Brenning, A. (2008): Statistical geocomputing combining R and SAGA: The example of landslide susceptibility analysis with generalized additive models. In: J. Böhner, T. Blaschke & L. Montanarella (eds.), SAGA – Seconds Out (= Hamburger Beiträge zur Physischen Geographie und Landschaftsökologie, vol. 19), 23-32.

Statistical geocomputing combining R and SAGA:

The example of landslide susceptibility analysis with generalized additive models

Alexander Brenning

Department of Geography, University of Waterloo, Ontario, Canada

Abstract

The integration of statistical software with geographical information systems is required to be able to efficiently combine the most powerful tools and techniques available in both environments. The RSAGA package, which provides access to SAGA GIS geoprocessing functions from within the R statistical data analysis environment, is a recent contribution to this endeavor. The present work gives an overview of the structure of the RSAGA package, and demonstrates its usefulness in the context of landslide susceptibility modeling with terrain attributes and generalized additive models (GAMs). The GAM is an extension of the generalized linear model (e.g. linear and logistic regression). It is able to model nonlinear relationships, but retains an interpretable additive structure. In the case study on landslide distribution in the Ecuadorian Andes, several local as well as catchment-related morphometric attributes are important, mostly nonlinear predictors of landslide occurrence. Other applications that can benefit from an integration of modern statistical computing techniques and GIS-based digital terrain analysis include pedometrics, Precision Agriculture, and species habitat studies.

Introduction

Spatial analysis using Geographical Information System (GIS) techniques on one side and the statistical analysis of environmental data on the other provide two different views of spatial data analysis problems that are too often separated by disciplinary and software-related barriers. However, both perspectives have much to contribute to spatial data analysis: GIS software provides a rich set of tools for spatial data manipulation, queries, and visualization, and statistical data analysis software can offer spatial and non-spatial techniques required for understanding data or applying predictive models. These models may range from traditional statistical ones such as linear regression to complex black-box models such as the support vector machine, a powerful, emerging machine-learning technique.

Given the fast progress in geographical information science and computational statistics, efficient ways of coupling GIS and data analysis software are required. This integration has been an important topic in recent years in the developer community of R (see e.g. Bivand, 2000; Brenning & van den Boogaart, 2001; Bivand et al., 2008), an open-source data analysis environment that is widely used in statistical

sciences (Ihaka & Gentleman, 1996). A wide variety of free R extensions ("packages") is now available that support geodata formats (e.g. packages shapefiles, rgdal, maptools) or spatial statistical techniques (e.g. packages gstat, spdep, nlme).

The objective of the present work is to introduce the R package RSAGA (Brenning, 2007a), which integrates SAGA functionality into R, in a case study of landslide susceptibility modeling that also exemplifies the need for a seamless integration of terrain analysis and statistical models. A similar link is currently being established to access geoprocessing tools of ESRI's ArcGIS from within R using the RPyGeo package and a Python interface (Brenning, 2007b).

This work is divided into six sections: The first two sections introduce the landslide distribution data, and the terrain analysis and statistical techniques suitable for their analysis. The next two sections give an overview of the structure of the RSAGA package, and outline its application in constructing generalized additive models of landslide susceptibility. Finally, analytical and predictive model results are presented and discussed, and some general conclusions are drawn.

Case study: Landslide susceptibility

Landslide susceptibility modeling on a regional scale can be considered a (soft) classification problem, where a measure of the likelihood of landslide occurrence has to be estimated. A variety of approaches has been applied for this purpose, mainly empirical ones (see the review by Brenning, 2005), but also physically based models (e.g. Montgomery & Dietrich, 1994), and cellular-automata models (Guthrie et al., 2008). Empirical models require a training data set – a landslide inventory – to be provided for model fitting, and especially for the assessment of flexible machine learning techniques it is also required to use spatially or temporally independent test data sets or a spatial cross-validation (Brenning, 2005). Among the predictors that are most widely used in landslide susceptibility modeling are terrain attributes (especially slope angle and contributing area), land use, and soil or lithological properties.

