CONTINENTAL-SCALE DIGITAL SOIL MAPPING USING EUROPEAN SOIL PROFILE DATA: SOIL pH

Hannes Isaak Reuter 1, Luis Rodriguez Lado 1, Tomislav Hengl 2 & Luca Montanarella 1

1Institute for Environment and Sustainability – Land Management and Natural Hazards Unit – European Commission – Joint Research Centre (JRC), TP 280, Via Fermi 2749, I-21027 Ispra (Italy)
2University of Amsterdam – Institut for Biodiversity and Ecosystem Dynamic, Nieuwe Achtergracht 166, 1018 WV Amsterdam (Netherlands)

Abstract: We used a compilation of 12,333 soil pH measurements from 11 different sources to create a quantitative map of estimated soil pH values across Europe using a geostatistical framework based on regression-kriging. We made use of 54 auxiliary variables in the form of raster maps at 1km resolution to explain the differences in the distribution of soil pH CaCl₂ and we added the kriged map of the residuals from the regression model. The goodness of fit of the regression model was satisfactory ($R^2_{adj} = 0.43$) and its residuals follow a Gaussian distribution. The map results agree with our previous knowledge about this soil property. The lowest values correspond to the soils developed on acid rock (granites, quartzite’s, sandstones, etc), while the higher values are related to the presence of calcareous sediments and basic rocks. The validation of the model also shows that our models is quite accurate ($R^2_{adj} = 0.56$). This shows the validity of regression-kriging in the estimation of the distribution of soil properties when a large and adequately documented number of soil measurements are available.

1 INTRODUCTION

Soil acidity influences chemical, physical and biological soil properties as well as plant growth and the quality of fresh water in direct and indirect ways (SCHETTFER et al. 1992). The acidification of freshwater and soils constitutes an important environmental concern in Europe and North America since the first symptoms of soil water and degradation due to the deposition of atmospheric acidifying compounds were observed in the 1950’s (ODEN 1968, ALCAMO et al. 1990) on lakes and forest ecosystems from northern areas in Europe. Although policies have already been adopted across Europe to reduce the emission level of these polluters into the atmosphere, the acidification process is still one of the main threats for European soils as recently debated around the “Thematic Strategy for Soil Protection“ in Europe. In addition to the atmospheric deposition of acidifying compounds, agricultural practices such as the use slurries and ammonium based fertilizers can contribute significantly to the acidification of soils (RASMUSSEN & ROHDE 1989). On the other hand, liming produces the opposite effect and corrects temporally the environmental negative effects of decreases in soil pH (INSAM & PALOJARVI 1995). It is clear that the effects of the acidification process do not only depend on intensity of the input of acidifying compounds but also on the sensitivity of soils to this degradation process. The same amount of acid inputs on a Distric Regosol in the north of Europe will produce more adverse consequences that the same acid amount on a calcic soil in the Mediterranean area. Thus, in order to determine the impact of an acidifying process we should previously determine the sensitivity of each type of soil to this threat.

According to ULRICH (1981, 1983) soils have a fixed sequence of buffering mechanisms in order to avoid pH reduction. As the acidification process progresses, these mechanisms enter into force sequentially, the buffering capacity of soils decreases progressively and only an abrupt pH descent is observed when each buffering mechanism is completely exhausted. In this sense, the value of the soil pH is an indicator of the buffer capacity of the soil (resistance of the soil to the acidification) and it can be used determine the risk in relation to this threat.
There is no consistent or harmonized soil profile database which would allow for the generation of an overview of pH values across Europe using digital soil mapping techniques. We therefore made an attempt to combine, harmonize and QA/QC check values from 11 available databases in order to create a quantitative map of estimated soil pH values across Europe. We are aware that these databases reflect biased sampling strategies, non precise coordinates, different sampling times (e.g. especially sensitive on agricultural soils) and many other factors which influence the predicted result. However, these databases represent a snapshot of available data on soil pH currently available in ESDAC.

This paper describes the major distribution of soil pH with a reported quantitative uncertainty in order to determine the extent of the risk of acidification across Europe.

