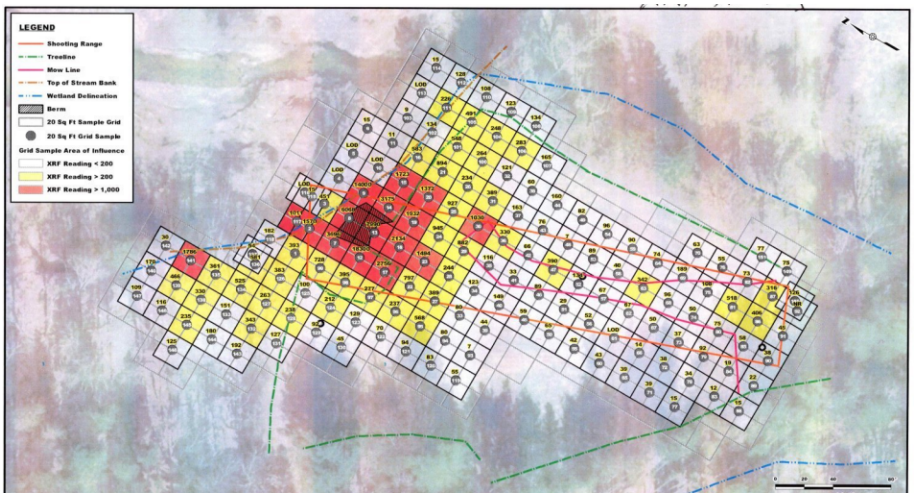
Final Project
XRF Sampling Accuracy: A Cost-Effective Method for Detecting Hazardous Lead Levels in Soil
Jared Winslow (jdw2218)
December 16, 2022

Introduction
X-Ray Fluorescence (XRF) is a method used in forensic science, art conservation, and the environmental sciences for ascertaining the chemical composition of ceramics, soil, and other materials. An object of interest is bombarded with gamma rays, and the resulting fluorescent rays radiating from this object are measured to determine the concentration of chemical compounds. In geochemistry, it can be used for the identification and investigation of various metals.

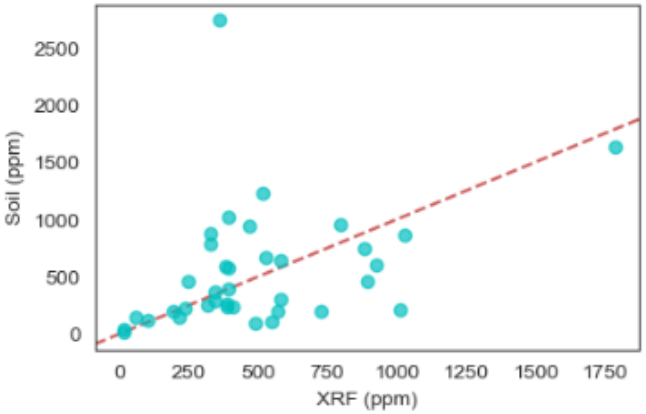
Recently, driven by an imminent lead hazard, a Massachusetts environmental clean-up agency has needed to assess XRF sampling's reliability. Can it be used to accurately detect elevated lead levels in the soil, thereby bypassing the cost- and time-intensive lab testing of excavated soil? This paper attempts to answer that question by providing geospatial statistical analysis of the cleanup site.

Incorporating the physical location of sampling can enhance XRF's predictive capabilities, perhaps enabling more accurate and confident predictions of lead levels. Therefore, the primary goal is to capture the geospatial dependence of measurements and leverage this structure to inform prediction.

Data, Exploratory Analysis, & Methods
At the time of writing, the agency has collected approximately forty (n = 39) physical soil samples, as well as approximately two hundred (m = ~150) XRF samples at the same locations as soil samples, and elsewhere. Both measurements are in parts per million (PPM). Soil and XRF samples were taken from locations that lie on an equidistant grid. For the purpose of statistical analysis, location data was taken from multiple maps with this grid by selecting an arbitrary center, and encoding every sample location with a relative distance from this center.

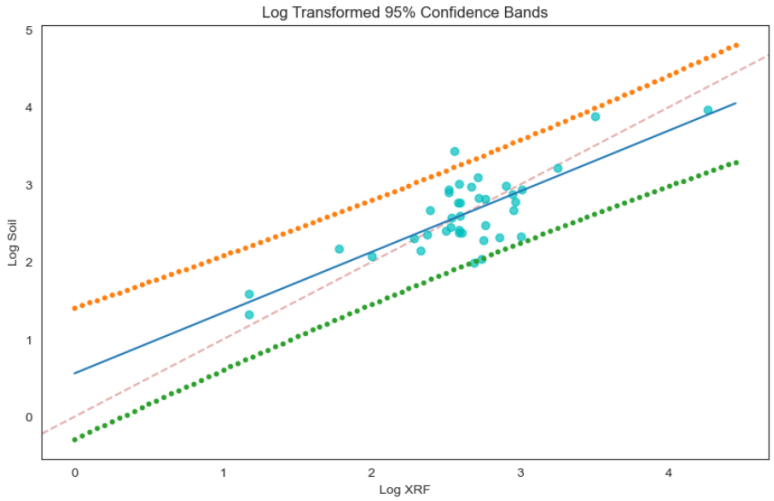


Exploratory analysis revealed both heavy spatial autocorrelation and correlation between XRF samples and soil samples. The two variables are both conditionally heteroskedastic given the other. However, transforming both with a logarithm made them conditionally homoskedastic. This result makes intuitive sense. While measurements get increasingly noisy as they become larger, the noise still scales uniformly as a *percentage* of the measurements. This and having only positive values are the exact conditions for a log transform. Performing a simple linear regression after the transform further verified linearity and homoscedasticity. Furthermore, the variances for the transformed XRF samples and soil samples with respect to location seem similar as well. Given these benefits, the log-transformed data will be used for modeling.