The present study uses a multitemporal landslide inventory from the Andes of Southern Ecuador that was created by Stoyan (2000). Landslides and their importance in controlling vegetation succession in the tropical mountain rainforests of the area were studied by Stoyan (2000), Wilcke et al. (2003), Bussmann (2004) and Lozano et al. (2005). Brenning (2005) compared different statistical and machine-learning techniques for landslide prediction using a different subset of Stoyan's inventory, however with an extended set of predictor variables.

The study area comprises 11.2 km², 0.94 km² of which were classified as mass movements in the transition between shallow translational landslides and debris flows (Figure 1). Landslide inventories were compiled for 1962, 1969, 1976, 1989, and 1998. Deforestation and subsequent grazing take place along the road from Loja to Zamora; this as well as the direct impact of the road on slope stability lead to increased landslide activity near the road. Distance from deforestation and distance from road were therefore used as predictor variables in addition to terrain attributes (see next section) and distance to past landslide locations. Note that all distance values above a certain cutoff value (100 m for distance from past landslides, 300 m otherwise) were reduced to that cutoff value to avoid overfitting.

In the present study, two landslide susceptibility models are fitted to Stoyan's (2000) landslide inventory of 1976, and evaluated based on the inventories of 1989 and 1998. This evaluation using "future" inventories is intended to avoid reporting overoptimistic error measures that would be obtained by measuring the performance on the training data itself (compare Brenning, 2005).

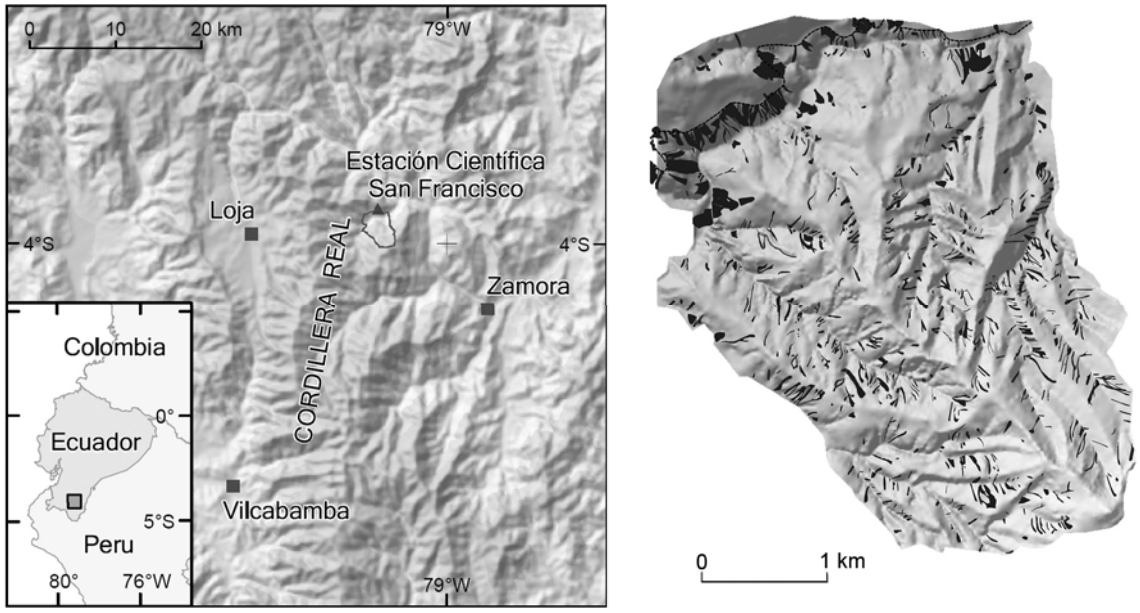

Figure 1: Location of the study area (left) and spatial distribution of all landslides observed between 1962 and 1998 (right).

Terrain Analysis and Statistical Methods

In addition to soil and geological controls on landslide susceptibility, for which accurate data is unavailable in many mountain areas, terrain attributes derived from digital elevation models (DEMs) have been identified as important influences on the spatial distribution of landslides (e.g. Montgomery & Dietrich, 1994; Atkinson & Massari, 1998; Ohlmacher & Davis, 2003). The importance of terrain attributes is related to the physical controls on slope stability (slope angle, upslope contributing area; Montgomery & Dietrich, 1994), but also to the fact that terrain attributes are proxies for a variety of processes and substrate properties (e.g. soil erosion and accumulation; soil moisture content, soil texture). The present work estimates landslide susceptibility based on selected terrain attributes (slope angle, plan and profile curvature, size and slope of the upslope contributing area) and land-cover related variables (distance to road, to deforestation, to past landslide locations). Terrain attributes are derived from a DEM created by Stoyan (2000) from topographic map data enhanced by air photo evaluation.

