Spatial Mapping of Soil Nutrients

Instructor: Dr. Mainak Thakur

Outline

- Problem Statement
- Dataset Description
- Deterministic Machine learning models for spatial Analysis
- Bottleneck of Traditional ML models
- Search for a better model in terms of speed and accuracy

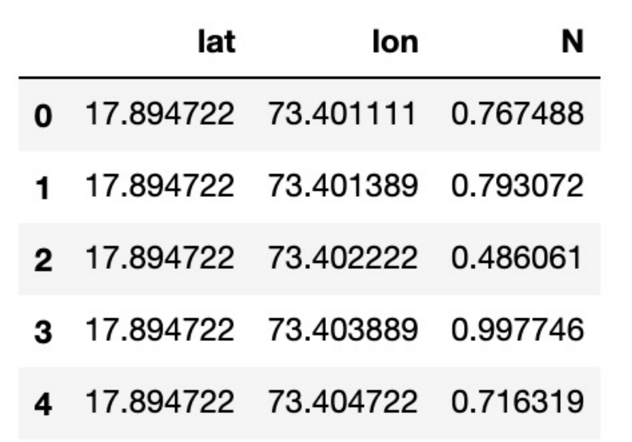
Problem Statement

- The knowledge of soil nutrients in a farmer’s land could help them use strategically plant crops and use fertilizers accordingly, but it’s infeasible to collect data at all points.
- Develop a Machine Learning model that can map the nutrients(N, P, K, OC) of a particular area.

Dataset Description

- Dataset consists of Soil nutrients of two districts (Pune & Ahmednagar) of Maharashtra.
- There are around **25k** data points for each of the following four nutrients.

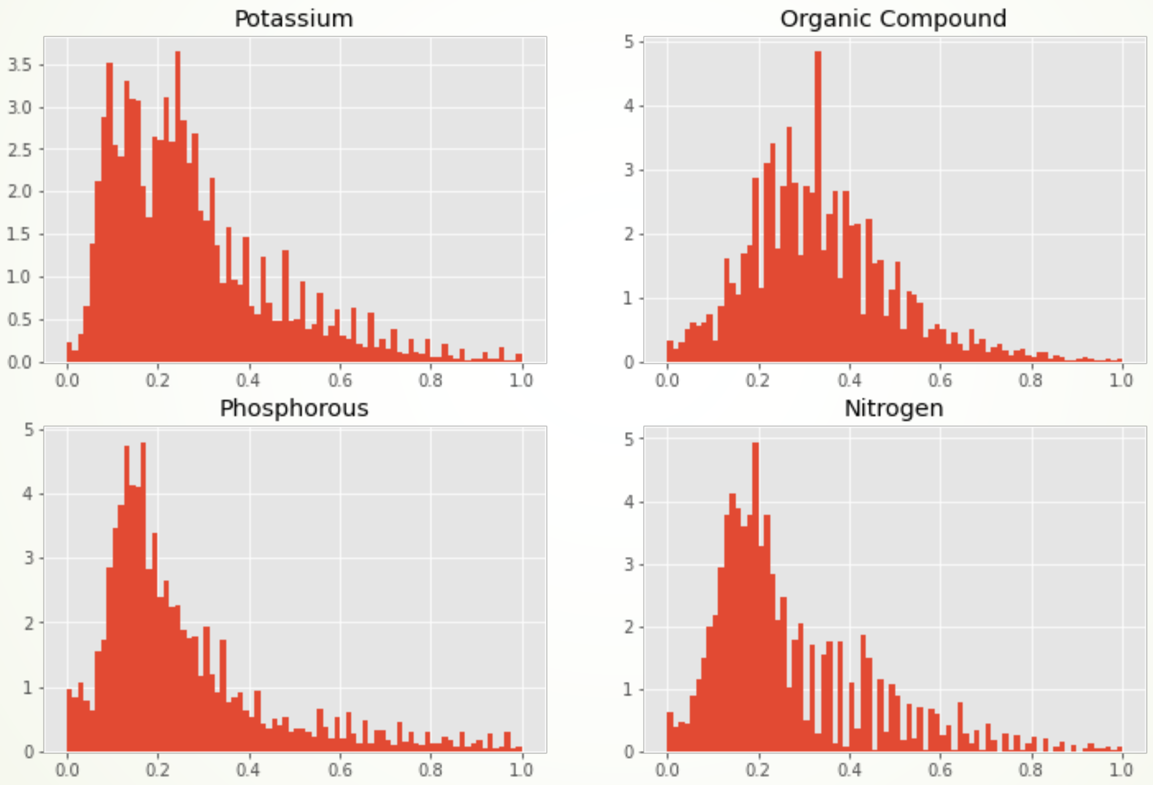
  - *Nitrogen, Phosphorus, Potassium* and *Organic Compounds*.
- CRS: Coordinate Reference System i.e., latitude and longitude.
- Datasets are normalized between 0 and 1.
- Nitrogen, Phosphorous, Potassium is measured in kg/ha and Organic Compound is measured in %unit.



