, 2009). The sampling of a balanced training set might in some situations be sufficient to compensate class-imbalance when using RF (e.g. Fusaro et al., 2009), but the optimal class distribution will generally depend on the specific method and studied case (Burez & Van den Poel, 2009).

For the analysis of class imbalance and the final accuracy assessment only the previously selected object metrics were used. 20% of each dataset were sampled randomly for training (Tr_{20}, Fig. 2 d1–2) and the remaining 80% were used as test set (Te_{80}, Fig. 2 d3). To estimate the class ratio in the training sample that leads to a balance of commission and omission errors an iterative procedure was implemented and tested, where Tr_{20} was split repeatedly into subsets for training ($train_{sub}$) and testing ($test_{sub}$, Fig. 2 d2). The parameter β_i was defined as the ratio of O_{LS} and O_{NLS} in the current $train_{sub}$, and changed systematically to approximate a target value β_n yielding a balance between user's and producer's accuracy on the $test_{sub}$. In each iteration 20% of the O_{LS} and β_i-fold number of O_{NLS} were sampled randomly from Tr_{20} to train a RF ($N = 500$) and assess the classification accuracies on the remainder $test_{sub}$ (Fig. 2 d2). The procedure started from a balanced class distribution ($\beta_i = 1$) and in each step β_i increased by 0.1 (Fig. 2 d2). The underlying assumption was that the estimated β_n could be applied to adjust the class-balance for the entire training set Tr_{20}, and would also yield balanced user's and producer's accuracies in the classification of the actual test set Te_{80}.

For each β_i the procedure was repeated ten times using replicates of $train_{sub}$ and $test_{sub}$ randomly sampled from Tr_{20}. Mean error rates and their standard deviations were calculated from ten runs, and in cases where the observed variance in the resulting learning curves were too high for the determination of a unique β_n the number of random replicates was increased (Section 4.3.1).

To assess the accuracy of the described RF framework, RFs ($N = 500$) were trained with β_n-adjusted subsets of Tr_{20}, and applied on the remainder 80% test sample (Te_{80}, Fig. 2 d3). At each test site the sample balancing and accuracy assessments were performed exemplarily at a fine, medium and small segmentation scale (10, 30, and 70), and compared to reveal the effects of the segmentation on the user's, producer's and overall accuracies.

4. Results and discussion

4.1. Effects of scale on variable importance and selection

In none of the examined cases the *OOB error* reduced if more than 77 object metrics were introduced, and on average only about one third of the pre-selected metrics were detected as useful. In most cases the *OOB error* remained rather stable or increased if all variables were used. Especially at the test sites Haiti and Barcelonnette many of the object metrics provided only minor further enhancements. This is reflected by flat parts of the respective curves in Fig. 4, where slight changes of the object characteristics can have a stronger impact on the position of the *OOB error* minima, which was the criterion for the model selection. Consequently, among all segmentation scales there is a high variability in the observed overall number of selected features (boxplots Fig. 4), which coincides with those flat parts of the curves. Larger segmentation scales generally yield fewer sample objects, and consequently the standard error of the *OOB error* estimate increased (Fig. 4). It should be considered that in situations where the number of samples becomes much smaller than the number of features, the feature selection method can deteriorate strongly (Diaz-Uriarte & Alvarez de Andres, 2006; Yu et al., 2006). However, this was not an issue in the present study because even at the largest segmentation scales the number of sample objects was at least twice the number of features.

Selecting the model with the lowest *OOB error* is a rather conservative strategy that may retain some redundant and partially correlated variables. However, it was suitable for the present study in order to retain all useful features and targeting a maximal predictive accuracy. For applications where the smallest set of features with causal relationships is important (e.g. Diaz-Uriarte & Alvarez de Andres, 2006) a further reduction might be desirable, but no further enhancements of the predictive accuracy can be expected.

Although the absolute number of selected object metrics strongly depended on the particular test site and segmentation scale, some features emerged as significant in most cases and should be further highlighted. Unsurprisingly, metrics related to spectral information resulted as the most important ones for all test cases and scales (Table 3). The band ratios and PC that depict the contrast between vegetated and non-vegetated areas ranked with a particularly high variable importance (*VI*). Object means of the slope and hillshade layers significantly reduced the error rates, but in most cases their relative importance decreased with larger segmentation scales (Fig. 5). Shape metrics displayed a rather contrary behavior (Fig. 6), and generally contributed little to the reduction of the error rates. Only for larger segmentation scales at Wenchuan and Messina, where the segments more closely approached the elongated shape of the landslides, shape metrics were selected by the selection procedure. They have been reported as useful after initial spectral classification steps (Martha et al., 2010; van der Werff & van der Meer, 2008), but provide little additional information within the tested sample-based framework.

However, the *VI* ranks of the most important spectral and textural metrics exhibited low variability among the different segmentation scales (σ in Table 3) and were not subject to a persistent trend. The topographically-guided *GLCM Con.*, *Cor.* and *Ent.* helped to reduce the *OOB error* at all tested sites and largely outperformed the rotation-invariant *GLCMs*. Furthermore, the topographically-guided *GLCM Con.* was apparently more efficient when derived from the higher resolution panchromatic channels. Both rotation-invariant and topographically-guided versions of *GLCM Mean* and *Stdv.* were frequently included in the selected models, but the rotation-invariant versions were in most cases ranked higher, indicating that the topographic control did not enhance the significance of *GLCM Mean* and *Stdv*. Although *GLCMs* have been previously adopted for landslide mapping (e.g. Martha et al., 2010) the proposed topographic control on their calculation provides significant enhancement (Fig. 4), and makes such object metrics potentially useful for the automated mapping of various geomorphological processes.

Although the optimal choice of the texture measures depends to a certain degree on the application, it is interesting to note that Clausi (2002) highlighted *Con.*, *Cor.* and *Ent.* as particularly useful *GLCM* derivatives for the recognition of sea ice, and Laliberte and Rango (2009) concluded that *Con.*, *Ent.* and *Stdv.* are the most suitable texture measures for rangeland mapping.

