MODELLING SOIL ACIDITY IN SWITZERLAND USING SPATIAL STATISTICS TOOLS

Andri Baltensweiler* & Stephan Zimmermann
Swiss Federal Institute for Forest, Snow and Landscape Research
CH-8903 Birmensdorf, Switzerland
*Corresponding author. E-mail: andri.baltensweiler@wsl.ch

Abstract

Soil pH can be used as a general guide for determining nutrient availability and therefore species that may grow on a given site. Within the framework of the Swiss National Forest Inventory, more than 1'000 soil profile samples have been described across Switzerland and samples have been analyzed regarding acidity.

The main objective of this study was to generate a top soil acidity map for forested areas in Switzerland applying two regression model approaches: the ordinary least squares regression (OLS) and the geographically weighted regression (GWR). Environmental factors of climate, topography and vegetation were used as predictors to build quantitative relationships. Besides these parameters, the OLS regression contained additionally nominal variables which describe mainly geomorphological and pedological entities. Because the GWR approach is able to represent spatial regimes, these nominal variables can be neglected in the GWR model.

The strength of the predictive relationships were moderate for both regression types, with adjusted R^2 values of about 0.44 to 0.47. Considering that topsoil pH values are highly variable, the model performance is satisfying on a nationwide scale.

1 Introduction

Spatial modelling of the distribution of tree species is an active area of research. Information on species and habitat distribution is needed for many purposes, in particular to assess and model the consequences of global change. (Geo)statistical models are developed to relate the geographical distribution of species to ecological parameters. The most common predictive variables are climate and derived topographic variables. Although soil properties are important to determine the occurrence of tree species, they are often neglected because accurate data are lacking.

In order to improve the soil data base, more than 1'000 soil profiles were sampled in forested areas across Switzerland. Besides pH, various chemical soil properties were analysed and morphological soil characteristics such as stone content, soil density, color, soil structure etc. were investigated. The main goal of this study was to generate a top soil acidity map for forested areas in Switzerland by developing a predictive topsoil pH model.

A common approach to build quantitative predictive soil property models is based on the five factors of soil formation as described by Jenny (1941):

$$S = f(c,o,r,p,t)$$

where

S = an individual soil produced as a function (f)
c = climate that has influenced soil development
o = organisms living within or upon soil
r = relief or topography upon which the soil has developed
p = parent material of which the soil has formed
t = time over which the factors have acted upon the soil

This function implies that a (given single) soil property can in principle be calculated by a predictive equation such as a simple linear model. Soil observation points are intersected with continuous spatial datasets of soil-related variables, a model is fitted to predict soil variables at the observation points, and then the model is used to predict the soil variables for the whole area of interest. The success of this approach will depend on a) having sufficient predictor variables observed everywhere, b) having enough soil observations (data points) to fit a relationship and c) having a function $f()$ to fit a good relationship between the soil and its environment (McBratney et al., 2003).
Geographic Information Systems are used to approximate the soil forming factors such as topography attributes.

2 Method

2.1 Study area

The study area was Switzerland which covers 41'000 km^2 in central Europe and ranges from 190 to 4600 m a.s.l. (Fig 1). Approximately 60% of the country is in the Alps, 30% in the Central Plateau and 10% in the Jura Mountains. Switzerland can be subdivided in three main geological units: the Jura Mountains, the Molasse Basin and the Alps. The Jura Mountains consist mainly of Mesozoic cover rocks (mainly calcareous bedrock), uncoupled from the European basement and folded in the late Miocene to early Pliocene times. The Molasse Basin is a thick prism of Oligocene and Miocene detrital sediments (mainly a mixture of calcareous and siliceous rocks). The Alps can be further subdivided in a so called Helvetic belt (mainly Jurassic and Cretaceous calcareous rocks besides Flysch formations), in a Penninic belt (with a great variety of calcareous and siliceous rocks) and in the Austroalpine nappes and the Southern Alps (also with a mixture of siliceous and calcareous rocks). In the inner part of the Alps, cristalline basement rocks are the main substratum for soil formation. The

mean annual temperature ranges from -10.5 to 12.5°C, and annual precipitation from 440 to 3000 mm (Zimmermann & Kineast, 1999).

