Creating Low-Cost Soil Maps for Tropical Agriculture using Gaussian Processes

Juan Pablo Gonzalez, Simon Cook, Thomas Oberthur, Andrew Jarvis, J. Andrew Bagnell and M. Bernardine Dias

Robotics Institute
Carnegie Mellon University
5000 Forbes Ave, Pittsburgh PA 15213, USA
{jgonzale, dbagnell, mbdias}@ri.cmu.edu

Center for Tropical Agriculture
(CIAT)
AA6713, Cali, Colombia
{s.cook, a.jarvis, t.oberthur}@cgiar.org

Abstract

Soil maps are essential resources to soil scientists and researchers in any fields related to soil, land use, species conservation, hunger reduction, social development, etc. However, creating detailed soil maps is an expensive and time consuming task that most developing nations cannot afford. In recent years, there has been a significant shift towards digital representation of soil maps and environmental variables that has created the field of predictive soil mapping (PSM), where statistical analysis is used to create predictive models of soil properties. PSM requires less human intervention than traditional soil mapping techniques, and relies more on computers to create models and predict properties. However, because most of the funds for soil research come from developed nations, the research in this field has mostly focused in temperate zones where these nations are located. The areas of the world with more needs in terms of hunger and poverty are mostly located in the tropics, and require different statistical models because of the unique characteristics of their weather and environment. This paper reports on collaborative work with a group of soil scientists from the International Center for Tropical Agriculture (CIAT) and a group of computer scientists from Carnegie Mellon University to develop statistical soil models for Honduras. The reported work leverages the knowledge of the soil science and computer science communities, and creates a model that contributes to the state of the art for PSM.

1 Introduction

The world is currently witnessing a growing demand for technological innovation to empower developing communities [Sachs, 2002]. Inspired by the current demand for advanced technology relevant to developing communities, this paper focuses on the topic of applying Machine Learning techniques to the problem of soil mapping in the tropics. Soil maps are essential resources to soil scientists and researchers in any fields related to soil, land use, species conservation, hunger reduction, social development, etc. However, creating detailed soil maps is an expensive and time consuming task that most developing nations cannot afford.

In recent years, there has been a significant shift towards digital representation of soil maps and environmental variables that has created the field of predictive soil mapping (PSM) [Scull*, et al.*, 2003]. In PSM, statistical analysis is used to create predictive models of soil properties, thus requiring less human intervention than traditional soil mapping techniques, and relying more on computers to create models and predict soil properties. However, because most of the relevant funding is provided by developed nations, soil research has mostly focused on temperate zones (where these nations are located). Thus, the results produces by this research has not been relevant to the tropics; one area of the world with more needs in terms of hunger and poverty. The tropics require significantly different statistical models because of the unique characteristics of their weather and environment.

The task of addressing PSM relevant to the tropics thus became the focus of a research partnership between technologists at TechBridgeWorld [Dias*, et al.*, 2005] at Carnegie Mellon University, and soil scientists from the International Center for Tropical Agriculture (CIAT). More specifically, the goal of the project was to develop statistical soil models for Honduras. The partnership was established as part of TechBridgeWorld's "V-Unit" program[1], and was constructed to leverage the knowledge of the soil science and computer science communities, and create a model that matches or advances the state of the art for PSM, with relevance to tropical countries.

1.1 Background

The International Center for Tropical Agriculture (CIAT) is a not-for-profit organization that conducts socially and environmentally progressive research aimed at reducing hunger and poverty and preserving natural resources in developing countries through partnerships with farmers, scientists, and policy makers. One of CIAT's current areas of interest is soil modeling, because the lack of accurate soil

[1] http://www.cs.cmu.edu/~vunit

data limits their ability to visualize catchment hydrology at a scale amenable to community-based management, target soil-sensitive crops confidently within new areas, and explain complex patterns of changing land use that underwrite landscape resilience. CIAT has done some research for soil modeling based on climate data alone [Corner, *et al.*, 2002]. The addition of elevation data (and derived features) as well as land-cover should significantly increase the accuracy of the prediction. Additionally, the use of newer data mining, modeling and prediction algorithms could make better use of the existing data.