Since substrate properties – including variables that vary in time such as soil wetness – are usually not known at all or not in detail, the combination of a variety of terrain attributes in empirical-statistical models is often more promising than the application of physically based model with insufficient data. Logistic regression is among the most widely used methods in landslide susceptibility modeling and many other applications where a “soft” discrimination between the presence and absence of a phenomenon is desired (e.g. Atkinson & Massari, 1998; Ohlmacher & Davis, 2003).

While logistic regression has proven to be competitive with some of the highly flexible machine-learning techniques that tend to overfit to spatially dependent data (Brenning, 2005), it is unable to model nonlinear relationships that are often found in landslide patterns. The generalized additive model (GAM; Hastie & Tibshirani, 1990; implemented in the R package gam: Hastie, 2006) extends the logistic and

linear regression models by introducing nonlinear transformations of the predictor variables while maintaining an interpretable additive structure, i.e. the effects of different predictors are still added up. The study of Brenning et al. (2007) is a recent example of the use of the GAM in geomorphological research based on terrain attributes as predictors.

The receiver operating characteristic (ROC) curve is suitable for assessing the predictive performance of soft classifiers independently of a specific decision threshold that could be used to delineate unstable areas. The ROC curve represents all possible combinations of sensitivities (fraction of true positive predictions) and specificities (fraction of true negative predictions) that can be achieved by a given soft classifier. The area under the ROC curve (AUROC) takes a value between 0.5 (no discrimination) to 1.0 (perfect discrimination), which summarizes the ability of a classifier to separate landslide and non-landslide locations.

The RSAGA Package

The RSAGA package provides R geocomputing functions that make use of the command line interface of SAGA GIS, saga_cmd.exe, to execute SAGA GIS modules. RSAGA is currently limited to the Windows platform. The RSAGA package consists of the following three general components.

1. The **RSAGA core** provides low-level access to the SAGA command line program through the `rsaga.geoprocessor` function, which converts R arguments into a command that is executed by the operating system. It also accesses SAGA module libraries and help files to give information on available modules (function `rsaga.get.modules`) and command line options (`rsaga.get.usage`). Most RSAGA users will however find the information they need in the R help files, unless they wish to use a SAGA module that has not yet been integrated in R.

2. The **RSAGA modules** part of the package implements a series of interface functions that allow the R user to run SAGA modules from within the accustomed R environment. Examples are functions such as `rsaga.local.morphometry` (or the derived `rsaga.slope`) and `rsaga.flow.accumulation`, which will be applied later in this work. These functions are often more convenient than the corresponding SAGA modules; as an example, flow accumulation algorithms can be chosen by their name or acronym (e.g. “mfd” for the multiple flow direction algorithm) as an alternative to the integer code required by SAGA itself.

3. The **RSAGA grid tools** are a compilation of additional useful functions for geocomputing and data input/output. This package section implements, for example, generic functions for applying user-defined functions on moving windows (`[multi.]focal.function`). This allows the user to implement application-specific terrain attributes (such as the wind shelter index for snow distribution modeling; Plattner et al., 2004) or experiment with new filters. The family of “pick” functions (e.g. `pick.from.ascii.grid`, `pick.from.shapefile`) furthermore allows to quickly add data from a variety of sources as new variables to a data frame in R. (It should be noted that these functions are currently rather inefficient for large data sets because they are fully implemented in R, which is an interpreter language.)

Landslide Susceptibility Modeling using R and RSAGA

Key parts of the R implementation of the present analysis are outlined in this section. The starting point is a set of ASCII grids representing the DEM, distance to road, the five landslide inventories, and distance to deforestation at the five time points. (The RPyGeo package of Brenning (2007b) could have been used to derive the distance grids from vector data, and to export the original ESRI rasters to ASCII format, but integration with ArcGIS is not the focus of this work.)

