Nat. Hazards Earth Syst. Sci., 14, 259–278, 2014
www.nat-hazards-earth-syst-sci.net/14/259/2014/
doi:10.5194/nhess-14-259-2014

Natural Hazards and Earth System Sciences
Open Access

Sample size matters: investigating the effect of sample size on a logistic regression susceptibility model for debris flows

T. Heckmann[1], **K. Gegg**[1], **A. Gegg**[2], **and M. Becht**[1]

[1]Physical Geography, Catholic University of Eichstaett-Ingolstadt, Ostenstr. 18, 85072 Eichstaett, Germany
[2]Statistics, Catholic University of Eichstaett-Ingolstadt, Ostenstr. 26, 85072 Eichstaett, Germany

Correspondence to: T. Heckmann (tobias.heckmann@ku.de)

Received: 21 May 2013 – Published in Nat. Hazards Earth Syst. Sci. Discuss.: 21 June 2013
Revised: 5 December 2013 – Accepted: 9 January 2014 – Published: 17 February 2014

Abstract. Predictive spatial modelling is an important task in natural hazard assessment and regionalisation of geomorphic processes or landforms. Logistic regression is a multivariate statistical approach frequently used in predictive modelling; it can be conducted stepwise in order to select from a number of candidate independent variables those that lead to the best model. In our case study on a debris flow susceptibility model, we investigate the sensitivity of model selection and quality to different sample sizes in light of the following problem: on the one hand, a sample has to be large enough to cover the variability of geofactors within the study area, and to yield stable and reproducible results; on the other hand, the sample must not be too large, because a large sample is likely to violate the assumption of independent observations due to spatial autocorrelation. Using stepwise model selection with 1000 random samples for a number of sample sizes between $n = 50$ and $n = 5000$, we investigate the inclusion and exclusion of geofactors and the diversity of the resulting models as a function of sample size; the multiplicity of different models is assessed using numerical indices borrowed from information theory and biodiversity research. Model diversity decreases with increasing sample size and reaches either a local minimum or a plateau; even larger sample sizes do not further reduce it, and they approach the upper limit of sample size given, in this study, by the autocorrelation range of the spatial data sets. In this way, an optimised sample size can be derived from an exploratory analysis. Model uncertainty due to sampling and model selection, and its predictive ability, are explored statistically and spatially through the example of 100 models estimated in one study area and validated in a neighbouring area: depending on the study area and on sample size, the predicted probabilities for debris flow release differed, on average, by 7 to 23 percentage points. In view of these results, we argue that researchers applying model selection should explore the behaviour of the model selection for different sample sizes, and that consensus models created from a number of random samples should be given preference over models relying on a single sample.

1 Introduction

Spatial modelling, i.e. finding and applying a model of the spatial distribution of some phenomenon, can be used for two slightly different purposes: first for regionalisation, i.e. the transfer of findings from the surveyed area to some larger region. In geomorphology, the methodological framework for regionalising the occurrence of a process or a landform (that is associated with the activity of geomorphic processes) is termed "predictive geomorphological mapping" (Luoto and Hjort, 2005). It can be helpful in reducing time, cost and, to some degree, subjectivity associated with area-wide geomorphological mapping (van Asselen and Seijmonsbergen, 2006). Second, models are applied to identify areas where the phenomenon might occur in the future, even or especially where there is no evidence of recent activity. The (spatial) probability of occurrence of an event forms an important factor of the hazard term in quantitative risk assessment, although for a complete formulation one also needs to consider the temporal probability and the magnitude–frequency relationship of events (Guzzetti et al., 2006a). However, spatial modelling includes some temporal aspects as well. Specifically for landslides, the most important underlying assumptions (see Pike et al., 2003, for more) are (i) that landslides

can occur and/or have occurred in the larger area wherever the conditions are equal or similar to those in the surveyed area and (ii) that future events will take place under conditions the same as or similar to those in the past (e.g. Fabbri et al., 2003).

In this study, we apply the method of multivariate logistic regression to the identification of potential debris flow initiation sites in a high mountain catchment; the spatial unit is the raster cell (as opposed to e.g. slope units; see Van Den Eeckhaut et al., 2009). Together with discriminant analysis (e.g. Baeza and Corominas, 2001), soft computing techniques – such as "weights of evidence" (Bonham-Carter, 1994; Neuhäuser and Terhorst, 2006) or "certainty factor" (e.g. Binaghi et al., 1998) – and artificial neural networks (e.g. Lee et al., 2003; Ermini et al., 2005; Liu et al., 2006), logistic regression belongs to the most frequently chosen approaches to spatial modelling of landslides (Atkinson et al., 1998; Ohlmacher and Davis, 2003; Beguería and Lorente, 2003; Brenning, 2005; Ayalew and Yamagishi, 2005; Beguería, 2006; Van Den Eeckhaut et al., 2006; Meusburger and Alewell, 2009; Van Den Eeckhaut et al., 2010; Atkinson and Massari, 2011; Ruette et al., 2011; Guns and Vanacker, 2012). Recently, some published studies dealt specifically with debris flow susceptibility models on the regional scale; for the identification of potential release areas, a range of different approaches has been used, including heuristic (Horton et al., 2008; Kappes et al., 2011; Fischer et al., 2012) and statistical ones (Heckmann and Becht, 2009; Blahut et al., 2010a, b). The so-delineated release areas can be used as starting points for models that predict the pathways, lateral extent, runout length and other relevant properties of debris flows (e.g. Blahut et al., 2010b; Kappes et al., 2011), which is important for hazard assessment and has also been used in geomorphological applications, for example research on sediment cascades (Wichmann et al., 2009; Heckmann and Schwanghart, 2013).

In order to use a model for prediction, a sample has to be drawn, and the model parameters of the population are estimated based on that sample. Sampling is essential, because event and non-event units show spatial autocorrelation (see Sect. 1.2), and dependent data lead too easily to the rejection of null hypotheses and the incorrect declaration of parameters as significant; Legendre (1993) explains this for ecological models (see also Van Den Eeckhaut et al., 2006). Using a stepwise approach, the predictor variables for an effective yet parsimonious model are selected from a set of candidate geofactors (Sect. 3.2.2). Brenning (2005) found that logistic regression with stepwise variable selection yielded the lowest error rates in his comparison of different statistical methods. Logistic regression was also the best single method in the comparative study by Rossi et al. (2010), and exhibited the highest area under the curve (AUC) for "fine slope units" (second rank in overall comparison) in Carrara et al. (2008), a study specifically referring to debris flows.

The choice of predictor variables will understandably depend on the sample (Guns and Vanacker, 2012), and it is also clear that the aim of every susceptibility model should be a reliable and reproducible prediction. This prediction should not depend too much on the sample that is taken in order to select the variables and estimate the model parameters. Several previous studies do not involve sampling at all (e.g. Ohlmacher and Davis, 2003; Ruette et al., 2011); i.e. they use all available data for estimating the model parameters. The majority of studies use only one single sample (e.g. Atkinson et al., 1998; Van Den Eeckhaut et al., 2006; Meusburger and Alewell, 2009), the size of which usually depends on the number or size of landslide initiation zones (see Sect. 1.1). Recognising the dependence of model results on the sample, Brenning (2005) takes 50 samples to compare error rates across different sample sizes and statistical methods. Beguería (2006) and Guns and Vanacker (2012) apply 50-fold replication in order to estimate the robustness of the modelling result with respect to sampling, and Van Den Eeckhaut et al. (2010) calculate an ensemble of 25 models from different samples of their data. Hjort and Marmion (2008) conduct repeat sampling to explore the influence of sample size on the predictive power of (among others) multiple logistic regression models for predictive geomorphological mapping.

The present study has two main foci that will be developed in detail in the following subsections. It is not the aim of our study to find out the best performing method for a debris flow susceptibility model (comparative studies of predictive models were carried out, for example, by Brenning, 2005; Marmion et al., 2008; Carrara et al., 2008; Vorpahl et al., 2012); we deliberately chose logistic regression for its widespread use, and for the relevant assumption of sample independence which we found to be frequently neglected in previous studies. First, we explore the sensitivity of stepwise model selection to sample size. Sections 1.1 and 1.2 will explain why the sample size must neither be too small nor too large. In this context, the main aim of the study is to investigate if an "optimal" sampling size can be found as a compromise between samples too small and too large. Second, we quantify the uncertainty inherent in a stepwise modelling approach, with respect to (i) the selection of geofactors, (ii) model parameters, and (iii) the spatial pattern of uncertainty in the resulting susceptibility map. This study aim will be developed in Sect. 1.3.

1.1 Constraints on sample size 1: why the sample must not be too small

In inferential statistics, confidence intervals are calculated for population parameters based on a sample; the width of the former depends, besides the desired confidence level, on the sample size. Small samples result in large standard errors and wide confidence intervals for the population parameters. In the case of regression parameters, small samples

cause the estimation to be uncertain, and there is a higher risk of coefficients being insignificant when the respective confidence interval includes zero. With respect to replicate sampling and model selection, it is expected that the diversity of models (and hence the dependence of the models on the sample) will be large in this case.

Moreover, in a large study area, a small sample is unlikely to cover the variability of geofactors, especially if several of them are part of the model. Here, a larger sample would include more information on the study area and would possibly provide a better model. There are rules of thumb that estimate the minimum sample size for a regression analysis on the basis of a constant (e.g. > 50), of the ratio of observations and predictor variables, or of a combination of the latter; such rules have been explored in light of significance, power and effect size, e.g. by Green (1991), who found "some support" for the rule of thumb $n_{\min} \geq 50 + 8\,m$, where $n_{\min}$ is the minimum sample size and m is the number of predictor variables.

In this study, when we speak of sample size, we always address a sample of "non-events", i.e. a sample of raster cells without debris flow initiation. If a random sample referred to all raster cells, including event and non-event cells, the number of event cells in the sample would certainly be smaller than in the original inventory. This would cause a loss of information particularly for those cells that represent the target of the modelling exercise; therefore, all initiation areas will be represented in the models and only the size of the non-event sample is varied in our investigation. Besides the non-event sample size, the relative sample size n_{rel} (i.e. the areal extent of the total sample divided by the size of the study area) will be reported.

1.2 Constraints on sample size 2: why the sample must not be too large

While it is intuitive that larger samples contain more information that can be used by the model, and the model might be better, there are several reasons why the sample size must not be too large either.

King and Zeng (2001) argue that the non-event sample size has to be kept as small as possible because of the disproportionate cost and effort of acquiring data for many variables and observations that are not related to the target phenomenon (event). Like in political science, the acquisition of observations is costly in ecology (with the application of regression models to the spatial prediction of species distribution). In this context, the complexity of the investigated systems is reflected in large numbers of predictors; moreover, the logistic difficulty of mapping the presence or absence of a species in large and remote areas should not be underestimated (see e.g. Stockwell and Townsend Peterson, 2002). An important justification for predictive geomorphological mapping (Luoto and Hjort, 2005) is that area-wide field mapping is time-consuming, difficult in remote or inaccessible areas, and may suffer from subjectivity (van Asselen and Seijmonsbergen, 2006; Hjort and Marmion, 2008). However, in contrast to the examples from political and ecological science, many if not most variables in predictive geomorphological mapping are easily derived from digital elevation models and remote sensing data; both are available globally, with ever-increasing accuracy and resolution. This does not change the effort required for mapping the target phenomenon ("events"), but the motivation for limiting sample size of non-events appears to be quite different, as it does not so much refer to the effort of data acquisition (quantity and quality). In order to limit the sample size and to mitigate the rare-events issue (see below), the literature suggests different ratios of event : non-event sample sizes, mostly without justifying the particular choice of this ratio. Instead of merely adopting one of these suggestions (which generally range from 1 : 1 to 1 : 10), our paper aims at an empirical analysis of sample dependence and performance of the susceptibility model as a function of sample size.

Other reasons for restricting sample size are overparameterisation and overfitting of the model (Hjort and Marmion, 2008, and references therein). Increasing sample sizes causes standard errors and confidence intervals in parameter estimation to decrease. In a significance-based stepwise model selection, very large samples are expected to facilitate the inclusion of more and more variables (risk of overparameterisation). Such inclusion of more information does not necessarily lead to better model performance; Stockwell and Townsend Peterson (2002) describes "plateaus" wherein new data add little to model performance. In some cases, inclusion of more data even causes worse performance, because a model fit to a very specific set of information may perform poorly on new data (risk of overfitting; see Stockwell and Townsend Peterson, 2002, and references therein). Brenning (2005), however, states that overfitting is "not a serious problem for logistic regression", contrary to machine-learning methods (cf. Petschko et al., 2014, and references therein).

