Updating legacy soil data for digital soil mapping

Bas Kempen, Dick J. Brus & Folkert de Vries
Alterra, Wageningen University and Research Centre, P.O. Box 47, 6700AA Wageningen, The Netherlands

Bas Engel
Biometris, Wageningen University and Research Centre, P.O. Box 47, 6700AA Wageningen, The Netherlands

ABSTRACT: Legacy soil point data stored in soil information systems are a valuable resource for digital soil mapping. For dynamic soil properties, however, these data may not represent the actual field conditions, which may hamper their utility for mapping exercises. Because collection of field data is a major cost component in soil mapping, updating legacy data might be an appealing alternative to collecting new data. In this paper we show how we updated the thickness of the peat layer for more than 3 000 soil profiles obtained from the Dutch soil information system with a statistical model. In addition, we illustrate how the uncertainty about the updated values can be taken into account for digital soil mapping.

1 INTRODUCTION

The national soil map of the Netherlands at scale 1:50 000 requires updating for 365 000 ha of peat soils. Intensive agricultural use and deep drainage in combination with relatively shallow peat layers have resulted in major changes in soil conditions since the 1:50 000 survey was completed in the early 1990s (the first map sheets date from the 1960s). Recent studies on the conditions of peat soils have shown that almost 50% of the area originally mapped as peat soils (peat layer $>$ 40 cm thick) changed to peaty soils (peat layer $<$ 40 cm thick) and that 50-60% of the mapped peaty soils are now mineral soils (de Vries et al. 2009; Kempen et al. 2009). The peat oxidation rate is estimated between 5–10 mm year^{-1} (Hoogland et al. 2012).

The 1:50 000 soil map is the main source of nationwide soil information in the Netherlands, and is used for a variety of environmental and agro-economic analyses in support of policy-making on daily basis. The Dutch national government recognizes the importance of good quality, up-to-date soil information and has commissioned an update for the peatlands of the national soil map. Updating with conventional methods, however, is not a viable option given the available resources. Updating will therefore rely on digital soil mapping (DSM), which recently has been shown to be an efficient alternative to conventional soil mapping in the Netherlands (Kempen et al. 2012). Nevertheless, this will be the first time that DSM will be made operational in a nationwide mapping project.

The Dutch soil information system *BIS* stores over 300 000 soil profile descriptions at point locations that were collected during surveys and research projects since the 1950s. These data are an important resource for DSM (Bui and Moran 2003; Carré et al. 2007), but may not properly represent actual field conditions, as soils change in time. This limits their utility for the calibration of prediction models for dynamic soil properties such as the thickness of the peat layer or the organic matter content. Since the collection of field data is the largest cost component of DSM (Kempen et al. 2012), updating legacy soil data—so that the most can be made out of existing data—can be an appealing alternative to collecting new field data.

In this paper we show an example of how soil property information from profile descriptions, in our case the thickness of the peat layer, can be updated. In addition we show how the uncertainty in the updated values can be taken into account through simulation. Updated soil profile data are after all 'soft' data. They are predictions and not actual measurements, and, in soil mapping, the associated prediction errors should be taken into account.

2 MATERIALS AND METHODS

2.1 Study area

The 68 000 ha study area comprises the glacial till plateau in the northern part of the Netherlands (Figure 1). The till plateau is dissected by a system of