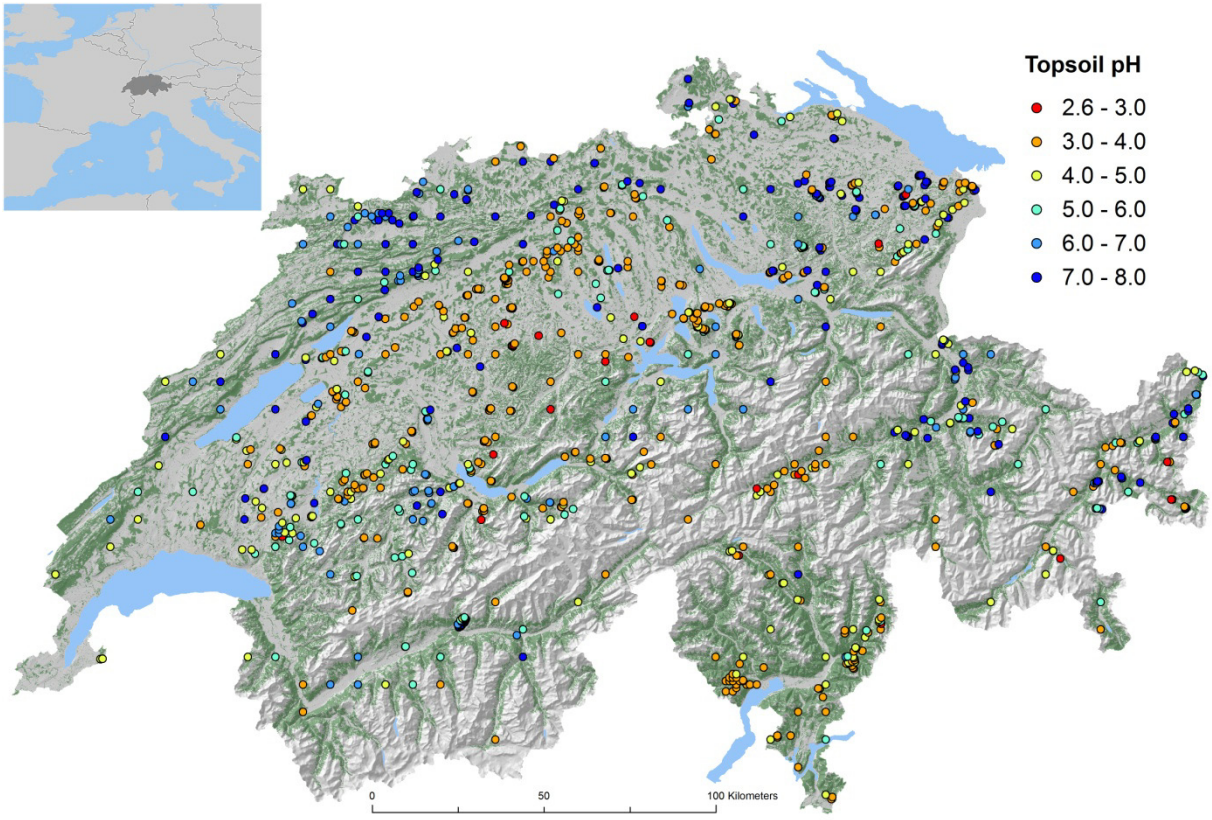

Figure 1. Soil acidity samples within forested areas (green).

2.2 Soil Acidity Data

In the forested area of Switzerland, on a 8 x 8 km grid 172 soil profiles were dug down to the slightly weathered bedrock. In a random sampling, further 865 soil profiles were opened. All 1037 soil profiles were described, sampled by genetic horizons and analyzed regarding pH-value. Soil pH was measured in a suspension of fine earth (dried at 60°C until constant weight, sieved by a 2 mm mesh) in 0.01 M $CaCl_2$ (1:2 solid-to-solution ratio). The "short-term precision" (standard deviation of replicates in the same measurement run, typically n=2) was < 0.2 pH, and the "long-term precision" (standard deviation of an internal reference soil, measured over 400 times between 1996 and 2010) was 0.05 pH.