	lat	lon	N
0	17.894722	73.401111	0.767488
1	17.894722	73.401389	0.793072
2	17.894722	73.402222	0.486061
3	17.894722	73.403889	0.997746
4	17.894722	73.404722	0.716319

Dataset source: https://soilhealth.dac.gov.in/

Distributions of Soil Nutrients



***Note**: Histograms are plotted after removing outliers and normalizing the data.*

Deterministic Models

Regression on coordinates

results for Nitrogen:

Dataset	Model	MAE	RMSE	R^2	
Train	Linear regression	0.12	0.16	0.08	
	Polynomial (deg=8)	0.10	0.14	0.26	
	SVR	0.11	0.15	0.15	
Test	Linear regression	0.12	0.16	0.07	
	Polynomial (deg=8)	0.11	0.15	0.25	
	SVR	0.11	0.16	0.13	

Why these models fail ??

- Assumptions behind the deterministic model is that if there is a physical process involved in an event, the outcomes can be best modeled by Deterministic processes. (Assuming that there is always a natural interference in all the phenomenon)

- Variability in spatial coordinates arises due to some random processes called Diffusion and diffusion is a random process. So, there is a meta-statistical argument that it is a stochastic process.

- One can realize that the distribution of soil nutrients merely depend on location and more on physical phenomenon and neighborhood.

So, How to apply statistics on these datasets?

Solutions:

- Feature Generation
- Model Selection

How to generate features?

- Split the dataset based on the regions.
- Data Augmentation.

What models to select?

- Model that can best model the stochastic process.
- Multivariate Analysis treating each point as an outcome of some random gaussian processes whose mean is given by the value of that nutrient.
- Plain geostatic models

Plain Geostatic Models

Kriging

It is analogous to regression but it takes the spatial covariance into account.

It is given by

$$\hat{z}(s_o) = \sum \lambda_i Z(s_i)$$

Where λ_i is calculated by covariance or semi-variogram.

$$\begin{bmatrix} C(\mathbf{s}_1,\mathbf{s}_1) & \cdots & C(\mathbf{s}_1,\mathbf{s}_n) & 1 \\ \vdots & & \vdots & \vdots \\ C(\mathbf{s}_n,\mathbf{s}_1) & \cdots & C(\mathbf{s}_n,\mathbf{s}_n) & 1 \\ 1 & \cdots & 1 & 0 \end{bmatrix}^{-1} \cdot \begin{bmatrix} C(\mathbf{s}_0,\mathbf{s}_1) \\ \vdots \\ C(\mathbf{s}_0,\mathbf{s}_n) \\ 1 \end{bmatrix} = \begin{bmatrix} w_1(\mathbf{s}_0) \\ \vdots \\ w_n(\mathbf{s}_0) \\ \varphi \end{bmatrix}$$