TechBridgeWorld is an initiative within Carnegie Mellon University that innovates and implements technology solutions to meet sustainable development needs around the world. Through strong collaborations with partners in developing communities, they explore and enhance the role of technology globally, focusing on two main principles: sharing expertise to create innovative and locally suitable solutions, and empowerment of indigenous populations to create sustainable solutions. Through these efforts TechBridgeWorld creates technology accessible and relevant to all[2].

1.2 Traditional soil maps

Currently, 68% of the countries of the world have soil maps at 1:1,000,000 or better [Nachtergaele, 1996]. However, these countries only represent 31% of the world's land surface. Most of the remaining 69% corresponds to developing countries. Even though there are ongoing efforts to create a world map at 1:1,000,000, at the current pace it would take 100 years to accomplish this task.

For those areas without detailed coverage, the best available soil maps date to 1974, when the Food and Agricultural Organization (FAO) soil map was published. This map provides worldwide coverage at 1:5,000.000 and is based on Soil Taxonomy [Staff, 1975], which classifies the soils in 12 main categories (*soil orders*) with subcategories. Fig. 1 shows the FAO world map for Honduras.

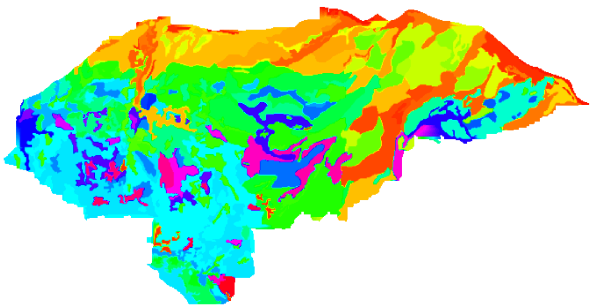

Fig. 1. FAO soil map for Honduras

The FAO soil map of the world is a valuable tool because of its coverage, but it has significant drawbacks: it was made with information and technology of 1960; since then, there have been significant changes in technologies such as GPS, remote sensing and geographic information systems (GIS). Another limitation, which is shared with traditional soil survey techniques, is the classification of soils as distinct categories.

Most modern soil scientists believe that it is more appropriate to model the soil as a combination of elements that vary continuously. At large scales, traditional soil maps are able to capture some of the characteristics of the soil, but at smaller scales, the attempt to classify the soil tends to fail, since soil attributes do not cluster perfectly: a cut on the basis of one attribute may split the variance of another attribute near its peak. The failure of traditional soil survey techniques to produce accurate results at smaller scales significantly limits the soil information available to programs that attempt to help small communities and that implement community-based management of resources.

Furthermore, traditional soil maps depend on subjective expert opinion which varies significantly depending on the person creating the maps and the soil classification used. The maps are therefore predominantly qualitative, and depend on poorly specified predictive models that are not updatable.

1.3 Predictive Soil Mapping for the Tropics

Statistical Soil Modeling is the development of statistical soil models for large areas based on soil samples and digital maps of environmental variables. It is also known in the literature as predictive soil mapping (PSM).

Recent scientific advances in soil-landscape modeling have demonstrated the power of predictive modeling of soil characteristics (including texture, moisture, pH, and some nutrients) at the fine scale. These advances are built on statistically defined relationships between observable features of the landscape as well as understanding of the physical processes and controls behind soil formation. At the same time, significant advances have been made in the availability of high resolution global data on many of the driving mechanisms of soil variability, especially terrain, climate and land-cover.