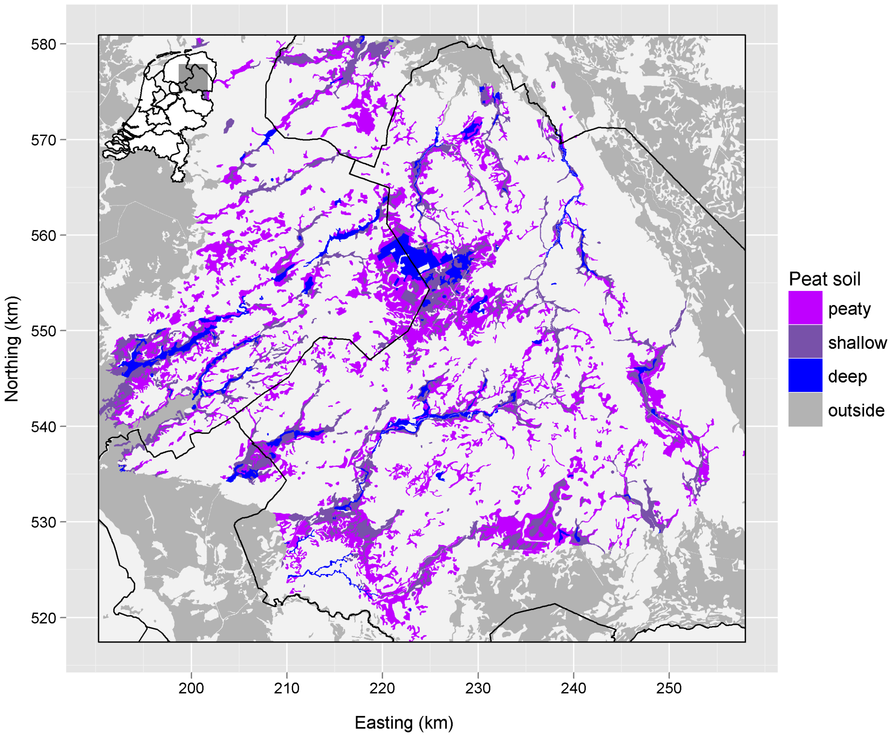

Figure 1: Extent of the peat soils in the study area according to the 1:50 000 national soil map. Three types of peat soils are distinguished: 'peaty' soils (0–40 cm peat), shallow peat soils (40–120 cm peat) and deep peat soils (>120 cm peat). The legend entry 'outside' indicates the extent of the peat soils outside the study area.

brook valleys that are filled with fen peat. Remains of the once vast highmoor bogs on the plateau are now reclaimed for agriculture. Figure 1 shows the extent of the peat soils in the study area according to the 1:50 000 soil map. The study area (partly) covers ten map sheets. The first was produced in 1967 and the last in 1995. Only deep peat soils (at least 40 cm of peat present and the peat layer extending deeper than 120 cm below the surface), shallow peat soils (at least 40 cm of peat present and the peat layer ending within 120 cm below the surface) and peaty soils are distinguished here.

2.2 Soil data

In 2007, 95 geo-referenced sampling sites (dating from 1955 to 1989) situated in the peatlands of the province of Drenthe were selected from *BIS* and revisited (Figure 2). Field sketches and recorded coordinates (note that these were recorded before the GPS era and are prone to errors) assisted the field pedologists in relocating the former sampling sites. Once a sampling site was relocated, the soil profile was described and classified from an auger bore observation. The newly obtained profile descriptions were screened before being used for the update exercise. For instance, locations with censored observations were excluded from the dataset because the annual decrease in peat thickness cannot be determined from these. Censored observations include observations where the peat layer exceeds the auguring depth (thickness cannot be determined) or observations where peat was absent at the time of revisiting (the decrease rate cannot be estimated). Also locations where the soil was strongly disturbed during the period between the original and new observation, or locations for which it was not possible to properly relocate the sample location (e.g. reference points used in field sketches were absent) were excluded from the dataset. After screening, 44 profiles remained that could be used to calibrate a statistical model that in turn can be used to update other, outdated soil profile descriptions in *BIS*.

BIS stores 5 715 soil profile descriptions situated in the study area that are eligible for updating (Figure 3). Most of these profiles are located in areas where large-scale soil surveys at scale 1:10 000 were carried out. Of these profiles 1 654 lacked a peat layer. Of the remaining profile descriptions, 809 dated from the period after 2004. These we consider 'recent' and were not be updated. In total 250 profiles contained censored observations on peat thickness (bottom of peat layer larger than auguring depth) and were excluded. After screening the point dataset, 3 002 profile descriptions of peat soils remained for updating.