After converting all ASCII grids to SAGA format with RSAGA by calling

```
rsaga.esri.to.sgrd(dir(pattern = glob2rx("*.asc")))
```

local morphometric parameters are calculated with the method of Zevenbergen & Thorne (1987), sinks are filled, and contributing area size and slope are determined with the multiple flow direction (MFD) algorithm (Quinn et al., 1991):

```
rsaga.local.morphometry("dem", out.slope = "slope",
        out.hcurv = "hcurv", out.vcurv = "vcurv",
        method = "poly2zevenbergen")
rsaga.sink.removal("dem", out.dem = "sdem", method = "fill")
rsaga.parallel.processing("sdem", out.carea = "carea",
        out.cslope = "cslope", method = "mfd")
```

The results are then converted back to ASCII format with the `rsaga.sgrd.to.esri` function.

To obtain a compact training sample with a sufficient number of landslide pixels, random samples of 500 landslide and 500 non-landslide pixels are drawn from each of the landslide inventory data sets. Without going into the details of the sampling, the terrain attributes are added to the table (data frame d) with the sampling locations using commands such as

```
d = pick.from.ascii.grid(d, “slope”)
```

This uses nearest neighbor interpolation, which is just fine since the sample locations coincide with grid nodes. Similar “pick” functions are available in RSAGA to conveniently match data from, for example, point shapefiles to R data frames using either nearest neighbor or kriging interpolation. The time-dependent grids (landslide inventories and deforestation distances) are matched to the samples in a similar way.

Spinograms (as produced by the R function `spineplot`) are a simple means for simultaneously plotting the empirical frequency of landslide occurrence conditional on a predictor variable, and the empirical frequency of that variable (Figure 2).

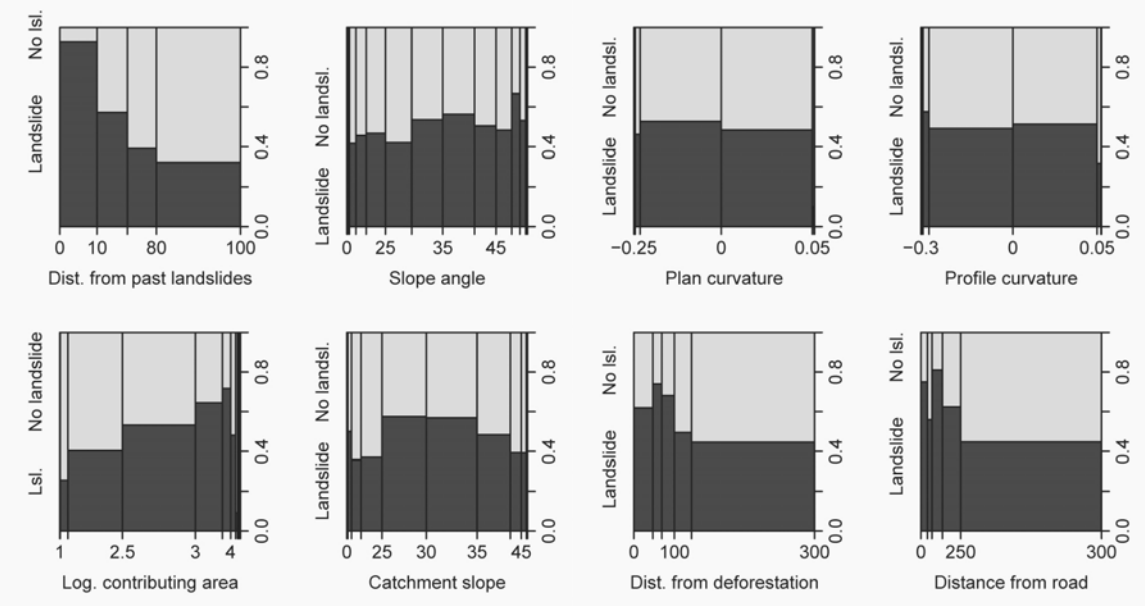

Figure 2: Empirical frequencies of landslide occurrence in the 1976 training data conditional on the predictor variables. Bar widths in these spinograms are proportional to the empirical frequency of the given interval of predictor values.