There is a significant amount of research in predictive soil mapping. For a thorough review of existing approaches to predictive soil mapping see references [Scull, *et al.*, 2003; Nachtergaele, 1996] and [Heuvelink and Webster, 2001]. However, most of the work in predictive soil mapping has been done for temperate zones, corresponding to North America, Europe and Australia. This is due in part to the fact that most of the funding for agricultural research is generated from these regions. Most of the developing world is, however, located in the tropics. This includes significant portions of Africa, Asia, and Central and South America. Very little research has been done in developing appropriate PSM techniques for the tropics, since there is not much funding, and there are few institutions doing research for this region. The tropics have very different climate patterns than temperate zones, therefore PSM models developed for

[2] http://www.techbridgeworld.org

North America, Europe or Australia cannot be directly applied.

Some of the unique climate characteristics of the tropics are the following: temperature stays almost constant during the year, and the main factor determining temperature is elevation. It is possible to find places with 100^{o} F temperatures year-long, but it is also possible to find snow covered places year-long. There are only two seasons, a wet season and a dry season. The duration of the day is also almost constant since the sun trajectory on the sky throughout the year does not vary much.

1.4 Existing Approaches

Most existing approaches to predictive soil mapping use a technique called Kriging [Krige, 1951; Matheron, 1962]. Ordinary Kriging is a form of weighted local spatial interpolation that uses a Gaussian model for the data. Its main drawbacks are the fact that it does not use knowledge of soil materials or processes, and that it requires a large number of closely-spaced samples in order to produce satisfactory results. There are extensions to this method that allow the use of ancillary data, but they are difficult (if not impossible) to extend to more than one ancillary variable.

Some of the most promising approaches to PSM are expert systems and regression trees[Corner*, et al.*, 2002]. Expert systems use expert knowledge to establish rule-based relationships between environment and soil properties. Often they do not use soil data to determine soil-landscape relationships, but some approaches do. Regression Trees are decision trees with linear models in the leaves. They create a piecewise linear representation of the predicted variable. Using this method Henderson [Henderson*, et al.*, 2005] obtained the best results in the literature, which are able to explain more than 50% of the variance of several soil properties such as pH, clay content and sand content.

2 Gaussian Processes for Predictive Soil Mapping

Based on data and resources availability relevant to the tropics, we chose Honduras as a case study. Honduras is a small tropical country (112,000 km^2) for which CIAT has a relatively good database of soil samples (2670 samples). In spite of its small size, Honduras has coastal and mountainous areas, elevations from 0 to 2870 meters, and temperatures from 10 to 30 degrees Celsius.

The goal was to model and predict variations in pH content, clay content and sand content in the topsoil. The input variables available for training and prediction were 32 terrain and climate-related variables such as elevation, slope, curvature, mean temperature, temperature ranges, mean precipitation, precipitation ranges, vegetation index, etc. Each one of these variables was as a digital map at resolutions varying from 90 m to 1 km.

We chose the approach of Gaussian Processes (GPs), a powerful, non-parametric regression technique with solid probabilistic foundations. The main advantages of GPs over other approaches is that they provide well defined confidence intervals, which are very important for soil scientists to assess the quality of the model; and that they allow the use of spatial interpolation and ancillary features to create the model.

GPs can be seen as a generalization of Gaussian distributions to function space, which is of infinite dimension. Even though they are not new, they have regained relevance as a replacement for supervised neural networks [MacKay, 1997; Gibbs, 1997]. GPs are equivalent to several other mathematical approaches including neural networks with infinite number of hidden units, radial basis functions with infinite number of basis functions, least squares support vector machines and kernel ridge regression.