2.3 Modelling

The thickness of the peat layer in soil profile descriptions stored in *BIS* was updated with the following

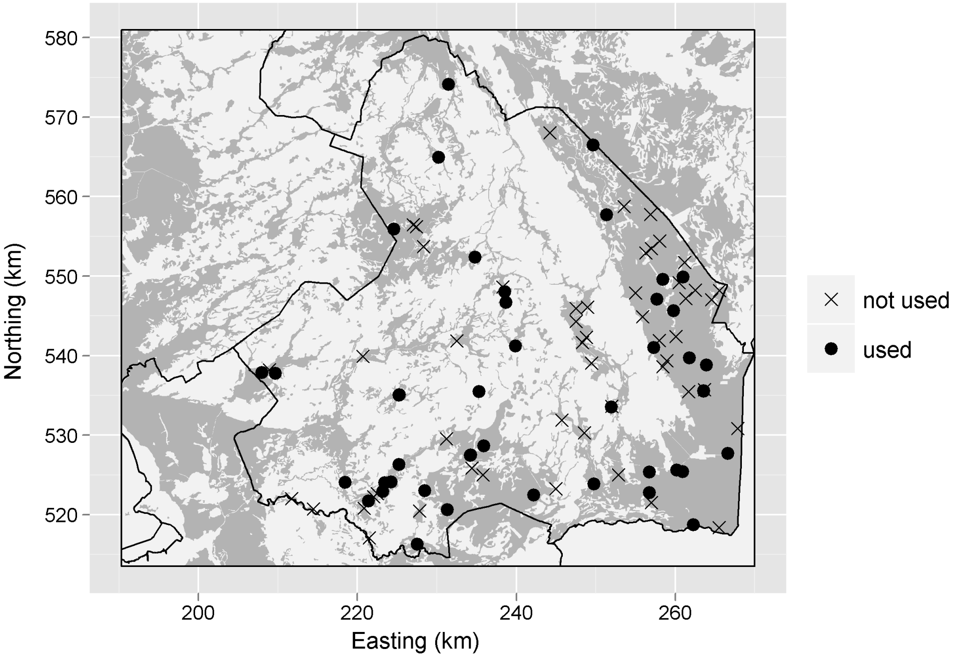

Figure 2: Locations of the 95 revisited sampling sites. The grey shaded area indicates the extent of the peat soils according to the 1:50 000 national soil map.

model:

$$z_{ti} = z_{0i} * p_i^t \tag{1}$$

with z_{ti} the thickness in soil profile i t years after the soil profile was described, z_{0i} the original thickness in this soil profile (as derived from the soil profile description), and p_i the proportion of the thickness of the peat layer in soil profile i that remains after one year. So p_i equals 1 minus the annual proportion of the peat layer that disappears through oxidation. The use of proportional annual decrease prevents the predicted decrease from being larger than the original thickness. The model is extended by the following sub-model for p_i

$$\begin{aligned} p_i &= \pi_i + \epsilon_i \\ logit(\pi_i) &= \mathbf{x}_i^{\mathrm{T}} \beta \\ E[\epsilon_i] &= 0 \\ Var[\epsilon_i] &= \sigma^2 \pi_i (1 - \pi_i) \\ Cov[\epsilon_i, \epsilon_j] &= 0 \quad \text{for} \quad i \neq j \end{aligned} \tag{2}$$

In words, a non-spatial generalized linear model (GLM) (McCullagh and Nelder 1989) was fitted, with a logit link function and residual variance proportional to $\pi(1-\pi)$.(σ^2 is the dispersion parameter). The model was fitted by maximum quasi-likelihood (Wedderburn 1974).

