Gaussian Processes for Statistical Soil Modeling of the Tropics

Juan Pablo Gonzalez  
*Carnegie Mellon University*

J. Andrew Bagnell  
*Carnegie Mellon University*

Simon Cook  
*Carnegie Mellon University*

Thomas Oberthur  
*Carnegie Mellon University*

Andrew Jarvis  
*Carnegie Mellon University*

*See next page for additional authors*

Follow this and additional works at: [http://repository.cmu.edu/robotics](http://repository.cmu.edu/robotics)
Gaussian Processes for Statistical Soil Modeling of the Tropics

Juan Pablo Gonzalez, J. Andrew Bagnell, Simon Cook, Thomas Oberthur, Andrew Jarvis and Mauricio Rincon

CMU-RI -TR-05-52

The Robotics Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213

June 2005

© 2002 Carnegie Mellon University

The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies or endorsements, either expressed or implied, of Carnegie Mellon University.
Abstract

Soil maps are essential resources to soil scientists and researchers in any fields related to soils, 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 research funds for soil research come from developed nations, the research in this field has mostly focused in temperate zones (where most developed 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.

Through to the v-unit/TechBridgeWorld initiative at Carnegie Mellon we were able to work with a group of soil scientists from the International Center for Tropical Agriculture (CIAT) and develop statistical soil models for Honduras. Thanks to this joint work, we were able 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.
## Contents

Abstract ......................................................................................................................................................... 2  
Contents.................................................................................................................................................................. 3  

1 Introduction .......................................................................................................................................... 4  
1.1 The International Center for Tropical Agriculture (CIAT) ............................................................. 4  
1.2 Traditional soil maps ...................................................................................................................... 4  
1.3 Predictive Soil Mapping for the Tropics ........................................................................................ 5  
1.4 Existing Approaches ......................................................................................................................... 6  

2 Gaussian Processes for Predictive Soil Mapping ............................................................................... 7  
2.1 Covariance function ......................................................................................................................... 7  
2.2 Learning the hyperparameters .......................................................................................................... 7  
2.3 Variable Selection ............................................................................................................................. 7  
2.4 Prediction ........................................................................................................................................... 8  

3 Results ........................................................................................................................................................ 9  
3.1 Accuracy Of Current Techniques ...................................................................................................... 9  
3.2 pH in Topsoil ...................................................................................................................................... 9  
3.3 Percentage of sand in topsoil ........................................................................................................... 12  
3.4 Percentage of clay in topsoil ........................................................................................................... 14  

4 Conclusions and Future Work .......................................................................................................... 17  
4.1 Impact ................................................................................................................................................ 17  
4.2 Future Work ....................................................................................................................................... 17  
4.3 Future TechBridgeWorld with CIAT .............................................................................................. 17  

References ................................................................................................................................................... 18  

Appendix A: Hyperparameters for each model ...................................................................................... 19  
Appendix B: Variable selection for each model ...................................................................................... 21  
Appendix C: Variation of predicted variable with each input ................................................................. 24
1 Introduction

Soil maps are essential resources to soil scientists and researchers in any fields related to soils, 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 has 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 research funds for soil research come from developed nations, the research in this field has mostly focused in temperate zones (where most developed nations are located). The areas of the world with more needs from 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.

Through to the v-unit/TechBridgeWorld initiative at Carnegie Mellon we were able to work with a group of soil scientists from the International Center for Tropical Agriculture (CIAT) and develop statistical soil models for Honduras. Thanks to this joint work, we were able 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.

1.1 The International Center for Tropical Agriculture (CIAT)

The International Center for Tropical Agriculture (CIAT) is a non-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. CIAT is one of 15 centers known as the Future Harvest centers, which are funded mainly by the 58 donor countries, private foundations, and international organizations that make up the Consultative Group on International Agricultural Research (CGIAR).

One of their current areas of interest is soil modeling. The lack of accurate soil data limits the CIAT's 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. 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.