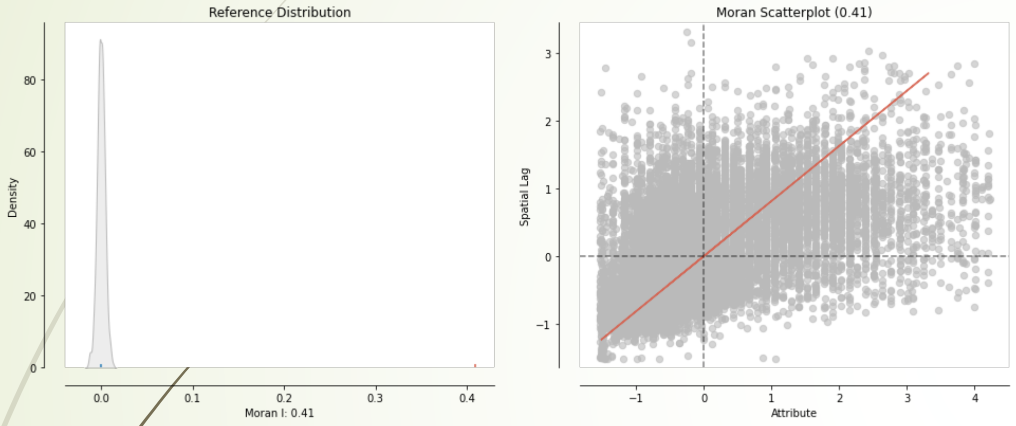
Source: A practical guide to geostatistical mapping by Tomislav Hengl

Kriging assumptions

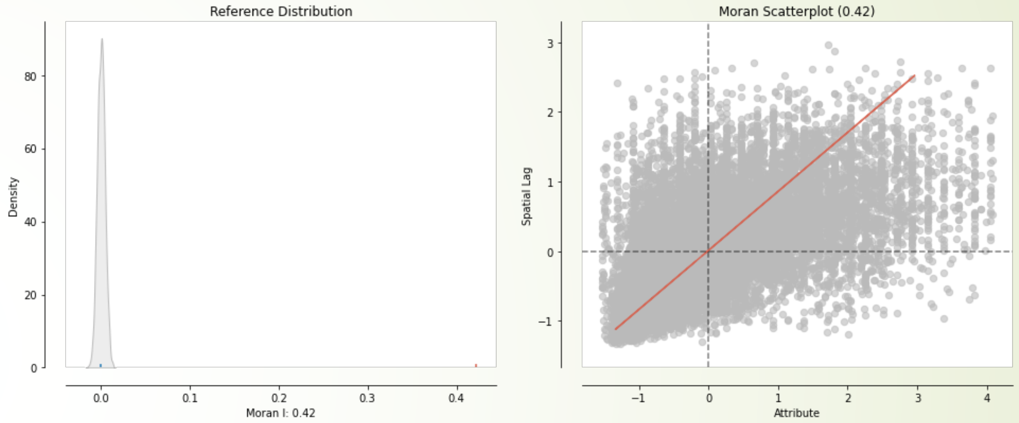
- Isotropy - spatial data only depends on the distance separating them.
- Spatial autocorrelation
- Stationarity - mean and variance is constant across the spatial field.

Autocorrelation (N, K, P, OC)