In order to generate a topsoil acidity map, the mean of all pH-values in a depth of 0 to 20 cm of each soil profile was calculated. The following calculations are all based on these mean values.

2.3 Environmental Predictors

Following Jenny's (1941) soil formation approach, all environmental predictors c,o,r,p,t should be included in the model. Because the factor time (t) is difficult to characterise, it have been neglected for this study.

Climate (c)

All climate variables (Table 1) refer to measurements from the national meteorological network for the period 1961 – 1990 (Zimmermann & Kienast, 1999). The parameters were derived from monthly mean temperature, precipitation and cloudiness. For the interpolation a digital terrain model with a resolution of 25 meters was used.

Actual evapotranspiration was simulated using the spatially distributed hydrological model PREVAH for the period 1981 - 2000 (Gurtz et al., 1999; Zappa, 2008). In PREVAH the Penman-Monteith formula was implemented which accounts for stomatal resistance for various vegetation classes (Monteith, 1965).

The potential global radiation was calculated using the solar radiation analysis equations form ArcGIS Desktop 9.3.

Organisms (o)

The main soil forming or altering organisms are vegetation or humans, although other organisms can have appreciable soil-modifying effect locally (Hole, 1981). In this study, three different predictors were used to describe the forest vegetation (Table 1).

The Normalized Difference Vegetation Index (NDVI) was used to model the vegetation cover and biomass. The NDVI is calculated as a ratio between measured reflectivity in the red and near infrared portions of the electromagnetic spectrum (Tucker & Sellers, 1986). We used a Spot5 Mosaic Scene which was radiometrically and atmospherically corrected.

The second predictor to describe the forest structure was derived from Light Detection and Ranging (LiDAR) data. Based on the LiDAR raw data points a Digital Surface Model (DSM) and a Digital Terrain Model (DTM) were interpolated (Hyyppä et al., 2000). Both datasets have a spatial resolution of 2 meters. The DSM represents all the visible elements of a terrain surface including vegetation, whereas the DTM represents the bare ground without any vegetation. These two datasets were used to obtain the tree heights (Magnussen, 1999; Heurich 2008).

In order to describe the type of the forest, a raster representing the degree of mixture of conifer and deciduous trees was used (Bundesamt für Statistik, 2001). The raster contained 3 classes: coniferous forest, mixed forest and deciduous forest. The raster was derived from 11 Landsat-5 TM scenes.

Relief or topography (r)

Using a DTM with a spatial resolution of 25m, the standard surface attributes such as slope, aspect, curvature, upstream flow length, flow direction, flow accumulation and topographic wetness index (TWI) were calculated (Table 1; Gallant & Wilson, 2000; Zimmermann & Kienast, 2000).

A topoindex was used to identify topographic exposure (ridge, slope, toe slope, etc) at various spatial scales, and to hierarchically integrate these features into a single raster (Zimmermann, 2010).

Parent material (p)

Parent material information was obtained from the Swiss Soil Suitability Map (SSSM) with a map scale of 1 : 200'000. This map shows soil-land units defined on the basis of geomorphological and pedological criteria which were assessed and aggregated in regard to their agricultural and forestry utilization potential. This was done through the appraisal of the pedological characteristics of the map units, dependent on the major soil groups. The map is strongly generalized.

Table 1. Predictors

Predictor	Spatial Resolution [m]
Climate	
Temperature, annual average (°C)	25 x 25
Temperature, July (°C)	25 x 25
Precipitation, year (mm)	25 x 25
Cloudiness (%)	25 x 25
Actual Evapotranspiration (mm)	500 x 500
Global Radiation (WH/m^2)	25 x 25
Organisms	
Normalized Difference Vegetation Index (-)	10 x 10
Vegetation height (m)	2 x 2
Degree of mixture of conifer and deciduous trees (4 classes)	25 x 25
Relief, Topography	
Height a. s. l.(m)	25 x 25
Slope (°)	25 x 25
Aspect (°)	25 x 25
Curvature (1/100 m)	25 x 25
Upstream flowlength (m)	25 x 25
Flow direction (-)	25 x 25
Flow acculmulation (number of cells)	25 x 25
Topographic wetness index (-)	25 x 25
Topoindex (-)	25 x 25
Parent material	
Jura mountains Plains of central plateau Moraines of the hilly country Molasse, partly covered by moraines Valleys of the Central Plateau Molasse, partly alterated by glaciers Landscape with Drumlins Eroded moraines of the hilly country Foothills of the Alps, mainly Molasse Foothills of the Alps, mainly Nagelfluh Valleys of the alps Prealpine scenery with Flysch (Bündnerschiefer) Bündnerschiefer in the upper Rhone valley and in Ticino Alpine scenery with limestone Alpine scenery with cristalline bedrock Valley scenery (southern part of the alps)	All variables originate from the Swiss Soil Suitability Map, scale 1 : 200'000