Modeling

Since the distribution of lead in soil is often clustered, capturing auto-correlation is an especially important model feature. Gaussian processes (GPs) can do this by fitting an infinite (computationally finite) function to data. In practice, this amounts to estimating a flexible curve according to certain restrictions. These restrictions come in the form of kernels / covariance functions, which specify the extent to which nearby points should be similar. GPs can be used in a Gaussian process regression (GPR) to model a generative process such that responses (XRF and soil samples) originate from auto-correlated input data (sample locations). The kernel captures auto-correlation within the multi-dimensional spatial input, with different kernel types expressing different types of auto-correlation decay. Beyond that, the two response variables are hypothesized to be correlated. To incorporate this correlation (which is different from the auto-correlation of the inputs), a second kernel is added through coregionalization. This multi-output addition to the model encapsulates soil samples' hypothesized dependence on XRF samples.



Specifically, there are two kernels and five parameters used in this Gaussian process regression, and each parameter is modeled as coming from a prior distribution. The exponential kernel, exponentiated quadratic kernel, and Matern52 kernel were used, with the results being discussed in the evaluation section. The five parameters are length scale (ℓ), amplitude (η), coregionalization similarity vector (W), coregionalization difference (κ), and single noise standard deviation for both GPs (σ).

The length scale determines the rate at which autocorrelation decays in the input spatial dimensions. The amplitude adjusts outputs to be on the correct scale. Together, the coregionalization similarity vector and coregionalization difference vector form a covariance matrix, which relates the two GPs XRF samples and soil samples. This combination of parameters capture the dependence between the two responses.

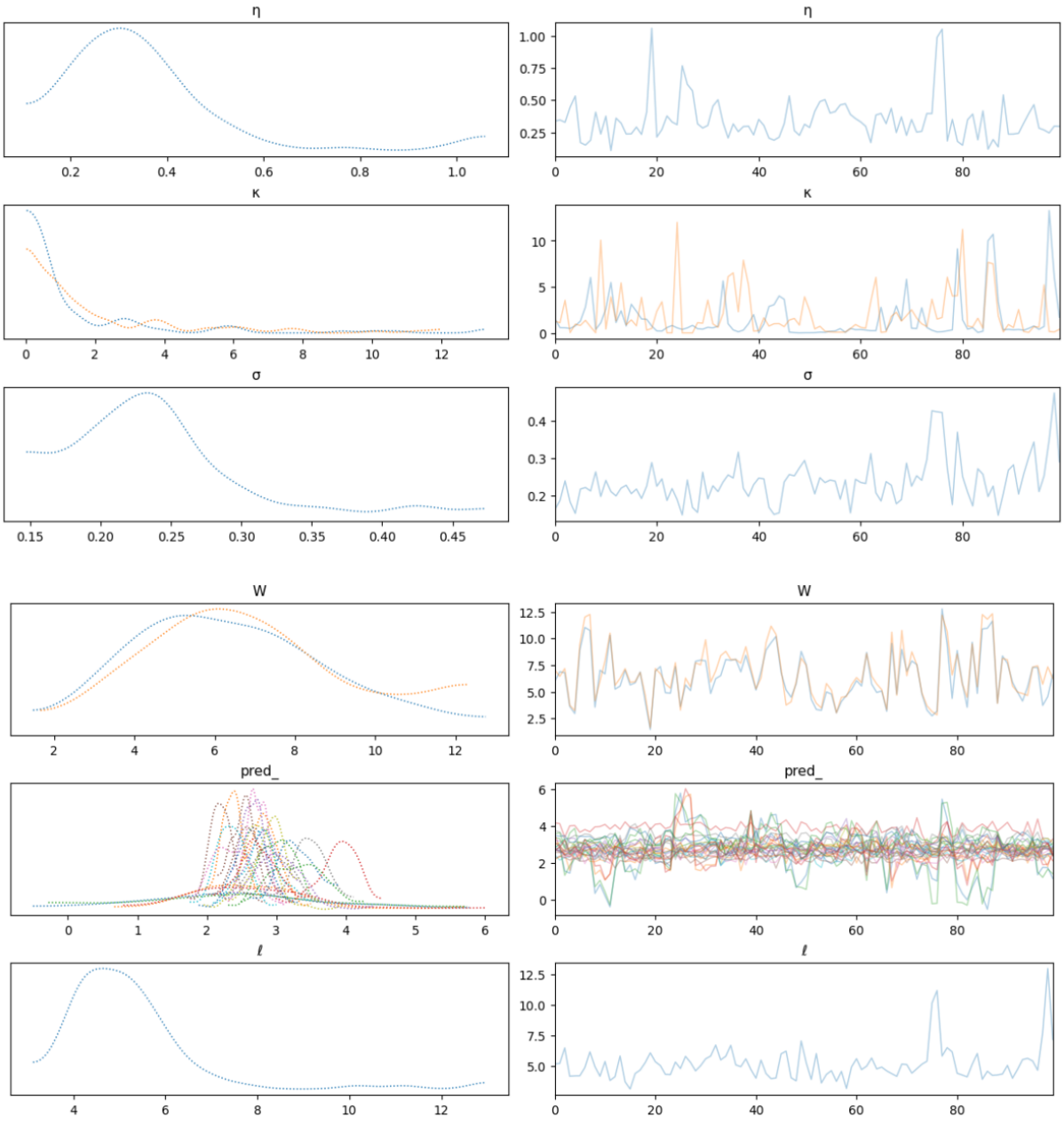
The length scale, coregionalization difference, and noise parameters are the important parameters, each given a noninformative prior in the form of a half Cauchy distribution. Length scale is a positive metric since we want the autocorrelation to be positive and is measured like a standard deviation. The coregionalization difference and noise parameters are similar. In terms of prior distributions, the half Cauchy is preferred over the inverse gamma (another distribution for modeling variances / standard deviations) because inverse gammas do not have a proper limiting posterior distribution as the parameters tend to zero. The coregionalization similarity parameter was given a normal prior because of its interpretation as a low-rank structure underlying a covariance matrix.