4.2. Effects of the feature reduction on the predictive accuracy

The *OOB errors* reported during the feature selection process (Fig. 4) are not suitable to assess the predictive accuracies of the models because (i) in a real case only a fraction of the O_{LS} would be available for training, (ii) the set of optimal features may differ among subpopulations (Diaz-Uriarte & Alvarez de Andres, 2006), and (iii) the overall *OOB error* does not inform about commission and omission errors.

Those facts motivated a further experiment in which the training sets included only 20% of all O_{LS} (number of O_{LS} in Table 4) and an

Table 3
The 20 object metrics with the highest average variable importance rank among all 15 tested scales and at each respective test site. The number of scales at which the variable has been selected (n_{sel}), and the standard deviation of the rank among all 15 scales (σ_{rank}), are provided as indicators for the stability of the variable importance.

Messina		Haiti		Wenchuan		Barcelonnette	
Feature	n_{sel}/σ_{rank}	Feature	n_{sel}/σ_{rank}	Feature	n_{sel}/σ_{rank}	Feature	n_{sel}/σ_{rank}
Red/NIR	15/0.0	Red/NIR	15/0.0	Red/NIR	15/0.0	Blue/Green	15/0.0
NIR	15/0.2	Slope	15/0.2	PC 2	15/0.0	PC 2	15/0.0
PC 1	15/0.4	Green/red	15/0.4	Red	15/0.5	Max. Diff.	15/1.0
Max. Diff.	15/1.2	Red	15/1.2	Green/red	15/1.0	Blue	15/0.8
$GLCM_{flow.dir.}$ Con. PAN	15/1.6	PC 1	15/1.6	Blue	15/0.5	PC 3	15/1.8
$GLCM_{flow.dir.}$ Cor. PAN	15/2.2	Blue	14/2.2	Blue/green	15/1.0	Slope	15/2.6
Blue/Green	15/3.9	Blue/green	15/3.9	Green	15/0.6	Hillshade	15/1.4
GLCM Cor. PAN	15/2.4	Green	14/2.4	PC 1	15/1.1	PC 1	9/2.9
GLCM Con. PAN	15/3.0	PC 2	14/3.0	Brightness	15/1.0	$GLCM_{flow.dir.}$ Con. Blue	9/2.4
GLCM Cor. Red	15/4.4	PAN	14/4.4	Slope	15/2.8	Brightness	9/2.1
GLCM Cor. Green	15/4.3	$GLCM_{flow.dir.}$ Con. PAN	15/4.3	PAN	15/1.9	Green/red	8/2.5
Blue	15/6.8	NIR	13/6.8	*GLCM* Con. PAN	15/1.3	$GLCM_{flow.dir.}$ Con. Red	9/3.4
Slope	15/4.4	Hillshade	12/4.4	*GLCM* Cor. Green	13/2.7	$GLCM_{flow.dir.}$ Cor. Blue	7/2.4
GLCM Cor. blue	15/4.9	$GLCM_{flow.dir.}$ Cor. PAN	13/4.9	*GLCM* Cor. blue	14/3.0	$GLCM_{flow.dir.}$ Con. Green	8/3.3
GLCM Con. blue	15/3.8	Max. Diff.	14/3.8	*GLCM* Con. blue	13/2.9	$GLCM_{flow.dir.}$ Con. red	9/3.4
PC 2	14/7.8	Brightness	13/7.8	*GLCM* Cor. red	12/4.0	Green	6/2.8
GLCM Con. red	15/3.9	*GLCM* Con. PAN	12/3.9	Max. Diff.	14/4.0	$GLCM_{flow.dir.}$ Con. green	7/2.8
$GLCM_{flow.dir.}$ Cor. blue	15/3.9	PC 3	11/3.9	$GLCM_{flow.dir.}$ Cor. PAN	13/2.4	Red	7/2.2
$GLCM_{flow.dir.}$ Con. blue	15/5.1	*GLCM* Cor. PAN	10/5.1	NIR	10/8.5	*GLCM* Con. blue	5/5.4
GLCM Con. green	15/4.7	$GLCM_{flow.dir.}$ Ent. blue	9/4.7	*GLCM* Con. green	11/3.7	$GLCM_{flow.dir.}$ Ent. blue	4/2.5

equal number O_{NLS}, which were randomly sampled from the entire population. The training sets consequently comprised between 3% (Barcelonnette) and 11% (Wenchuan) of the datasets, while the classification accuracies were assessed on the remaining test sets in terms of correctly classified objects.

RFs (N = 500) were trained and tested using once all object features and once only the previously selected feature subsets (Section 4.2). As expected the F-measures, which are the harmonic means of user's and producer's accuracies, indicated a generally lower predictive power than the *OOB errors*, but also enhanced accuracies if only the previously selected object metrics were used (Fig. 7). Especially for the cases Messina and Barcelonnette, with rather low overall accuracies, the feature reduction enhanced the F-measures by up to 5%.

More importantly, the test revealed that a balanced training sample did not provide balanced user's and producer's accuracies, and the RF overestimated the landslide area in all cases (Fig. 7). It could be argued that for hazardous processes such as landslide an over-detection might be easier to accept than omission. However, uncertainties in landslide inventories propagate forward into susceptibility assessment (e.g. Galli et al., 2008), and a high error of commission would lead to unrealistic overestimates of the associated hazard and risks. Under the assumption that in a real case it might be an acceptable additional labor to provide further O_{NLS} samples, the

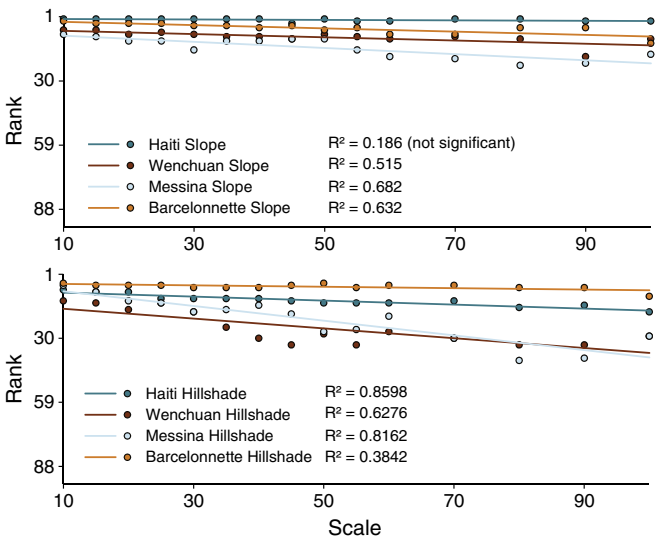


Fig. 5. Relationships between the VI-ranks of slope and hillshade and the segmentation scales. Linear regression lines fitting the data series show the overall trends, and their significance was tested at $p<0.05$ level.