2.4 Statistical Methods

In soil science many different forms of prediction functions have been used to find relationships between the soil and its environment. Examples include multiple linear regression (MLR), generalized linear models (GLMs), generalized additive models (GAMs) and regression classification trees (McBratney et al., 2003). Geographically Weighted Regression (GWR) is a recent approach and is relatively rarely applied in physical geography (Miller et al., 2007).

In this study we used the Ordinary Least Squares (OLS) Regression and the Geographically Weighted Regression (GWR). Both regression methods are implemented in ArcGIS Desktop 9.3.

OLS has been used widely in prediction of soil attributes because of the easiness and wide availability. The predictors are usually continuous variables. However, qualitative factors or nominal variables can also be integrated. OLS provides a global model of the variable or process and creates a single regression equation to represent that process. More information of the OLS regression can be found e.g. in Hastie et al. (2001) or in the ESRI ArcGIS Desktop Help.

GWR provides a local model of the variable or process by fitting a regression equation to every feature in the dataset. It constructs these separate equations by incorporating the dependent and explanatory variables of features falling within the bandwidth of each target feature. The shape and size of the bandwidth is dependent on user input for the Kernel Type, Bandwidth Method, Distance, and Number of Features (Fotheringham et al., 2002; ESRI ArcGIS Desktop Help, 2009).

3 Results and Discussion

3.1 Ordinary Least Square Regression

To analyze the relationship between the topsoil pH and the predictor variables and to find multicollinearity among the predictors, scatterplot and correlation matrices were used. In general, the correlation coefficients between pH and the predictors were low. The highest correlation coefficient was found between pH and slope (0.22). In order to identify the most relevant predictor variables a stepwise variable selection was performed using the statistics software R ver. 2.7.0 (R Development Core Team, 2009). The Akaike Information Criterion (AIC) determined the stopping point (i.e., number of variables included):

$$AIC = n \log\left(\frac{RSS}{n}\right) + 2p$$

where *n* is the number of observations, RSS is the model residual sum of squares, and *p* is the number of parameters. The minimum of the AIC is commonly used, as in this case, to identify a parsimonious model that has both low error and few parameters (Hastie et al., 2001).

An (ideal) requirement for linear regression is that the dependent variable is normal distributed (Draper & Smith, 1998). In many soil studies, however, the variables show skewed non-normal distributions, which is reflected in the residuals. To account for the normality requirements the soil acidity data were log-transformed prior to the regression.

The best OLS model found regarding AIC (-78.7) and adjusted R^2 (0.44) consists of 11 continuous variables (height a.s.l., slope, topoindex, wetness index, NDVI, vegetation height, yearly precipitation, cloudiness, actual evapotranspiration, global radiation, x coordinate of the sample location) and 2 nominal variables. The nominal variables include 17 dummy variables for parent material and 1 dummy variable for presence of conifers trees.

The coefficients confirmed the expected signs of the relationship between dependent and independent variables.