The generalized additive model (GAM) is built using a stepwise variable selection starting from the null model and using the Akaike Information Criterion (AIC) as selection criterion. The scope of the model is set up to allow the GAM to choose in the variable selection between (a) including a variable as a nonparametrically transformed predictor consuming two equivalent degrees of freedom, (b) including it as a linear predictor, or (c) omitting it. The AIC penalizes the model size and therefore guarantees that the final model is compact. The R implementation is as follows:

```
library(gam)
gam.scope = list(
    ~ 1 + distroad + s(distroad,2),
    ~ 1 + I(distdeforest > 0) + distdeforest + s(distdeforest,2),
    ~ 1 + log.carea + s(log.carea,2),
    ~ 1 + cslope + s(cslope,2),
    ~ 1 + hcurv + s(hcurv,2),
    ~ 1 + vcurv + s(vcurv,2),
    ~ 1 + slope + s(slope,2) )

fit0 = gam(slides ~ 1, data = train, family = binomial)

fit = step.gam(fit0, scope = gam.scope,
        direction = "both", trace = TRUE)

pred = predict(fit, newdata = test, type = "response")
```

In addition to this model, which will be referred to as GAM1, a second model GAM2 is constructed that is offered an additional variable, the distance from past landslides (i.e. from the previous inventory) as a candidate predictor in stepwise variable selection. The nonlinear transformations both GAM models are shown in Figures 3 (GAM1) and 4 (GAM2).

In R, fitted statistical models can conveniently be applied to data available as R tables ("data frames"). A simple approach for applying fitted models to stacks of raster files representing the predictor variables is however not available. RSAGA provides an interface for applying focal functions on moving windows to raster data sets. This approach is currently being extended to be able to deal with model predictions on grids in the next RSAGA release (version 0.9-1). The following R statement outlines how this will be resolved (implementation details may still change); the resulting prediction map is shown in Figure 6:

```
multi.focal.function(
    in.grids = c("slope", "hcurv", "vcurv", "carea",
                 "cslope", "distroad", "distdeforest"),
    out.varnames = "pred",
    fun = grid.predict, control.predict = list(type = "response"),
    fit = fit, trafo = my.trafo )
```

In detail, the `grid.predict` function passed on to `multi.focal.function` is responsible for calling the `fit` object's actual `predict` method; since the raw variables are often modified before going into a model as predictors (example: log-transformation of the contributing area variable), the `trafo` argument allows for user-defined variable transformations as defined in the `my.trafo` function that will be automatically applied to the grid values taken from the `in.grids`.

Results and Discussion

Both models, GAM1 and GAM2 (Figures 3 and 4, respectively), share the $\log_{10}$-transformed upslope contributing area (`log.carea`) and the profile curvature (`vcurv`) as nonlinear terms; the transformations are practically the same in both models, with an obvious nonlinearity in `log.carea` and a nearly linear behavior for `vcurv`. GAM1 (Figure 3) furthermore uses linear terms for distance from road and slope angle; as expected, steep slopes and areas close to the road are more likely to be unstable in GAM1, and this is especially true where the upslope contributing area is "not too small" (at least a few hundred square meters) and where the profile curvature is convex (i.e. positive).

The second model (GAM2; Figure 4) indicates that landslides are very likely to occur at or near past landslide locations (within <30 m distance), which relates to the reactivation or lateral growth of existing landslide scarps. Since landslide density was always highest near the road and near deforestation areas, the past-landslide variable makes it unnecessary to use the road and deforestation related variables. A steep upslope contributing area (>20° slope angle) is an additional risk factor with a nonlinear influence, as well as a "not too small" contributing area and a convex profile curvature.

The predictive performance of both models on the 1989 and 1998 data sets, i.e. in the prediction of "future" landslides, is visualized in Figure 5 as ROC curves. AUROC values range between 0.67 and 0.77, which is reasonable for landslide susceptibility models evaluated on independent test data sets (compare Brenning, 2005). The GAM2 model, which uses the distance to past landslide locations as a predictor variable, achieves higher AUROC values than GAM1. This is due to the relatively high sensitivities (true positive rates) achieved by GAM2 at low false positive rates (high specificities), which is related to reactivated landslides that can easily be detected based on distance from past landslides.