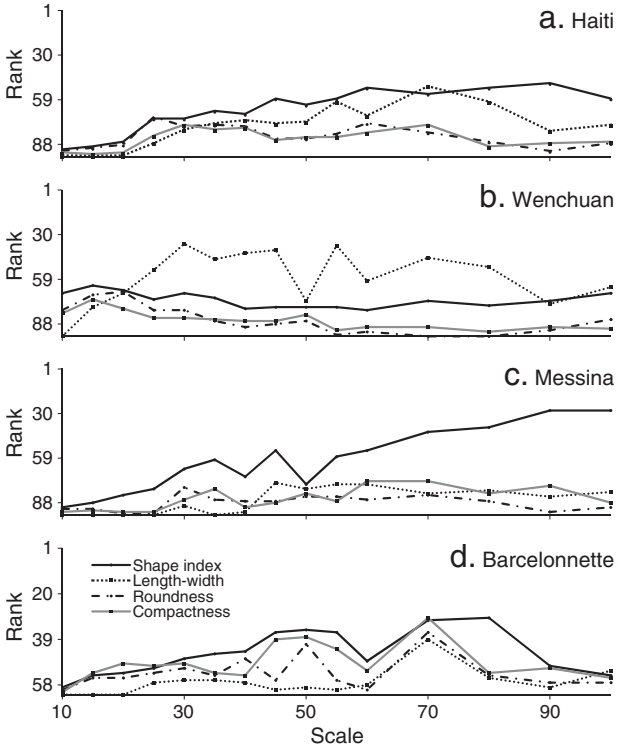


Fig. 6. Dependency of the variable importance of shape-metrics on the segmentation scale for the different test sites. Especially with small scale factors (<25) the segmentation did not generate landslide objects and non-landslide objects with distinguishable shapes. Only the length–width ratio (b, between scale 30 and 80) and the shape index (c, scale>55) had some impact on the accuracy.

Table 4
Final accuracy assessment for all test sites at three exemplary segmentation scales. Accuracies show the average performance of RFs (N = 500), trained with 20% of the O_{LS} and β_n–fold amount of O_{NLS}, applied to the test set Te_{80}.β_o is the original class-ratio of the entire population. The mean accuracies and their standard deviations were calculated over 50 randomly resampled replicates of Tr_{20}. The best results for each test site are indicated with bold numbers.

	Scale	$\beta_n(\beta_o)$	User's accuracy [%]	Producer's accuracy [%]	F_{area} [%]	F_{obj} [%]	β_n-adjusted Tr_{20}		
							O_{LS}	O_{NLS}	% of all objects
Haiti	10	3.0 (5.8)	**88.8 ± 0.1**	**85.7 ± 0.2**	**87.1 ± 0.1**	89.7 ± 0.1	4512	13536	11.7
	30	2.3 (4.2)	82.8 ± 1.2	87.1 ± 0.9	84.9 ± 0.7	88.3 ± 0.3	564	1297	12.8
	70	2.6 (4.0)	88.5 ± 1.1	72.4 ± 1.3	79.6 ± 0.7	88.5 ± 0.5	149	387	14.3
Wenchuan	10	2.7 (3.4)	**81.3 ± 0.1**	**81.1 ± 0.1**	**81.2 ± 0.1**	80.5 ± 0.1	6535	17645	17.0
	30	2.5 (3.0)	81.2 ± 0.4	77.1 ± 0.5	0.791 ± 0.2	80.3 ± 0.2	570	1425	17.4
	70	2.0 (2.6)	77.7 ± 0.9	75.3 ± 1.1	76.5 ± 0.6	79.9 ± 0.6	125	250	16.5
Messina	10	1.8 (4.2)	**72.9 ± 0.3**	**74.6 ± 0.2**	**73.7 ± 0.1**	73.0 ± 0.1	6135	11043	10.8
	30	1.9 (4.1)	69.0 ± 1.2	60.9 ± 0.9	64.7 ± 0.4	59.2 ± 0.4	663	1260	11.3
	70	1.9 (3.7)	64.3 ± 2.0	59.8 ± 1.3	62.0 ± 0.8	60.5 ± 1.1	125	238	11.9
Barcelonnette	10	4.7 (9.5)	**77.8 ± 1.0**	**78.0 ± 0.5**	**77.9 ± 0.4**	76.5 ± 0.2	1810	8507	10.8
	30	5.5 (11.5)	74.7 ± 2.1	75.9 ± 1.8	75.2 ± 1.0	67.4 ± 0.8	237	1304	10.1
	70	4.9 (12.1)	63.3 ± 5.6	88.6 ± 2.3	73.3 ± 3.5	65.3 ± 2.7	46	226	8.9

next section of this paper examines a procedure to balance user's and producer's accuracy.

4.3. Accuracy assessment

The balancing of under- and over-detection and the final accuracy assessment (Fig. 2 d1–3) were performed at three exemplarily selected scales (10, 30, and 70) with the previously selected features and in a scenario where 20% of the data would be available for training. The datasets were split (Fig. 2 d1) into a training subset (Tr_{20}), used for the estimation of the class balance and the classifier construction, and a testing subset for the final accuracy estimate (Te_{80}).