2.1 Covariance function

The idea with Gaussian processes is to put a prior in the probability of the interpolating function given the data. Since this prior is Gaussian, a GP is defined by its covariance function. The covariance function and its hyperparameters define the family of functions that can be chosen by the GP for interpolating the data. The covariance function selected was the squared covariance with a linear term as shown below:

$$C(\mathbf{x}_i, \mathbf{x}_j) = \theta_1 \exp\left[-\frac{1}{2}\sum_{l=1}^{L}\frac{(x_i^{(l)} - x_j^{(l)})^2}{r_l^2}\right] + \theta_2 + \theta_3\delta_{ij} + \sum_{l=1}^{L}\sigma_w^2 x_i^{(l)} x_j^{(l)}$$

where

L number of inputs
l l^{th} input
θ_1 vertical scale
r_l length scale
θ_2 bias
θ_3 output noise
σ_w linear term

2.2 Learning the hyperparameters

The covariance function depends on a set of hyperparameters that need to be determined. The best way to determine the hyperparameters is to learn them from the data. We would like to maximize the likelihood of a prediction given the training data and the parameters. We used a modified version of the NetLab matlab toolbox to accomplish this.

2.3 Variable Selection

One of the main drawbacks of Gaussian processes is that they are computationally intensive to train, since each iteration of the training algorithm requires inverting an NxN matrix, where N is the number of samples (2670 in our case). In order to keep training time low and to prevent overfitting we decided to use a small training set: 20% of available soil samples. 60% of the samples were used for

validation, and the remaining 20% were used as an independent test set. We used greedy search for the most promising variables based on the R^2 score* of the validation set. When several variables had similar R^2 values, we asked the soil scientists at CIAT to select the variable they thought was most important to include in the model. We continued adding variables until the R^2 score of the model stopped improving. With this configuration it takes approximately 27 hours to select variables and create each model. This process only takes place once, unless new variables become available and they need to be added to the model.

After the variables are selected, we use the training and validation data to create a new model, and run the hyperparameter learning procedure starting with the hyperparameters that performed best in the small training sets.

2.4 Prediction

Once a model is chosen, the next step is to use that model to generate soil maps for an area of interest. In order to do this, features from digital maps of the area are used as the inputs to the model, therefore creating a predicted map for a soil component. We generated maps for pH, sand content, and clay content in the topsoil of Honduras. Even though the prediction stage of GPs is much faster than the training stage, the much larger amount of points for which a prediction is required make the process very computationally intensive. With the current implementation, using a Pentium 4 @1.8GHz, it takes 21ms to generate the prediction for one location. The time required to generate a map depends on the size of the map and its resolution. For Honduras (112,000 km^2), it takes 40 minutes to generate a map with 1km grid size, 3.4 days with 90m grid size and 30 days with 30m grid size. If we were to generate a map of Africa it would take 7.2 days, 2.4 years and 22 years respectively. However, this assumes that all the calculations take place on a single computer, which is not likely to be the case. If multiple computers are available, each one could process a much smaller area therefore reducing the total time required proportionally to the number of computers available.

3 Results

3.1 Accuracy of Current Techniques

In order to understand the significance of the results achieved, it is important to be aware of the accuracy of current techniques for soil mapping. According to the soil scientists at CIAT, a rule of thumb is that a soil survey is good if the map units have the right soil more than 50% of the time. Most measurements have a variability of 20% or more between laboratories [Nachtergaele, 1996] and most quantitative prediction methods explain less than 10% of variation. The most important exception is the results from Henderson in Australia which explain up to 50% of the variance of pH in soil and are the motivating force behind the current effort for PSM at CIAT.

3.2 pH in Topsoil

pH in Topsoil was the variable that produced the best results. Two different models were created: one that includes the *x* and *y* location of the samples as variables (i.e.: uses spatial interpolation), and one that does not. The model that uses spatial interpolation performed better, but the one that does not use it gives better insight into the driving factors for pH determination.

The variables found to be relevant for the model with spatial interpolation were *x* and *y* (spatial location of the sample) and P5 (maximum temperature of warmest month). The R^2 for this model is 0.4544 (for the test data). In this case, the model can explain approximately 45% of the variance in the data. From a Computer Science or Engineering perspective, this number seems very low. However, for soil prediction and from a Soil Science perspective, it is a great achievement comparable to be the best results published in the literature.