1.2 Traditional soil maps

Currently, 68% of the countries of the world have soil maps at 1:1,000,000 or better[1]. However, these countries only represent 31% of the world’s land surface. Most of the 69% remaining 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[2], which classifies the soils in 12 main categories (soil orders) with subcategories.

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 is more appropriate to model the soil as 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.

Furthermore, traditional soil maps depend on subjective expert opinion, which varies significantly from 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. Figure 1 shows the FAO world map for Honduras.

![Figure 1. FAO soil map for Honduras](image)

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. 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 provides from these regions. 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 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°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.

Most of the developing world is located in the tropics. This includes significant portions of Africa, Asia and Central and South America.
1.4 Existing Approaches

Most existing approaches to predictive soil mapping use a technique called Kriging. 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 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. 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[3] 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.

For a thorough review of existing approaches to predictive soil mapping see references [3] and [5].
2 Gaussian Processes for Predictive Soil Mapping

Gaussian Processes (GPs) are a powerful, non-parametric technique with solid probabilistic foundations. They can be seen as a generalization of Gaussian distributions to function space, which is of infinite dimension. Even tough they are not new, they have regained relevance as a replacement for supervised neural networks[6][7].

GPs are equivalent to several other mathematical approaches: 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

The main advantages of GPs over other approaches is that provide well defined confidence intervals, which are very important for soil scientists to assess the quality of the model; and that they allow spatial interpolation and ancillary features to be used to create the model.

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:

\[
C(x_i, x_j) = \theta_1 \exp \left[ -\frac{1}{2} \sum_{l=1}^{L} \left( \frac{x_i^{(l)} - x_j^{(l)}}{\sigma_l} \right)^2 \right] + \theta_2 + \theta_3 x_i \cdot x_j + \sum_{l=1}^{L} \sigma_l^2 x_i^{(l)} x_j^{(l)}
\]

where

- \( L \) number of inputs
- \( l \) \( l^{th} \) input
- \( \theta_1 \) vertical scale
- \( \sigma_l \) length scale
- \( \theta_2 \) bias
- \( \theta_3 \) output noise
- \( \sigma_n \) 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 data. 60% of the available data was used for validation, and the remaining 20% was used as an independent test set. We used greedy search for the most promising variables based on the R² score of the validation set. When several variables had similar R² values, we asked the soil scientists at CIAT to select the most variable they thought was more 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.

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 best performing hyperparameters for 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 amount of time required to generate a map depend on the size of the map and its resolution. For Honduras (112,000 km²), 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 and most quantitative prediction methods explain less than 10% of variation. The most important exception are the results from Henderson in Australia, which 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.

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 vs 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. The variables found to be relevant for this model 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). R^2 is a measure of the percentage of variance that the model explains. 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.

Figure 3 shows the predicted pH maps for Honduras, and the 67% confidence interval (1-sigma). Most of the predictions have a 67% confidence interval of about 0.5 pH units, which was considered very good by the soil scientists at CIAT.

Figure 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.

Figure 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.
Figure 2. Model performance for pH in topsoil

Figure 3. Predicted map of pH in topsoil and 67% confidence interval
Figure 4. Model performance for pH in topsoil without spatial interpolation

Figure 5. Predicted map of pH in topsoil and 67% confidence interval, without using spatial interpolation
3.3 Percentage of sand in topsoil

Percentage of sand in topsoil was the second best performing variable. However, the results were significantly inferior with respect to the results for pH in Topsoil. As with the previous variable, 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 with X and Y performed much better than the one without.

Figure 6 shows the performance of the model created for percentage of sand in topsoil. The variables used by the model are x, y (spatial location of the sample), zdem (elevation), mean NDVI (normalized vegetation index), intra-year variation of NDVI, geology class of parent material P13 (precipitation of wettest period), P19 (precipitation of coldest quarter) and P14 (precipitation of driest period). The $R^2$ for this model is 0.2350 (for the test data), which is significantly lower than for the pH model. Still, the model explains 23% of the variance in the data, which is better than most of the existing published results in the soil mapping literature.