The most serious reason for limiting the sample size is spatial autocorrelation. Logistic regression generally requires few assumptions to be met; the most important are (i) the independence of observations and (ii) uncorrelated independent variables. While violations of the second assumption can be avoided by testing for multicollinearity and excluding variables (see Sect. 3.2.1), the first assumption proves to be critical when dealing with spatial data. Geofactors tend to have very similar values in a close neighbourhood, a property called spatial autocorrelation. If several observations from nearby sites are included in a model, the independence assumption will not hold. In the case of the generalised linear modelling approach adopted in this study, the maximum likelihood method that is used to estimate the model parameters strictly requires the observations to be independent (e.g. Hosmer and Lemeshow, 2000). Atkinson and Massari (2011) explain that (spatial) autocorrelation of the geofactors causes the model residuals to be spatially autocorrelated (which is not acceptable as model residuals have to

be uncorrelated), and that this may lead to "incoherent significance estimates for the parameters" (see also Brenning, 2005). Consequently, such incoherent estimates compromise both significance-based model selection and the assessment of parameter importance that is based on the latter.

In previous studies applying logistic regression to landslide susceptibility analysis, the problem of stochastically dependent samples has frequently been ignored (e.g. by using all available data instead of a sample; see above). In some instances, the risk of autocorrelation is dealt with for events only, as geofactors tend to be homogeneous (and consequently strongly autocorrelated) on landslide terrain (Atkinson and Massari, 2011). However, the independence assumption refers to all observations of the dependent variable (Hosmer and Lemeshow, 2000; Van Den Eeckhaut et al., 2006), in our case to the occurrence and non-occurrence of debris flow initiation. As the geofactors used as independent variables are supposed to be associated with the dependent variable, we argue that the degree of autocorrelation of these geofactors should be accounted for in the sampling procedure. In order to mitigate the issue of spatial autocorrelation, some authors choose one raster cell for each landslide source area on a systematic basis. Atkinson et al. (1998) and Van Den Eeckhaut et al. (2006), for example, use the centre of each landslide source area. Similarly, Vanwalleghem et al. (2008) use the centre of each topographic depression, and the centre of each gully in their study predicting the spatial distribution of closed depressions and gullies under forest. Different authors draw samples of source areas on different grounds; besides spatial autocorrelation, Atkinson et al. (1998) explain their approach with the aim of preventing model bias towards larger landslides – in a full sample of events, more data would enter the model from larger source areas than from smaller ones. Beguería and Lorente (2003) use one raster cell for each debris flow initiation zone because the raster size (10 m) of the data in their study corresponds to the size of a typical debris flow scar. All approaches have in common that they prevent a contiguous (and hence potentially strongly spatially autocorrelated) sample of hundreds of landslide initiation cells from entering the model. Spatial autocorrelation has also been accounted for in model validation (Brenning, 2005). However, as Atkinson and Massari (2011) point out, autocorrelation in the geofactors is frequently not adequately accounted for in the regression model. While the latter study proposes an autologistic model (see also Brenning, 2005), we will try to warrant independence of observations through the choice of an adequate sampling size (see Sect. 3.3.2): as the number of sampled raster cells in a finite study area increases, the average distance between those cells will decrease, and finally the independence assumption will no longer hold given the spatial autocorrelation of the geofactors.

Normally, a logistic regression model is fit to a sample where the ratio of event : non-event cases is approximately 1 : 1. Then, the so-called cutoff, i.e. the value of the model result that discriminates between event and non-event, equals 0.5. King and Zeng (2001) explain that the number of non-events should be typically 2–5 times higher than that of events. In this case, the cutoff needed to translate the model result to a classification (event or non-event) would need to be adjusted accordingly. Because the ratio of event : non-event spatial units (not only raster cells but also lumped spatial units; Beguería and Lorente, 2003) usually is by far smaller, a bias towards small probabilities arises.[1] This problem has been addressed by the development of "rare events logistic regression" (King and Zeng, 2001). Besides endogenous stratified sampling (a sampling strategy that includes all events plus a random sample of non-events), these authors propose corrections for the intercept and for the estimated probabilities. Rare-events logistic regression was applied in landslide susceptibility modelling by Van Den Eeckhaut et al. (2006) and Guns and Vanacker (2012). In many studies, endogenous stratified sampling has been adopted, and the authors chose event : non-event ratios of 1 : 1 (e.g. Brenning, 2005; Meusburger and Alewell, 2009; Van Den Eeckhaut et al., 2010), 1 : 2 (Wang and Sassa, 2005), 1 : 5 (Van Den Eeckhaut et al., 2006), or 1 : 10 (Beguería and Lorente, 2003; Beguería, 2006; Guns and Vanacker, 2012). Finally, Atkinson et al. (1998), who use only the central cell of each landslide as the event sample, sample as many non-event cells as required in order to attain the ratio of landslide to non-landslide area.

In our study, we adopt stratified random sampling by a random sample of one cell for each debris flow initiation zone, and a random sample of non-event cells. The size of the latter is then varied in order to explore the effect on stepwise model selection; hence, we do not pre-select an event : non-event ratio. Rare-event correction according to King and Zeng (2001) is not applied.

1.3 Uncertainty: model selection, parameters, spatial patterns

The result of the investigations motivated in the previous subsections is a suitable sample size reaching a compromise between sample sizes too small and too large. The aims of this procedure can be summarised as follows: first, a stable model selection that is a low diversity of geofactors remaining in the repeat stepwise selection; second, the independence of the sample (i.e. avoiding spatial autocorrelation). Even with an optimised sample size in that respect, the selection of predictor variables will still depend on the specific sample. As different predictor variables, with their distinct spatial structure, will be part of the model when the procedure is repeated with a different sample, the spatial pattern of the resulting susceptibility map will also differ from time to time; the predictive power of the model might be different as well.

[1] In our study areas, the ratio of release area cells to the total study area is 1 : 200 and 1 : 500, respectively.

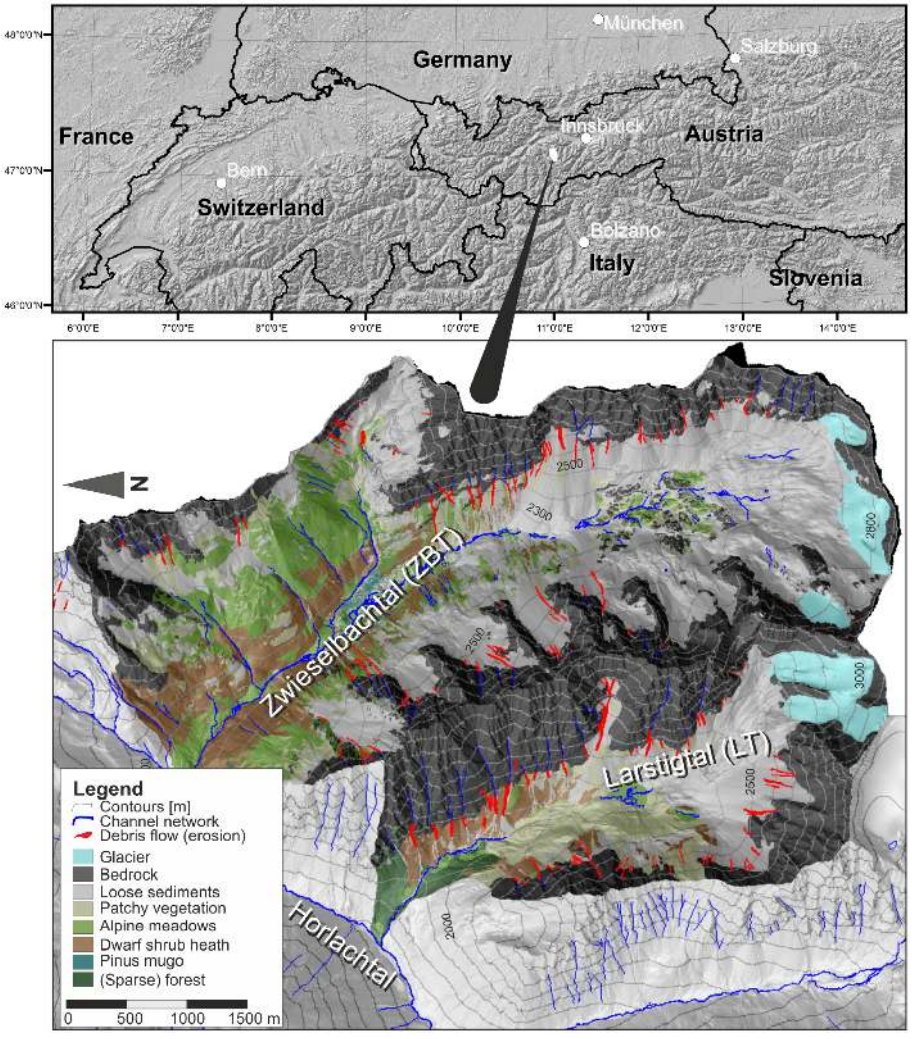

Fig. 1. Overview of the study areas.

The second main goal of this study is to elucidate three aspects of this uncertainty: (i) geofactors and how often they are included after stepwise selection, (ii) the range of model parameters estimated for the replications, and (iii) the spatial distribution of differences in the estimated susceptibility. This is important because, in the majority of studies employing sampling for model calculation, only one sample is taken, and no account is given of uncertainty beyond the standard errors of the parameters. On the other hand, most studies involving repeat sampling (e.g. Brenning et al., 2005; Beguería, 2006; Van Den Eeckhaut et al., 2010; Guns and Vanacker, 2012) concentrate on the set of geofactors, the parameters and the predictive ability of the models, and do not investigate how this affects the spatial distribution of susceptibility. Only rarely has the spatial distribution of model uncertainty been addressed using multiple replication approaches (e.g. Guzzetti et al., 2006b; Luoto et al., 2010; Petschko et al., 2014).

2 Study area

This study has been conducted in two adjacent subcatchments of the Horlachtal, a tributary of the Oetztal, located in the Austrian Central Alps (Stubai Alps). The two valleys, the Zwieselbachtal (ZBT, ca. 19 km^2) and the Larstigtal (LT, ca. 7 km^2), strike approximately S–N and have a typical trough cross-section. Due to their adjacency, they are similar in their natural characteristics. Figure 1 shows the location and an overview of the catchments. The most important properties of the study areas are listed in Table 1; the Horlachtal and its subcatchments are described in more detail by Rieger (1999) and Geitner (1999).

The lithology of both valleys is dominated by gneiss and mica schist; metamorphic granites can also be found. Pleistocene glaciations have shaped the valleys and are evidenced by glacial landforms (e.g. moraines, cirques, roches moutonnées). Glacial cirques are concentrated on the east-facing

Table 1. Selected properties of the study areas.

Property	LT	ZBT
Area [km^2]	7.04	18.77
Elevation [m a.s.l.]	1770–3287	1903–3188
Slope, mean ± std.dev. [°]	35.6 ± 14.31	31.6 ± 13.78
Slope, range [°]	0.01–82.05	0.01–83.12
Roughness, mean ± std. dev.	0.13 ± 0.12	0.11 ± 0.11
Roughness, range	0–0.90	0–0.86
Inventory [no. of events]	64	81
Mapped release area [m^2]	33 400	37 875
Land cover [%]		
Glacier	5.8	2.4
Bedrock	46.6	29.1
Unvegetated scree	25.6	38.3
Patchy vegetation	10.9	5.9
Alpine meadows	1.9	14.1
Dwarf shrub heath	5.6	9.9
Dwarf mountain pine (*Pinus mugo*)	–	0.2
Woodland	3.6	0.1

valley sides, whereas the west-facing valley sides are marked by extensive scree slopes. Currently, the two catchments are formed primarily by fluvial and gravitational processes such as rock falls and debris flows. Sediment transfer through the catchments is limited as the valleys consist of largely disconnected subsystems (at least with respect to the transport of coarse sediment; see Heckmann and Schwanghart, 2013) separated by alluvial reaches of the Zwieselbach and Larstig creeks, respectively. These reaches are located immediately upstream of the terminal moraines of the Little Ice Age and of the particularly well-preserved terminal moraines of the Egesen stadial (corresponding to the Younger Dryas, ca. 11 to 12 ka BP, recent datings for the European Alps are listed by Ivy-Ochs et al., 2008).