Optimization & Inference

For this project, I learned PyMC and used it to perform Bayesian inference. To approximate the model's intractable posterior, I used the The No-U-Turn Sampler (NUTS), which is a modified Hamiltonian Monte Carlo algorithm with the number of steps planned ahead of time.

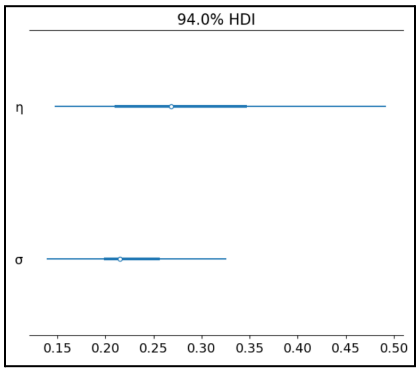
To speed up convergence, appropriate initial values for hyper parameters were chosen. There were two main scales at which initial values needed to be chosen: input spatial scale and output response scale. To determine a reasonable initial value for the input data, the standard deviation of the distances between every pair of location points was calculated. For the output data, the standard deviation of the response was calculated. A multiplier was applied, and those values acted as the initializations.

Below are trace plots, showing both the estimated posterior distributions of parameters and held-out test data (left) as well as Markov chain convergence (right). The predictive posterior distributions on held-out data (pred_) differ by the location they are sampled from, which verifies that the model predicts a different reading for each location. For instance, one is centered at 4 (red), while another is centered at 2 (brown).



Parameter Results

The length scale parameter ℓ has a posterior mean of 5.23. The noise parameter has a mean of 0.24, which is quite small compared to the response values which range from 1 to 4. Estimating some of the parameters requires some reformatting. Once there are estimates for W and κ, these can be taken together to find the correlation between the two GPs. Take the outer product of W and add κ to retrieve the covariance matrix. Standardizing gives the correlation



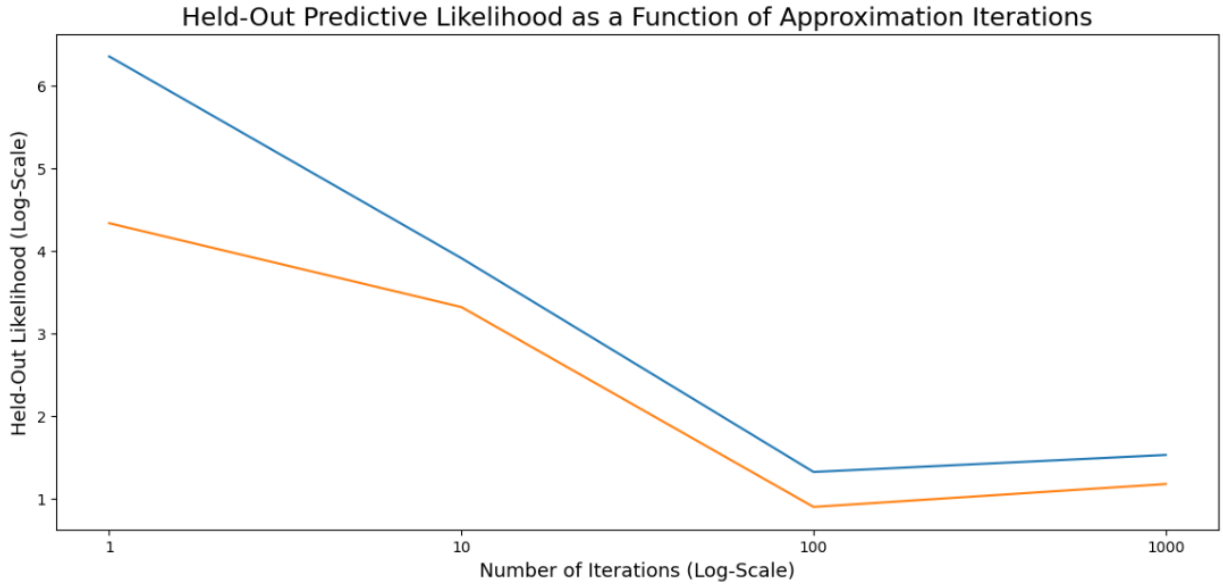
matrix. By performing this operation on all the samples, we can get both posterior means of these parameters and credible intervals. The estimate for the correlation between the two GPs is 0.94 and the 95% credible interval for this correlation is (.81, .99). See table 1 in the appendix for more values, as well as graphs 1 and 2.

Put these pieces of information together, and it can be seen that the two response variables are quite similar as hypothesized, which means that a single noise parameter for both GPs is justified. Furthermore, this noise parameter is estimated to be quite small, showing a good fit between the GPs and the log response data for both XRF samples and soil samples.

Evaluation

To evaluate the model, two types of prediction were done: 1) predicting both XRF sample and soil sample values from location alone and 2) predicting soil sample values from location and XRF sample values. For the first type, predictions were made at both i) locations where there was held-out test data and ii) all locations, even locations with no observed response data.