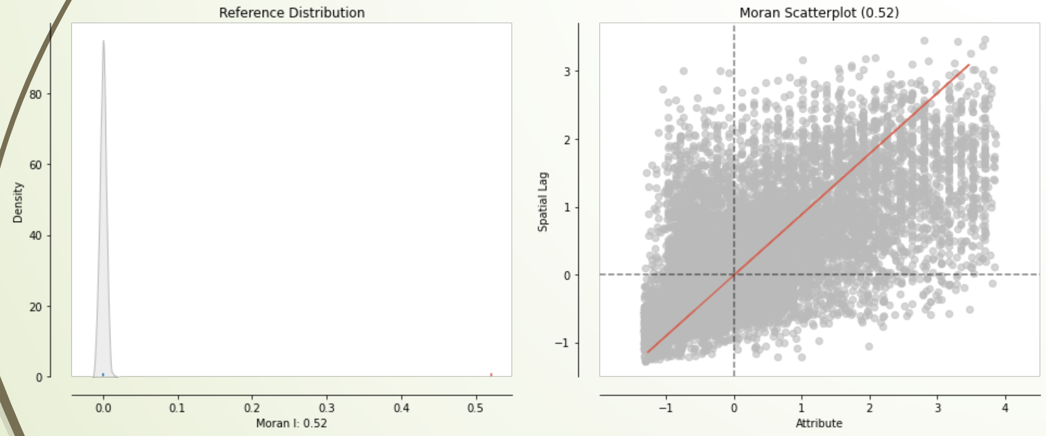
P = 0.001



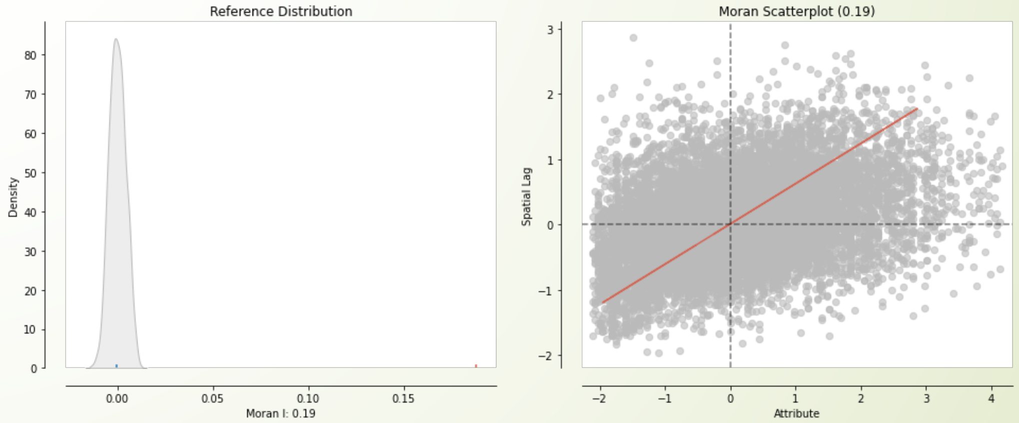
P = 0.001