The OLS Tool in ArcGIS automatically tests for heteroskedasticity (inconsistence of residual variance) and non-stationarity (regional variation of independent variable). The Koenker's studentized Bruesch-Pagan test indicated that our model violated the homoskedasticity assumption and it revealed non-stationarity. ArcGIS computes standard errors that are robust in regard to these problems. We therefore consulted the robust probabilities to determine the significance of the explanatory variables. Redundant variables identified by the variance inflation factor were removed. The residuals were normally distributed. Finally, the OLS model was controlled for spatial autocorrelation of the regression residuals. The Moran's I statistic (Mitchell, 2005) showed that the residuals were highly clustered (Moran's Index = 0.25, p = 0.0, Z = 5.47). To map clusters of over and under predictions, we applied the Getis-Ord Gi* Hotspot Analysis with a default bandwidth of 22 kilometers. Most of the clusters, independently whether cold- or hotspots, were found in the Molasse Basin (Figure 2). As described in section 3.1, the Molasse Basin consists of either calcareous or siliceous sediments which impact the soil pH in opposite ways. Because the Swiss Soil Suitability Map is strongly generalized and does not indicate the genesis of the Molasse, the corresponding sample points cannot be differentiated according to their parent material. Therefore the spatial regime of the geology is insufficiently represented in the Molasse Basin and leads to clusters of over and under prediction.

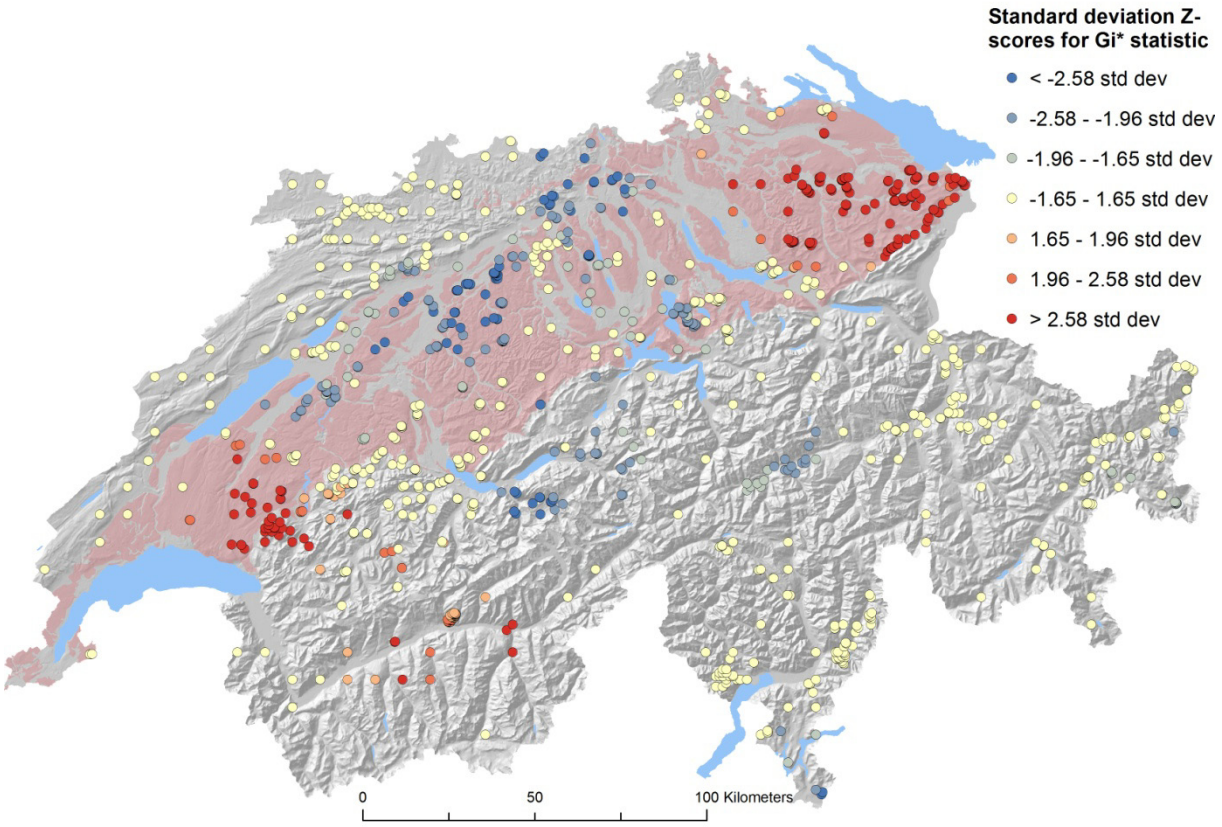

Figure 2. Z-Scores of the OLS regression residuals. The blue points indicate over prediction , the red points under prediction. The area in red are Molasse sediments.