Figure 7 shows the predicted map of percentage of sand in topsoil for Honduras. Most of the predictions have 67% confidence intervals of approximately 15%. The map is able to predict sand contents from about 15% to about 70%. Even though the quantitative measurements of the model ($R^2$) are not very good, the map has great predictive value for the soil scientists at CIAT.

Figure 8. Model performance for percentage of sand in topsoil without spatial interpolation shows the performance of the model created for percentage of sand in topsoil, without using spatial interpolation. The variables used by the model are mean NDVI (normalized vegetation index), terrain feature type at 9km, geology class of parent material, P12 (annual precipitation), zdem (elevation), intra-year variation of NDVI, and P13 (precipitation of wettest period). The $R^2$ for this model is 0.1026 (for the test data), which is significantly lower than model with spatial interpolation. The model seems to have very limited predictive value, explaining only 10% of the variance in the data.

Figure 9 shows the predicted map for this model. Even though the $R^2$ for this model is much lower than for the previous one, the confidence intervals for most measurements are very similar, at around 15%. The predicted range is smaller, at 25 to 70%, but it seems to perform better in some regions such as the coastal areas (upper right), predicting a higher concentration of sand.

Figure 6. Model performance for percentage of sand in topsoil without spatial interpolation
Figure 7. Predicted map of percentage of sand in topsoil and 67% confidence interval

Figure 8. Model performance for percentage of sand in topsoil without spatial interpolation
3.4 Percentage of clay in topsoil

Percentage of clay in topsoil was the worst-performing of the models that used spatial interpolation (according to the $R^2$ metric). However, it performed better than the model for sand when no spatial interpolation was used.

Figure 10 shows the performance of the model. The variables used were $x$, $y$, geology class of parent material and $P_{16}$ (precipitation of wettest quarter). The $R^2$ for this model was 0.1667, somewhat lower than the one for sand.

Figure 11 shows the predicted map for this model. The confidence intervals are approximately 10% for most predictions, but the predicted range is more limited than for the sand model. The predicted range is now between 15 and 45%. Still, the map has some predictive value that was not expected given the low $R^2$ obtained for this model.

Figure 12 shows the performance of the model when no spatial interpolation is used. The variables used were $P_{13}$ (precipitation of driest period), geology class of parent material, $P_2$ (Mean diurnal temperature range), $P_{19}$ (Precipitation of coldest quarter), mean NDVI, and $P_4$ (Temperature seasonality). The $R^2$ for this model was 0.1403, slightly lower than the one using spatial interpolation, and somewhat better than the one for sand without spatial interpolation.

Figure 13 shows the predicted map for this model. The predicted range is between 15 and 40%, and the confidence intervals are about 10% for most predictions.
Figure 10. Model performance for percentage of clay in topsoil

Figure 11. Predicted map of percentage of clay in topsoil and 67% confidence interval
Figure 12. Model performance for percentage of sand in topsoil, without spatial interpolation

Figure 13. Predicted map of percentage of clay in topsoil and 67% confidence interval, without 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 it is feasible, but we were able to match or improve 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 factors with spatial interpolation, allow for complete coverage of an area and enable continued improvement.

By applying computer science techniques to other fields, and by working together with scientists from these fields, we were able to achieve much more than them or us alone would have in the limited time frame of the project. The v-unit/TechBridgeWorld initiative 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 feasibility of large scale predictive soil mapping for the developing world. They will use the results obtained to pursue the creation of a worldwide alliance to create new soil maps of the world.

From the point of view of the computer scientists at CMU, this work provided a unique opportunity to apply computer science knowledge to the developing world. It shows that the knowledge acquired at CMU 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 work 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 used in the models. This could indicate that more effort is required in calculating the derived variables and ensure that we are getting the most of them. 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 see the impact they have in the models.

Another important step to perform 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 make a direct comparison of them.

Finally, the group at CIAT would like to organize an international workshop to assess viability of worldwide coverage using predictive soil mapping.