Debris flows in both study areas can be termed slope-type debris flows of type 2 according to Zimmermann et al. (1997). Events of this type initiate on scree slopes following failure that is caused by positive pore water pressure in the course of intense rainfall, and by progressive erosion. This is often the case at the base of rock walls where debris flow formation is triggered by the so-called "firehose effect" (Johnson and Rodine, 1984) which describes concentrated flux of water out of the rock face onto the talus. Slope-type debris flows can be regarded as a transport-limited process; thus their frequency is primarily controlled by hydroclimatic events (Bovis and Jakob, 1999). In the study area, rain intensities of around 20 mm within half an hour have been reported to trigger debris flows (Becht, 1995; Rieger, 1999), while Zimmermann et al. (1997) suggest regional intensity-duration thresholds of about 11 mm per hour. The threshold is comparatively low, which has been attributed to the low mean annual precipitation (Hagg and Becht, 2000) of ca. 1000 mm (Becht, 1995).

Vegetation primarily consists of dwarf shrub heath, alpine meadows and pioneer vegetation. At elevations of > 2300–2500 m, bedrock and scree are predominant. In general, more than 60 % of the study area are completely lacking vegetation cover.

3 Data and methods

3.1 Data and data preparation

3.1.1 Debris flow inventory

Like every statistical approach, logistic regression requires an inventory of targets (here: a map of debris flow initiation areas) for the dependent variable, and maps of (potentially) influencing factors as independent variables, hereafter referred to as geofactors. The dependent variable (here: debris flow initiation) is observed as a binary variable (1: presence; 0: absence). The debris flows inventory of the Zwieselbachtal and Larstigtal catchment was compiled using orthophoto and field maps (Thiel, 2013), updating an earlier inventory for which debris flows had been surveyed using a total station (Rieger, 1999). It contains 81 events within the Zwieselbachtal and 64 events within the Larstigtal. Debris flows areas are represented by polygon features (which had to be converted to raster format for the pixel-based approach of this study), and divided into three zones related to geomorphic activity: erosion (indicated by incision), transition (indicated by a channelised reach accompanied by levées) and the depositional lobe(s). Conceptually, as the susceptibility map specifically aims at predicting potential initiation zones, the event samples for the regression models should be taken from the erosional zones, preferably from the uppermost part as the latter represents the area where events typically started (and probably will also initiate in the future). The strategy of using only the detachment zone of a mass movement for susceptibility modelling has been advocated by several workers (see for example Van Den Eeckhaut et al., 2006; Heckmann and Becht, 2009); Magliulo et al. (2008), however, report that this restriction does not automatically lead to better results. The initial idea of manually setting one raster cell for each debris flow initiation zone was discarded, because placing this raster cell in the channelised part would introduce a bias towards larger catchment areas and concave plan curvature. Therefore, a GIS procedure was used to select, for each debris flow erosional zone, the area that is higher than the P75 percentile of elevation, i.e. the uppermost 25 %. The raster cells belonging to the initiation zone of each debris flows are coded with an ID, allowing for a stratified random sampling of one cell per debris flow event for each regression model.

Guzzetti et al. (2012) discuss the importance of landslide inventory maps and report on advantages, limitations and new methodological developments. With respect to susceptibility mapping, the quality of the underlying inventory is a

limiting factor for the reliability of predictive models (e.g. Ardizzone et al., 2002). While fresh landslides are readily detected, post-event modifications such as human impact (e.g. ploughing), land cover change, erosion and landslide reactivation etc. can hamper the identification of landslides and thus jeopardise the completeness of the inventory (Bell et al., 2012, e.g., analyse persistence and change of landslide morphology depending on age). For debris flows in our study area, however, we argue that the risk of false negatives, i.e. the risk of an incomplete inventory due to overlooked debris flow scars, is small: the activity of debris flows tends to persist once it has started, because an incision enhances and sustains the convergence of surface runoff. Due to the transport-limited conditions of debris flow initiation in our study area, this is supposed to hold for a long time, until either sediment storage is depleted or slope gradient has become too low. Conversely, debris flow deposits are frequently modified by renewed activity, and less pronounced depositional lobes can lose contrast on aerial photos due to progressive weathering (see e.g. Heckmann et al., 2008). Human activities that could potentially modify the appearance of debris flow scars are completely absent in the relevant regions of our study area.

3.1.2 Digital terrain model

Before model selection (see Sect. 3.2.2), geofactors conceptually related to debris flow initiation have been pre-selected. Debris flow initiation is related to (i) the availability of mobile debris, (ii) steep slopes, and (iii) large amounts of water, typically provided by intense rainfall. Not all influencing factors in these three groups (material, relief, water) can be directly measured or calculated; many of them, however, can be derived from a DEM, either directly or as proxies. Although geological and land cover maps were available, we tried to use only geofactors that can be derived from high-quality digital elevation models (DEMs) in order to test the feasibility of DEM-based modelling. Such high-quality DEMs are increasingly available for large parts of the world.

For the derivation of several topographical parameters used as geofactors for the regression models, we used a raster DEM with a resolution of 1 m that was interpolated from an airborne lidar survey in the year 2006. For most applications, and for the modelling itself, the original DEM (DEM1) was resampled to a raster resolution of 5 m (DEM5). Apart from saving memory and computing time, the resampling smoothes the DEM so that very fine scale topography is no longer contained in the resulting DEM5. This effect is desired, as debris flow initiation is not expected to result from microscale topography.

Information on *available sediment* is usually provided by land cover and/or geological maps. The former mainly contain information on vegetation that might in some cases stabilise soils and sediments. The latter focus on different types of bedrock. In this study, the "available sediment" group is represented by one single geofactor (roughness class). This geofactor is derived from a cluster analysis of slope (DEM5; see below) and roughness. Roughness was calculated as the "vector ruggedness measure" (Sappington et al., 2007) on the DEM1 within a moving window of radius 5 m, and the result was resampled to the same resolution and extent as the DEM5 using the nearest-neighbour approach. The comparatively small radius was chosen to capture the roughness of surfaces rather than the roughness induced by landforms, e.g. by gullies. The cluster analysis yields two clusters closely representing (i) bedrock and (ii) areas covered by sediments. For the Zwieselbachtal, this unsupervised classification could be validated with a very detailed land cover map created from orthophoto imagery; the ϕ coefficient of the mapped vs. the DEM-based classification was 0.78. The reason for the satisfactory fit is the characteristic fine-scale roughness [2] of bedrock areas that can easily be discerned on a shaded relief map, together with the existence of a sharp threshold of slope (resembling the angle of internal friction) above which an area cannot be covered by unconsolidated scree. Leaving out the information on land cover/vegetation is not expected to be decisive in our case study, because the study areas are only sparsely covered with vegetation, mostly grass, and forest is widely missing, at least in the areas relevant for debris flow genesis.

Relief parameters were derived from the DEM5 using the algorithm of Zevenbergen and Thorne (1987) implemented in SAGA GIS (www.saga-gis.org). As slope stability, especially for scree, is a function of *slope*, this parameter is expected to be very important for debris flow initiation. As both valley axes have a north to south orientation (resulting in a strong bias towards east- and west-facing slopes), and as the physical role of *aspect* cannot be described unambiguously, it was not included in the analysis. *Plan and profile curvatures* were derived with the same algorithm as slope, but from a DEM5 smoothed with a moving window mean filter with a radius of 10 m. This was deemed necessary because of the extremely noisy character of fine-scale curvature. Medium-scale curvature based on a DEM that retains details on the typical spatial scale of channels within the rock faces and talus cones (that are both prone to and indicative of debris flow activity) is expected to be a better proxy variable for convergent flow of water (plan curvature) and changes in flow velocity (profile curvature).

Relief parameters related to the local catchment area are derived from the DEM5 as proxies for the *availability of water* for debris flow initiation. We calculated the *specific catchment area* (SCA) as the local flow accumulation per unit contour length using a multiple-flow-direction algorithm (Freeman, 1991). Heavy rainfall on steep bedrock slopes is expected to be converted almost entirely to Hortonian

[2] as the roughness is derived from the DEM1, the cluster analysis can make use of sub-grid-scale roughness for the classification of DEM5 raster cells

overland flow; on talus slopes bordering steep rock faces, this runoff can cause the initiation of debris flows, especially where it enters the talus in a channelised manner ("firehose effect"; see e.g. Johnson and Rodine, 1984; Coe et al., 2008). However, if the sediment is coarse grained, large amounts of water are expected to infiltrate; this leads to a decrease of hydrological connectivity, and at least to an attenuation of the increase of runoff with increasing catchment size. Therefore, we re-calculated the catchment area, accumulating only bedrock cells in the roughness class map instead of every DEM5 raster cell. The modified SCA map hence refers to the size of the bedrock catchment draining into each raster cell.

3.2 The susceptibility model

Multivariate logistic regression (Hosmer and Lemeshow, 2000) forms part of the family of generalised linear models (GLMs); in contrast to ordinary linear models, a function of the expected value of a response variable is modelled by a linear combination of continuous or discrete predictor variables. In logistic regression, the response variable is binary (Bernoulli distribution); here, it takes the values 0 (no debris flow initiation) and 1 (debris flow initiation). The response function is the logit transform of the probability $p \in]0, 1[$ that the response variable takes the value 1:

$$f(p) = \text{logit}(p) = \ln \frac{p}{(1-p)}. \quad (1)$$

Since the logit is within the interval $]-\infty, \infty[$, it can be modelled as a linear combination of predictor variables $X_1 \ldots X_n$:

$$f(p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n, \quad (2)$$

where β_0 is the intercept and $\beta_1 \ldots \beta_n$ are the model parameters. These are estimated using a maximum likelihood approach.

The spatial data are generated and managed in SAGA GIS, including the derivation of relief parameters (Sect. 3.1.2); for the statistical analysis, they can be directly read from the SAGA native data format using the RSAGA package (Brenning, 2013) for the statistical software R (R Development Core Team, 2012). Logistic regression is then performed using the glm and stepAIC functions of the MASS package (Venables and Ripley, 2002). For reasons explained in the Introduction, we estimate the model parameters for a sample (the size of which we will try to optimise in this study) of event (debris flow initiation) and non-event cells; sampling is also performed in R. The resulting susceptibility maps are written back to SAGA data format for visualisation and further spatial analysis. They contain the probability that the dependent variable takes the value 1, i.e. that debris flow initiation will take or has taken place.

3.2.1 Multicollinearity analysis

Besides sample independence, an important prerequisite for the application of GLM is the absence of *multicollinearity*, i.e. that the predictor variables are not correlated with each other. In order to test for multicollinearity, we applied the vif function of the car package (Weisberg and Fox, 2010) to a full model (i.e. including all geofactors described in Sect. 3.1), yielding the variance inflation factors (VIF) of each geofactor. Although no binding rules exist for their interpretation, several authors who conduct a multicollinearity analysis apply a very strict threshold of 2, above which variables are considered multicollinear and are excluded from the model (e.g. Van Den Eeckhaut et al., 2006, 2010; Guns and Vanacker, 2012). However, the most common rule of thumb is reported to be the "rule of 10" (using VIF = 10 as a threshold for severe multicollinearity), and the use of strict thresholds of VIF appears to be questionable (O'brien, 2007). The analysis of VIFs yields values of 1.18 and 1.47 for the two curvature variables, and 1.77 for SCA. Roughness and slope have VIFs of 2.06 and 2.76, respectively, which is only slightly above the threshold used in other studies, so we decided to keep all candidate variables.

3.2.2 Stepwise selection of predictor variables

An important task in susceptibility modelling is model building, i.e. the *selection of the independent variables* (geofactors). In Sect. 3.1, several candidate variables are described that conceptually explain the spatial distribution of debris flow initiation. Model building is achieved in this study through an automatic stepwise variable selection (function stepAIC; Venables and Ripley, 2002). Starting from a full model, i.e. a model including all variables, variables are removed (or re-included) in order to minimise the Akaike information criterion (AIC; Akaike, 1973) which is calculated from the likelihood function of the model and the number of predictor variables. The AIC penalises for the number of predictor variables; i.e. it increases with the number of variables, and it decreases with a larger likelihood function indicating a better model. Hence, although there is no theoretical justification of the AIC (Sachs and Hedderich, 2006), this procedure is suitable in practice for selecting a parsimonious model, i.e. a best-fit model using as few variables as possible (Brenning, 2005). The results of stepwise logistic regression have often been used to rank the controlling factors by importance (e.g. Van Den Eeckhaut et al., 2006). While we assume that the methodological framework of our study would also be suitable for the assessment of sample size effects in such investigations (Guns and Vanacker, 2012, e.g., suggest a "robust detection of controlling factors" based on repeated sampling and stepwise model selection), the latter are not the aim of our present study.