4.3.1. Estimates of β_n from the training samples (Tr_{20})

For all cases we observed a strong over-prediction of landslide areas if a class-balanced training sample was employed. The over-prediction problem was more pronounced for Messina and Barcelonnette, where already visual examination of the images suggested a higher class-overlap than in the two other areas. In controlled experiments such a behavior of classifiers has been explained by a higher density of positive examples in the class-overlap region (e.g. García et al., 2007).

Nevertheless, the iterative increase of β_i described in Section 3.3.2 (Fig. 2 d1), which corresponds to a relative increase of O_{NLS} in the training sample ($train_{sub}$), was an efficient strategy to adjust the balance of user's and producer's accuracies in the test sets ($test_{sub}$). At all test sites the estimated ratios of β_n (Fig. 8) resembled solutions that were a trade-off between the natural class-distribution (Table 4) and a completely balanced sample. The highest β_n estimates were obtained for the Barcelonnette dataset, where also the over prediction problem was most prominent.

Larger segmentation scales generally lead to a smaller number of image objects, and the 20% benchmark for the proportion of training data consequently translated into a reduced number of training samples over constant areas. Fig. 8 shows that the reduced number of sample objects resulted in an increasingly large variability in the test set accuracies, and yielded larger uncertainties in the estimation of β_n. It can be demonstrated that in such cases an increased number of random replications for each β_i still led to smoother converging curves with one unique crossing (Fig. 8). However, it should also be stressed that the estimation of β_n still only provides an intelligent guess on the design of the training sample for the classification of "unknown" image objects. The efficiency of the estimated β_n, to generate an RF with balanced user's and producer's accuracies was examined for the test set Te_{80} as described in the final section of this paper.

4.3.2. Estimation of the accuracy on the test set (Te_{80})

The majority class (O_{NLS}) in the training sample (Tr_{20}) was under-sampled according to the estimated ratio β_n. A RF (N = 500) was constructed from the β_n-adjusted Tr_{20} and applied to the remainder test set Te_{80} to assess the efficiency of the β_n estimate and the overall

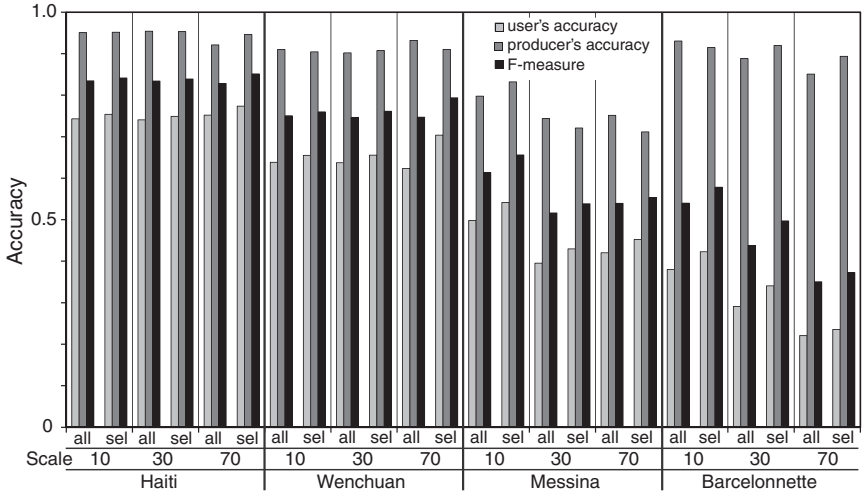


Fig. 7. Comparison of the accuracies of correctly classified objects before (all) and after (sel) variable selection at the different test sites and three exemplarily selected segmentation scales. 20% of all O_{LS} and an equal number O_{NLS} are used for the training of a RF (N = 500).