2 MATERIAL AND METHODS

Soil Database creation: Eleven soil profile databases which are available in the European Soil Data Center (ESDAC) have been used. ESDAC is the single focal point for soil information as one of the 10 data centers in Europe for Environmental Data (EEA, EUROSTAT and JRC). The main task is to work together with national and regional soil data centers to provide timely, quality controlled added value products for policy and research needs. The top layer of each soil profile stored in the following datasets has been obtained: ICP-Forest (FSS), soil profiles from the ecopedological map of Italy, FOREGS (SALMINEN et al. 2007), Spade (HIEDERER et al. 2006), Soveur (NACHTERGAELLE et al. 2002), WISE (BATJES 2002), Galicia (RODRIGUEZ LADO 2008), Danube-SIS (Bavaria, Slovenia and Bulgaria) and Puglia. Data were projected to the INSPIRE complaint Lambert Equal Area (LAEA) projection, if not already available in that projection. Secondly, the measurement of pH in the different databases were performed in different solutions (H2O, KCl, CaCl2). For each database we adjusted if necessary to report the pH in CaCl2 solution. The following linear pedotransfer functions were generated based on values reported for topsoil sampling locations from the WISE global soil database:

\[\text{pH}_{\text{CaCl2}} = 0.9761 \times \text{pH}_{\text{H2O}} - 0.427 \quad (R^2 = 0.92, n=1997)\]

\[\text{pH}_{\text{CaCl2}} = 1.0572 \times \text{pH}_{\text{KCl}} + 0.123 \quad (R^2 = 0.90, n =377)\]

We tested the validity of the global dataset derived pedotransfer function (pH_{KCl}=0.87 pH_{H2O}, R^2 =0.93) by comparing it against the fit obtained from pH data measured in KCl and H2O from the Galicia DB (pH_{KCl}=0.84 pH_{H2O}, R^2 =0.75, n = 414) and observed a good agreement.

Duplicate detection: We define a duplicate soil sampling location as a point which is within 1,000 m of another point. Some of the datasets used in the analysis have insufficient precision in their geo-location, allowing for false positives. We identified duplicates, and clustered these points with a minimum distance of 1,000 m to obtain single independent regions. From the total dataset of 12,333 sampling points, 2,093 were within 1,000 m of a neighboring point, resulting in 847 clusters.

Clustering of the duplicates occurred mainly in regions where denser sampling occurred (e.g. Italy had 571 duplicates out of 5,114 records) however even random sampling designs such as the 813 points extracted from the FOREGS DB contained two sampling locations within 1,000 m of neighboring records.

The average standard deviation of adjusted pH_{CaCl2} of all 847 clusters was 0.39 with a maximum standard deviation of 2.54, whereas the average difference between the minimum/maximum pH_{CaCl2} at each cluster location was 0.6 with a maximum difference of 4.1 reached at sampling location in Galicia. Minimum was in both cases 0.0. Therefore we can specify the measurement/ scale error for our analysis of ~0.5 unit pH_{CaCl2}. 

Hamburger Beiträge zur Physischen Geographie und Landschaftsökologie – Heft 19 / 2008
We were interested to see if the chosen resolution was appropriate with respect to the pH\textsubscript{CaCl\textsubscript{2}} point data. Coarsest, finest and recommended cell sizes were determined following Hengl (2006) and were ~200 m, ~1 km and ~4.7 km respectively. The 0.5 probability to meet the next point was reached at ~3 km (0.95 at 30 km), for 2 points at ~5 km (0.95 at 43 km), for all points at 1,000 km (0.95 probability at 2,700 km).
Geostatistical mapping: We have used regression-kriging to estimate the pH\textsubscript{CaCl\textsubscript{2}} values in soils of Europe. For a more in-depth description of the data preparation/treatment please refer to RODRIGUEZ et al. (2008), the general process of regression kriging is described in detail in HENGL (2007).

Firstly we create a linear regression model for the measured pH values against a number of auxiliary environmental variables and then we interpolate the residuals of this regression model by ordinary kriging. The final map is an additive combination of both models. This technique allows to take into account the boundaries of some environmental features that can highly influence the distribution of the studied soil property in the final map of estimates, and thus to obtain more realistic predictions.

The original dataset of observations (12,333 records) was divided using a random sampling function into a “model” dataset, that includes the 80% of the samples and a “validation” dataset, containing the remaining 20% of the samples, that was using for validation of the model. We used the “model” points dataset to build the multiple linear regression model. The derived regression equation has than be applied on the standardized 1 km resolution 56 auxiliary raster grids. The results have been aggregated to 5 km resolution. The auxiliary variables are either directly influencing soil pH or serve as a proxy for a factor:

- **Topography:** The 1 km DEM was derived from the SRTM30 V2 dataset obtained from the Jet Propulsion Laboratory. The SRTM DEM was used to derive a slope map, the Topographic Wetness Index (MOORE et al. 1991) and total incoming solar insolation (CONRAD 2001) using SAGA 2.0.