Stepwise procedures can be applied as a backward selection, as in this study (and e.g. in Brenning, 2005; Ruette

et al., 2011), but also as a forward selection (Beguería, 2006; Meusburger and Alewell, 2009; Atkinson and Massari, 2011). Menard (2002) explains that backward selection is in some cases superior to the forward procedure. Note that the stepwise procedure used here and in Brenning (2005) differs from other studies where the decision of keeping or dropping predictor variables is based on the significance of model improvement (e.g. Beguería, 2006; Meusburger and Alewell, 2009; Guns and Vanacker, 2012), not on an information criterion. Recently, alternative approaches for model selection have been proposed (e.g. Calcagno and Mazancourt, 2010); they will be tested in future research.

3.2.3 Model validation

It has been stressed that a modelling study without proper validation is useless (Chung and Fabbri, 2003). Many studies in susceptibility modelling use spatial or temporal cross-validation (space or time partition; cf. Chung and Fabbri, 2003) within the same study area; i.e. the data are split either systematically or randomly into training and test data sets according to their location or time of occurrence (Chung and Fabbri, 2003; Beguería, 2006). Here, we estimate model parameters based on samples drawn from the Zwieselbachtal catchment, and apply the resulting models to the neighbouring Larstigtal catchment. Hence, training and test areas are completely independent. For each model run, the predictive ability is evaluated using receiver operating curves (ROCs) or prediction-rate curves sensu Chung and Fabbri (2003), plotting true-positive against false-positive rates. The advantage of ROCs is that they yield a threshold-independent measure of predictive ability; in our case, we do not have to define a threshold of modelled landslide probability below which we do not recognise susceptibility. Additionally, as a single measure of predictive ability, the AUC is calculated (Hosmer and Lemeshow, 2000; Beguería, 2006); this parameter falls in the range $[0.5, 1]$, where 0.5 is equivalent to random prediction and 1 to a perfect prediction.

3.3 Exploring the effect of sample size

In the Introduction, we have argued why the sample size should be neither too small nor too large. Here, we describe (i) how the effect of sample size on the diversity of models is explored, and (ii) how we constrain the upper limit of sample size.

3.3.1 Sample size and model diversity

For small sample sizes, the geofactor composition of the resulting model depends extremely on the random sample, because small samples cannot sufficiently cover the diversity of geofactors within the study area. We hypothesise that with increasing sample size the diversity of relevant models (selected by the stepwise procedure) first decreases towards a plateau that can be explained with the overall variability of geofactors in the study area; when the sample size approaches the size of the study area, the variability of models will eventually decrease to zero. Such a behaviour is similar to the dependence on sample size of the predictive power of predictive geomorphological models that has been explored by Hjort and Marmion (2008).

We analyse model diversity by repeating the stepwise model selection with 1000 independent samples of a given sample size. Such a high number of replications is novel compared to existing studies that employ multiple samples; we chose the number of 1000 because we noticed in first experiments that the model diversity assessment was too unstable with a lower number of replications (e.g. between 25 and 50 in the studies of Brenning, 2005; Beguería, 2006; Guns and Vanacker, 2012). Sample size varies between $n = 50$ and $n = 5000$ non-event raster cells; together with the sample of $n = 81$ initiation areas in the ZBT area, the samples cover between 0.02 and 0.68 % of the study area (ZBT). Specifically, a stratified sampling scheme has been adopted; one single raster cell is randomly selected from each debris flow initiation zone, and the sample size of non-event cells (from the area outside of the mapped initiation zones) is varied. The choice of non-event sample sizes in relation to event sample size ranges from ca. 1 : 1.6 to ca. 60 : 1, thus including the recommendations of King and Zeng (2001) and the alternatives chosen in landslide susceptibility studies, e.g. 5 : 1 (Van Den Eeckhaut et al., 2006) or 10 : 1 (Beguería, 2006; Guns and Vanacker, 2012).

For each of the 1000 samples, the geofactors that remain in the "best" model (with respect to the AIC) after stepwise selection are saved in a table. Each geofactor is evaluated by the percentage of models which it was part of (cf. Guns and Vanacker, 2012). The set of selected geofactors for one sample defines a "model species" (if, for example, the geofactors A, B and D are selected from the candidate geofactors A, B, ... E, the species of the resulting model is ABD). The term model species was used in order to highlight the similarity of the proposed method for model diversity assessment with investigations of biodiversity in ecology. Theoretically, $k_{max} = 2^g - 1$ different model species can exist if g candidate geofactors are available for model selection, and if the resulting model has to contain at least one geofactor. The diversity of the 1000 replicate models calculated for each sample size is evaluated using three measures: (i) the number k of different model species ("species richness"); (ii) the Shannon diversity index H, also known as Shannon information entropy; and (iii) the Simpson index D.

The Shannon index was developed in information theory (Shannon, 1948) and has been widely applied in ecology as an index measure of biodiversity (e.g. Magurran, 2004). In geomorphology, it has been used to assess the uncertainty of drainage routing and watershed delineation (Schwanghart and Heckmann, 2012). In our study, it is calculated as

$$H = -\sum_{i=1}^{k} p_i \cdot \ln(p_i), \quad (3)$$

where $i = 1 \ldots k$ represents the ith of k different model species, and p_i is the probability of occurrence of the ith species, estimated by n_i/N, the proportion of the ith model species found in N individual stepwise modelling runs.

The log-transformed Simpson index (Simpson, 1949) has been developed for measuring biodiversity; it is considered superior to the H as it is independent of sample size (Magurran, 2004). It is calculated as

$$D = -\ln \sum_{i=1}^{k} \frac{n_i \cdot (n_i - 1)}{N \cdot (N - 1)}, \quad (4)$$

where n_i is the absolute frequency of the ith model species and N is the number of individual models (here: 1000).

H and D combine the number of different model species (species richness) and their relative frequency (relative "abundance") in one single number: a large diversity associated with a high species richness (k different terms have to be summed up for H and D, respectively) and/or an even distribution of model species across the 1000 samples. Conversely, diversity is low when there is only a small number of different species, and/or one or few species strongly dominate. Shannon's entropy has been interpreted in terms of the "average surprise a probability distribution will evoke" (see e.g. Thomas, 1981, p. 7). The result of a stepwise selection with a sample size for which low diversity (low H) has been measured is not expected to be surprising, because one or few species have a very high probability of occurrence. We hypothesise that the diversity of model species, and the degree of surprise with which we see one particular outcome of the selection given the results of 1000 models, will reflect the sample dependence of the stepwise selection. Therefore, we propose the "model diversity" as a measure of model quality in terms of reproducibility; similarly, Petschko et al. (2014) recently proposed a "thematic consistency" index that assesses variable-selection frequencies in model replications and is based on the Gini impurity index.

3.3.2 Sample size and spatial autocorrelation

In our study, the spatial autocorrelation of a data set is explored with the empirical semivariogram, which is typically used for geostatistical interpolation techniques such as Kriging (Webster and Oliver, 2007). It is derived from point measurements by evaluating the semivariance of values of a variable (geofactor) for pairs of points separated by a specific distance. One important property of the semivariogram is the range; points separated by a distance below this range are autocorrelated. Brenning (2005) uses the range of the empirical correlogram of the residuals of a logistic regression model (180 m in his study) to constrain the minimum distance between training and test data points in spatial cross-validation. Similarly, we estimate the range parameter of the variogram of each geofactor to constrain the sample size: we argue that the average distance between raster cells in the (non-event) sample should not fall within the autocorrelation range(s) of the geofactors included in the model in order to keep the non-event sample as uncorrelated or independent as possible. As the average distance implies that some points in the sample will be closer neighbours, we concede that this strategy minimises spatial autocorrelation rather than preventing it.

Assuming a set of randomly distributed points (here: raster cells), the average distance $\hat{d}$ to the nearest neighbour can be estimated by Eq. (5):

$$\hat{d} = \frac{1}{2 \cdot \sqrt{\rho}}, \quad (5)$$

(Clark and Evans, 1954) where ρ is the density of the sample, i.e. the sample size n divided by the study area (here: the number of raster cells within the study area multiplied by 25 m^2, the area of each cell). For each study area, $\hat{d}$ is calculated as a function of n and used to estimate the upper boundary for the "suitable sample size". Instead of using the highest autocorrelation range (i.e. that of the geofactor with the most far-reaching spatial autocorrelation) as a crisp, absolute upper limit of sample size, we take into account $\hat{d}(n)$ as it progressively falls below the autocorrelation range of more and more geofactors, and we regard the corresponding n as progressively less acceptable. An upper limit is finally reached when the smallest autocorrelation range from the set of geofactors is undercut.

Figure 2 shows the empirical geofactor semivariograms and the practical range parameter (i.e. the range where 95 % of the sill is reached) of the fitted variogram models. Depending on the geofactor, spherical and exponential models were used. It is obvious that some geofactors, e.g. slope, are autocorrelated on multiple scales. In these cases, the lower range is used; however, it appears that a sample which is independent with respect to all geofactors is not possible.

3.4 Variability of model results

The investigations described in the previous sections have the aim of quantifying and reducing the dependence of the results on the sample while maintaining sample independence. Once a suitable sample size is estimated, we investigate the variability of model results – both quantitatively and with respect to its spatial distribution. In order to do so, we repeat 100 times the sampling, model selection, fitting and application for the Zwieselbachtal area, creating a stack of 100 gridded susceptibility maps of the whole study area. The median of 100 probabilities in each raster cell is taken as a consensus model (Marmion et al., 2009) and the final susceptibility map. The interquantile range IQR90 = $p_{0.95} - p_{0.05}$, which encompasses 90 % of the modelled susceptibility values as a non-parametric measure of dispersion, quantifies the uncertainty caused by sampling and stepwise model selection.

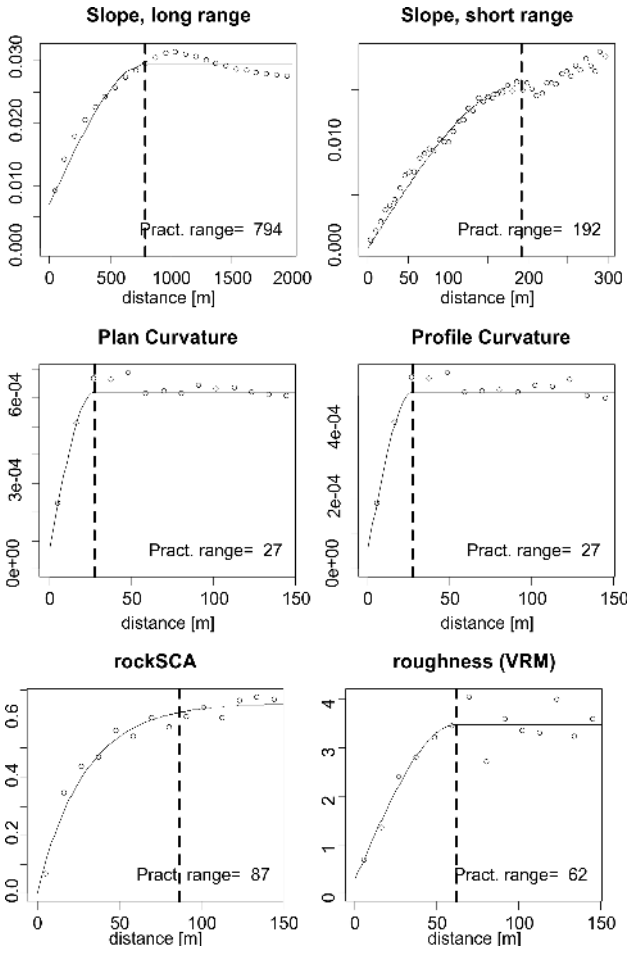

Fig. 2. Empirical variograms of geofactors used in this study. Note that slope is autocorrelated at different spatial scales.

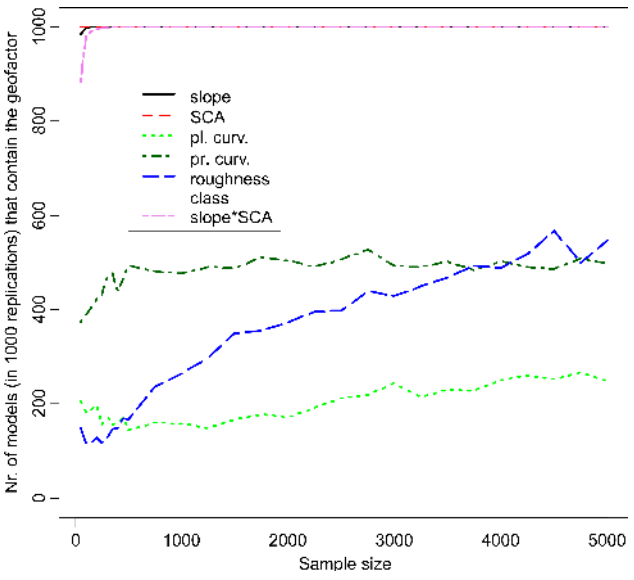

Fig. 3. Overview of the six geofactors and their contribution to 1000 models of different sample sizes. The *y* axis denotes the number of models for which the respective geofactor was selected. "slope*SCA" signifies the interaction term of the two variables slope and specific catchment area.