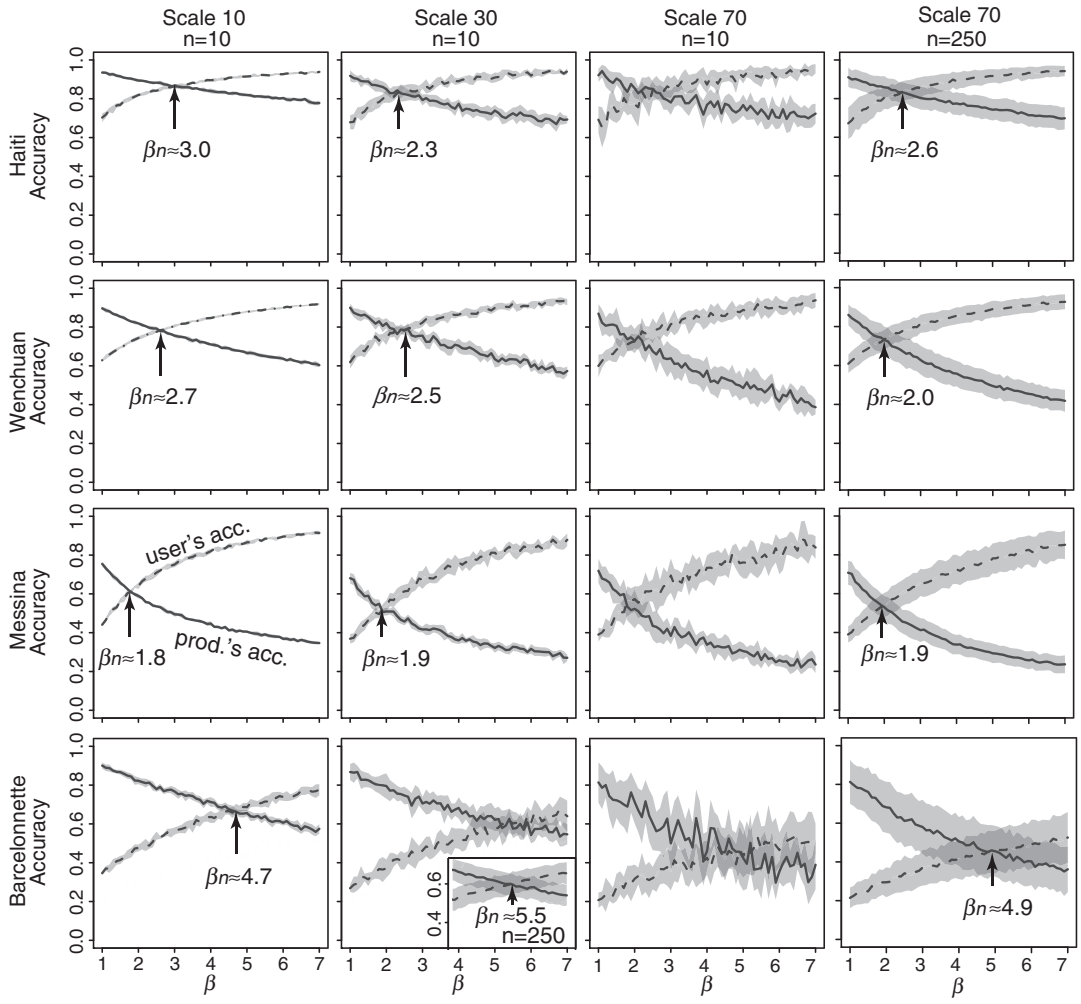


Fig. 8. Estimates of the class balance (β_n) in the training sample that lead to a balance of the mean user's (dashed black line) and mean producer's accuracies (solid black line). Accuracies are expressed in terms of image objects. The means of the accuracies for each β were calculated from 10-fold random replicate runs (n = 10). The grey margins show the corresponding standard deviations. For learning curves with high variance additionally figures from 250 random replicate runs (n = 250) are presented.

accuracy (Fig. 2 d3). The accuracy was assessed in terms of objects (Table 4, F_{obj}) and, to provide a final accuracy estimate for the entire approach, furthermore by comparing the classified areas with the landslides and non-landslide areas in the manually elaborated inventories (Table 4, user's accuracy, producer's accuracy, F_{area}). Each test was repeated with 50 β_n-adjusted randomly sampled replicates of Tr_{20}. The means and standard deviations of the achieved accuracies were calculated from the 50 runs and are displayed in Table 4. Although it did not solve the problem entirely, the strategy provided a significantly better balance between user's and producer's accuracies than could be achieved with the natural class distribution or an *ad hoc* balanced training sample (Fig. 7).

The accuracies in terms of correctly predicted area generally decreased for larger segmentation scales. At the test sites Haiti and Wenchuan this must be attributed to an increasing misfit between segmented object boundaries and the reference inventory leading to greater impurities within mixed objects. This means that the misclassified area increased due to a stronger generalization of the segments with a larger scale factor, the predictive accuracy of the RF (expressed by F_{obj}, Table 4) remained nearly constant among the different scales.

Conversely, for Messina and Barcelonnette F_{obj} was consistently lower than F_{area} (Table 4), and the classifier performance decreased significantly with larger scale factors. The comparatively higher areal accuracy can be explained by the fact that the average size of correctly classified objects was greater than those of misclassified objects. Spectral confusion and hence the importance of additional textural and topographic features was higher for the classification of the datasets from Barcelonnette and Messina (Table 3, Fig. 4). Leaving such features unconsidered during the segmentation may contribute to a higher class-overlap and a consequently lower F_{obj} at larger scales.

The general observations for the Messina test site confirm once more that omission is an especially likely error for the mapping of debris flows (Barlow et al., 2006; Lu et al., 2011), due to a high probability of occlusions in the local topography and beneath the remaining vegetation. At the Barcelonnette site most of the spectrally very similar badlands (Fig. 1 d) were successfully distinguished (Fig. 9 d) through the combination of spectral, textural and morphological features. Spatial clustering of missed areas at the crown and the toe of the landslide (Fig. 1 d, Fig. 9 d) indicates that such a landslide complex might be still better treated as a multi-class problem.

In summary, the RF classifier provided relatively high accuracies of up to 87% for the test sites Haiti and Wenchuan, while in the case of Messina the best model reached an accuracy of 73%. Those figures 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., 2011; Martha et al., 2010). Though the quantities of employed samples are not always explicitly mentioned (Barlow et al., 2006; Nichol & Wong, 2005), all proposed solutions depend on the availability of some sort of training data. Once the samples are provided, the framework