- **Geology:** We used the digital map of main geologic surface units in Europe (PAWLEWICZ et al. 2003). The original legend was reduced to 10 classes according to the genetic nature of each unit: (1) Granites, rhyolites and quartzites; (2) Paleozoic schists, phyllites, gneisses and andesites; (3) Shales and sandstones; (4) Mesozoic Ultramafic, basic phyllites, schists, limestones and evaporates; (5) Jurassic, Triassic and Cretaceous calcareous rocks; (6) Cenozoic serpentinites, gabros and sand deposits; (7) Tertiary basanites and andesites; (8) Neogene and Paleogene calcareous rocks; (9) Quaternary limestones and basaltic rocks; and (10) Other Ultramafic and undefined rocks.

- **EVI remote sensing images:** Monthly averaged MODIS images of the Enhanced Vegetation Index EVI at 1 km resolution for the period 01/01/2004 to 31/12/2006 were obtained from the MODIS Terra imagery at the Earth Observing System Data Gateway. We performed a Principal Component Analysis on the 19 complete mosaics and used the first five resulting components.

- **Image of Lights at Night:** The lights at night image for the year 2003 was obtained from the Defense Meteorological Satellite Program (http://www.ngdc.noaa.gov/dmsp/), which measures night-time light emanating from the earth’s surface at 1 km resolution. This map is a proxy for urbanisation and is now increasingly used for quantitative estimation of global socioeconomic parameters as well as for human population mapping (SUTTON 1997, SUTTON et al. 1997, DOLL et al. 2000).

- **Distance to infrastructure:** The map of distances to roads, airports and utility lines was calculated using the distance operation in ILWIS and the GIS layers from the GISCO database of the European Commission (http://eusoils.jrc.it/gisco_dbm/dbm/home.htm).

- **Cumulative Earthquake’s magnitude:** The cumulative earthquake’s magnitude map was calculated by using the 90,000 registered earthquakes in period 1973-1994. These measurements were recorded in the Global Seismology point database (http://earthquake.usgs.gov/eqcenter/). We used the logarithmic measure of the “size” of an earthquake and then rasterized the point map to the 10 km grid by using the point density operation in ILWIS. This operation sums all earthquake magnitudes observed within a 10 km grid and gives a cumulative map of earthquake activity.

- **Land use:** We used the Corine Land Cover 2000 map of Europe generalized to a 1 km grid. For Switzerland, we used the Corine Land Cover from 1990 since no updated information was available. The CLC1990 classes for this country were adjusted to those described in the CLC2000 and both datasets were merged together and aggregated to 1 km resolution. The original 44 classes were simplified to 8 classes: (1) urban infrastructures; (2) agriculture; (3) forest; (4) natural vegetation; (5) beaches; (6) ice bodies, (7) wetlands and (8) water bodies. Class 5 (beaches) disappeared in the process of upscaling from 100m to 1km. Additionally data from the Global Land Cover classification (FRITZ et al. 2003) has been used for areas where no Corine classification has been available.
• **Land forms:** The land forms were calculated by a modified method proposed by IWAHASHI & PIKE (2007). Basically it combines the topographic variables slope gradient, surface texture and local convexity to create 16 classes of landforms from steep to gentle landforms with fine/coarse texture and low/high convexity.

• **Climatic variables:** Mean annual temperature and accumulated precipitation maps were obtained from the very high resolution raster layers created by HIJMANS et al. (2005) on a global scale at 1 km grid resolution. The annual potential evapotranspiration (PET) was calculated from monthly temperature data using the THORNTHWAITE method (1948). Runoff was calculated as the difference between annual accumulated precipitation and annual potential evapotranspiration.

• **Alkalinity release rates due to the weathering of primary minerals in soils:** The release of alkalinity by weathering of primary minerals in soils was calculated using the “Simple Mass Balance Method” as described in RODRIGUEZ-LADO et al. (2007).

• **Atmospheric deposition of contaminants:** We also used the estimated annual deposition and emission rates of cadmium, lead and mercury in Europe for year 2004 calculated within the European Monitoring Evaluation Programme (http://webdab.emep.int/). The original 50 km grids were downscaled to 1 km grids using ordinary kriging.

We further masked out non-soil surfaces such as water bodies (rivers, lakes, sea etc.) and permafrost areas. A consistent European wide water mask, indicating the percentage of water area inside a 1 km pixel, has been created based on the NASA SRTM V2 SWDBB dataset (RABUS et al. 2003), the CORINE land use classification, lakes contained in the GISCO data, base water reflection by the use of Image2000 dataset (DE JAGER et al. 2006), the GSHHS - Database (WESSEL & SMITH 1996), and the Global Lakes and Wetlands Database (LEHNER & DÖLL 2004).

