Updating the Dutch soil map using soil legacy data: a multinomial logistic regression approach

B. Kempen[1], G.B.M. Heuvelink[2], D.J. Brus[3], and J.J. Stoorvogel[4]

[1]Wageningen University, P.O. Box 47, 6700 AA Wageningen, The Netherlands. bas.kempen@wur.nl
[2]Wageningen University, P.O. Box 47, 6700 AA Wageningen, The Netherlands. gerard.heuvelink@wur.nl
[3]Alterra, P.O. Box 47, 6700 AA Wageningen, The Netherlands. dick.brus@wur.nl
[4]Wageningen University, P.O. Box 47, 6700 AA Wageningen, The Netherlands. jetse.stoorvogel@wur.nl

Introduction

The Dutch national soil map at scale 1:50,000, which was completed in the early 1990s after more than three decades of mapping, is gradually becoming outdated. Large-scale changes in land use and management that took place after the field surveys have had a great impact on the soil. For instance, oxidation of peat soils has resulted in a substantial decline of these soils within the Netherlands. A field quick scan in 2004 on the status of the map units with thick peat soils (peat layer >40 cm) showed that 47% of the area mapped as thick peat soils during the 1:50,000 survey was deformed into another soil type. Unfortunately this soil type was not recorded. For thin peat soils (peat layer <40 cm) estimates of deformed area range between 60 and 70%. Deformation of thick peat soils to thin peat soils and of thin peat soils to mineral soils has important consequences for soil management.

The aim of this research was to test whether the soil map of the province of Drenthe (2680 km^2) can be updated without additional fieldwork, by exploiting existing point data of soil type and the dependence of soil type on spatially exhaustive high-resolution environmental ancillary variables. For this purpose the 96 map units of the 1:50,000 soil map of Drenthe were aggregated to ten map units: four peat soil map units and six mineral soil map units (Figure 1, top). Multinomial logistic regression (MLR) was used to quantify the relationship between the ancillary variables and soil type, using a large dataset (>16,000) of recently obtained point observations on soil. A comparison of recorded soil groups at observation locations with mapped soil groups, shows that these match for 55% of the observations. For thin peat soils this is only 20-30% and for thick peat soils 40-45%, showing that there is ample room for improvement of the soil map.

Model development

The focus of this study was on model building. This is perhaps the most critical and difficult step in the digital soil mapping process and requires careful attention. A framework for building logistic regression models was taken from the literature (Hosmer and Lemeshow, 1989) and adapted for digital mapping of soil variables. The model building framework consists of eight steps:

1. Definition of a conceptual model of pedogenesis;
2. Collection of explanatory variables from available environmental ancillary data;
3. Univariate analysis and selection of candidate explanatory variables;
4. Multivariate analysis of selected candidate explanatory variables;
5. Evaluation of adequacy of the multivariate model(s);
6. Checking the assumption of linearity in the logit;
7. Checking for interactions between explanatory variables;
8. Statistical and visual assessment of the final model;

We fitted separate models for the ten map units. Pedological expert knowledge was used throughout the model building process to ensure that the final MLR models are not only statistically sound but also pedologically plausible. The models were subsequently applied to the whole of Drenthe on a 25m grid. The model outcome is a probability distribution of ten soil groups at each location. The predicted soil group is the soil group with the largest probability. Shannon entropy was used to quantify the uncertainty of the updated soil map.

Results

Figure 1 (bottom) shows the updated soil map for the province of Drenthe. Changes are most dramatic for the peat map units. The area with peat soils declined with 34% (33,525 ha) compared to the aggregated 1:50,000 soil map (the reference map). Only 45%, 20% and 30% of the soils mapped as *mP*, *PY* and *mPY* are predicted as such. Roughly 60% of the soils mapped as thin peat soils (*PY*, *mPY*) is predicted to be deformed to mineral soils. Ten percent of the thick peat soils (*P*, *mP*) is predicted to be deformed to mineral soils and 29% is predicted to be deformed to thin peat soils. The area with podzols increased with almost 40,000 ha as a result of peat soil deformation. Theoretical purity, which is the mean of the maximum probability at each node of the prediction-grid (Brus et al., 2008), of the updated map was 67%, which is 12% larger than that of the reference map (Table 1). The entropy of the updated map is smaller than that of the reference map (Table 1). This indicates that the updated soil map gives an overall more accurate and less uncertain prediction of soil group than the reference map. The prediction uncertainty is largest for the area corresponding to the map units of the reference map representing thin peat soils, followed by earth soils and thick peat soils (Table 1 and Figure 2). For calculation of the theoretical purity and entropy of the reference map, the frequency distribution of the observed soil groups within the map units of the reference map were used.