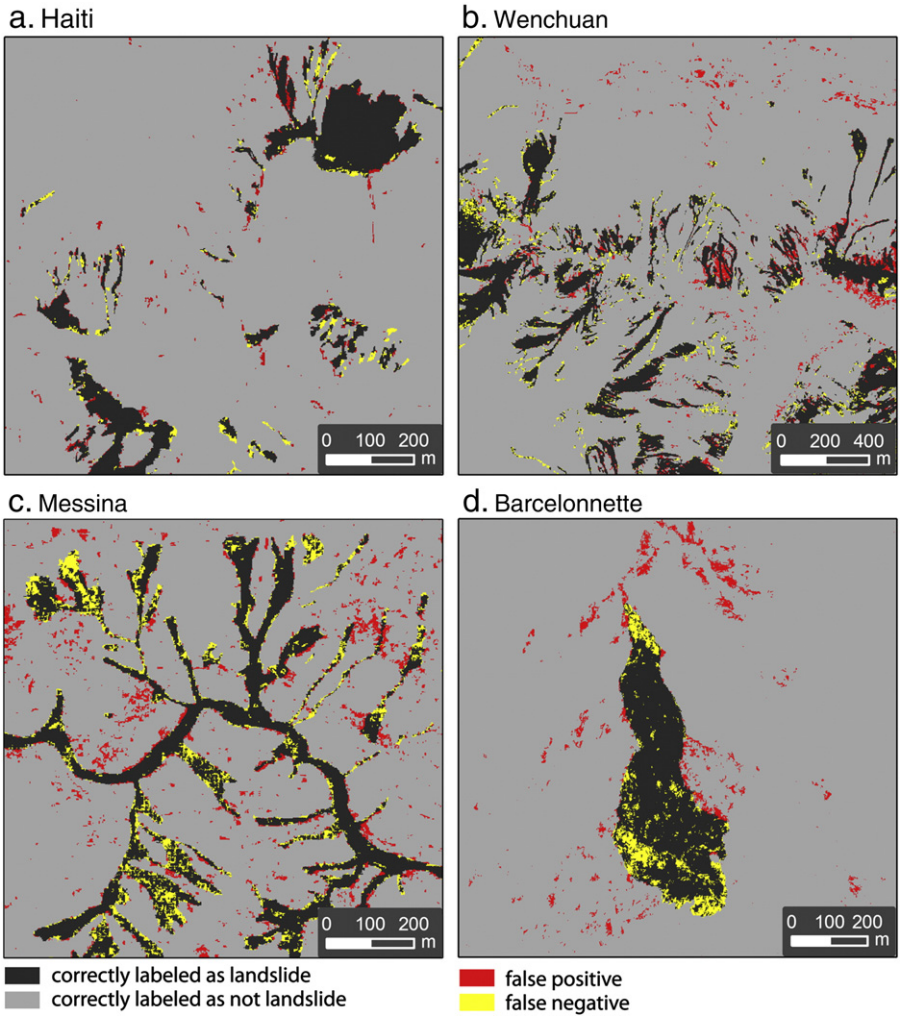


Fig. 9. Results with a segmentation scale of 10, after feature selection and balancing of the error rates as indicated in Table 4 at a, Haiti b, Wenchuan c, Messina and d, Barcelonnette. Correctly classified areas include the samples used for training.

presented in this paper has the potential to run fully automated with different image types, and liberates the user from the selection of appropriate features and thresholds. At each of the four test sites a medium resolution DEM, one VHR resolution image, and the reported numbers of training objects were sufficient for an efficient performance of the RF classifiers. In most practical situations such kind of data will be available, and the described algorithms may provide a generic approach to map the overall affected area more efficiently before site and data specific tasks, such as the classification of landslide types (Barlow et al., 2006; Martha et al., 2010), are targeted. In situations where more data (e.g. pre-event imagery) are available the proposed framework is suitable to accommodate a large variety of additional datasets and object metrics, which may be used to further increase the mapping accuracies.

In order to differentiate individual landslides and provide map products with less dispersed class distributions (Fig. 9) the current architecture still needs enhancements. This is closely related to the observed fact that high-level features such as shape are better exploited on larger scales (Section 4.1). The design of a hierarchical algorithm that robustly and efficiently incorporates sample data and relevant features in the classification, and delineation of image objects among a number of different scales, remains a major challenge, and with potential benefits for many remote sensing applications.

It also has to be noted that at this point we only explored the technical aspects of the supervised framework in relatively small test areas. A detailed analysis of the impact of sample quality and quantity provided by different users and over larger areas was beyond the scope of the present study, while research in this direction is certainly desirable before an operational use of the technique.

5. Conclusions

Previously proposed methods for object-oriented mapping of landslides from VHR images are highly reliant on manual thresholding and a subjective selection of suitable features, making it difficult to adapt them to new locations and datasets. To overcome such issues this study investigated the use of image segmentation and the Random Forest framework for feature selection and image classification. A variety of VHR remote sensing images and different landslide processes was analyzed with the RF data-mining technique to evaluate useful image object metrics, the influence of the segmentation scale, and the consequences of class-imbalance.

Although the optimal set of object metrics varies considerably from case to case, a number of spectral, topographic, and textural features are generally useful. Rotation-invariant and topographically-guided GLCMs provide complementary information to distinguish affected from non-affected areas, while topographically-guided GLCM derivatives introduced in this paper provide more significant enhancements. They also appear potentially useful for the automated mapping of other geomorphological processes such as gully erosion

(Shruthi et al., in press) or fluvial sediments. The range of potentially useful object metrics for landslide mapping and other applications seems still not fully exploited, and data mining techniques such as RF are valuable tools to ease feature selection for machine learning, or to guide experts during the elaboration of knowledge-driven rule sets. Our results indicate that feature reduction leads to an improved image classification, but also that not all significant features can be fully exploited with one particular segmentation scale.

Class-imbalance and class-overlap caused severely imbalanced error rates at all test sites. An iterative scheme to estimate a compensating class balance for the training data was found to enhance substantially the balance of user's and producer's accuracies. In the presented setup, accuracies between 73% and 87% were achieved when 20% of the total area was provided for training.

In the short term further enhancements are certainly possible through the integration of ancillary datasets such as pre-event imagery or the exploration of additional object metrics. More research is needed to optimize the segmentation process, which at present is based on spectral information solely. An initial sample-based estimate of the variable importance might thereby be an interesting tool to decide which further layers should be included in the segmentation. The processing time for a small test area can be streamlined to a few hours on a standard desktop PC, while for larger areas the RFs can easily be implemented for parallel processing, and the scale factor may provide an interesting parameter to trade between accuracy and processing time.

Acknowledgments

The work described in this paper was 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.

References

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.

Barlow, J., Martin, Y., & Franklin, S. E. (2003). Detecting translational landslide scars using segmentation of Landsat ETM + and DEM data in the northern Cascade Mountains, British Columbia. *Canadian Journal of Remote Sensing*, *29*, 510–517.

Benz, U. C., Hofmann, P., Willhauck, G., Lingenfelder, I., & Heynen, M. (2004). Multi-resolution, object-oriented fuzzy analysis of remote sensing data for GIS-ready information. *ISPRS Journal of Photogrammetry and Remote Sensing*, *58*, 239–258.

Blagus, R., & Lusa, L. (2010). Class prediction for high-dimensional class-imbalanced data. *BMC Bioinformatics*, *11*, 523.

Blaschke, T. (2010). Object based image analysis for remote sensing. *ISPRS Journal of Photogrammetry and Remote Sensing*, *65*, 2–16.

Booth, A. M., Roering, J. J., & Perron, J. T. (2009). Automated landslide mapping using spectral analysis and high-resolution topographic data: Puget Sound lowlands, Washington, and Portland Hills, Oregon. *Geomorphology*, *109*, 132–147.