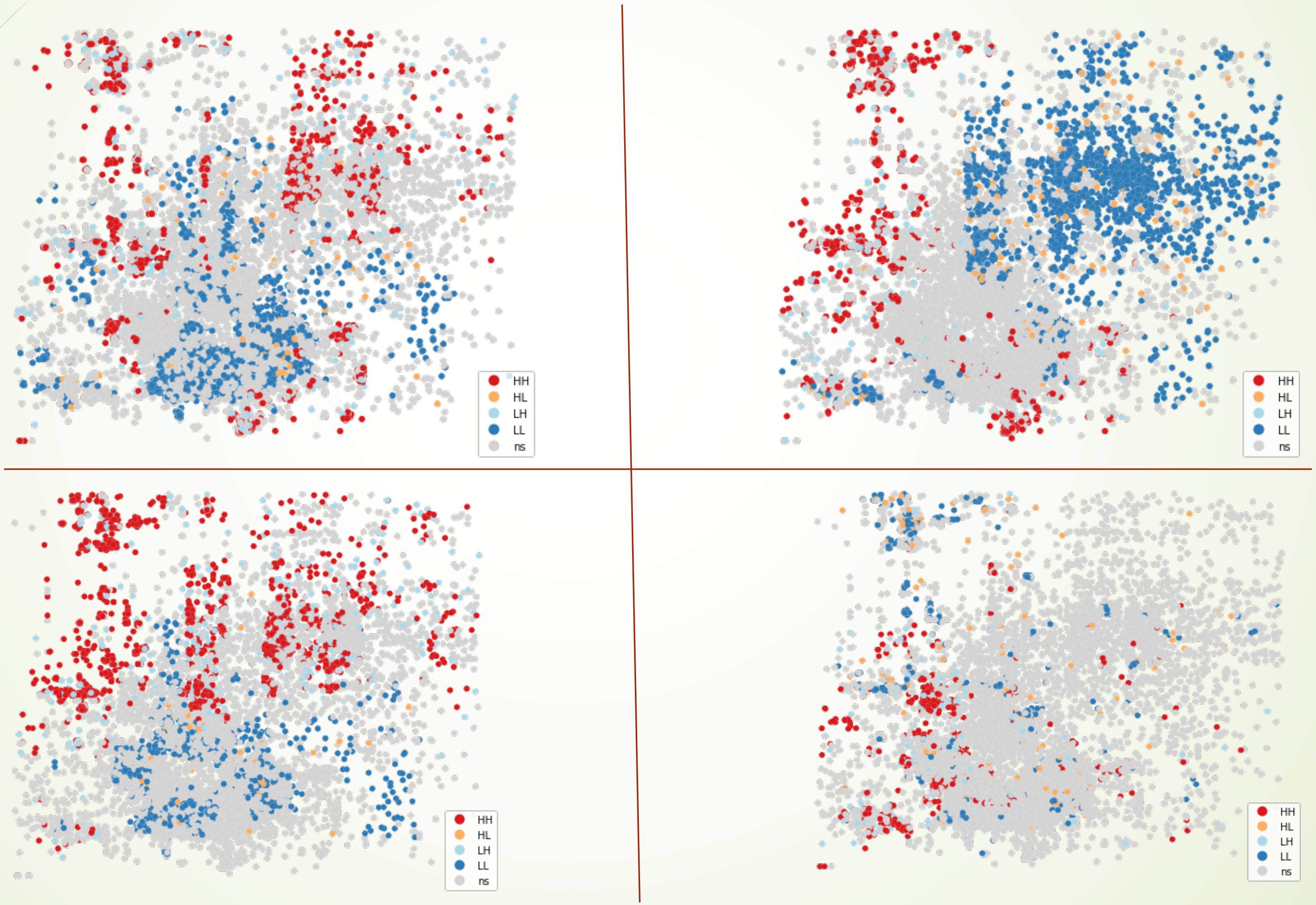
P = 0.001



P = 0.001



Local Autocorrelation (N, K, P, OC)



Stationarity (N, K, P, OC)