Fig. 2 shows the performance of the model for the training set (80%) and the test set (20%). The figure on the left shows the comparative performance of the model vs. a mean predictor. The *x* coordinate is the bound, in pH units, and the *y* coordinate is the percentage of the predictions that fit within the predicted value +/- the bound. For example, 95% of the predictions will fall within 1 pH unit of the predictions for the training set. This number is slightly lower for the independent test set (92%) and much lower for a mean predictor (80%). The figure on the right shows actual values versus predicted values. In an ideal case, both would be the same (solid, green line), but in practice there will always be dispersion around the *y* axis. The more dispersion, the worse the model is.

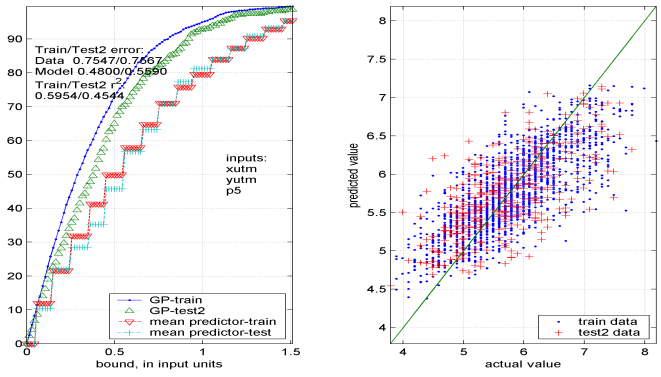

Fig. 2. Model performance for pH in topsoil

Fig. 3 shows the predicted pH maps for Honduras, and the 67% confidence interval (1-sigma). Most of the

* R^2, or coefficient of determination is a measure of the percentage of variance that a model explains.

predictions have a 67% confidence interval of about 0.5 pH units, which was considered very good by the soil scientists at CIAT.

Fig. 4 shows the performance of the model created for pH when no spatial interpolation is used. The variables used by the model are P5 (Maximum temperature of warmest month), P2 (Mean diurnal temperature range), P16 (Precipitation of wettest quarter), and geology class of parent material. The R^2 for this model is 0.3652 (for the test data), which is significantly lower than for the previous model, but is still considered useful.

Fig. 5 shows the predicted pH maps and the 67% confidence intervals when no spatial interpolation is used. Most of the predictions now have a 67% confidence interval of about 0.6, which is still satisfactory.