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.

Breiman, L. (1996). Bagging Predictors. *Machine Learning*, *24*, 123–140.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*, 5–32.

Burez, J., & Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. *Expert Systems with Applications*, *36*, 4626–4636.

Carr, L., & Rathje, E. (2008). The use of remote sensing to identify landslides caused by the 2004 Niigata-ken Chuetsu earthquake in Japan. *6th International Workshop on Remote Sensing for Disaster Management Applications. Pavia, Italy*.

Cascini, L., Fornaro, G., & Peduto, D. (2010). Advanced low- and full-resolution DInSAR map generation for slow-moving landslide analysis at different scales. *Engineering Geology*, *112*, 29–42.

Civil-Protection-Sicily (2010). Landslide and mud food emergency Messina province, Itlay, October 1st 2009. In A.f.a.f.t.E.U.S. Fund (Ed.), *Regione Siciliana — Presidenza, Dipartimento dellla Protezione Civile* (pp. 83).

Clausi, D. A. (2002). An analysis of co-occurrence texture statistics as a function of grey level quantization. *Canadian Journal of Remote Sensing*, *28*, 45–62.

Cruden, D. M., & Varnes, D. J. (1996). Landslides types and processes. In A. K. Turner & R.L. Schuster (Eds.), *Landslides: Investigation and Mitigation* (pp. 36–75). Transportation Research Board, National Academy of Sciences: Washington D.C.

Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random Forest for classification in ecology. *Ecology*, *88*, 2783–2792.

Danneels, G., Pirard, E., & Havenith, H. -B. (2007). Automatic landslide detection from remote sensing images using supervised classification methods. *Geoscience and Remote Sensing Symposium. Barcelona*. Spain: IGARSS.

Debella-Gilo, M., & Kääb, A. (2011). Sub-pixel precision image matching for measuring surface displacements on mass movements using normalized cross-correlation. *Remote Sensing of Environment*, *115*, 130–142.

Denil, M., & Trappenberg, T. (2010). Overlap versus imbalance. In A. Farzindar & V. Kešelj (Eds.), *Advances in Artificial Intelligence* (pp. 220–231). Heidelberg: Springer Berlin.

Dey, V., Zhang, Y., & Zhong, M. (2010). A review on image segmentation techniques with remote sensing perspective. In W. Wagner & B. Székely (Eds.), *ISPRS TC VII Symposium — 100 Years ISPRS* (pp. 31–42). Vienna, Austria: IAPRS.

Diaz-Uriarte, R. (2010). *varSelRF: Variable selection using random forests. R package version 0.7-2*.

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

Drăguţ, L., Tiede, D., & Levick, S. R. (2010). ESP: A tool to estimate scale parameter for multiresolution image segmentation of remotely sensed data. *International Journal of Geographical Information Science*, *24*, 859–871.

Eberhard, M. O., Baldridge, S., Marshall, J., Mooney, W., & Rix, G. J. (2010). The MW 7.0 Haiti earthquake of January 12, 2010. *USGS/EERI Advance Reconnaissance Team report* (pp. 58). : U.S. Geological Survey Open-File Report.

Forsythe, K. W., & Wheate, R. D. (2003). Utilization of Landsat TM and digital elevation data for the delineation of avalanche slopes in Yoho National Park (Canada). *Geoscience and Remote Sensing, IEEE Transactions on*, *41*, 2678–2682.

Fusaro, V. A., Mani, D. R., Mesirov, J. P., & Carr, S. A. (2009). Prediction of high-responding peptides for targeted protein assays by mass spectrometry. *Nat Biotech*, *27*, 190–198.

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

García, V., Mollineda, R., Sánchez, J., Alejo, R., & Sotoca, J. (2007). When overlapping unexpectedly alters the classimbalance effects. In J. Martí, J. Benedí, A. Mendonça, & J. Serrat (Eds.), *Pattern Recognition and Image Analysis* (pp. 499–506). Heidelberg: Springer Berlin.

Gislason, P. O., Benediktsson, J. A., & Sveinsson, J. R. (2006). Random Forests for land cover classification. *Pattern Recognition Letters*, *27*, 294–300.

Gorum, T., Fan, X., van Westen, C.J., Huang, R.Q., Xu, Q., Tang, C., & Wang, G. (in press). Distribution pattern of earthquake-induced landslides triggered by the 12 May 2008 Wenchuan Earthquake. *Geomorphology*. doi:10.1016/j.geomorph.2010.12.030.

Guo, B., Damper, R. I., Gunn, S. R., & Nelson, J. D. B. (2008). A fast separability-based feature-selection method for high-dimensional remotely sensed image classification. *Pattern Recognition*, *41*, 1653–1662.

Hall-Beyer, M. (2007). GLCM Tutorial. http://www.fp.ucalgary.ca/mhallbey/tutorial.htm [15 May 2010].

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

Hayes, G. P., Briggs, R. W., Sladen, A., Fielding, E. J., Prentice, C., Hudnut, K., Mann, P., Taylor, F. W., Crone, A. J., Gold, R., Ito, T., & Simons, M. (2010). Complex rupture during the 12 January 2010 Haiti earthquake. *Nature Geoscience*, *3*, 800–805.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *Ieee Transactions On Knowledge And Data Engineering*, *21*, 1263–1284.

Hervás, J., Barredo, J. I., Rosin, P. L., Pasuto, A., Mantovani, F., & Silvano, S. (2003). Monitoring landslides from optical remotely sensed imagery: The case history of Tessina landslide, Italy. *Geomorphology*, *54*, 63–75.

Hervás, J., & Bobrowsky, P. (2009). Mapping: Inventories, susceptibility, hazard and risk. In K. Sassa & P. Canuti (Eds.), *Landslides — Disaster Risk Reduction* (pp. 321–349). Heidelberg: Springer.