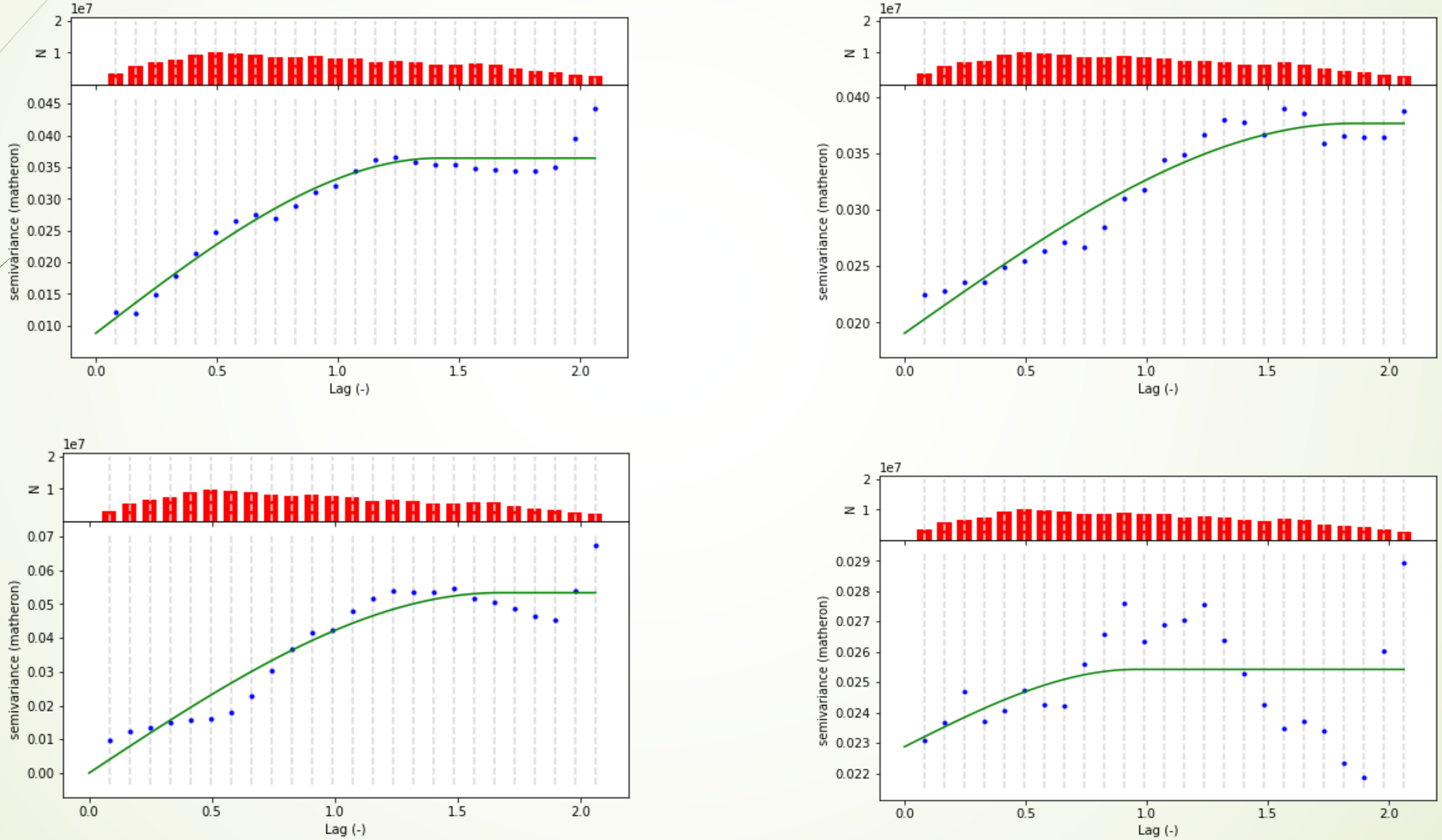


Fig: Semi-variogram

Kriging Result on test set

Nutrient	MAE	RMSE	R^2
Nitrogen(N)	0.09	0.14	0.37
Potassium(K)	0.09	0.13	0.40
Phosphorous(P)	0.09	0.14	0.41
Organic Compound(OC)	0.11	0.14	0.14

Kriging Interpolation (N, k, P, OC)