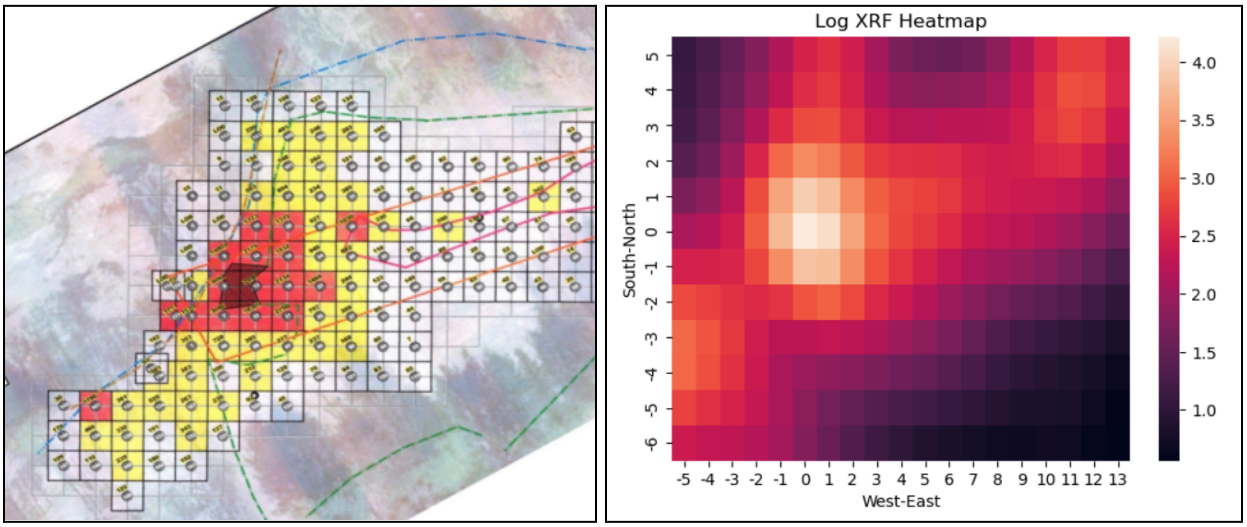
The first type of prediction was used to calculate the held-out predictive likelihood. For GPs, the likelihood reduces down to mean squared error (MSE). As can be seen below, the negative likelihood (and hence the error) decreases as a function of the number of iterations. Transforming the held-out values and predictions back to the original scale gives a similar result.



On the trace plot a page up, the predictions are different for different sample locations. Graph 3 in the appendix further reflects this fact. In fact, the predicted values in graph 3 with the highest credible intervals are those location points farthest away from the others. The model appropriately downgrades its confidence the farther you are away from the majority of points.

The second part of the first type of prediction, predicting all locations regardless of knowledge of XRF and soil response values, enabled the most widespread prediction. By predicting all locations, it can be seen how the model is able to generalize. Furthermore, GPs are smooth, so visualizing their predicted value at all points illuminates trends.

Below, the resemblance to the original map is striking. Not only is the GP able to smoothly predict different measurement levels at different locations, but the actual curve that it has estimated matches the data remarkably well.



The above predictions are made using the Matern52 kernel. This kernel was chosen after first attempting prediction with both the exponential kernel (graph 6) and the exponentiated quadratic kernel (graph 7). While models using both of these kernels fit the data reasonably well, both do worse than Matern52. The exponential kernel is relatively rough, and the exponentiated quadratic is relatively smooth. The Matern52 kernel acted as a compromise between the two kernels, outperforming both of them on heatmaps and held-out predictive likelihood.

How do the predictions compare for XRF vs soil samples? The difference between the predictions for XRF sampling and soil sampling can be seen visually (graph 4). While the two predicted GP curves are extremely similar, they differ nonuniformly. Additionally, transforming the log data back to the original scale results in a similar but less interpretable picture (graph 5).

Finally, for the second type of prediction, the model used XRF sample levels in its training data for locations with held out soil sample data. Calculating the held-out predictive likelihood on the soil sample data resulted in small errors. The predictive likelihood at each number of iterations was roughly half that of the model using only location-based data. Therefore, the model was able to successfully use XRF sample values to bolster its predictions for soil sample values at those locations.

Conclusion

There are multiple evidence points in support of using XRF sampling to predict lead levels in the soil. First, XRF samples and soil samples are correlated, and the correlation between their GPs is even stronger. The correlation between their GPs is the correlation that accounts for spatial auto-correlation, which is a more robust metric. Second, the noise around the GPs is relatively small, meaning the model has relatively good fit. Third, the predicted curves for XRF samples and soil samples are very similar. The difference between the two can be exploited to make better predictions. Fourth, the negative held-out predictive likelihood is smaller when the XRF data at specified locations is given to the model to predict soil sample data at those same locations. Overall, there is evidence for predicting soil lead levels using XRF and even stronger evidence for differentiating samples based on location. In the future, more research could be done on similar geospatial datasets to verify XRF's use as a proxy for soil sampling in detecting elevated levels of lead.

References

1. Andrew Gelman. 2006. Bayesian Analysis, 1, Number 3, pp. 515–533. Prior distributions for variance parameters in hierarchical models.