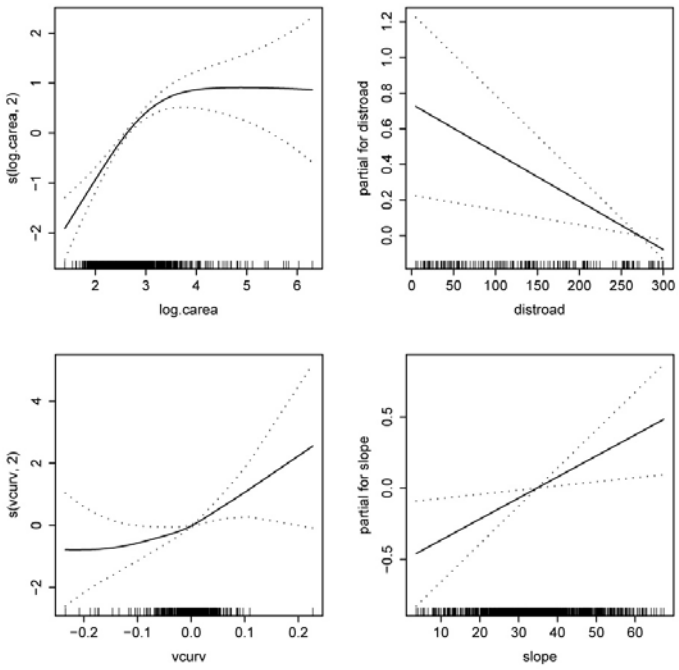

Figure 3: Transformations of predictor variables in the generalized additive model that *does not use* distance to past landslide locations as a predictor variable.

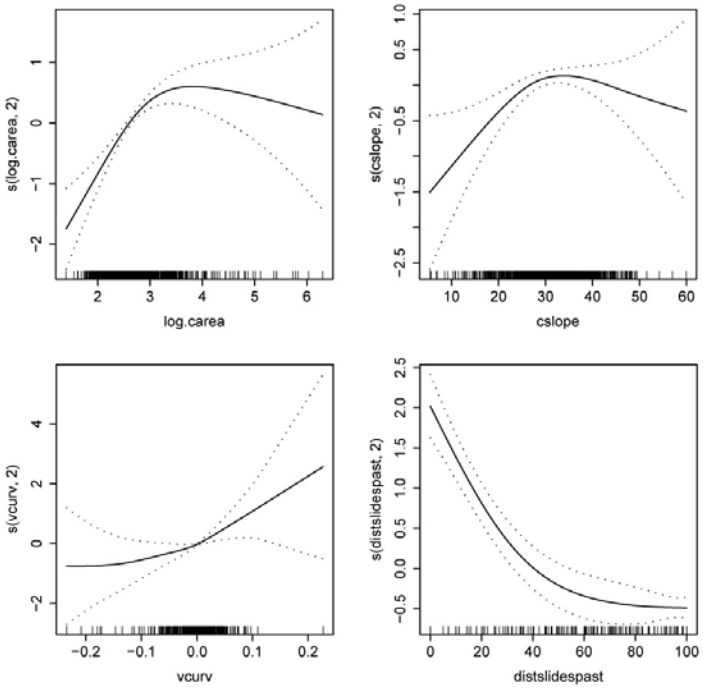

Figure 4: Transformations of predictor variables in the generalized additive model that *uses* distance to past landslide locations as a predictor variable.