Model predictions were validated using an independent dataset with 150 observations collected with stratified random sampling (De Gruijter et al., 2006). Actual map purity, which is the areal fraction correctly predicted, of the updated map was estimated as 58.1%, which is 6.1% larger than that of the reference map (P = 0.039) (Table 1). This increase is mainly attributed to the 11% purity increase of the peat map units (P = 0.069). The discrepancy between theoretical and actual map purity is likely explained by the clustered distribution of the profile observations used to calibrate the MLR models. Roughly 96% of the observations are located in four areas, which comprise 10% of the total area of Drenthe. Apparently the modelled relationships between observed soil groups and predictors are not so easily extended from the four areas to the entire province. An explanation for this might be the strong human influence on soil formation and spatial distribution of soil groups.

Conclusions

We showed that existing, recent soil data in combination with high-resolution environmental ancillary data, can be used to update the Dutch 1:50,000 soil map. Although an independent validation showed that the global map quality of the updated map was larger than that of the existing 1:50,000 soil map, the purity gain was not very large. Updating proved to have more effect for the peat map units than for the mineral map units. Caution should be taken when soil observations for model calibration are clustered within the survey area and there is strong human influence on soil development. Modeled relationships between clustered soil observations and environmental data might not be so easily extended across the survey area in such situation.

Multinomial logistic regression proved to be a simple, suitable method for mapping categorical soil variables. The main limitation is that it ignores spatial autocorrelation during coefficient estimation and spatial prediction. Methods that incorporate spatial autocorrelation exist for binomial logistic regression (Dorman et al., 2007) but not for MLR. Incorporation of spatial autocorrelation in MLR would be a challenging task for the future.

There is certainly room for more improvement of the updated map although this will require additional fieldwork, mainly because soil spatial distribution is strongly influenced by human activities, which cannot be easily represented by environmental data. This is especially true for the peat map soils. However, the results of the digital soil mapping exercise, the maps with predicted soil group and prediction uncertainty, can be used to make future fieldwork more efficient. Fieldwork can be targeted at areas with large prediction uncertainty such as the map unit representing thin peat soils. These soils are also the soils most heavily affected by deformation of the peat layer.

References

Brus, D.J., Bogaert, P. and Heuvelink, G.B.M., 2008. Bayesian Maximum Entropy prediction of soil categories using a traditional soil map as soft information. European Journal of Soil Science, 59(2): 166-177.

De Gruijter, J.J., Brus, D.J., Bierkens, M.F.P. and Knotters, M., 2006. Sampling for natural resource monitoring. Springer, 328 pp.

Dormann, C.F., McPherson, J.M., Araujo, M.B., Bivand, R., Bolliger, J., Carl, G., Davies, R.G., Hirzel, A., Jetz, W., Kissling, W.D., Kuhn, I., Ohlemuller, R., Peres-Neto, P.R., Reineking, B., Schroder, B., Schurr, F.M. and Wilson, R., 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. Ecography, 30(5): 609-628.

Hosmer, D.W. and Lemeshow, S., 1989. Applied Logistic Regression. 1st edn. John Wiley & Sons, New York, 307 pp.