2. Matthew D. Hoffman, Andrew Gelman. 2014. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo .

3. Alexandra M. Schmidt, Alan E. Gelfand. 2003. A Bayesian coregionalization approach for multivariate pollutant data.

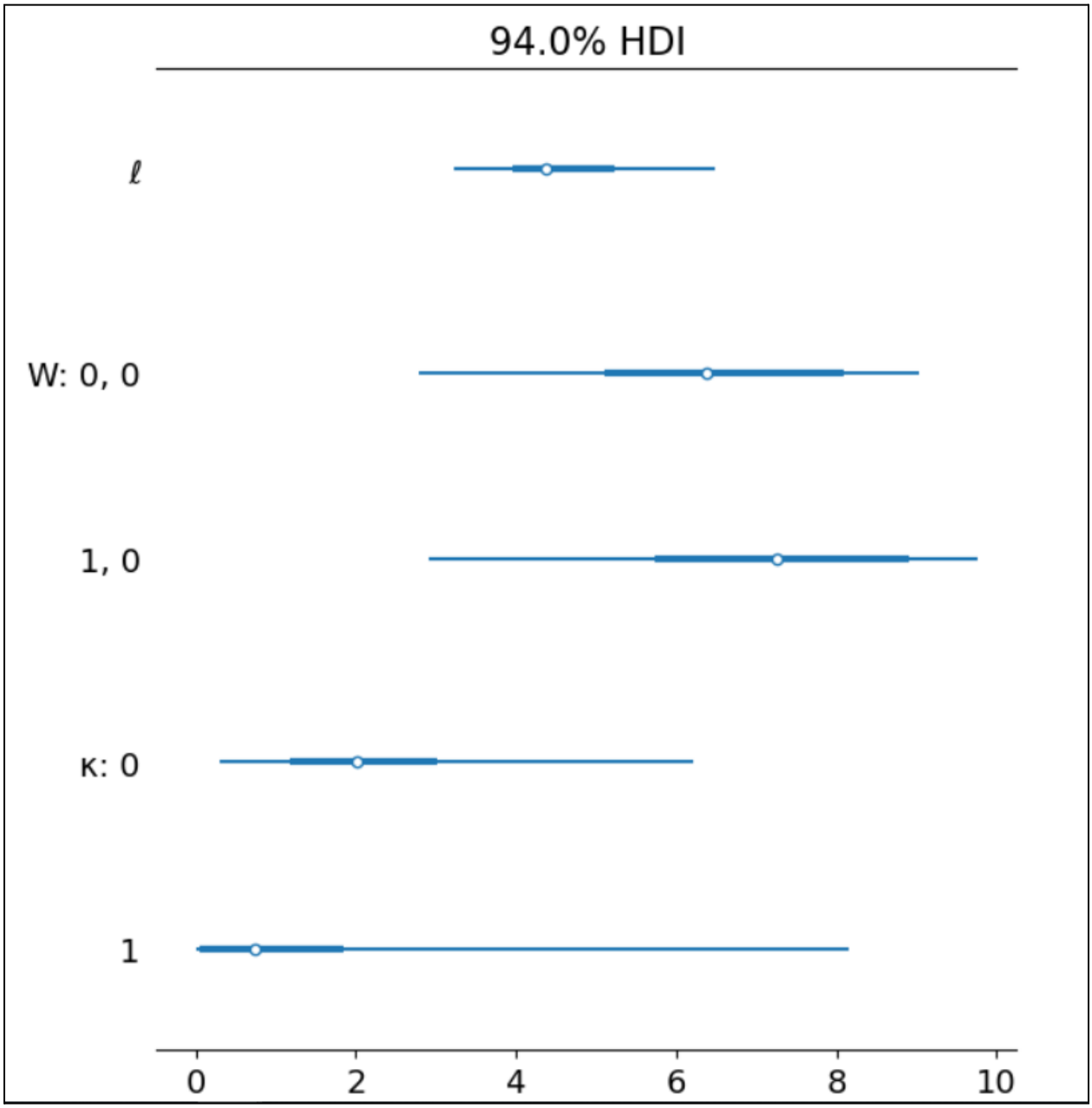
4. Artem Smirnov. 2022. Spatial prediction of soil pollutants with multi-output Gaussian processes.