As this measure is calculated for each raster cell, the respective map can be used to visualise the spatial distribution of model uncertainty (not with respect to the true probability, but with respect to model variability). In addition, the distribution of the parameter coefficients of the 100 models, and their predictive power (ROCs and AUC; see Sect. 3.2.3) can be displayed and analysed.

4 Results and discussion

4.1 Investigation of sample size effects

Before we approach the question of an optimal range of sample sizes, we take a look at the results of model selection as a function of sample size. Specifically, Fig. 3 shows, for each geofactor, the number of models that retained this geofactor after the AIC-based selection procedure. The six geofactors that were eligible for model selection were slope, SCA, the interaction of the previous two factors (denoted "slope*SCA" in Fig. 3), the roughness category which distinguishes bedrock from debris-mantled slopes, and the two curvature variables. While roughness and profile curvature gradually increase their membership with larger sample sizes (roughness starting from only ca. 15 % of the replications), the interaction term slope*SCA quickly attains 100 % (i.e. all of the 1000 samples lead to models containing this variable) even with small samples. Here, it is important to mention that interaction terms may only be part of a model if their marginals (here: slope and SCA) are also contained. This is the case, as the given variables are contained in all models, irrespective of sample size. The proportion of models containing the geofactor plan curvature is very low, starting with about 20 % and only slightly increasing in larger samples.

If the "success" of a geofactor in the model selection procedure is a measure of its importance, then the most important variables are slope, SCA, the interaction of slope and SCA, and profile curvature. The importance of roughness and plan curvature is low, but the number of models containing roughness surpasses that of models containing plan curvature even at sample sizes below 1000. These findings are consistent with previous work on (slope-type) debris flow susceptibility: Heckmann and Becht (2009) and Wichmann et al. (2009), for example, use slope, land cover, and a variable called the CIT index (Montgomery and Foufoula-Georgiou, 1993). The latter is calculated as the specific catchment area times the squared tangent of slope. The interaction term slope*SCA used in our study can be interpreted physically (mathematically, it is the product of the two geofactors) as

the compound topographic index indicating stream power (Moore et al., 1991); in this index, catchment area and slope serve as proxies for the abundance and energy of surface runoff. In comparing several models (discriminant analysis and logistic regression) Carrara et al. (2008) observed that factors relating to slope gradient, land cover, availability of detrital material, and active erosional processes best described debris flow initiation. The most frequent model species in our study include geofactors that represent these categories.

Figure 4 evaluates the diversity of models selected by the AIC-based procedure as a function of sample size. The diversity is expressed as the number of model species (i.e. models defined by a given combination of geofactors) in 1000 samples (centre panel), and is quantified using the Shannon and Simpson diversity measures (bottom panel). The number of model species declines exponentially to reach a stable minimum of 8 species at a sample size of $n = 1000$. Even for the largest sample size in our analysis ($n = 5000$), differences between the 1000 samples result in as many as 8 different model species. The diversity measures show a local minimum at $n = 300$ and $n = 350$, respectively; for these sample sizes ($n_{rel} = 0.05\,\%$), the number of model species is higher, but the distribution of the 1000 models across this number of species is more uneven – i.e. few species make up the lion's share of the selections – and the rest is represented only by a few cases. For larger sample sizes, model diversity slightly increases again and reaches a more or less stable value. Sample sizes much larger than 5000 ($n_{rel} > 0.68\,\%$, not shown) lead to a decrease of the diversity indices; when the sample size approaches the size of the population (i.e. the complete study area), the stepwise procedure of course yields only one model species, and the diversity indices attain their absolute minimum (0). The plateau of the diversity measures is also reflected in the model composition shown in Fig. 3 where all geofactors (except roughness) exhibit only slight changes with sample sizes larger than ca. 1000 ($n_{rel} = 0.15\,\%$).

We interpret the minimum of the diversity indices as a minimum of the dependence of model selection on the sample and therefore the corresponding sample size (300–350) as a data-based recommendation for our case study. Such a strategy is, in our opinion, better than the adoption of arbitrary recommendations with respect to event : non-event ratios, absolute, or relative sample sizes. The sample size of 300–350 non-event cells corresponds to a ratio of event : non-event of 1 : 3.7 to 1 : 4.3, which is approximately consistent with the 1 : 5 ratio used by Van Den Eeckhaut et al. (2006) and with the recommendation (1 : 2–1 : 5) given by King and Zeng (2001). It is also in the range of the ratio of event to non-event cells in our study areas (about 1 : 500 in ZBT, 1 : 200 in LT), a ratio that has been used by Atkinson et al. (1998). Considering Green's rule of thumb (Green, 1991) reported in the Introduction (Sect. 1.1), the six candidate geofactors in our case study would require a minimum sample size of ca. 100. Hjort and Marmion (2008), who investigate

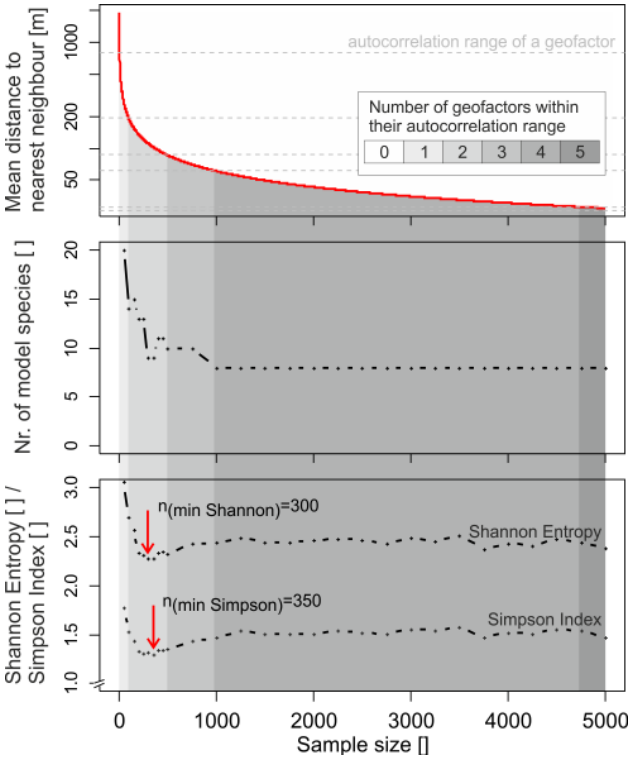

Fig. 4. Mean distance between neighbouring sample points (top panel), number of model species in 1000 samples (center panel), and two model diversity measures (bottom panel) as a function of sample size. Shades of grey denote the degree to which the raster cells in a sample of size n lie, on average, within the autocorrelation range of geofactors. Red arrows indicate the sample sizes for which the Shannon and Simpson indices reach a local minimum, respectively.

the predictive power of different models estimated with different sample sizes, state that a "level of robust predictions" is attained, with all statistical techniques, at a sample size of $n = 200$.

The local minima do not appear to be always present, depending on the choice of geofactors and the study area used for model selection (not shown). However, there is always at least a conspicuous knickpoint in the empirical diversity diagram where an increase in sample size does not lead to a significant reduction of model diversity. The analysis of the LT data, for example, shows a plateau, not a local minimum, of model diversity, and this is only reached between $n = 1000$ and $n = 2000$ ($n_{rel} = 0.38$ and $0.74\,\%$), a sample size which is already becoming problematic with respect to spatial autocorrelation (see next paragraph). The LT is smaller than the ZBT and has a smaller number of debris flows but a higher debris flow density (events per square kilometre); hence there does not appear any conspicuous relationship of the existence and location of plateaus or local minima, absolute or relative sample size, and the aforementioned study area properties. The investigation of these problems is left

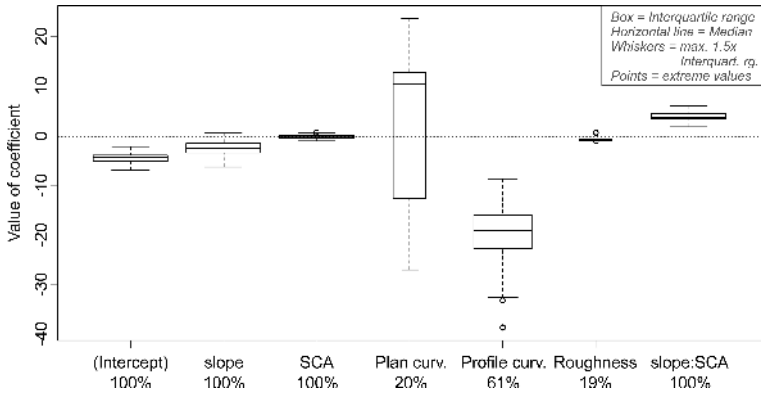

Fig. 5. Distributions of model coefficients estimated from 100 random samples ($n = 350$ non-event cells) in the ZBT area. The percentages below the parameter name refer to the proportion of the 100 models that contain the respective geofactor.

open to future research, employing a systematic analysis of multiple study areas with different sizes, characteristics, and debris flow densities.

In Sect. 3.3.2, we proposed the mean distance between sampled locations in relation to ranges of spatial autocorrelation as an upper constraint of sample size. Figure 4 (top panel) shows the expected mean distance between nearest neighbours as a function of sample size (see Sect.3.3.2). Additionally, the horizontal dashed lines indicate the autocorrelation ranges of the geofactors mentioned above (cf. Fig. 2). As the red curve intersects the autocorrelation ranges of more and more geofactors, the sample of the corresponding size is more and more likely to violate the independence assumption. The decreasing suitability of larger samples to this end is visualised across the whole Fig. 4 through darker shades of grey. The optimal sample sizes indicated by the red arrows in the bottom part of the diagram belong to a range of sample sizes that are within the autocorrelation range of one single geofactor only. In this case, it is the "large-scale" range of slope (ca. 800 m, slope is autocorrelated also at smaller spatial scales with a range of ca. 200 m; see Fig. 2). We consider this only a minor violation of the independence assumption, so that the sample size recommended above remains optimal also with respect to the spatial autocorrelation issue that has been raised in Sect. 1.2.

While the typical scale of application of landslide susceptibility models is in the order of (many) tens to thousands of square kilometres, our study took place in a comparatively small study area. Considering the small size and the associated homogeneity of our study area with respect to the statistical and spatial distribution of geofactors, we add a note of caution to the interpretation of our findings. First, we expect the necessary sample size to be larger in more heterogeneous areas, and we expect a larger variability of model selection and model coefficients. One possibility of dealing with large, heterogeneous study areas has recently been proposed by Petschko et al. (2014), who partition their study area in sub-areas based on lithological properties that are related to landslide activity. Second, the assessment of spatial autocorrelation from variograms of the geofactors is much less straightforward in larger, heterogeneous areas. For example, different ranges of autocorrelation could exist for the same geofactor in different (sub-)regions of the study area, which calls into question the existence of a single sample size (and the associated average distance between sample points) below which the autocorrelation issue is mitigated. However, we are confident that our observation of a local minimum or plateau in model diversity will apply also at larger spatial scales (see, for example, Hjort and Marmion, 2008; Guns and Vanacker, 2012). Moreover, we uphold the general recommendation to investigate, through repeated sampling with different sample sizes, the behaviour of parameter selection in order to explore a suitable (small) sample size that both minimises sample dependence and facilitates a robust parameter selection.

4.2 Model results

4.2.1 Model parameters

In this section, the results of the procedure described in Sect. 3.4 are evaluated. Figure 5 shows the distribution of the estimated coefficients for each of the geofactors. Additionally, the percentage below the parameter name gives the proportion of models that contained the respective geofactor after stepwise selection. The coefficients were estimated using 100 independent random samples of $n = 81 + 350$ (event + non-event sample) in the ZBT area. The geofactors slope, SCA, and their interaction are part of every model, followed in decreasing order by profile curvature, plan curvature, and roughness class. The spread of the coefficients is low for most of the geofactors, with the exception of the two curvature parameters. The coefficient for plan curvature has the largest range, and it takes positive and negative values, which makes the interpretation very difficult; this is probably caused by the fact that the random sampling of event cells

from the upper erosional zones in the debris flow inventory will select locations not only in the centre of channelised debris flow paths (with highly concave plan curvature) but also at the boundary of these areas, which are highly (plan) convex. Conversely, the profile curvature coefficient is strictly negative, which means that a concavity in the long profile increases the probability of debris flow initiation. The explanation for this finding is a morphological one: the typical locations of debris flow initiation (facilitated by the firehose effect; see Fig. 1) at the contact of steep rock faces and the corresponding talus cones are marked by large negative (i.e. concave) profile curvatures.