2.4 Simulations

Our final aim is to use the updated soil profile descriptions to map the actual thickness of the peat layer in the study area with the linear mixed model with parameters estimated by residual maximum likelihood (Lark et al. 2006). These predictions will subsequently be used to update the soil class of the peat map units of the Dutch national soil map. Updated profile descriptions contain ‘soft’ observations on peat thickness, because a model is used to predict the actual thickness. This means that the observations are not error-free, or at least do not have negligible error (recent observations of peat thickness were assumed to be error-free). The uncertainty in the updated point data should be accounted for when these data are used for mapping. For this purpose, we suggest using simulated values of p_i.

To simulate values for p_i, a beta(a, b) distribution was used. This probability density function is only positive on [0,1], a useful property for simulating proportions. The expectation of this distribution is $a/(a+b)$, and the variance is $(ab)/[(a+b+1)(a+b)^2]$. By choosing $\hat{\pi}_i(1-\hat{\sigma}^2)/\hat{\sigma}^2$ for a and $(1-\hat{\pi}_i)(1-\hat{\sigma}^2)/\hat{\sigma}^2$ for b, the expectation and variance equal $\hat{p}_i$ and $\hat{\sigma}^2\hat{\pi}_i(1-\hat{\pi}_i)$, respectively. Simulated values for p_i were raised to power t and multiplied by z_{0i} (Eq. 1) to obtain simulated values for the actual thickness z_{ti}. Note that spatial independence was assumed when simulating the peat thickness at the point observation locations.

Mapping of the actual thickness of the peat layer is then repeated as many times as the number of simulations, each time with a different simulated value of

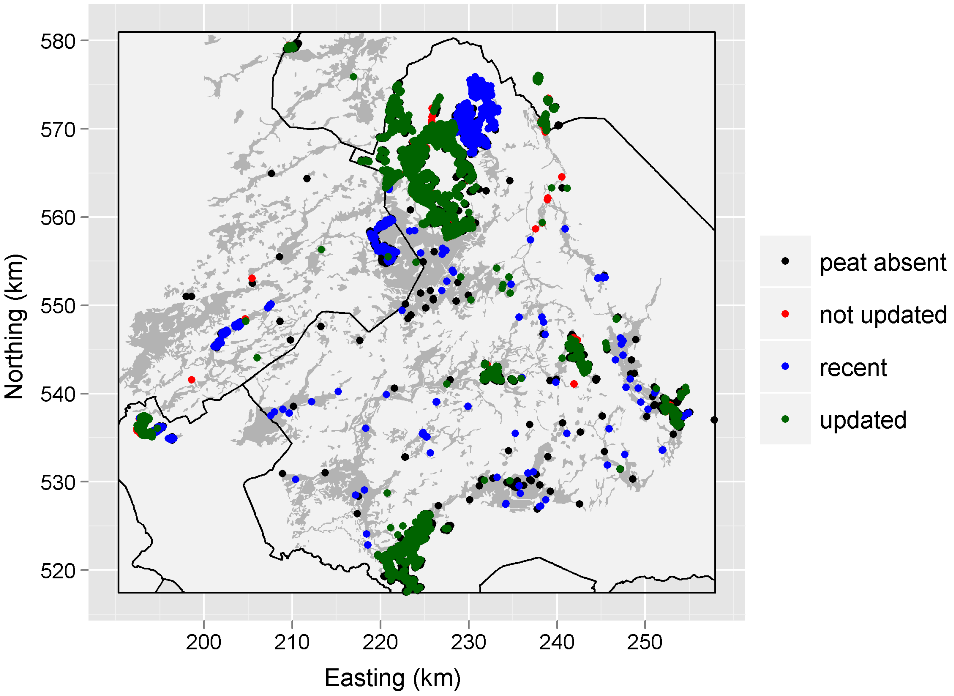

Figure 3: Locations of the sampling sites obtained from the Dutch soil information system that are located within the peat soils of the study area according to the 1:50 000 national soil map.

peat thickness at the updated observation locations.