Appendix

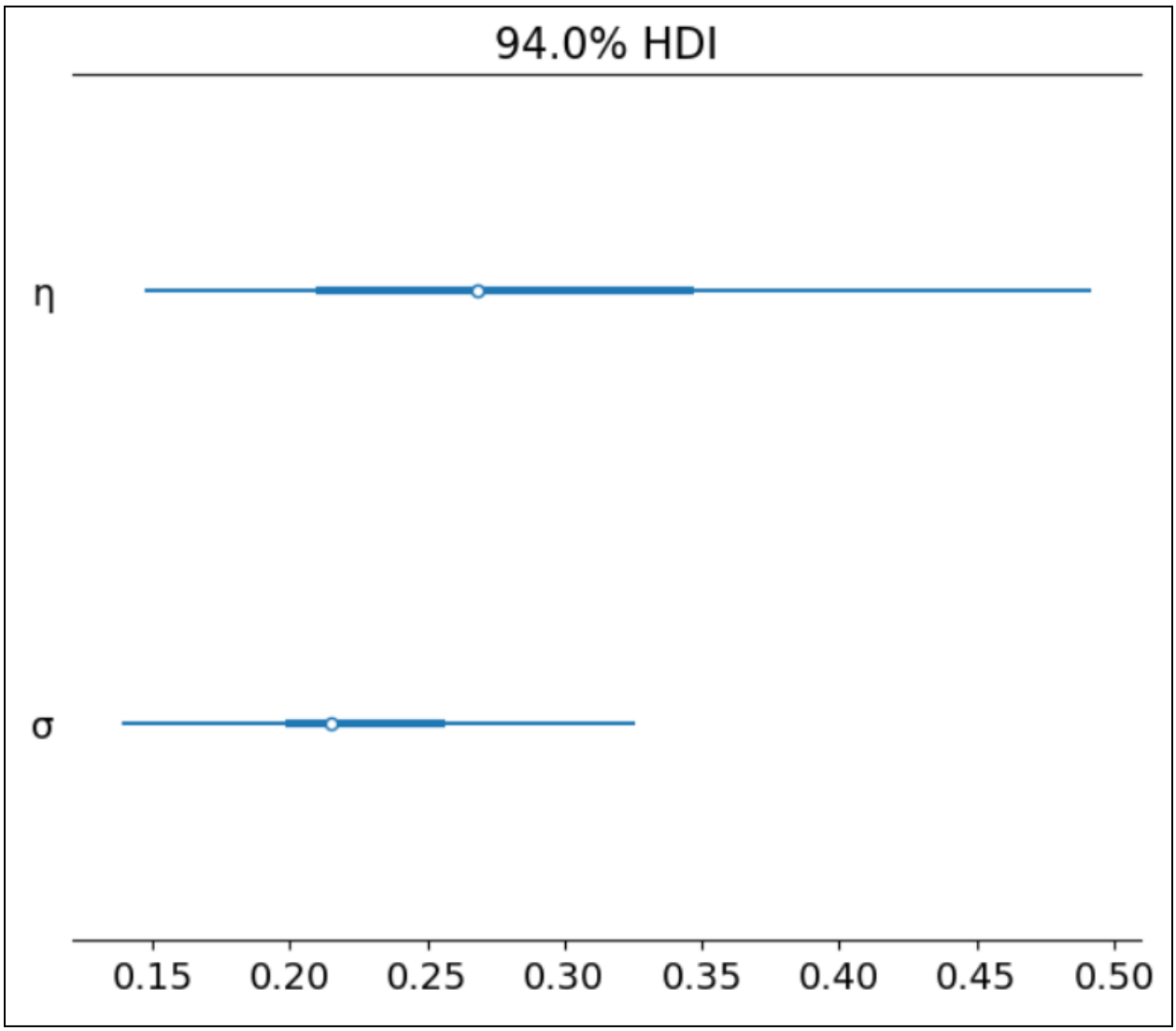
Table 1

	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
ℓ	5.23	1.72	3.27	7.63	0.09	0.06	375.72	414.48	NaN
η	0.32	0.15	0.12	0.60	0.01	0.00	706.98	732.15	NaN
W[0,0]	6.64	2.21	2.65	11.03	0.09	0.07	605.57	612.08	NaN
W[1,0]	6.69	2.23	2.38	10.69	0.09	0.07	623.51	617.52	NaN
κ[0]	2.55	3.19	0.00	7.76	0.14	0.10	313.57	365.81	NaN
κ[1]	2.51	3.04	0.00	7.89	0.10	0.07	498.85	174.28	NaN
σ	0.24	0.05	0.16	0.33	0.00	0.00	504.39	436.72	NaN