3.2 Geographically Weighted Regression

One strategy to deal with non-stationarity is to incorporate regional variation into the regression model such as the GWR (ESRI ArcGIS Desktop Help, 2009). The OLS regression model served as a starting point to build the GWR regression. The categorical variables describing parent material were excluded from the model. These dummy variables represent spatial regimes and because explanatory variable coefficients can vary in GWR, these spatial effects can be represented by the model itself.

Although all explanatory variables from the OLS model had a variance inflation factor less than 3.5, global or local multicollinearity prevented GWR from solving the equations. Thus we omitted variables which tend to have multicollinearity like temperature and height a.s.l. or variables which only have a very smooth variation in value like the interpolated climate variables. The best GWR model regarding AIC (-35.1) and adjusted R^2 (0.47) contained four continuous variables (height a.s.l., slope, topoindex, vegetation height) and one nominal variable which described the presence or absence of conifer trees, respectively. Because the spatial distribution of the soil sample points varied within the study area, we specified an adaptive Gaussian kernel type where the spatial context is a function of specified number of neighbors. The selection of the bandwidth, which controls the size of the kernel, was determined using the Akaike Information Criterion (AIC). Minimizing the AIC provides a trade-off between the goodness-of-fit and degrees of freedom (Fotheringham et al., 2002).

The GWR output table reports the most important diagnostic values (Table 2). 94 neighbors were used in the estimation of each set of coefficient. This means that each local estimation was based on about 9% of the data.

Table 2. GWR diagnostic values

Name	Value
Neighbours	94
Residual Squares	41.64
Effective Number	194.32
Sigma	0.2223
AICc	-35.10
R^2	0.57
R^2 Adjusted	0.47

Applying Moran's I statistics (Mitchell, 2005) showed that there is no significant clustering of the residuals (Moran's Index = 0.04, p = 0.32, Z = 1.0). The Hotspot Analysis using the Getis-Ord Gi* statistic confirmed that there was no significant over or under prediction (Figure 3).