3 RESULTS AND DISCUSSION

We tried several predictors, including soil class, thickness of the aerated peat layer and land cover, to model the proportional annual decrease π_i, but no predictor was significant (p-values were between 0.14–0.46). We therefore assumed that π_i is constant in space. The logit-linear model then equals $logit(\pi_i) = \beta_0$. The estimated coefficient for the intercept was 4.091 and the dispersion parameter σ^2 was 0.012. Following the model specification in Eq. 2, π_i was estimated as 0.984. This means that the average proportional annual decrease in peat layer thickness equals 1.6%. Using the estimated values for π_i and σ^2 the parameters a and b of the beta distribution equaled 82.0 and 1.4, respectively. These parameters were used to obtain 10 000 simulations of π_i at each sampling location. From these we computed 10 000 actual peat thicknesses, making use of Eq. 1.

Fig. 4 shows the frequency distributions of simulated peat thicknesses at two sampling sites. The sites were sampled in 2004 (left) and 1983 (right). Simulations reflect the peat thickness in 2011. At both sites the initial peat thickness was 105 cm. The two plots shows that the uncertainty about the p_i, and thus about the actual peat thickness, increases with increasing age of the profile description. The simulated average at the 2004 site is 94 cm, with a minimum of 42 cm and a maximum of 105 cm, whereas at the 1983 site the average is 70 cm, with a minimum of 2 cm and a maximum of 105 cm.

Fig. 5 shows a scatter-plot of the updated versus the initial peat thickness for the updated point observations that are grouped by year of observation. This figure shows several properties that are inherent to the logit-linear model we used. First, the absolute decrease of the peat thickness becomes larger when the initial thickness increases. Second, the effect of age on the predicted actual thickness diminishes when the initial peat thickness becomes smaller. This also implies that the absolute annual decrease becomes smaller in time and, because we used a proportional model, the updated thickness will always be greater than zero cm (i.e. the thickness approaches zero asymptotically). The former might be plausible since the most resistant parts of the peat layer will tend to accumulate. The latter, however, is less realistic since the peat layer will eventually completely disappear, evidenced by our observations, through oxidation or through incorporation of peat remains in the mineral soil material for example by ploughing (a large part of the northern peatlands is cultivated). This might result in an over-estimation of the remaining peat layer thickness at the older sampling sites.

For reasons of simplicity we assumed that the data were spatially uncorrelated for calibration of the GLM as well as for simulation of peat thickness at the data points. For calibration this assumption might not affect the results too much, since most calibration sites are spaced far apart from each other (Figure 2). Furthermore, 44 data points are not enough

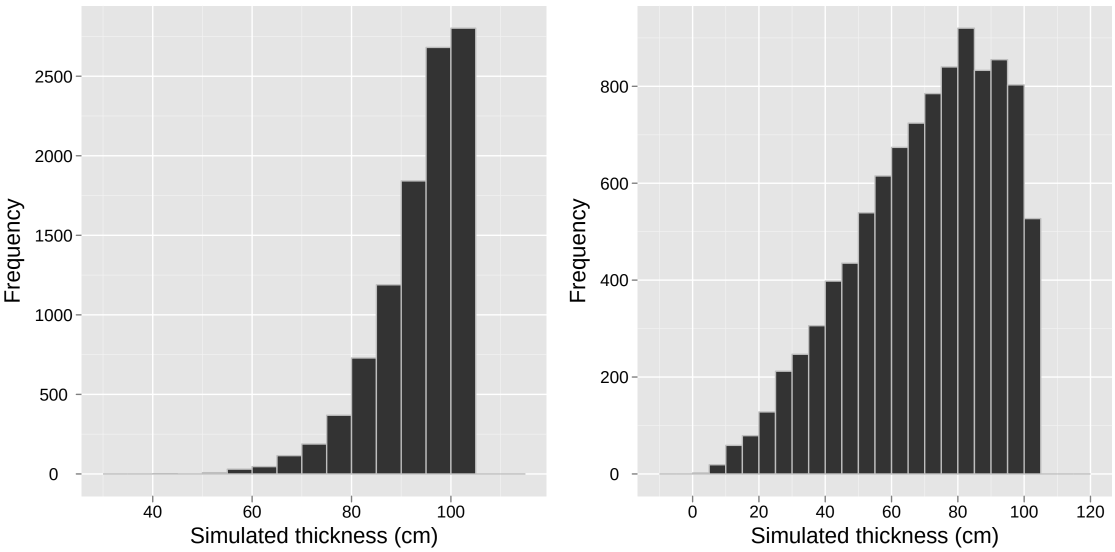

Figure 4: Examples of the frequency distribution of the simulated peat thickness at two sites for 2011. The sites were sampled in 2004 (left) and 1983 (right). The initial peat thickness was 105 cm at both sites.