The mostly negative coefficients for slope and SCA are difficult to interpret, as one would expect that the probability of debris flow initiation would increase with steeper slopes and with larger catchment areas. However, this problem appears to be only a mathematical one, as the interaction term of slope and SCA is present in the model. Therefore, the coefficient of slope (alone) models the effect of slope where SCA is zero (and vice versa); the coefficient for the interaction term is positive, indicating higher probabilities with steep slopes and large catchment areas, which is conceptually correct. The interaction term plays an important role in the model: without it, the positive relationship of SCA with debris flow release causes the modelled susceptibility to increase even in the comparatively flat valley bottoms. Under these conditions, slope-type debris flows cannot occur; Rickenmann and Zimmermann (1993) report starting zone slopes for type 2 debris flows (that type which occurs in our study areas) between 26.5 and 38°, with catchment sizes of up to 1 km^2; Takahashi (1981) gives a lower threshold for debris flow initiation of 15°. Generally, there appears to be a trend that the minimum slope angle required for debris flow release decreases with larger catchment areas (Rickenmann and Zimmermann, 1993; Heinimann et al., 1998; Horton et al., 2008), so there is, besides the stream power index (cf. Sect. 4.1), one more theoretical justification for including the interaction of slope and SCA.

4.2.2 Susceptibility maps

The previous analyses have shown the dependence of models found through AIC-based model selection on the respective sample and its size. The spatial pattern of a model result (here: the susceptibility map containing the debris flow initiation probability) depends on the spatial pattern of the geofactors that form part of the model. Figure 6 shows a section of the susceptibility map that can be seen as a consensus model (see Marmion et al., 2009) as every raster cell contains the median of 100 model predictions, the coefficients of which have been summarised in the previous section (Fig. 5). Susceptibility in both valleys has been predicted using the model estimated with ZBT data only. The whole map is part of the supplementary material of this paper. On the map, debris cones are highlighted by yellowish to reddish colours

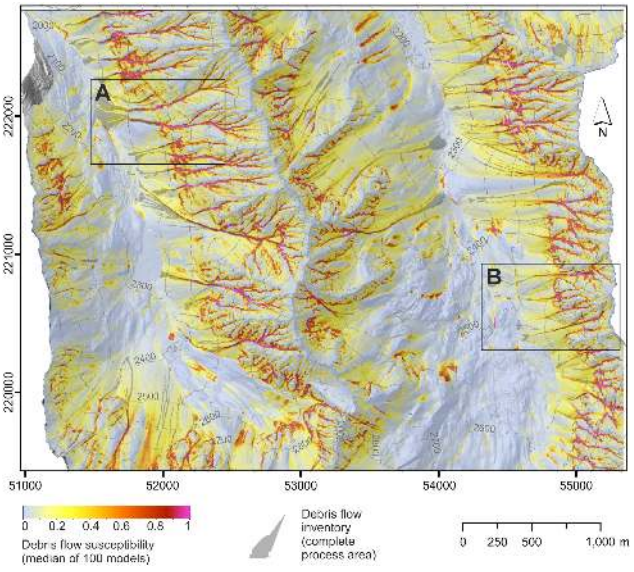

Fig. 6. Part of the susceptibility map (for full extent, see Supplement) of the ZBT and LT areas. The susceptibility values represent a model ensemble, specifically the median value of 100 models estimated from 100 random samples ($n = 350$ non-event cells) in the ZBT area. Insets A and B refer to map sections in Fig. 7.

indicating medium to high probability of debris flow release. The distal parts of the cones are characterised by lower (if any) susceptibility, while their apices and channel-like portions of the upslope area show the highest values. Most of the valley floor and most steep parts of the rockwalls have very low to zero susceptibility. This can be seen in detail in the upper row of Fig. 7; virtually all mapped debris flows (including not only the depositional lobes, but the whole process area) have high to very high susceptibility values in their upper part, and it can be stated that the spatial pattern of debris flow occurrence appears to be reproduced well by the model.

This visual validation also reveals problems. The zones of highest susceptibility, indicated by violet colours, extend very far upslope along very steep channel-like features within the rockwalls. Many of these locations appear to be too steep for debris to accumulate (one of the preconditions for debris flow generation); for this problem, we offer two explanations: first, an analysis of slope values within the mapped starting zones (see Sect. 3.1.1) reveals that ca. 75 % of slope values within the initiation areas are within a physically meaningful range (below ca. 40°), while the remaining values clearly speak against the accumulation of debris in these locations. This can be attributed in part to mapping errors (Ardizzone et al., 2002) where the upper portion of a debris flow area is spuriously extended into very steep bedrock channels that are in part poorly identifiable on aerial imagery. Another source of this error, probably to a lesser degree, is a mismatch in the exact location of the rockwall–talus contact between the DEM (which is decisive for the model) and

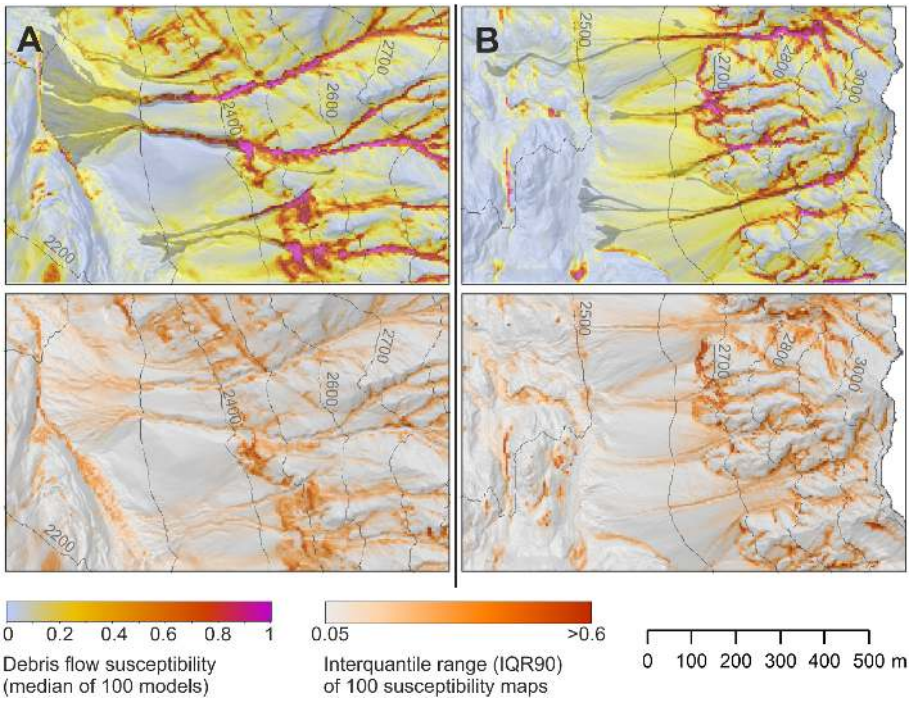

Fig. 7. Map sections (for full extent, see Supplement) from the ZBT **(B)** and LT **(A)** areas. The maps show the susceptibility map (see Fig. 6) and a map of the IQR90 calculated from the model ensemble. The latter map represents the uncertainty of the susceptibility map that is due to the sampling process.

the aerial photo. Second, a linear modelling approach is not capable of modelling complex non-linear relationships such as the one of slope and debris flow release: conceptually, susceptibility should increase, starting from some minimum slope, up to a maximum and then decrease again. The susceptibility then reaches zero at slope gradients that are prohibitive for the formation and persistence of sediment storage that is needed for debris flow generation. The GLM approach, however, only handles monotonic relationships between independent and dependent variables, e.g. an increase of susceptibility with slope. Problems of this kind could be solved by using other approaches, for example the weights of evidence, certainty factor, or generalised additive models (GAM; see e.g. Hjort and Luoto, 2011).

A novel output of our model replication exercise is the quantification of the variation in model results and the assessment of its spatial distribution. The model uncertainty addressed here is due to the sampling and model selection procedure only. For each raster cell of the susceptibility map, we computed not only the median but also the interquantile range (IQR90) between the $p_{0.95}$ and $p_{0.05}$ quantiles; the corresponding map can be seen in the supplementary material and in Fig. 7, bottom row. In the whole study area, the IQR90 has a highly positively skewed distribution that ranges from 0.0 to 0.98. It has a mean of 0.081; i.e. debris flow release probability predicted by the 100 models varies by 8 percentage points, on average. In the ZBT area (that was used to estimate the models) this value equals 0.073, while in the LT area it is slightly higher (0.103). For samples taken according to the "1 : 1 event to non-event" rule ($n = 81$ non-event cells, $n_{\mathrm{rel}} = 0.022\,\%$), the average IQR90 is 0.190 (ZBT), 0.230 (LT) and 0.200 (total study area). The expected variability is consistently higher for smaller samples, and when a model is applied to a different area. The latter can be explained with the effect of extrapolation beyond the range of geofactors in the respective training area.

Generally, the lowest uncertainty is found for both the lowest and the highest susceptibility values, an observation also reported by Guzzetti et al. (2006b). On the uncertainty maps, the largest standard deviations occupy spatially coherent areas along the zones of high susceptibility, and additionally in considerable portions of the valley bottom where the slope gradient is low. In some places, the spatial pattern of uncertainty is consistent with the fact that profile curvature is included in only about 60 % of the models; here, zones of high curvature (both concave and convex) are characterised by high IQR90 values. Such zones of high uncertainty may generally occur where a high (or low) predicted susceptibility relies on one parameter only that is not part of all models. In our opinion, the map adds information to the susceptibility map that can be useful for its interpretation.

4.2.3 Validation

The variability of model parameters and predictions is also reflected in the validation. A first qualitative validation is done by visually inspecting the susceptibility map (here: the median of 100 models, Figs. 6 and 7). Each model is quantitatively validated by means of a ROC (see Sect. 3.2.3) using data from the Larstigtal (LT) only; hence, the data used to estimate the model parameters (from the ZBT area) and the validation data are completely independent, and the corresponding diagram represents a "prediction curve" (Chung and Fabbri, 2003). Split-sample validation approaches such as cross-validation, spatial and temporal partitions (Chung and Fabbri, 2003) do not warrant such independence when, for example, subsets of the same inventory are used to estimate model parameters and to validate the resulting model in one study area.

Figure 8 (top panels) shows the prediction curves for the 100 models, and the distribution of the corresponding area under the curve (AUC). The 100 curves are located quite close to each other, and there are no conspicuous extreme outliers. The AUC reaches 0.83, on average; the predictive ability of a model calculated in the LT area and applied to the ZBT (not shown) is even higher, with AUC = 0.9. In total, the observed AUCs are within the range of many published studies (e.g. 0.69–0.8: Ruette et al., 2011; 0.84: Ayalew and Yamagishi, 2005; 0.89–0.93: Van Den Eeckhaut et al., 2010) and can be regarded as satisfying. The different performance of the ZBT model in the LT area and vice versa is an interesting fact. This could be caused by different characteristics of the study areas, related to a different range, and different spatial and statistical distributions of the geofactor values. The two neighbouring areas, however, are regarded as very similar and homogeneous. Heckmann and Becht (2009) investigated the transferability of a debris flow susceptibility model among different study areas and reported that the predictive power of models is largely independent of the degree of similarity of training and test area; their model approach (certainty factor), however, strongly differs from logistic regression. Besides computational and conceptual differences, continuous geofactors such as slope are classified using the same scheme in all study areas. Conversely, in our study, a different range of geofactors in training and test areas could lead to different coefficients and different model performance due to extrapolation. Another reason for the different performance could be the different debris flow density. In order to determine the controls of model performance, future research will have to use a larger number of different study areas with different debris flow densities. The methodological framework for the assessment of model variability and performance proposed here is considered useful for such investigations.