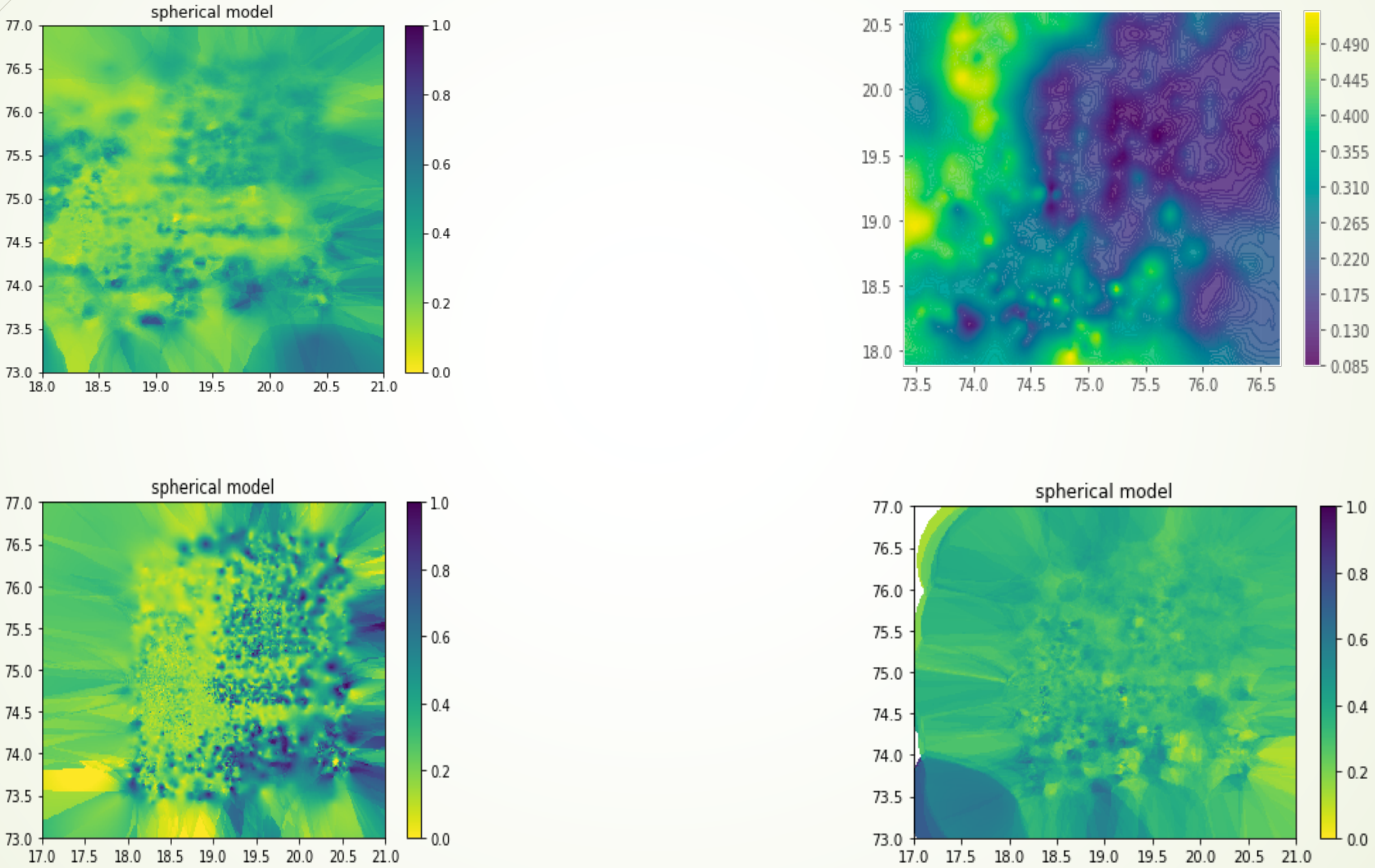
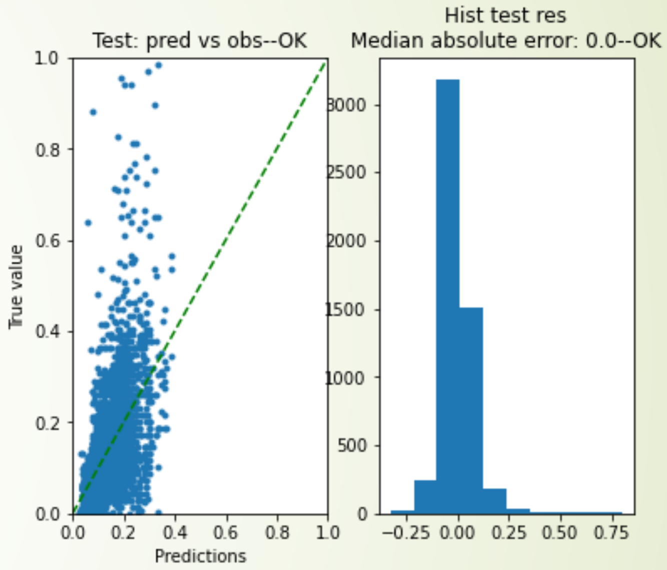


Fig: Kriging interpolation on 500x500 grid (neighborhood points)

Observations

- Though it was able to capture most of the relations, it still fails to acknowledge all the variability among the neighborhood points.
- Kriging is computationally very expensive operation.
- Nutrients like OC has negligible stationarity, that doesn’t suit the assumptions of kriging interpolation.
- Also, it can be observed by looking at the residual that this model does not generalize well and also the residual distribution is highly spread around the mean.



What’s next?

Let’s try Feature Generation techniques on Deterministic model.