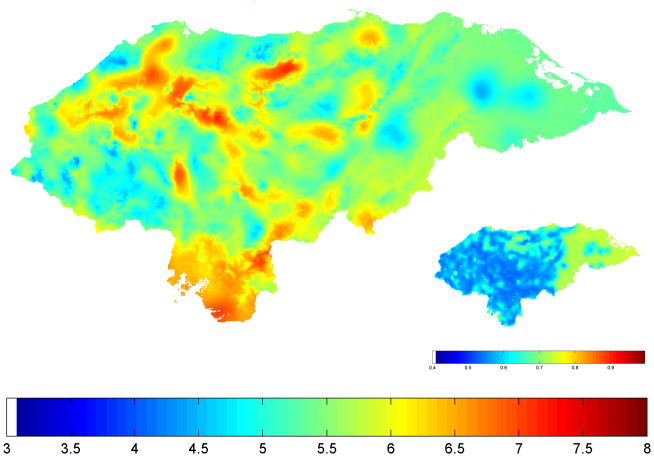

Fig. 3. Predicted map of pH in topsoil and 67% confidence interval

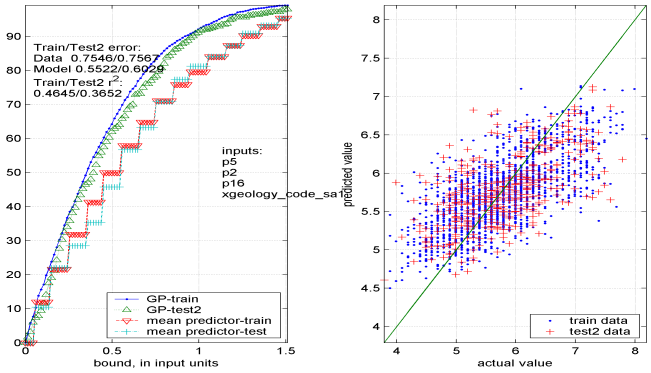

Fig. 4. Model performance for pH in topsoil without spatial interpolation

3.3 Sand and Clay Content in Topsoil

The models for sand and clay content didn't have as good performance as those for pH in topsoil. While the results using spatial interpolation were acceptable and still comparable to some existing approaches, these results had more limited predictive value. The R^2 for sand was 0.2350 (with spatial interpolation) and 0.1026 (without spatial interpolation). For clay, R^2 was 0.1667 (with spatial interpolation) and 0.1403 (without spatial interpolation).

There are several possible causes for the reduced performance of the sand and clay models. One of the most plausible explanations is that the clay and sand content are not as spatially correlated as pH, therefore requiring higher resolution input variables to accurately predict their variations.

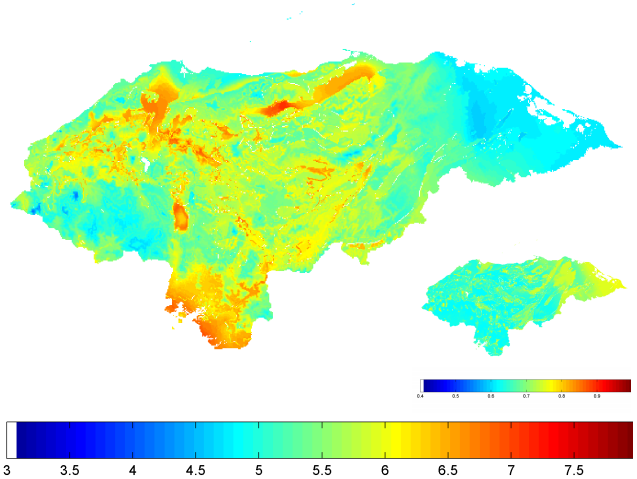

Fig. 5. Predicted map of pH in topsoil and 67% confidence interval, without using spatial interpolation

4 Conclusions and Future Work

4.1 Impact

We have shown the feasibility of performing predictive soil mapping for the tropics by using Gaussian processes. Not only is it feasible, but we were able to contribute to the state of the art in predictive soil mapping. Gaussian processes are an excellent technique for predictive soil mapping, since they produce quantitative predictions with solid confidence intervals, combine pedogenic[3] factors with spatial interpolation, allow for complete coverage of an area and enable continued improvement.

By applying computer science and AI techniques to other fields, and by working together with scientists from these fields, we were able to achieve much more than either group alone would have achieved in the limited time frame of the project. TechBridgeWorld enabled this joint work, which brought state-of-the-art machine learning algorithms to a scientific community that would be otherwise limited to off-the-shelf solutions to their statistical problems.

From the point of view of the soil scientists at CIAT, this work provided them with invaluable insight on the

[3] Pedogenic: related to soil-building processes occurring within the soil

feasibility of low-cost, large-scale, predictive soil mapping for the developing world.

From the point of view of the computer scientists at Carnegie-Mellon University, this work provided a unique opportunity to apply computer science knowledge to the developing world. It shows that this knowledge can be applied to many fields that are beyond military, space and industrial applications. And it shows that a short-time effort can be very productive if it is applied in the right place, at the right time, with the right partners.