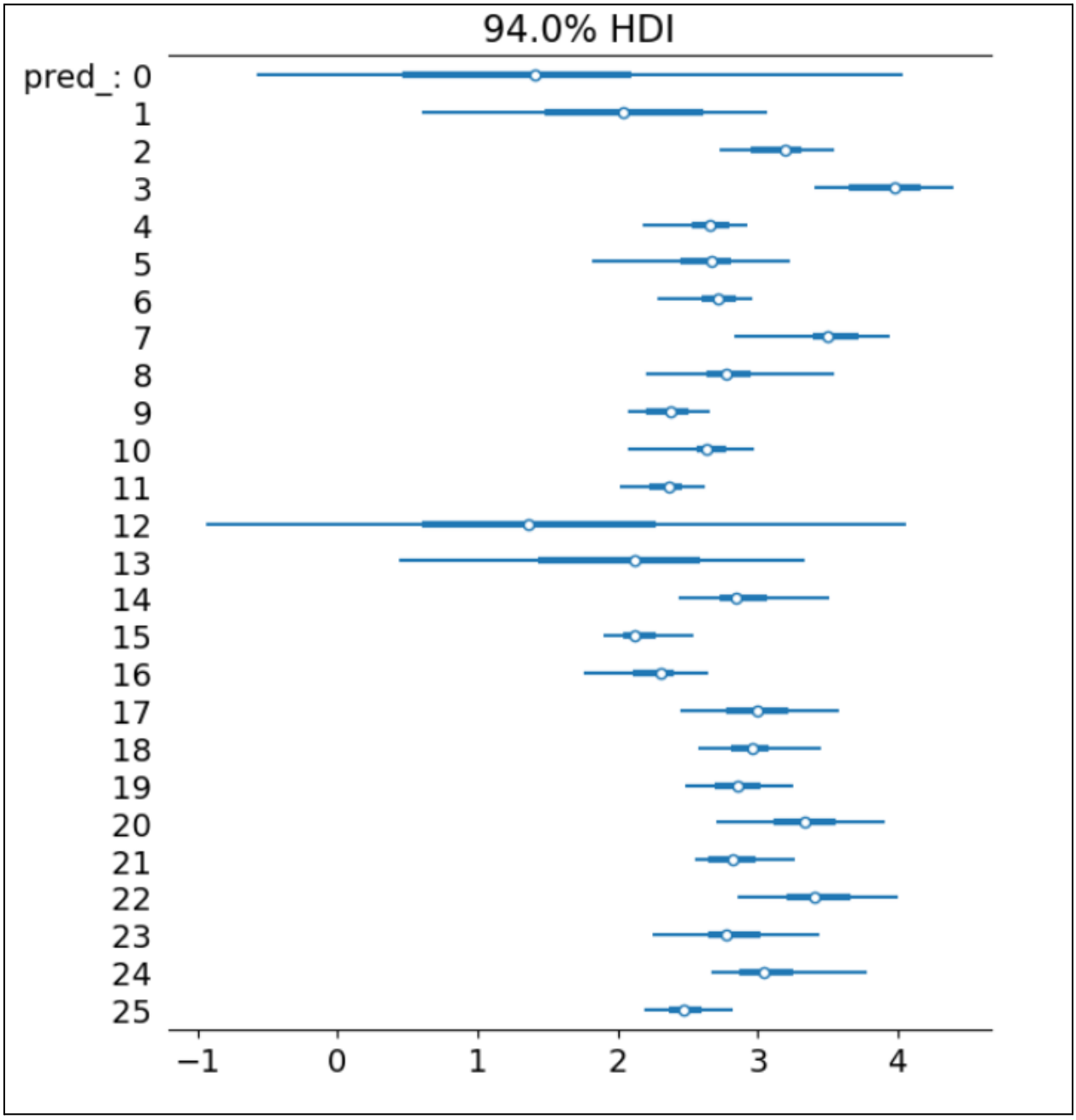
Graph 1



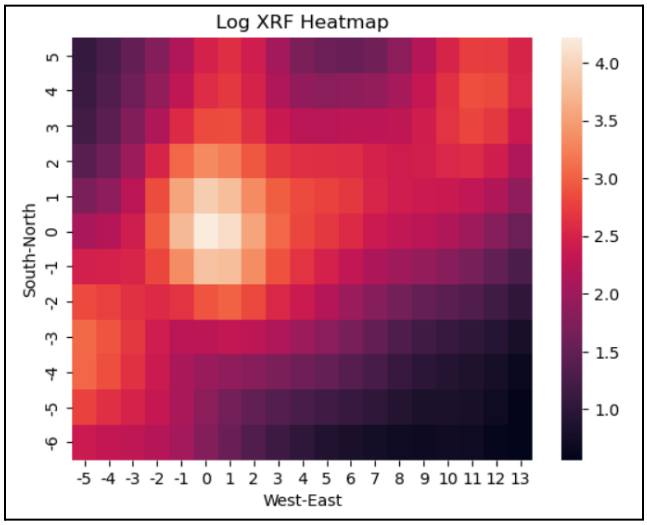
Graph 2

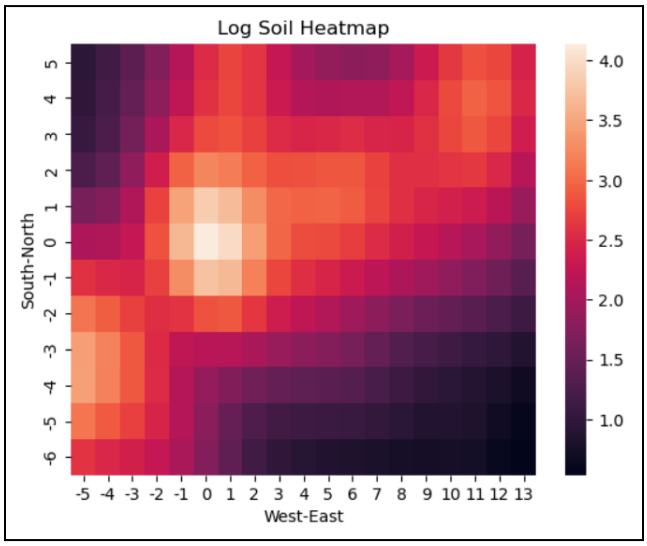


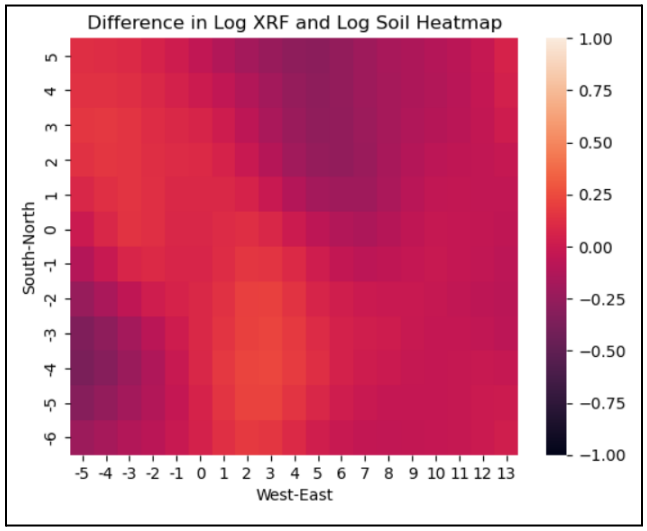
Graph 3


94.0% HDI
pred_: 0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
−1
0
1
2
3
4