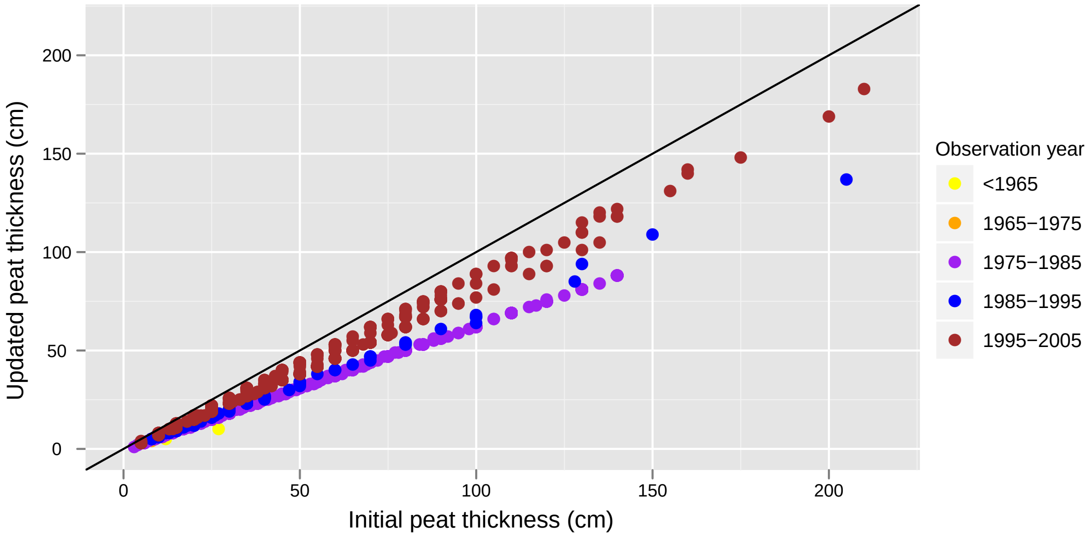

Figure 5: Initial versus updated peat thickness for 3 002 soil profiles, grouped by the year of observation.

to quantify the spatial correlation structure properly (Webster and Oliver 1992). For simulation, however, the assumption might not hold since we have a large dataset of closely spaced points. This is thus an issue that merits further attention.

We think that an important reason for the absence of significant predictors is the limited number of calibration sites. After screening less than half of the revisited sampling sites remained for calibration. For instance, we expected the thickness of the aerated peat layer to have an effect on p (Hoogland et al. 2012). With the fitted model, the proportional annual decrease of a peat layer with a shallow water table is equal to the decrease in a peat layer of the same thickness with a deep water table. This is unrealistic.

The model predictions, as well as its assumptions (e.g. proportionate annual decrease, use of a beta distribution to quantify model uncertainty), were not validated. Validation data were not available. Additional relocation of former sampling sites with acceptable accuracy to obtain such data proved to be hard. Yet, we realize that the lack of a validation study is a weak point. In addition to the proportionate model, we explored the use of a log-linear GLM. In this model the absolute annual decrease in peat thickness was an exponential function of the thickness of the aerated peat layer. Annual decrease values were simulated from a gamma distribution. Although this model was successfully calibrated, predicted and simulated values were implausible and unrealistic. A model for the proportional annual decrease in combination with a beta distribution for simulation at least gave us (physically) realistic predictions and simulations. Furthermore, in a recent study on the subsidence of peat soils in a Dutch coastal area, a proportional decrease of peat layer thickness was used as well (Hoogland et al. 2012). These authors found an oxidizing peat fraction of 1.05% year^{-1}. This comparable, albeit somewhat smaller, than the 1.6% year^{-1} found here. The fact that our calibration data are located in areas with deeply drained, cultivated peat soils, whereas the data used by Hoogland et al. (2012) are located in a fen peat area that is used for pasture, might explain the slightly larger fraction found in this study.