Hervás, J., & Rosin, P. L. (1996). Landslide mapping by textural analysis of ATM data. *11th Thematic Conference on Applied Geologic Remote Sensing* (pp. 394–402). Las Vegas, USA.

Jenson, S. K., & Domingue, J. O. (1988). Extracting topographic structure from digital elevation data for geographic information system analysis. *Photgrammetric Engineering & Remote Sensing*, *54*, 1593–1600.

Joyce, K. E., Dellow, G. D., & Glassey, P. J. (2008). Assessing image processing techniques for mapping landslides. *IEEE International Geoscience and Remote Sensing Symposium* (pp. 1231–1234). Boston, USA.

King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis*, *9*, 137–163.

Kjekstad, O., & Highland, L. (2009). *Economic and social impacts of landslides* (pp. 573–587).

Laliberte, A., & Rango, A. (2009). Texture and scale in object-based analysis of subdecimeter resolution unmanned aerial vehicle (UAV) imagery. *IEEE Transactions on Geoscience and Remote Sensing*, *47*, 761–770.

Lawrence, R. L., Wood, S. D., & Sheley, R. L. (2006). Mapping invasive plants using hyperspectral imagery and Breiman Cutler classifications (randomForest). *Remote Sensing of Environment*, *100*, 356–362.

Liaw, A. (2010). *randomForest: Breiman and Cutler's random forests for classification and regression, Version 4.5-36*.

Lu, P., Stumpf, A., Kerle, N., & Casagli, N. (2011). Object-oriented change detection for landslide rapid mapping. *Geoscience and Remote Sensing Letters, IEEE* (pp. 701–705).

Malet, J. P. (2003). Les 'glissements de type écoulement' dans les marnes noires des Alpes du Sud. Morphologie, fonctionnement et modélisation hydro-mécanique. *Institut de Physique de Globe* (pp. 364). Strasbourg: Université Louis Pasteur.

Martha, T., Kerle, N., van Westen, C., Jetten, V., & Kumar, V. (in press). Segment optimisation and data-driven thresholding for knowledge-based landslide detection by object-oriented image analysis. *IEEE Transaction in Geoscience and Remote Sensing*. doi:10.1109/TGRS.2011.2151866.

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.

Moine, M., Puissant, A., & Malet, J. -P. (2009). Detection of landslides from aerial and satellite images with a semi-automatic method. Application to the Barcelonnette basin (Alpes-de-Haute-Provence, France). In J. -P. Malet, A. Remaître, & T. Bogaard (Eds.), *International Conference 'Landslide Processes'* (pp. 63–68). Strasbourg, France.

Nadim, F., Kjekstad, O., Peduzzi, P., Herold, C., & Jaedicke, C. (2006). Global landslide and avalanche hotspots. *Landslides*, *3*, 159–173.

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.

Nicodemus, K., Malley, J., Strobl, C., & Ziegler, A. (2010). The behaviour of random forest permutation-based variable importance measures under predictor correlation. *BMC Bioinformatics*, *11*, 110.

Owen, A. B. (2007). Infinitely imbalanced logistic regression. *Journal of Machine Learning Research*, *8*, 761–773.

Pesaresi, M., Gerhardinger, A., & Kayitakire, F. (2008). A robust built-up area presence index by anisotropic rotation-invariant textural measure. *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, *1*, 180–192.

Petley, D. (2009). On the occurrence of fatal landslides in 2008. *Geophysical Research Abstracts*, *11* (EGU2009-9030).

R-Development-Core-Team (2009). *R: A language andenvironment for statistical computing — Version 2.10.0*. Vienna, Austria: R Foundation for Statistical Computing.

Rau, J. -Y., Chen, L. -C., Liu, J. -K., & Wu, T. -H. (2007). Dynamics monitoring and disaster assessment for watershed management using time-series satellite images. *Geoscience and Remote Sensing, IEEE Transactions on*, *45*, 1641–1649.

Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, *23*, 2507–2517.

Shruthi, B.V.M., Kerle, N., & Jetten, V. (in press). Object-based gully feature extraction using high resolution imagery. Geomorphology.

Tang, H., Jia, H., Hu, X., Li, D., & Xiong, C. (2010). Characteristics of landslides induced by the great Wenchuan earthquake. *Journal of Earth Science*, *21*, 104–113.

Trimble (2011). *eCognition developer 8.64.0 reference book*. München: Germany Trimble Documentation.

Van Coillie, F. M. B., Verbeke, L. P. C., & De Wulf, R. R. (2007). Feature selection by genetic algorithms in object-based classification of IKONOS imagery for forest mapping in Flanders, Belgium. *Remote Sensing of Environment*, *110*, 476–487.

Van Den Eeckhaut, M., Vanwalleghem, T., Poesen, J., Govers, G., Verstraeten, G., & Vandekerckhove, L. (2006). Prediction of landslide susceptibility using rare events logistic regression: A case-study in the Flemish Ardennes (Belgium). *Geomorphology*, *76*, 392–410.

van der Werff, H. M. A., & van der Meer, F. D. (2008). Shape-based classification of spectrally identical objects. *ISPRS Journal of Photogrammetry and Remote Sensing*, *63*, 251–258.

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.

Watts, J. D., Lawrence, R. L., Miller, P. R., & Montagne, C. (2009). Monitoring of cropland practices for carbon sequestration purposes in north central Montana by Landsat remote sensing. *Remote Sensing of Environment*, *113*, 1843–1852.

Whitworth, M., Giles, D., & Murphy, W. (2005). Airborne remote sensing for landslide hazard assessment: A case study on the Jurassic escarpment slopes of Worcestershire, UK. *Quarterly Journal of Engineering Geology & Hydrogeology*, *38*, 285–300.

Woodcock, C. E., & Strahler, A. H. (1987). The factor of scale in remote sensing. *Remote Sensing of Environment*, *21*, 311–332.

Yu, W., Wu, B., Huang, T., Li, X., Williams, K., & Zhao, H. (2006). Statistical methods in proteomics. In H. Pham (Ed.), *Handbook of Engineering Statistics* (pp. 623–636). London: Springer.