4.2 Future Work

Even though the results exceeded the expectations for the work, there is much to be done in the future. One of the main negative results obtained was that none of the variables derived from recently-acquired 90-m elevation maps were relevant to the final models. This could indicate that more effort is required in calculating the derived variables and ensure that we are taking full advantage of the information they provide. Other groups that have worked in PSM have devoted significant efforts to generating derived variables. There are also a few variables used by other groups in the literature that were not available for this project, especially hyperspectral imagery. An important next step would be to obtain these variables and evaluate the impact they have on the models. Another important step would be to compare the results obtained with the leading approach: regression trees. Because of the time constraints of this project, the comparison between the two approaches could not be carried out. However, it would be very important to use the same data set with both approaches and make a direct comparison between them.

This project opened an array of possibilities for joint work between TechBridgeWorld and CIAT. There are a number of projects in which CIAT researchers need expertise in statistical methods, machine learning or computer vision.

Some of the areas for possible collaborative work are:

- Monitoring and management of agricultural fields and natural resources from low cost flying platforms using Computer Vision
- Generation of digital elevation maps from low-cost flying platforms
- Automated image mosaicing
- Segmentation of individual tree crowns
- Detection and monitoring of diseases in plants
- Development of weather insurance schemes for small-holder farmers in developing countries
- Species/crop distribution modeling for targeting conservation and identifying new opportunities for farmers
- Temporal analysis of land cover data

As in this project, without partnership with groups such as TechBridgeWorld they would be limited to off-the-shelf solutions to their problems. A better solution would be for AI researchers to see these problems as new domains for which new or improved algorithms should be developed.

Acknowledgments

The authors would like to thank Manuela Veloso for her support of the v-unit initiative and her input at different stages in the development of this project.

References

[Sachs, 2002] Jeffrey Sachs. Science, Technology & Poverty: Five Ways to Mobilize Development in Low-income Countries. IAEA Bulletin, Making a Difference, Volume 44, Number 1, 2002.

[Scull, *et al.*, 2003] P. Scull, J. Franklin, O. A. Chadwick, and D. McArthur. Predictive soil mapping: a review. *Progress in Physical Geography,* vol. 27, pp. 171-197, 2003.

[Dias, *et al.*, 2005] M. B. Dias, G. A. MillsTettey, and J. Mertz. The TechBridgeWorld initiative: broadening perspectives in computing technology education and research. *Proceedings of the international symposium on Women and ICT: creating global transformation,* 2005.

[Corner, *et al.*, 2002] R. J. Corner, R. J. Hickey, and S. E. Cook. Knowledge Based Soil Attribute Mapping In GIS: The Expector Method. *Transactions in GIS,* vol. 6, pp. 383--402, 2002.

[Nachtergaele, 1996] F. O. Nachtergaele. From the Soil Map of the World to the Global Soil and Terrain Database. AGLS Working Paper. FAO. Rome, 1996.

[Staff, 1975] Soil Survey Staff. *Soil taxonomy: a basic system of soil classification for making and interpreting soil surveys*. Washington, DC: US Department of Agriculture Soil Conservation Service, 1975.

[Heuvelink and Webster, 2001] G. B. M. Heuvelink and R. Webster. Modelling soil variation: past, present, and future. *Geoderma,* vol. 100, pp. 269-301, 2001.

[Krige, 1951] D. G. Krige. A statistical approach to some basic mine valuation problems on the Witwatersrand. *Journal of the Chemical, Metallurgical and Mining Society of South Africa,* vol. 52, pp. 119--139, 1951.

[Matheron, 1962] G. Matheron. *Traite de geostatistique appliquee*: Editions Technip, 1962.

[Henderson, *et al.*, 2005] B. L. Henderson, E. N. Bui, C. J. Moran, and D. A. P. Simon. Australia-wide predictions of soil properties using decision trees. *Geoderma,* vol. 124, pp. 383-398, 2005.

[MacKay, 1997] D. J. C. MacKay. Gaussian Processes: A Replacement for Supervised Neural Networks. *Lecture notes for a tutorial at NIPS,* 1997.

[Gibbs, 1997] M. N. Gibbs. Bayesian Gaussian Processes for Regression and Classification. 1997.