Areas with permanent ice cover have been detected using the mean annual EVI derived from the MODIS 1 km images obtained for the years 2003 and 2004. In this case, ice cover, water bodies and bare rock areas were detected based on a negative VI index. The total soil-cover area for these 26 countries was estimated to be 4,217,241 km².

Each class within the categorical variables (geology, land use and land forms) was transformed to binary raster layers. Later all the auxiliary raster layers were standardized and finally converted to 54 Principal Components raster maps by Principal Component Analysis in ENVI v4.3 in order to minimize collinearity between variables. The kriging of the residuals from the linear model was done directly in 5 km blocks. In addition we also obtained a measurement of the estimation errors associated to the kriging interpolation method.

Filtering the database of observations and building the regression model were performed in the statistical environment ‘R v.2.6.0″. The Principal Component Analysis of the standardized auxiliary variables was done using the image processing software ENVI. The raster linear model and the final regression kriging map were done in SAGA-GIS 2.0.

### 3 Results and Discussion

**Descriptive statistics:** The database used in this study is a compilation of samples collected in 11 different surveys (Fig. 2). There is a high variability of pH values across Europe, ranging from 0.09 to 1340. About 51% of the samples present pH ≤ 10 while 41% present values higher than 50 ppb.

We observe that the big survey campaigns across Europe (Soveur, Spade and FOREGS) present similar boxplot diagrams. However, the differences in pH are very evident when comparing regional campaigns on acidic soils like “Galicia” (parent materials mainly of granitic nature) and on calcareous soils like that in “Puglia” (mainly limestones).

The higher pH measurements (pH =9) correspond to soils in Puglia (Italy) and high pH values (pH > 8) were also observed mainly in Bulgaria. Very low pH values (pH < 3) were observed in Bavaria (Germany), in the sandy soils from the “Landes” department in France, and in Belgium and UK.
Geostatistics of the original data: The semi-variogram of the original data shows a moderate spatial dependency with a range of around 50-100 km (half of the total variance) and a slight linear trend up to 500 km (Fig. 3). Further analysis showed that this trend is due to a slight anisotropy in the east direction, whereas the north direction shows a periodic pattern. The semivariance shows two additional nested structures, which can be observed with a range of 600 km (reaching variance 1.6) and 800 km (reaching the total variance ~1.85). These two additional structures are due to anisotropy in the north-south direction, whereas the west-south direction shows only a prolonged trend as described before. Automated approaches fitting a semi-variogram based on the range and total semivariance exist (see Hengl 2007). However, as we discovered these would have missed the 50-100 km scale variation structures.
Linear Model results: The regression model obtained explains 43% of the variability ($R^2_{adj}=0.433$) and was significant at the $p<0.05$ level. Fifty one principal components contributed significantly ($p<0.05$) to the model (Tab. 1).

The residuals of the regression model showed spatial structure and they were incorporated to the final model by ordinary-kriging using the software ISATIS V8.1 (GEOVARIANCES 2008). A spherical variogram was fitted to the observed semivariances with a nugget of 0.55, a range of 20,000 and a sill of 1.22. The nugget/sill ratio (0.45) indicates that the residuals have a good spatial dependence.

Tab. 1: Regression coefficients for the multiple linear model.

<table>
<thead>
<tr>
<th>VARIABLES</th>
<th>Estimate</th>
<th>VARIABLES</th>
<th>Estimate</th>
<th>VARIABLES</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>1.20E+03</td>
<td>[PCA18]</td>
<td>4.83E+01</td>
<td>[PCA36]</td>
<td>-7.49E-03</td>
</tr>
<tr>
<td>[PCA1]</td>
<td>2.28E-02</td>
<td>[PCA19]</td>
<td>1.26E+00</td>
<td>[PCA37]</td>
<td>8.70E-02</td>
</tr>
<tr>
<td>[PCA2]</td>
<td>2.69E-04</td>
<td>[PCA20]</td>
<td>-1.52E-01</td>
<td>[PCA38]</td>
<td>-1.16E-01</td>
</tr>
<tr>
<td>[PCA4]</td>
<td>-2.82E-01</td>
<td>[PCA22]</td>
<td>-5.82E-01</td>
<td>[PCA40]</td>
<td>2.88E-02</td>
</tr>
<tr>
<td>[PCA8]</td>
<td>-8.06E-01</td>
<td>[PCA26]</td>
<td>2.06E-02</td>
<td>[PCA45]</td>
<td>-1.14E-01</td>
</tr>
<tr>
<td>[PCA11]</td>
<td>8.02E-02</td>
<td>[PCA29]</td>
<td>2.38E-01</td>
<td>[PCA48]</td>
<td>5.07E-02</td>
</tr>
<tr>
<td>[PCA12]</td>
<td>2.43E-01</td>
<td>[PCA30]</td>
<td>8.85E-03</td>
<td>[PCA50]</td>
<td>1.65E-02</td>
</tr>
<tr>
<td>[PCA16]</td>
<td>4.30E-01</td>
<td>[PCA34]</td>
<td>-1.39E-01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[PCA17]</td>
<td>-8.86E-01</td>
<td>[PCA35]</td>
<td>2.49E-01</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