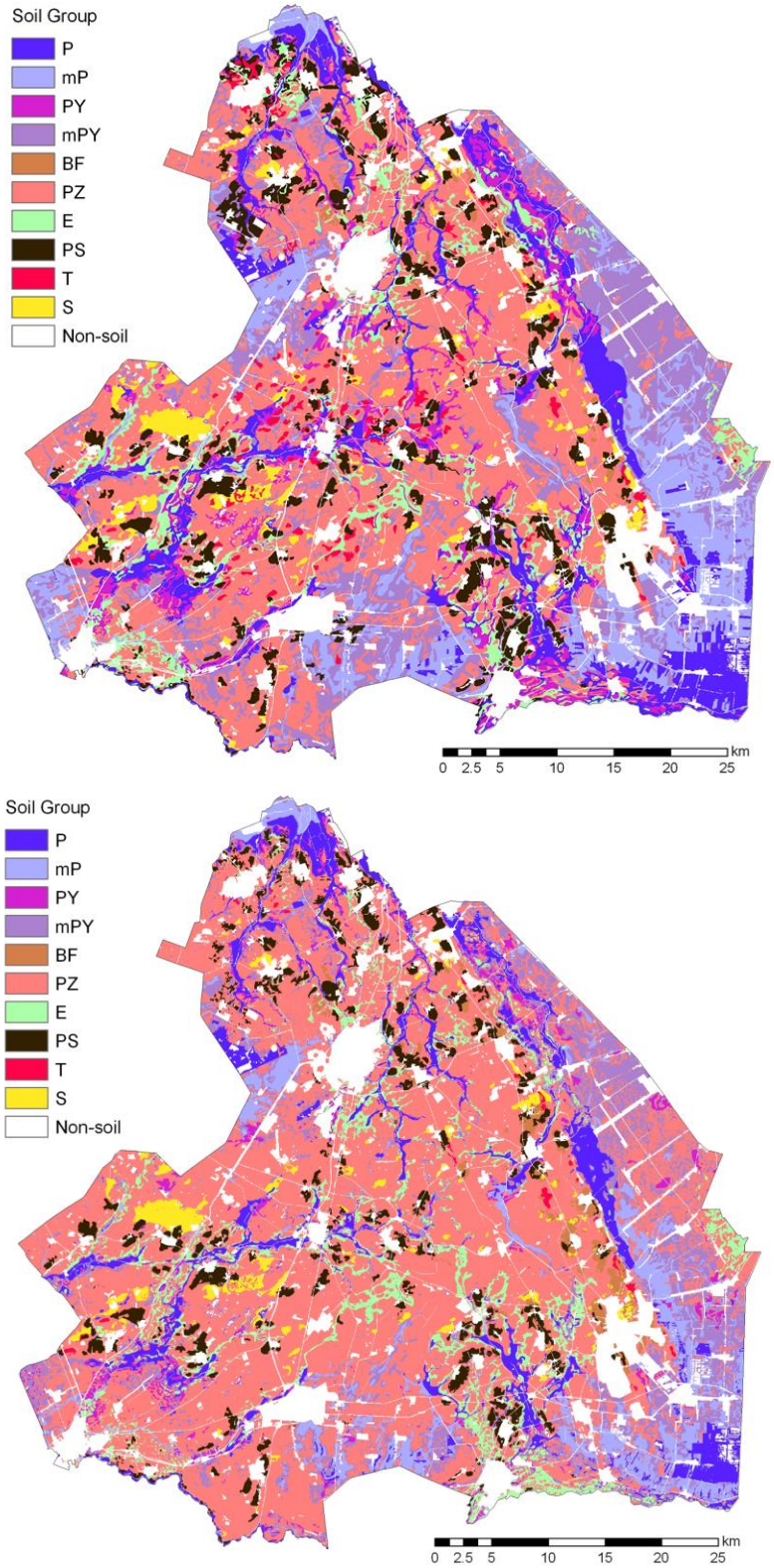

Figure 1. Reference soil map of the province of Drenthe (top) and updated soil map (bottom). Abbreviations: *P* = thick peat soils with peaty topsoil; *mP* = thick peat soils with mineral topsoil; *PY* = thin peat soils with peaty topsoil; *mPY* = thin peat soils with mineral topsoil; *BF* = brown forest soils; *PZ* = podzols; *E* = dark hydromorphic earth soils; *PS* = plaggen soils; *T* = glacial till soils; *S* = Sandy vague soils (soils with limited or no signs of soil formation).