Interestingly, the sample size did not influence the predictive ability of the model ensemble – both $n = 81$ and $n = 350$ have very similar mean AUC values. However, the smaller sample size leads to a much larger spread of the different

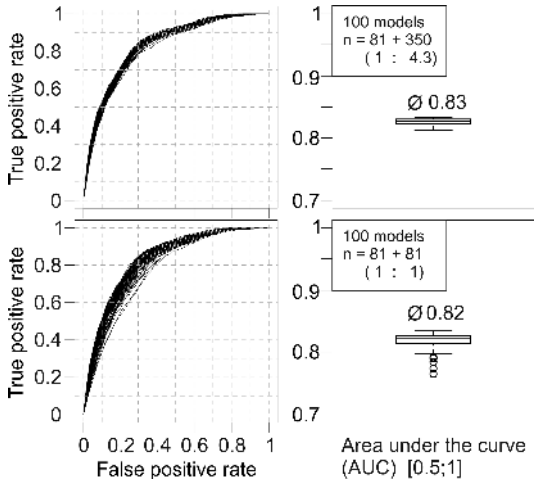

Fig. 8. Evaluation of the predictive ability of 100 models (top panels: $n = 350$ non-event cells, bottom panels: $n = 81$ non-event cells) by means of the area under the curve. As the model training (ZBT) and validation area (LT) are independent, the diagrams on the left represent prediction curves (Chung and Fabbri, 2003).

prediction curves and consequently also of the AUC values. In our case, a single sample of events and non-events at a ratio of 1 : 1 (see, for example, Brenning, 2005; Meusburger and Alewell, 2009) could have resulted in a good model (AUC 0.84) but also in a comparatively poor one (AUC 0.75), although the expected AUC is approximately the same. We deduce from our results a recommendation to create susceptibility maps from model ensembles, because they are supposed to yield a more reliable result on the one hand and give an estimation of (sample-induced) uncertainty on the other. Similarly, Marmion et al. (2009) propose "consensus models"; in their study, results from different predictive modelling approaches are combined using several methods, among them the median that was used in our study to combine the results of 100 models generated with the same method, but from independent random samples.

5 Conclusions

In this paper, we investigated the effect of sample size on a logistic regression model with a parameter selection procedure that is based on an information criterion (AIC). The case study aims at predicting the spatial distribution of slope-type debris flow release zones in the Larstigtal (LT) and Zwieselbachtal (ZBT) catchments in the Austrian Central Alps.

The procedure of random sampling and model selection was replicated 1000 times for different samples between $n = 50$ and $n = 5000$ non-event raster cells. For each candidate geofactor, the number of models it was part of after stepwise model selection was recorded. The diversity of models as a function of sample size was determined using the number of different models and two diversity indices (Shannon

Entropy and Simpson diversity index). In our case study, model diversity decreased with increasing sample size and reached a local minimum at $n = 300$–350, before it slightly increased again to a stable level. In some cases, no local minima were detected, but model diversity always reached a plateau on which even much larger samples could not improve (= decrease) model diversity. While we were unable to discern a dependence of local minima or plateaus on properties of the debris flow inventories and/or study areas, we recommend exploring the behaviour of model selection and diversity dependent on sample size in order to determine an optimised sample size. The latter is constrained by the range of spatial autocorrelation found in variogram analyses for each geofactor.

Most importantly, our results show that, even with large sample sizes (that will progressively violate the independence assumption), there will still be a variety of different models and, hence, also diverse model results depending on the sample. We argue that single-sample studies run the risk of accidentally yielding a poor model, and therefore strongly advocate the calculation of multiple models based on independent random samples; the results of these models are used (i) to construct a consensus susceptibility map (in our case study, we used the median of 100 models on each raster cell) and (ii) to investigate, both statistically and spatially, the variation in model results caused by the sampling and model selection procedure. In our study, the median of 100 models was used as the consensus model, and variation was quantified using the IQR90 interquantile range as a non-parametric dispersion measure. The latter was clearly influenced by sample size (less variation for larger samples) and study area (more variation in LT if the ZBT model was applied). Predictive power of the models was measured using receiver operating curves (area under the curve); all models yielded satisfying results that are in the range of other published landslide susceptibility models. Sample size did apparently not influence the average predictive power of the model ensemble, but smaller samples increased the range of AUC and hence also the proportion of comparatively poor models.

Supplementary material related to this article is available online at http://www.nat-hazards-earth-syst-sci.net/14/259/2014/nhess-14-259-2014-supplement.pdf.

Acknowledgements. Without the cooperation of several colleagues this work could not have been published. The authors would like to thank Dieter Rieger, who mapped debris flows in the LT and the ZBT, for making his data available; Markus Thiel for mapping recent events; and the TIWAG (Tiroler Wasserkraft AG, Innsbruck/Austria) for providing the digital elevation model and orthophotos. The SAGA modules used in the study were written by Olaf Conrad and Volker Wichmann. Special thanks go to the SAGA User Group for advancing the development of SAGA GIS, and to Alexander Brenning for developing the RSAGA package (Brenning, 2008) that made data exchange between SAGA GIS and R really smooth. This study formed part of a research project funded by a DFG grant to Tobias Heckmann and Michael Becht (HE5747/1-1 and 2), which is gratefully acknowledged. Last but not least, we wish to thank Mélanie Bertrand, two anonymous referees, and the handling editor for their comments that greatly helped to improve the clarity of the manuscript.

Edited by: A. Günther
Reviewed by: M. Bertrand and two anonymous referees

References

Akaike, H.: Information Theory and an Extension of the Maximum Likelihood Principle, in: Proceedings of the Second International Symposium on Information Theory, edited by: Petrov, B. N. and Csaki, F., Akademiai Kiado, Budapest, 267–281, 1973.

Ardizzone, F., Cardinali, M., Carrara, A., Guzzetti, F., and Reichenbach, P.: Impact of mapping errors on the reliability of landslide hazard maps, Nat. Hazards Earth Syst. Sci., 2, 3–14, doi:10.5194/nhess-2-3-2002, 2002.

Atkinson, P. M. and Massari, R.: Autologistic modelling of susceptibility to landsliding in the Central Apennines, Italy, Geomorphology, 130, 55–64, doi:10.1016/j.geomorph.2011.02.001, 2011.

Atkinson, P. M., Jiskoot, H., Massari, R., and Murray, T.: Generalized linear modelling in geomorphology, Earth Surf. Proc. Land., 23, 1185–1195, 1998.

Ayalew, L. and Yamagishi, H.: The application of GIS-based logistic regression for landslide susceptibility mapping in the Kakuda-Yahiko Mountains, Central Japan, Geomorphology, 65, 15–31, 2005.

Baeza, C. and Corominas, J.: Assessment of Shallow Landslide Susceptibility by Means of Multivariate Statistical Techniques, Earth Surf. Proc. Land., 26, 1251–1263, 2001.

Becht, M.: Untersuchungen zur aktuellen Reliefentwicklung in alpinen Einzugsgebieten, vol. 47 of Münchener Geographische Abhandlungen A, Geobuch, München, 1995.

Beguería, S.: Validation and Evaluation of Predicitve Models in Hazard Assessment and Risk Management, Nat. Hazards, 37, 315–329, 2006.

Beguería, S.: Changes in land cover and shallow landslide activity: A case study in the Spanish Pyrenees, Geomorphology, 74, 196–206, doi:10.1016/j.geomorph.2005.07.018, 2006.

Beguería, S. and Lorente, A.: Landslide hazard mapping by multivariate statistics: comparison of methods and case study in the Spanish Pyrenees, The Damocles Project Work, Contract No. EVG1-CT-1999-00007, Instituto Pirenaico de Ecologia, Zaragoza, Spain, 2003.

Bell, R., Petschko, H., Röhrs, M., and Dix, A.: Assessment of landslide age, landslide persistence and human impact using airborne laser scanning digital terrain models, Geograf. Ann. A, 94, 135–156, doi:10.1111/j.1468-0459.2012.00454.x, 2012.

Binaghi, E., Luzi, L., Madella, P., Pergalani, F., and Rampini, A.: Slope Instability Zonation: A Comparison between Certainty Factor and Fuzzy Dempster–Shafer Approaches, Nat. Hazards, 17, 77–97, 1998.

Blahut, J., Horton, P., Sterlacchini, S., and Jaboyedoff, M.: Debris flow hazard modelling on medium scale: Valtellina di Tirano, Italy, Nat. Hazards Earth Syst. Sci., 10, 2379–2390, doi:10.5194/nhess-10-2379-2010, 2010a.

Blahut, J., van Westen, C. J., and Sterlacchini, S.: Analysis of landslide inventories for accurate prediction of debris-flow source areas, Geomorphology, 119, 36–51, 2010b.

Bonham-Carter, G.: Geographic Information Systems for Geoscientists, vol. 13 of Computer Methods in the Geosciences, Pergamon, Elsevier Science Publications, Kidlington, Tarrytown NY, Tokyo, 1994.

Bovis, M. J. and Jakob, M.: The role of debris supply conditions in predicting debris flow activity, Earth Surf. Proc. Land., 24, 1039–1054, 1999.

Brenning, A.: Spatial prediction models for landslide hazards: review, comparison and evaluation, Nat. Hazards Earth Syst. Sci., 5, 853–862, doi:10.5194/nhess-5-853-2005, 2005.

Brenning, A.: Statistical geocomputing combining R and SAGA: The example of landslide susceptibility analysis with generalized additive models, in: SAGA – Seconds Out, vol. 19 of Hamburger Beiträge zur Physischen Geographie und Landschaftsökologie, edited by: Böhner, J., Blaschke, T., and Montanarella, L., Institute for Geography, University of Hamburg, Hamburg, Germany, 2008.

Brenning, A.: RSAGA: SAGA Geoprocessing and Terrain Analysis in R, R package version 0.93-6, http://CRAN.R-project.org/package=RSAGA, last access: 4 November 2013.

Brenning, A., Gruber, S., and Hoelzle, M.: Sampling and statistical analyses of BTS measurements, Permafrost Periglac. Process., 16, 383–393, doi:10.1002/ppp.541, 2005.

Calcagno, V. and Mazancourt, C. D.: glmulti: An R Package for Easy Automated Model Selection with (Generalized) Linear Models, J. Stat. Softw., 34, 1–29, 2010.

Carrara, A., Crosta, G., and Frattini, P.: Comparing models of debris-flow susceptibility in the alpine environment, Geomorphology, 94, 353–378, doi:10.1016/j.geomorph.2006.10.033, 2008.

Chung, C.-J. F. and Fabbri, A.: Validation of Spatial Prediction Models for Landslide Hazard Mapping, Nat. Hazards, 30, 451–472, 2003.

Clark, P. and Evans, F.: Distance to Nearest Neighbor as a Measure of Spatial Relationships in Populations, Ecology, 35, 445–453, 1954.

Coe, J. A., Kinner, D. A., and Godt, J. W.: Initiation conditions for debris flows generated by runoff at Chalk Cliffs, central Colorado, Geomorphology, 96, 270–297, doi:10.1016/j.geomorph.2007.03.017, 2008.

Ermini, L., Catani, F., and Casagli, N.: Artificial Neural Networks applied to landslide susceptibility assessment, Geomorphology, 66, 327–343, 2005.

Fabbri, A. G., Chung, C. J. F., Cendrero, A., and Remondo, J.: Is Prediction of Future Landslides Possible with a GIS?, Nat. Hazards, 30, 487–499, 2003.

Fischer, L., Rubensdotter, L., Sletten, K., Stalsberg, K., Melchiorre, C., Horton, P., and Jaboyedoff, M.: Debris flow modeling for susceptibility mapping at regional to national scale in Norway, in: Landslides and engineered slopes, edited by: Eberhardt, E. B., CRC Press, Leiden, 723–729, 2012.

Freeman, G. T.: Calculating catchment area with divergent flow based on a regular grid, Comput. Geosci., 17, 413–422, 1991.

Geitner, C.: Sedimentologische und vegetationsgeschichtliche Untersuchungen an fluvialen Sedimenten in den Hochlagen des Horlachtales (Stubaier Alpen, Tirol): Ein Beitrag zur zeitlichen Differenzierung der fluvialen Dynamik im Holozän, vol. 31 of Münchner Geographische Abhandlungen, Geobuch-Verlag, Diss., München, 1999.

Green, S. B.: How Many Subjects Does It Take To Do A Regression Analysis, Multivar. Behav. Res., 26, 499–510, doi:10.1207/s15327906mbr2603_7, 1991.

Guns, M. and Vanacker, V.: Logistic regression applied to natural hazards: rare event logistic regression with replications, Nat. Hazards Earth Syst. Sci., 12, 1937–1947, doi:10.5194/nhess-12-1937-2012, 2012.

Guzzetti, F., Galli, M., Reichenbach, P., Ardizzone, F., and Cardinali, M.: Landslide hazard assessment in the Collazzone area, Umbria, Central Italy, Nat. Hazards Earth Syst. Sci., 6, 115–131, doi:10.5194/nhess-6-115-2006, 2006a.