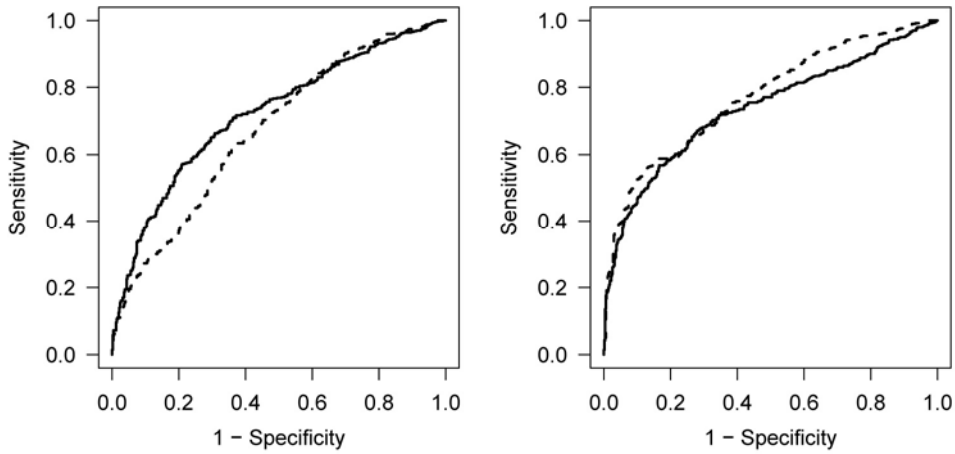

Figure 5: ROC curves of landslide susceptibility predictions generated by the generalized additive models trained on the 1976 landslide inventory (left: GAM1, not using distance to past landslides as predictor variable; right: GAM2, including this variable). ROC curves correspond to predictions for 1989 (solid line; AUROC = 0.72 for GAM1 and 0.74 for GAM2), and 1998 (dashed line; AUROC = 0.67 for GAM1 and 0.77 for GAM2). See also Figure 6.

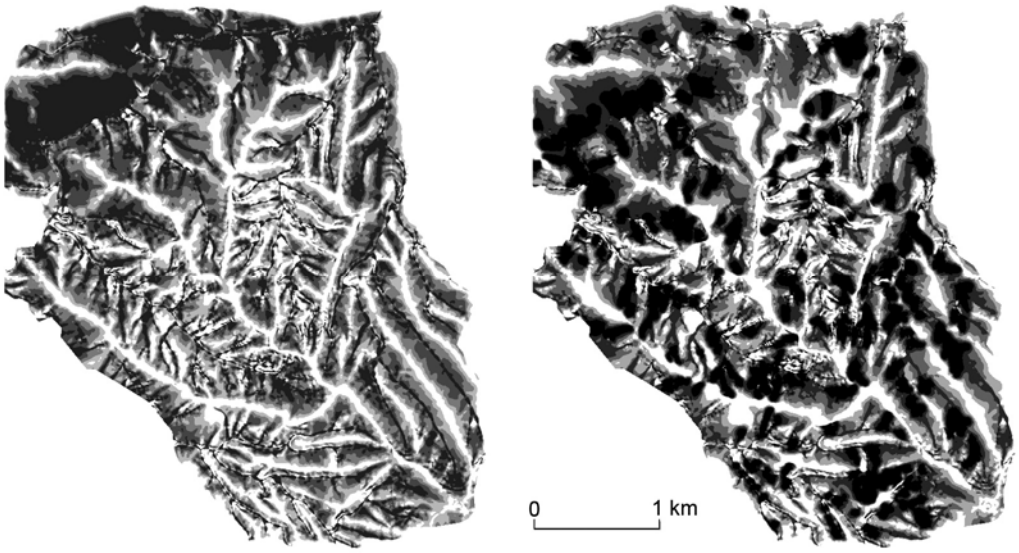

Figure 6: Landslide susceptibility map for 1989 predicted by the generalized additive models GAM1 (left, *not* using distance to past landslides as predictor variable) trained on the 1976 landslide inventory, and GAM2 (right, using the distance variable). See also Figure 5.

Conclusions

The present application of terrain analysis methods and generalized additive models to landslide susceptibility analysis provides an example of how GIS-based tools can be combined with powerful statistical models. The access to SAGA GIS modules through its command line interface facilitates the integration with R as implemented in the RSAGA package (Brenning, 2007a). Drawbacks of the present implementation are the need for frequent file conversions between SAGA and ASCII grid formats and the platform dependence (currently only Windows) because of the direct call to the command line interface program. It will be necessary to synchronize RSAGA with major steps in SAGA development, and to provide some level of forward and backward compatibility with SAGA GIS in the implementation and documentation of RSAGA.

In an applied context, terrain attributes are widely used not only in geomorphology and glaciology (e.g. Atkinson & Massari, 1998; Plattner et al., 2004; Brenning et al., 2007, and the present study), but also in pedometrics (e.g. Böhner & Selige, 2006) and Precision Agriculture (e.g. Brenning et al., 2008). These are some of the fields of research and application of terrain analysis and environmental statistics that may benefit from a better integration of GIS and statistical software, or from an integration of SAGA GIS and R in particular. The RSAGA package is a contribution to this integration that is hopefully useful to many spatial modelers.