To improve the model we propose to collect additional monitoring data on peat decrease for calibration of the model. Existing sampling sites where the thickness of the peat layer is observed in the past are less suitable for use as calibration. For many of these sites the precision of the registered geographical coordinates is unsatisfactory. We strongly recommend to install a new monitoring network for this purpose. Monitoring locations must be marked in the field so that these can be relocated exactly. Installation of a new monitoring network also has the advantage that we take into account potential predictors in selecting the calibration locations. For calibration it is advantageous when we have locations with a large spread for the potential predictors. In case of a linear model, it is optimal to have locations near the minimum and the maximum of the predictors. Installation of a soil monitoring network would also facilitate validation of the model predictions and assumptions.

4 CONCLUSIONS

We were able to update the thickness of the peat in a large set of legacy point data using a statistical model for the proportional annual decrease. In addition we showed how to handle the uncertainty about the updated thicknesses and how to use this information to obtain simulations of the actual thickness. This greatly enhanced their utility for digital soil mapping of dynamic soil properties.

Despite that updating existing data can be an efficient alternative to collecting new data, implementation proved to be challenging. Relocating sampling sites with acceptable precision was difficult and adequacy of the data obtained from the revisited locations was limited. Less than half of the collected data could be used to calibrate the model. Since we expect that soil mapping will evolve towards mapping the dynamics of soil conditions, quantifying rates of change from soil observations will become increasingly important. Installation of a soil monitoring network is therefore strongly recommended.

REFERENCES

Bui, E. N. and C. J. Moran (2003). A strategy to fill gaps in soil survey over large spatial extents: An example from the Murray-Darling Basin of Australia. *Geoderma 111*(1-2), 21–44.

Carré, F., A. B. McBratney, and B. Minasny (2007). Estimation and potential improvement of the quality of legacy soil samples for digital soil mapping. *Geoderma 141*(1-2), 1–14.

de Vries, F., J. P. Lesschen, J. J. H. van den Akker, A. M. R. Petrescu, J. van Huissteden, and I. van den Wyngaert (2009). Bodemgerelateerde emissie van broeikasgassen in Drenthe; de huidige situatie. Technical Report 1859, Alterra.

Hoogland, T., J. J. H. van den Akker, and D. J. Brus (2012). Modeling the subsidence of peat soils in the Dutch coastal area. *Geoderma 171-172*, 92–97.

Kempen, B., D. J. Brus, G. B. M. Heuvelink, and J. J. Stoorvogel (2009). Updating the 1:50,000 Dutch soil map using legacy soil data: A multinomial logistic regression approach. *Geoderma 151*(3-4), 311–326.

Kempen, B., D. J. Brus, J. J. Stoorvogel, G. B. M. Heuvelink, and F. de Vries (2012). Efficiency comparison of conventional and digital soil mapping for updating soil maps of a cultivated peatland. *Submitted to Soil Sci Soc Am J.*

Lark, R. M., B. R. Cullis, and S. J. Welham (2006). On spatial prediction of soil properties in the presence of a spatial trend: The empirical best linear unbiased predictor E-BLUP with REML. *European Journal of Soil Science 57*(6), 787–799.

McCullagh, P. and J. Nelder (1989). *Generalized Linear Models*. Chapman & Hall/CRC.

Webster, R. and M. A. Oliver (1992). Sample adequately to estimate variograms of soil properties. *European Journal of Soil Science 43*(1), 177–192.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the GaussNewton method. *Biometrika 61*, 439–447.