4.3 Future TechBridgeWorld with CIAT

This project opened an array of possibilities for joint work between TechBridgeWorld and CIAT. There are a number of projects in which they need expertise in statistical methods, machine learning or computer vision. As in this project, without TechBridgeWorld they would be limited to off-the-shelf solutions to their problems. Since off-the-shelf solutions are usually several years old, and the main strength of the group at CIAT is not in the areas mentioned, their ability to tackle the problems would be seriously limited.

Some of the areas in which we could work together in the future are:

- Monitoring and management of agricultural fields and natural resources from low cost flying platforms using Computer Vision
- Generation of digital elevation map 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

References


Appendix A: Hyperparameters for each model

pH in topsoil

Experiment: 554, PHW1 vs. inputs. Training set= 82%

out_variable = PHW1
variables = ['XUTM' 'YUTM' 'PH' ]

Final hyperparameters:
in_params = [ 0.1414 -1.3439 4.3123 3.5009 -1.9544 -0.8364 -1.3607 ]
Train/Test2 error:
Data: 0.7567/0.7567
Model: 0.4800/0.5500
Train/Test2 r^2:
0.5954/0.4544
bias:
1.151939
noise:
0.260834 (std = 0.51072)

Lengthscale:
XUTM: 0.115770 (11067.51)
YUTM: 0.173286 (11188.19)
P5: 2.560498 (6.40)

Linear coefficients:
XUTM: -0.039257
YUTM: -0.133932
P5: 0.276685

pH in topsoil, no X,Y

Experiment: 504, PHW1 vs inputs. Training set= 82%

out_variable = PHW1
variables = ['PH' 'P2' 'P16' 'XGeology_Code_SA1' ]

Final hyperparameters:
in_params = [-0.1648 -0.9890 1.6712 2.1778 -3.1989 3.5034 -3.7036 -1.8056 ]
Train/Test2 error:
Data: 0.7546/0.7567
Model: 0.5522/0.6029
Train/Test2 r^2:
0.4645/0.3652
bias:
0.840664
noise:
0.371960 (std = 0.60989)

Lengthscale:
P5: 0.433610 (1.08)
P2: 0.365855 (0.32)
P16: 4.058509 (843.96)
XGeology_Code_SA1: 0.173480 (0.57)

Linear coefficients:
P5: 0.185107
P2: 0.128562
P16: -0.159723
XGeology_Code_SA1: -0.081503

Sand content in topsoil

Experiment: 654, SA1 vs inputs. Training set= 82%

out_variable = SA1
variables = ['XUTM' 'YUTM' 'ZDEM' 'mean_ndvi' 'intra_var' 'XGeology_Code_SA1' 'P13' 'P19' 'P14' 'P13']

Final hyperparameters:
in_params = [-0.0829 5.0725 0.8620 1.8612 1.3229 0.9330 0.0655 -3.2179 -3.2194 0.0184 0.0115 -3.2189 -0.4414 3.8994 ]
Train/Test2 error:
Data: 14.9129/14.4163
Model: 11.5649/12.6090
Train/Test2 r^2:
0.3986/0.2350
bias:
0.920486
noise:
159.578757 (std = 12.63245)

Lengthscale:
XUTM: 0.649868 (62143.46)
YUTM: 0.394312 (25394.81)
ZDEM: 0.516106 (217.08)
mean_ndvi: 0.627201 (11.36)
intra_var: 0.967771 (0.13)
XGeology_Code_SA1: 4.997542 (16.44)
P13: 5.001308 (332.08)
P19: 0.990839 (144.56)
P14: 0.994271 (16.79)
P13: 5.000011 (332.00)

Linear coefficients:
XUTM: 1.010179
YUTM: 0.248507
ZDEM: 0.291923
mean_ndvi: -0.908423
intra_var: -0.590805
Sand content in topsoil, no X, Y

**Experiment:** 604, SA1 vs inputs. Training set= 82%

**out_variable = SA1**

variables = ['mean_ndvi', 'XFeat_1km_9_SA1', 'P12', 'XGeology_Code_SA1', 'intra_var', 'P13']