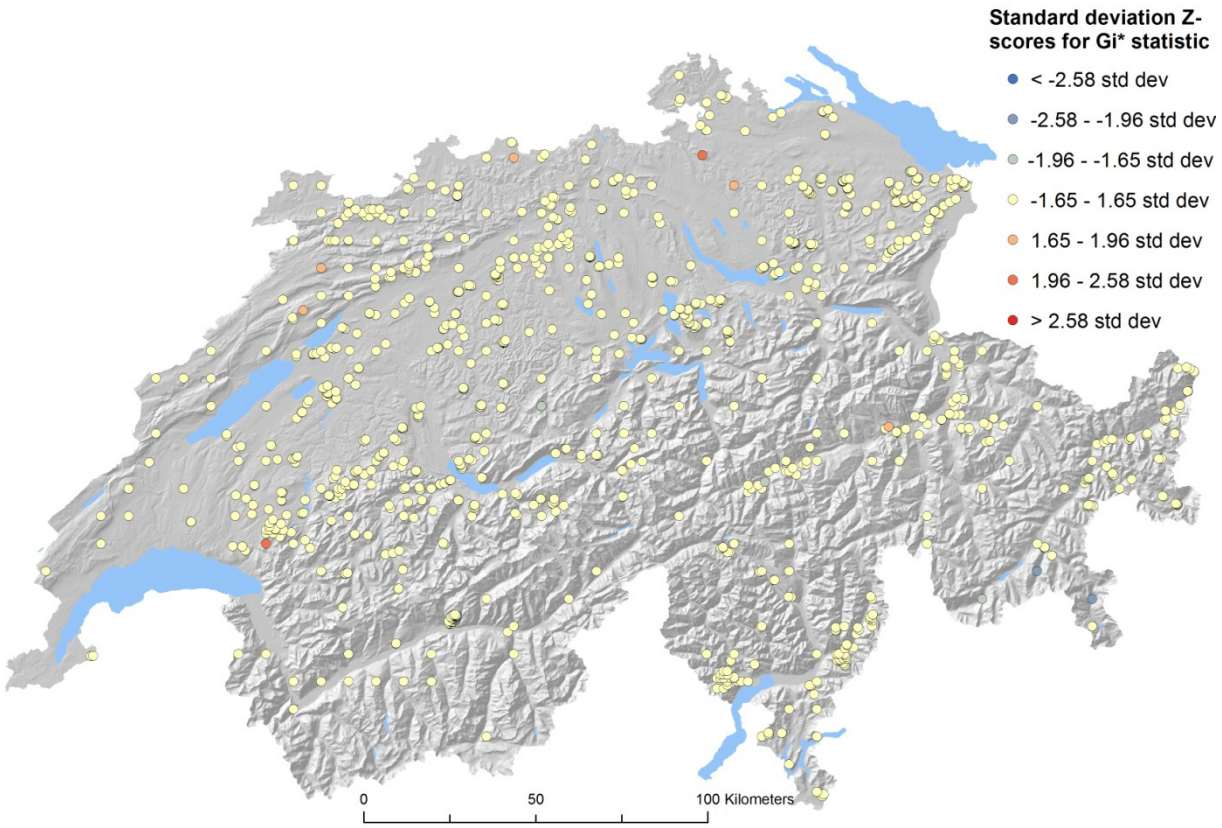

Figure 3. Z-Scores of the GWR regression residuals.

3.3 Model Comparison

In order to evaluate the model performances, we calculated diagnostic indicators (Table 3). It is obvious that the GWR model performed better than the reduced OLS, which both contained identical variables.

Table 3. Regression diagnostic values

Model	AIC	R^2	Adjusted R^2	Residual sum of square	Root mean square error
OLS full	-78.7	0.45	0.44	53.1	0.2263 (1.25*)
OLS reduced	344.1	0.14	0.13	83.6	0.2848 (1.33*)
GWR	-35.1	0.57	0.47	41.6	0.2000 (1.22*)

*in pH units

The AIC values indicated that the full OLS model, which contained seven additional variables (cf chapter 4.1), performed better than the GWR model. All other indicators showed the contrary, but were expected to do so given the difference in degree of freedom between the two models. However, the OLS full regression residuals were positively autocorrelated and therefore violated OLS assumptions. Autocorrelation of the residuals indicates that the standard errors are underestimated and the correlation coefficient often indicates a significant relationship between variables when in fact there is none (Clifford et al., 1989).

4 Conclusions

Although Geographically Weighted Regression was primarily developed for econometric analyses, it is increasingly applied in physical geography. In the present study we used this approach to model soil acidity on a nationwide scale. Soil pH is a key factor for plant species distribution and is therefore indispensable for vegetation modelling.

Parent material, which is one of the soil formation factors, is represented in this study by the Swiss Soil Suitability Map (SSSM). These nominal variables had a major influence on the OLS model. Without these variables, the OLS regression modeled only a weak relationship between the dependent and independent variables whereas the full OLS regression performed better. However, autocorrelation of the residuals indicated that important information was still missing. As shown by the Getis-Ord Gi* Hot Spot Analysis most of the clusters of over/under predictions were within the Molasse Basin. The nominal variables which represented the Molasse sediments are too generalized and do not reflect the spatial regime of the parent material adequately.

The GWR model in this study did not violate an underlying assumption and therefore is more reliable than the OLS model. An adjusted R^2 value of 0.47 is satisfying considering that pH of topsoils are highly variable. GWR models are able to represent local regimes because the explanatory variable coefficients can vary in space. Therefore spatial regime data like the SSSM are not required in GWR models. This can be a great benefit in soil modeling because nationwide parent material data are often not available or are not detailed enough.

The results of this study are acceptable but have to be cross-validated to ensure the model performance. Furthermore we plan to test different model approaches like generalized linear models (GLMs) or generalized additive models (GAMs). These models are more appropriate regarding non-normal response distribution and non-linear relationships between dependent and independent variables (Bishop & Minasny, 2005).