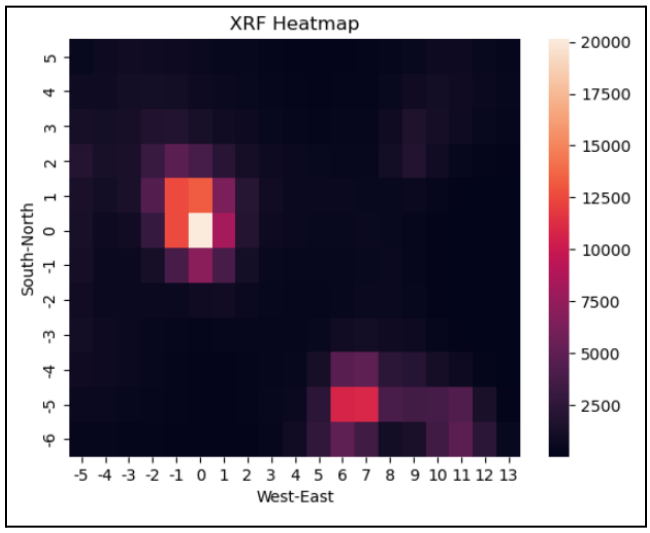
Graph 4: Matern52 - Log Scale

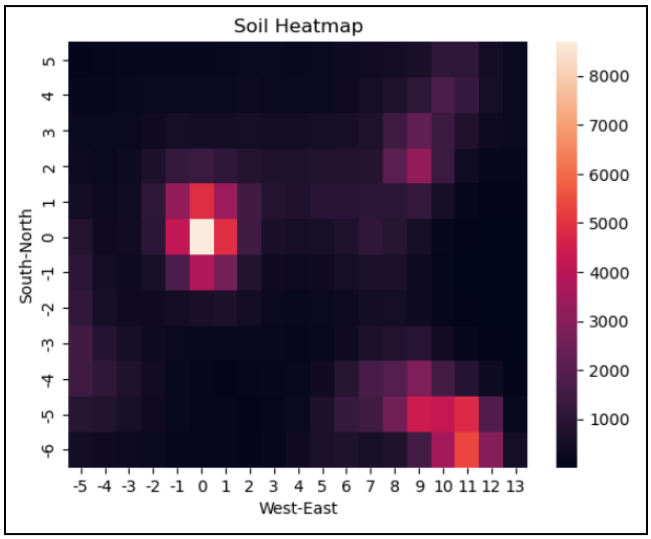


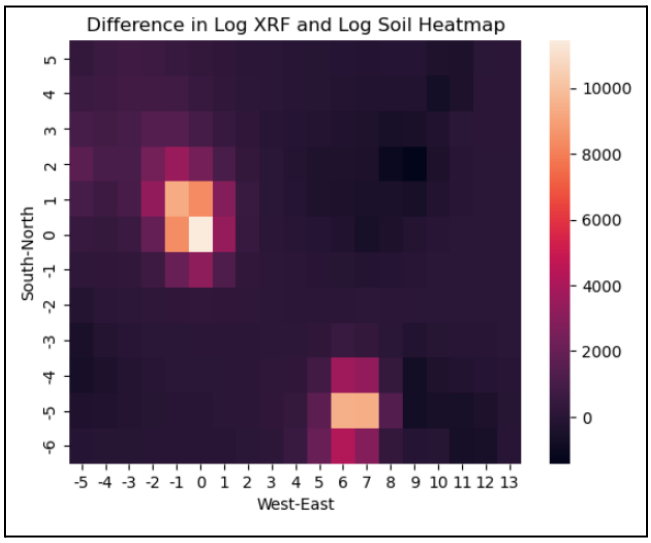




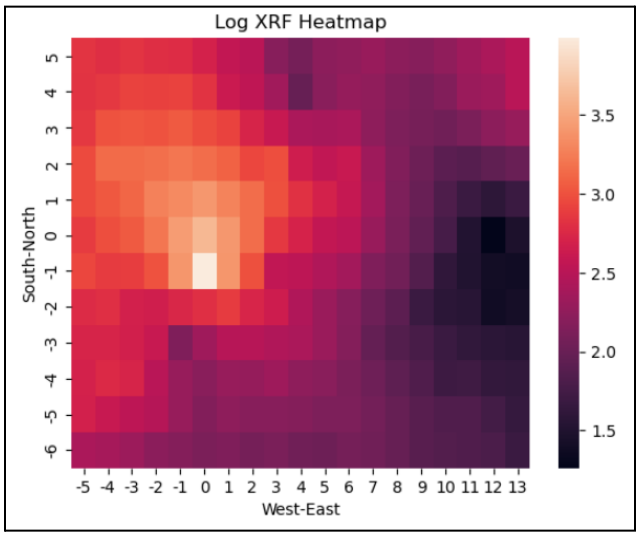
Graph 5: Matern52 Kernel - Original Scale

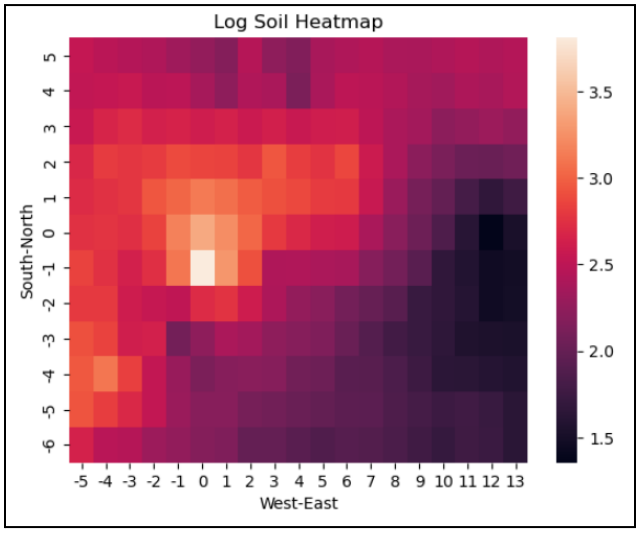


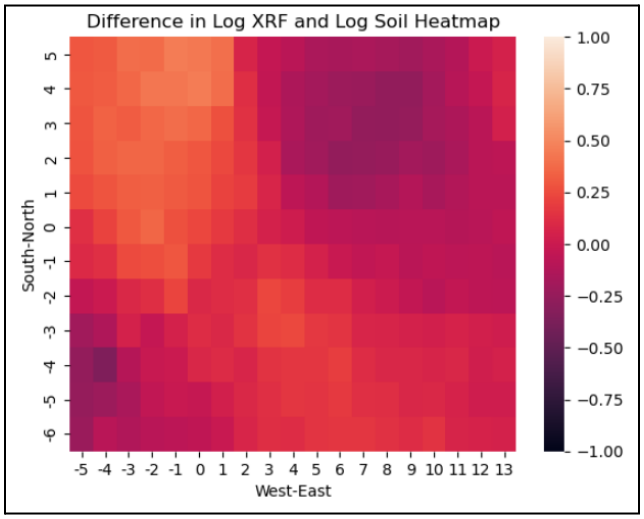




Graph 6: Exponential Kernel - Log Scale







Graph 7: Exponentiated Quadratic Kernel - Log Scale