As suspected, the following map (Fig. 4) shows that the spatial distribution of soil pH is highly dependent on the nature of the parent material. We can observe low pH values in the granitic areas all over the Hesperic massif (Portugal and north of Spain), in the Vosges mountains, in the Pyrenees, and in the shallow soils from Scandinavia, mainly developed on acid materials. The higher pH values ($pH > 7$) are mainly present in the sedimentary areas of the Mediterranean countries (Spain south of France, Italy, Albania and Greece) because of the calcareous nature of the parent material. We also observe differences due to land use patterns and large scale climatic differences (e.g. the Mediterranean area versus Scandinavia). These results might be useful in guiding further research - for example – as to why specific stream water catchments in the Czech republic showed differences in pH time series in periods of increasing acidic deposition (VESELY et al. 2002).

According to our computations 16.7% of the territory has pH values lower than 4.2 and only 1.9 % of the area present values of pH > 8. The higher pH estimates are located throughout all European countries, mainly located close to major cities, or in arid areas with intensive agriculture area (southeast of Spain) or Deltas. The distribution of pH values in relation to land use classes (Fig. 5) showed that forest soils have lower pH values, while agricultural and urban areas present the higher mean values probably due to liming and the influence of the dissolution of cement in buildings.
Fig. 4: Estimated values of pH_{CaCl2} for the EU27 MS and some adjacent countries.
Validation: The accuracy of the model was assessed by comparing the pH measurements and their estimates in the 2,362 samples from the validation dataset. We performed a linear regression between both variables to check their relationship and we obtained a significant correlation between them ($R^2_{adj}=0.56$, $\alpha=0.05$; Fig. 6). The equation of the regression model writes:

$$pH_{measured} = 1.34 + 0.779 * pH_{predicted}$$

Most of the samples fall within the limits of the 95% prediction bands. This relationship demonstrates that in general, our model slightly underestimates the real pH measurements. The Root Mean Squared Error of the predictions is low (RMSE=0.9). A Student's t-test for paired samples ($p < 0.05$) shows that there are not significant differences between measurements and the predictions. Our model does show a good fit in the prediction of soil pH values.

**Fig. 5:** Estimated pH$_{CaCl_2}$ boxplots for different major land uses

**Fig. 6:** Plot of measured vs. estimated pH in the validation dataset.
4 CONCLUSION

The study revealed that it was possible in a limited time and data frame to predict soil pH$_{Cl2}$. This might serve as a first proxy to the sensitivity of soils to the process of acidification. Our linear model was satisfactory and the residuals showed clear auto-correlation structure. If new data are available to be included in the DSM process, the procedure can be recomputed and uncertainties reduced.

Any further assessment of sensitivity of soils to acidification has to take into account not only the pH but the soil organic matter content, soil texture and Al and Fe contents (KOPTSIK & ALEWELL 2007) as well as the sulphate sorption capacity of soils. One possible approach would be to generate a buffering capacity map using Digital soil mapping techniques, therefore providing an additional data set to the already existing critical load acidity maps (HETTELINGH et al. 1991, 1993, 1995). However, the author’s first impressions are that such DSM approaches would need a dedicated sampling campaign as well as standardized measurements for these kinds of parameters that are generally missing in the databases used in the study.

A further geostatistical exercise would include simulations to predict probability density functions of the pH$_{Cl2}$ distribution. A disadvantage of our study is the lack of a temporal component. Most of the observations in the generated database contain no information on the time or timeframe of the sampling. If such information were available, maps for different times/time steps could be created, allowing results to be compared with changes in time as provided by e.g. de SCHRIJVER et al. (2006) or FÖLSTER et al. (2003).

REFERENCES