Acknowledgements

The author thanks R. Stoyan, Erlangen, for providing the landslide inventory and additional thematic data from the ECSF area, and the SAGA GIS developer team for creating and maintaining this excellent geocomputing environment.

References

Atkinson, P. M. and R. Massari (1998): Generalised linear modeling of susceptibility to landsliding in the Central Apennines, Italy, Computers & Geosciences, 24, 373-385.

Bivand, R. S. (2000): Using the R statistical data analysis language on GRASS 5.0 GIS data base files. Computers & Geosciences, 26: 1043-1052.

Bivand, R. S., E. J. Pebesma & V. Gómez-Rubio (2008): Applied spatial data analysis with R. Springer.

Böhner, J., and T. Selige (2006) Spatial prediction of soil attributes using terrain analysis and climate regionalisation, in SAGA – Analyses and Modelling Applications, eds. J. Böhner, K. R. McCloy and J. Strobl, Göttinger Geographische Abhandlungen, 115: 13-28 and 118-120.

Brenning, A. (2005): Spatial prediction models for landslide hazards: review, comparison and evaluation. Natural Hazards and Earth System Sciences, 5(6): 853–862.

Brenning, A., M. Grasser & D. Friend (2007): Statistical estimation and generalized additive modeling of rock glacier distribution in the San Juan Mountains, Colorado, USA. Journal of Geophysical Research, 112: F02S15.

Brenning, A. (2007a): RSAGA: SAGA geoprocessing and terrain analysis in R. R package version 0.9-0.

Brenning, A. (2007b): RPyGeo: ArcGIS geoprocessing in R via Python. R package version 0.9-0.

Brenning, A., H. Piotraschke & P. Leithold (2008): Geostatistical analysis of on-farm trials in Precision Agriculture. Proceedings, GEOSTATS 2008, Eighth International Geostatistics Congress, December 1-5, 2008, Santiago, Chile, 6 pp. (accepted).

Guthrie, R. H., P. J. Deadman, A. R. Cabrera & S. G. Evans (2008): Exploring the magnitude-frequency distribution: a cellular automata model for landslides. Landslides, 5: 151-159.

Hastie, T. (2006): gam: generalized additive models. R package version 0.98.

Hastie, T. & R. Tibshirani (1990): Generalized additive models. CRC Press, Boca Raton, Florida.

Ihaka, R. & R. Gentleman (1996): R: a language for data analysis and graphics. Journal of Computational and Graphical Statistics, 5: 299-314.

Lozano, P., R. W. Bussmann & M. Küppers (2005): Landslides as ecosystem disturbance – their implications and importance in South Ecuador. Lyonia – A Journal of Ecology and Application, 8(1): 67-72.

Montgomery, D. R. & W. E. Dietrich (1994): A physically based model for the topographic control on shallow landsliding. Water Resources Research, 30(4): 1153-1171.

Ohlmacher, G. C. & J. C. Davis (2003): Using multiple logistic regression and GIS technology to predict landslide hazard in northeast Kansas, USA. Engineering Geology, 69: 331-343.

Plattner, C., L. N. Braun & A. Brenning (2004): The spatial variability of snow accumulation on Vernagtferner, Austrian Alps, in Winter 2003/2004. Zeitschrift für Gletscherkunde und Glazialgeologie, 39: 43-57.

Quinn, P.F., K. J. Beven, P. Chevallier & O. Planchon (1991): The prediction of hillslope flow paths for distributed hydrological modelling using digital terrain models. Hydrological Processes, 5: 59-79.

R Development Core Team (2008): R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Wilcke, W., H. Valladaraz, R. Stoyan, S. Yasin, C. Valarezo & W. Zech (2003): Soil properties on a chronosequence of landslides in montane rain forest, Ecuador. Catena, 53: 79-95.

Zevenbergen, L. W. & C. R. Thorne (1987): Quantitative analysis of land surface topography. Earth Surface Processes and Landforms, 12: 47-56.