**final hyperparameters:**

in_params = [0.2806 5.2131 0.9563 -3.2208 -3.2170 0.5258 0.0168 -3.2173 -0.3648 2.1717]

Train/Test2 error:

Data 14.8333/14.3487
Model 13.4789/13.5924

Train/Test2 r^2:

0.1743/0.1026

bias: 1.323985
noise: 183.653649 (std = 13.55189)

**lengthscale:**

mean Ndvi 0.619923 (10.34)
XFeat_1km_9_SA1 5.004900 ( 4.32)
XGeology_Code_SA1 4.995334 (16.42)
P12 0.748003 (276.70)
intra_var 0.951377 (0.13)
P13 4.996137 (329.90)
vertical scale: 8.773429

linear coefficients:

mean Ndvi -1.329373
XFeat_1km_9_SA1 0.465161
XGeology_Code_SA1 1.879462
P12 -0.381605
intra_var -0.750048
P13 -1.323558
vertical scale: 0.694355

Clay content in topsoil

**Experiment:** 754, CL1 vs inputs. Training set= 82%

**out_variable = CL1**

variables = ['XUTM', 'YUTM', 'Geology_Code', 'P16']

**final hyperparameters:**

in_params = [-0.0301 4.7231 1.9283 1.0973 -0.1593 0.0280 -0.7034 3.3708]

Train/Test2 error:

Data 12.1955/11.3255
Model 10.4334/10.3302

Train/Test2 r^2:

0.2681/0.1680

bias: 0.970348
noise: 112.516514 (std = 10.60738)

**lengthscale:**

XUTM 0.381307 (36462.41)
YUTM 0.577729 (37207.37)
Geology_Code 1.082908 ( 6.07)
P16 0.986098 (168.11)
vertical scale: 26.101799

linear coefficients:

XUTM -0.552225
YUTM -0.491606
Geology_Code 0.167477
P16 0.791702
vertical scale: 0.494900

Clay content in topsoil, no X, Y

**Experiment:** 704, CL1 vs inputs. Training set= 82%

**out_variable = CL1**

variables = ['P13', 'XGeology_Code_SA1', 'P2', 'P19', 'mean_ndvi', 'P4', 'P2']

**final hyperparameters:**

in_params = [0.2078 4.7290 -1.0324 -1.2979 0.2249 0.1773 -0.4783 0.9329 1.3486 0.2713 2.8471]

Train/Test2 error:

Data 12.1955/11.3255
Model 10.4058/10.5010

Train/Test2 r^2:

0.2720/0.1403

bias: 1.231002
noise: 113.184496 (std = 10.63882)

**lengthscale:**

P13 1.675487 (111.26)
XGeology_Code_SA1 1.221370 ( 6.29)
P2 0.931022 (0.84)
P19 0.551544 (133.51)
mean Ndvi 1.279147 (23.01)
P4 0.485021 (124.45)
P2 0.509502 (0.48)
vertical scale: 17.238056

linear coefficients:

P13 1.601474
XGeology_Code_SA1 -1.315301
P2 0.676301
P19 0.564605
mean Ndvi 0.546458
P4 -0.438042
P2 0.476801
vertical scale: 1.311705
Appendix B: Variable selection for each model

Figure A 1. Variable selection for pH in topsoil

Figure A 2. Variable selection for pH in topsoil, no X,Y
Figure A 3. Variable selection for sand content in topsoil

Figure A 4. Variable selection for sand content in topsoil, no X, Y
Figure A 5. Variable selection for clay content in topsoil

Figure A 6. Variable selection for clay content in topsoil, no X, Y
Appendix C: Variation of predicted variable with each input

Figure A 7. Variation of pH in topsoil with each input

Figure A 8. Variation of pH in topsoil (no X,Y) with each input
Figure A 9. Variation of sand content in topsoil with each input
Figure A 10. Variation of clay content in topsoil with each input

Figure A 11. Variation of clay content in topsoil with each input, no X,Y