Guzzetti, F., Reichenbach, P., Ardizzone, F., Cardinali, M., and Galli, M.: Estimating the quality of landslide susceptibility models, Geomorphology, 81, 166–184, doi:10.1016/j.geomorph.2006.04.007, 2006b.

Guzzetti, F., Mondini, A. C., Cardinali, M., Fiorucci, F., Santangelo, M., and Chang, K.-T.: Landslide inventory maps: New tools for an old problem, Earth-Sci. Rev., 112, 42–66, doi:10.1016/j.earscirev.2012.02.001, 2012.

Hagg, W. and Becht, M.: Einflüsse von Niederschlag und Substrat auf die Murauslösung in Beispielgebieten der Ostalpen, Z. Geomorphol., 123, 79–92, 2000.

Heckmann, T. and Becht, M.: Invesigating the transferability of a statistical disposition model for slope-type debris flows, Erdkunde, 63, 19–33, 2009.

Heckmann, T. and Schwanghart, W.: Geomorphic coupling and sediment connectivity in an alpine catchment – Exploring sediment cascades using graph theory, Geomorphology, 182, 89–103, doi:10.1016/j.geomorph.2012.10.033, 2013.

Heckmann, T., Haas, F., Wichmann, V., and Morche, D.: Sediment Budget and Morphodynamics of an Alpine Talus Cone on Different Timescales, Z. Geomorphol., 52, 103–121, 2008.

Heinimann, H., Hollenstein, K., Kienholz, H., Krummenacher, B., and Mani, P.: Methoden zur Analyse und Bewertung von Naturgefahren, vol. 85 of Umwelt-Materialien, BUWAL – Bundesamt für Umwelt, Wald und Landschaft, Bern, 1998.

Hjort, J. and Luoto, M.: Novel theoretical insights into geomorphic process-environment relationships using simulated response curves, Earth Surf. Proc. Land., 36, 363–371, doi:10.1002/esp.2048, 2011.

Hjort, J. and Marmion, M.: Effects of sample size on the accuracy of geomorphological models, Geomorphology, 102, 341–350, 2008.

Horton, P., Jaboyedoff, M., and Bardou, E.: Debris flow susceptibility mapping at a regional scale, in: Proceedings of the 4th Canadian Conference on Geohazards, edited by: Locat, J., Perret, D., Turmel, D., Demers, D., and Leroueil, S., Presse de l'Université Laval, Laval and Québec, 2008.

Hosmer, D. W. and Lemeshow, S.: Applied logistic regression, 2nd Edn., Wiley, New York, 2000.

Ivy-Ochs, S., Kerschner, H., Reuther, A., Preusser, F., Heine, K., Maisch, M., Kubik, P., and Schlüchter, C.: Chronology of the last glacial cycle in the European Alps, J. Quaternary Sci., 23, 559–573, 2008.

Johnson, A. M. and Rodine, J. R.: Debris flow, in: Slope instability, edited by: Brunsden, D. and Prior, D. B., Wiley, Chichester, 257–361, 1984.

Kappes, M. S., Malet, J.-P., Remaître, A., Horton, P., Jaboyedoff, M., and Bell, R.: Assessment of debris-flow susceptibility at medium-scale in the Barcelonnette Basin, France, Nat. Hazards Earth Syst. Sci., 11, 627-641, doi:10.5194/nhess-11-627-2011, 2011.

King, G. and Zeng, L.: Logistic regression in rare events data, Polit. Anal., 9, 137–163, 2001.

Lee, S., Ryu, J.-H., Min, K., and Won, J.-S.: Landslide Susceptibility Analysis Using GIS and Artificial Neural Network, Earth Surf. Proc. Land., 28, 1361–1376, 2003.

Legendre, P.: Spatial Autocorrelation: Trouble or New Paradigm?, Ecology, 74, 1659–1673, 1993.

Liu, Y., Guo, H. C., Zou, R., and Wang, L. J.: Neural network modeling for regional hazard assessment of debris flow in Lake Qionghai Watershed, China, Environ. Geol., 49, 968–976, 2006.

Luoto, M. and Hjort, J.: Evaluation of current statistical approaches for predictive geomorphological mapping, Geomorphology, 67, 299–315, 2005.

Luoto, M., Marmion, M., and Hjort, J.: Assessing spatial uncertainty in predictive geomorphological mapping: A multi-modelling approach, Comput. Geosci., 36, 355–361, doi:10.1016/j.cageo.2009.07.008, 2010.

Magliulo, P., Di Lisio, A., Russo, F., and Zelano, A.: Geomorphology and landslide susceptibility assessment using GIS and bivariate statistics: a case study in southern Italy, Nat. Hazards, 47, 411–435, 2008.

Magurran, A. E.: Measuring Biological Diversity, Blackwell Science Ltd., Oxford, 2004.

Marmion, M., Hjort, J., Thuiller, W., and Luoto, M.: A comparison of predictive methods in modelling the distribution of periglacial landforms in Finnish Lapland, Earth Surf. Proc. Land., 33, 2241–2254, 2008.

Marmion, M., Hjort, J., Thuiller, W., and Luoto, M.: Statistical consensus methods for improving predictive geomorphology maps, Comput. Geosci., 35, 615–625, doi:10.1016/j.cageo.2008.02.024, 2009.

Menard, S. W.: Applied logistic regression analysis, 2nd Edn., Sage Publications, Thousand Oaks, 2002.

Meusburger, K. and Alewell, C.: On the influence of temporal change on the validity of landslide susceptibility maps, Nat. Hazards Earth Syst. Sci., 9, 1495–1507, doi:10.5194/nhess-9-1495-2009, 2009.

Montgomery, D. R. and Foufoula-Georgiou, E.: Channel network source representation using digital elevation models, Water Resour. Res., 29, 3925–3934, 1993.

Moore, I., Grayson, R., and Ladson, A.: Digital terrain modelling: A review of hydrological, geomorphological, and biological applications, Hydrol. Process., 5, 3–30, 1991.

Neuhäuser, B. and Terhorst, B.: Landslide Susceptibility Assessment Using Weights-of-Evidence Applied on a Study Site at the Jurassic Escarpment of the Swabian Alb (SW-Germany), Geomorphology, 86, 12–24, 2006.

O'brien, R. M.: A Caution Regarding Rules of Thumb for Variance Inflation Factors, Quality Quant., 41, 673–690, doi:10.1007/s11135-006-9018-6, 2007.

Ohlmacher, G. C. and Davis, J. C.: Using multiple logistic regression and GIS technology to predict landslide hazard in northeast Kansas, USA, Eng. Geol., 69, 331–343, doi:10.1016/S0013-7952(03)00069-3, 2003.

Petschko, H., Brenning, A., Bell, R., Goetz, J., and Glade, T.: Assessing the quality of landslide susceptibility maps – case study Lower Austria, Nat. Hazards Earth Syst. Sci., 14, 95–118, doi:10.5194/nhess-14-95-2014, 2014.

Pike, R. J., Graymer, R. W., and Sobieszczyk, S.: A simple GIS model for mapping landslide susceptibility, in: Concepts and Modelling in Geomorphology, edited by: Evans, I., Dikau, R., Tokunaga, E., Ohmori, H., and Hirano, M., Terrapub, Tokyo, 185–197, 2003.

R Development Core Team: R: A Language and Environment for Statistical Computing, http://www.R-project.org/ (last access: January 2013), 2012.

Rickenmann, D. and Zimmermann, M.: The 1987 debris flows in Switzerland: documentation and analysis, Geomorphology, 8, 175–189, 1993.

Rieger, D.: Bewertung der naturräumlichen Rahmenbedingungen für die Entstehung von Hangmuren, Möglichkeiten zur Modellierung des Murpotentials, vol. 51 of Münchener Geographische Abhandlungen A, Geobuch, München, 1999.

Rossi, M., Guzzetti, F., Reichenbach, P., Mondini, A. C., and Peruccacci, S.: Optimal landslide susceptibility zonation based on multiple forecasts, Geomorphology, 114, 129–142, doi:10.1016/j.geomorph.2009.06.020, 2010.

Ruette, J. v., Papritz, A., Lehmann, P., Rickli, C., and Or, D.: Spatial statistical modeling of shallow landslides – Validating predictions for different landslide inventories and rainfall events, Geomorphology, 133, 11–22, doi:10.1016/j.geomorph.2011.06.010, 2011.

Sachs, L. and Hedderich, J.: Angewandte Statistik: Methodensammlung mit R, 12. Aufl. Edn., Springer, Berlin, 2006.

Sappington, J., Longshore, K., and Thompson, D.: Quantifying Landscape Ruggedness for Animal Habitat Analysis: A Case Study Using Bighorn Sheep in the Mojave Desert, J. Wildlife Manage., 71, 1419–1426, 2007.

Schwanghart, W. and Heckmann, T.: Fuzzy delineation of drainage basins through probabilistic interpretation of diverging flow algorithms, Environ. Modell. Softw., 33, 106–113, doi:10.1016/j.envsoft.2012.01.016, 2012.

Shannon, C. E.: A mathematical theory of communication, Bell Syst. Tech. J., 27, 379–423, 1948.

Simpson, E. H.: Measurement of diversity, Nature, 163, 688, doi:10.1038/163688a0, 1949.

Stockwell, D. and Townsend Peterson, A.: Effects of sample size on accuracy of species distribution models, Ecol. Modell., 148, 1–13, 2002.

Takahashi, T.: Estimation of potential debris flows and their hazardous zones: Soft countermeasures for a disaster, J. Nat. Disaster Sci., 3, 57–89, 1981.

Thiel, M.: Quantifizierung der Konnektivität von Sedimentkaskaden in alpinen Geosystemen, Ph.D. thesis, Catholic University of Eichstätt-Ingolstadt, Eichstätt, 2013.

Thomas, R.: Information Statistics in Geography, vol. 31 of Concepts and techniques in modern geography, Geo Abstracts, Norwich, Norfolk, 1981.

van Asselen, S. and Seijmonsbergen, A. C.: Expert-driven semi-automated geomorphological mapping for a mountainous area using a laser DTM, Geomorphology, 78, 309–320, 2006.

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

Van Den Eeckhaut, M., Reichenbach, P., Guzzetti, F., Rossi, M., and Poesen, J.: Combined landslide inventory and susceptibility assessment based on different mapping units: an example from the Flemish Ardennes, Belgium, Nat. Hazards Earth Syst. Sci., 9, 507–521, doi:10.5194/nhess-9-507-2009, 2009.

Van Den Eeckhaut, M., Marre, A., and Poesen, J.: Comparison of two landslide susceptibility assessments in the Champagne–Ardenne region (France), Geomorphology, 115, 141–155, 2010.

Vanwalleghem, T., van den Eeckhaut, M., Poesen, J., Govers, G., and Deckers, J.: Spatial analysis of factors controlling the presence of closed depressions and gullies under forest: Application of rare event logistic regression, Geomorphology, 95, 504–517, doi:10.1016/j.geomorph.2007.07.003, 2008.

Venables, W. N. and Ripley, B. D.: Modern applied statistics with S, 4th Edn., Springer, New York, 2002.

Vorpahl, P., Elsenbeer, H., Märker, M., and Schröder, B.: How can statistical models help to determine driving factors of landslides?, Ecol. Modell., 239, 27–39, doi:10.1016/j.ecolmodel.2011.12.007, 2012.

Wang, H. B. and Sassa, K.: Comparative evaluation of landslide susceptibility in Minamata area, Japan, Environ. Geol., 47, 956–966, doi:10.1007/s00254-005-1225-2, 2005.

Webster, R. and Oliver, M. A.: Geostatistics for environmental scientists, 2nd Edn., Wiley, Chichester, 2007.

Weisberg, S. and Fox, J.: An R companion to applied regression, Sage Publications, Incorporated, Los Angeles, London, New Delhi, Singapore, Washington, D.C., 2010.

Wichmann, V., Heckmann, T., Haas, F., and Becht, M.: A new modelling approach to delineate the spatial extent of alpine sediment cascades, Geomorphology, 111, 70–78, 2009.

Zevenbergen, L. and Thorne, C.: Quantitative Analysis of Land Surface Topography, Earth Surf. Proc. Land., 12, 47–56, 1987.

Zimmermann, M., Mani, P., Gamma, P., Gsteiger, P., Heiniger, O., and Hunziker, G.: Murganggefahr und Klimaänderung – ein GIS-basierter Ansatz, Schlussbericht NFP 31, vdf Hochschulverlag AG, Zurich, 1997.