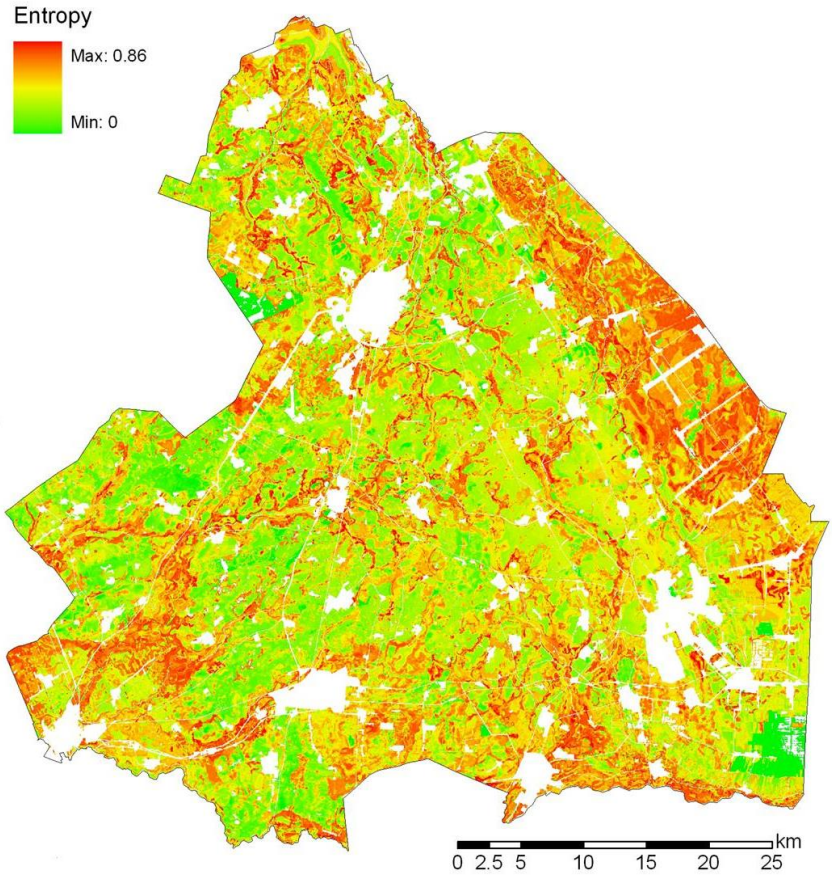

Figure 2. Theoretical entropy of the model predictions on soil group: the larger the entropy the more uncertain the prediction. The maximum value for the entropy is 1.

Table 1
Theoretical purity, mean entropy and estimated actual purity for the two soil maps. The number in brackets is the estimated standard error and *n* is the number of validation observations used to estimate the actual map purity.

		Updated map			Reference map		
	n	Theoretical purity	Entropy	Actual purity	Theoretical purity	Entropy	Actual purity
Global	*150*	0.671	0.40	0.581 (0.040)	0.545	0.53	0.521 (0.037)
Map unit[*]							
P	15	0.638	0.42	0.452 (0.410)	0.409	0.66	0.452 (0.140)
mP	15	0.629	0.43	0.310 (0.114)	0.459	0.59	0.256 (0.076)
PY	9	0.500	0.54	0.395 (0.165)	0.189	0.72	0.046 (0.046)
mPY	22	0.500	0.54	0.495 (0.112)	0.297	0.62	0.359 (0.111)
BF	4	0.826	0.21	0.131 (0.131)	0.597	0.49	0.262 (0.000)
PZ	55	0.786	0.34	0.727 (0.061)	0.781	0.40	0.727 (0.061)
E	10	0.582	0.47	0.422 (0.144)	0.425	0.62	0.554 (0.174)
PS	11	0.655	0.38	0.752 (0.150)	0.516	0.56	0.652 (0.104)
T	4	0.788	0.24	0.975 (0.025)	0.254	0.36	0.000 (0.000)
S	5	0.769	0.24	0.641 (0.321)	0.618	0.40	0.641 (0.321)
Peat	61	0.568	0.48	0.425 (0.065)	0.351	0.64	0.315 (0.057)
Mineral	89	0.751	0.35	0.698 (0.050)	0.689	0.44	0.674 (0.050)

[*] Note that this refers to the area corresponding to the map units of the reference soil map. For each of these map units a MLR model was build.