ArcGIS Desktop 9.3.1 contains very useful and important statistical tools for analyzing spatial distributions, patterns, processes, and relationships. Some of the analyses, as e.g. the stepwise variable selection, were conducted in the R software (The R Project for Statistical Computing, http://www.r-project.org). R provides a wide variety of statistical and graphical techniques. However, all the data calculations for the independent variables have to be primarily performed in GIS software as in ArcGIS for example. Therefore a data transfer between the two software packages is necessary which can be cumbersome sometimes. With the use of Python as a conduit, the R functionality can be integrated in the ArcGIS environment. On the ESRI resource center, there is an example of how to use R in ArcGIS Desktop.

Literature:

Bishop, T.F.A. & Minasny B. (2005) Digital soil-terrain modeling: the predictive potential and uncertainty. In: Grunwald, S. (Ed.). Environmental Soil-Landscape Modeling. Taylor & Francis, New York.

Bundesamt für Statistik (2001) Waldmischungsgrad der Schweiz. http://www.bfs.admin.ch (accessed May 2010).

Clifford, P., Richardson, S. & Hemon, D. (1989) Assessing the significance the correlation between two spatial processes. Biometrics, 45: 123-134.

Draper, N. & Smith, H. (1998) Applied regression analysis, 3rd ed. Wiley, New York.

Fotheringham, A.S., Brunsdon, C. & Charlton, M. (2002) Geographically Weighted Regression: The analysis of spatially varying relationships. Wiley & Sons, Chichester.

Gallant, J.C. & Wilson J.P. (2000) Primary Topographic Attributes. Terrain Analysis: Principles and Applications. Wilson J.P, Gallant, J C. New York, John Wiley and Sons: 51-85.

Gurtz, J., Baltensweiler, A. & Lang, H. (1999) Spatially distributed hydrotope-based modelling of evapotranspiration and runoff in mountainous basins. Hydrological Processes, 13, 2751–2768.

Hastie, T., Tibshirani, R. & Friedman, J. (2001) The elements of statistical learning: Data mining, inference, and prediction. Springer, New York.

Heurich, M. & Thom, F. (2008) Estimation of forestry stand parameters using laser scanning data in temperate, structurally rich natural European beech (*Fagus sylvatica*) and Norway spruce (*Picea abies*) forests. Forestry, Vol. 81, No. 5.

Hole, F.D. (1981) Effects of animals on soil. Geoderma 25, 75-112.

Hyyppä, J., Pyysalo, U., Hyyppä, H. & Samberg, A. (2000) Elevation accuracy of laser scanning-derived digital terrain and target models in forest environment. 20th EARSEL Symposium and Workshops Dresden, Germany, 14–17 June, 2000 9 pp. Available at: http://las.physik.uni_oldenburg.de/projekte/earsel/4th_workshop.html#proceedings .

Jenny, H. (1941) Factors of soil formation, a system of quantitative pedology. McGraw-Hill, New York.

Magnussen, S., Eggermont, P., LaRiccia, V.N. (1999) Recovering Tree Heights from Airborne Laser Scanner Data. Forest Science 45 (3).

Mc Bratney, A.B., Mendonca Santos M.L., & Minasny B. (2003) On digital soil mapping. Geoderma 117, 3-52.

Miller, J., Franklin J. & Aspinall R. (2007) Incorporating spatial dependence in predictive vegetation models Ecological Modelling, Volume 202, Issues 3-4, 10 April 2007, Pages 225-242.

Mitchell, A. (2005) The ESRI guide to GIS analysis. Volume 2: Spatial Measurements & Statistics. ESRI Press Redlands (CA).

R Development Core Team (2009) The R Project for Statistical Computing, Vienna, Austria: http://www.r-project.org.

Tucker C. J. & Sellers P. J. (1986) Satellite remote sensing of primary vegetation. International Journal of Remote Sensing, 7: 1395-1416.

Zappa, M (2008) Objective quantitative spatial verification of distributed snow cover simulations—an experiment for the whole of Switzerland. Hydrological Sciences, 53(1).

Zimmermann, N. E. & Kienast, F. (1999) Predictive mapping of alpine grasslands in Switzerland: species versus community approach. Journal of Vegetation Science, 10, 469-482.

Zimmermann, N. E. (2010) Topographic Position Mapping Routines. http://www.wsl.ch/staff/niklaus.zimmermann/programs/aml.html#4 (accessed May 2010).