Feature Generation

- Append 100 nearest neighbor points and corresponding distance along the columns.
- This assumption is based on the notion that the nearer points are likely to be correlated with each other than the distant points.
- It was observed by the other researchers that nearest neighbor feature generation can capture better spatial relationship than simple coordinates based regression.

Deterministic Models (After augmentation)

Regression on coordinates

results for Nitrogen:

Dataset	Model	MAE	RMSE	R^2
Train	Linear regression	0.09	0.13	0.38
	SVR	0.09	0.13	0.38
	KNN Regressor	0.08	0.12	0.46
Test	Linear regression	0.09	0.14	0.37
	SVR	0.09	0.14	0.36
	KNN Regressor	0.10	0.14	0.31

What’s next now?

Still not the best fit. Huh!!!! No relief yet.

Let’s look closely at the scatter plot of the data.

- Data points are widely scattered over the entire region.
- Multiple clusters is there, means one particular distribution is not enough to fit all the data.

Starter-

Is there a model that can fit the data in such a way that each realization of a stochastic process is a random variable of a gaussian distribution centered at that point.

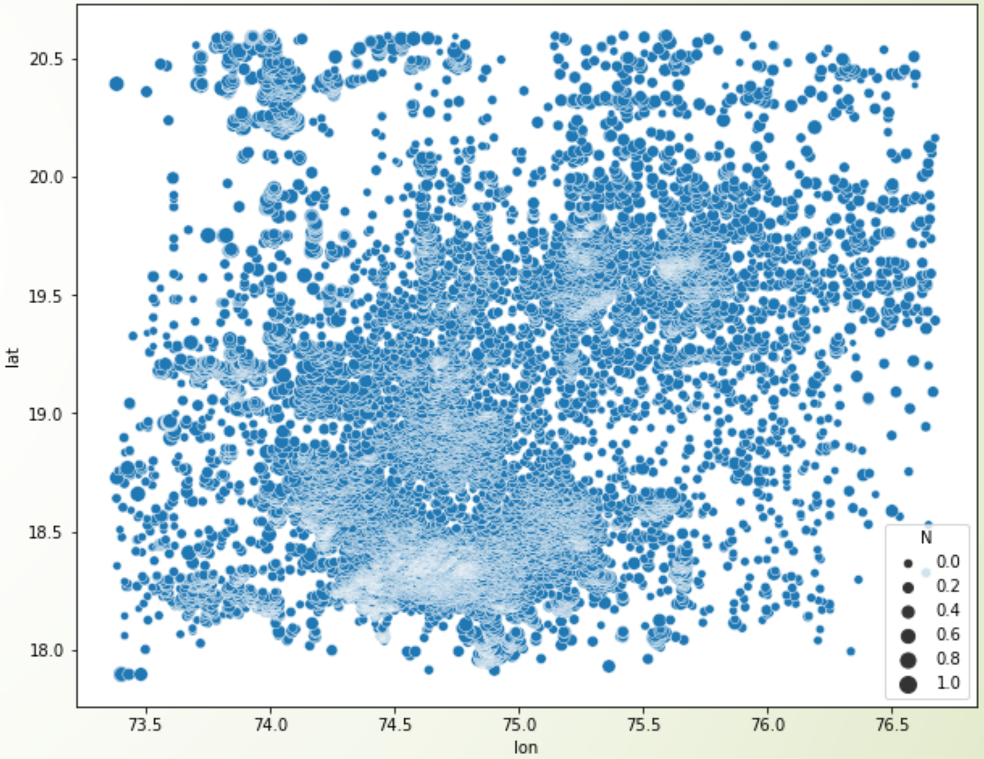


Fig: Scatter plot of Nitrogen(N)

Thinking statistically

- Once such process is **Gaussian process**. It assumes that any set of real valued random variable is clustered around it's mean (multivariate gaussian) and there can be n such clusters. In simple word, every finite set of those random variables has a multivariate gaussian distributions.
- It is a generalization of univariate normal distributions in higher dimension.
- It involves the calculation of covariance matrix. Nearer points have high covariance than the farther points and is measured by some similarity function.

Gaussian Process

Talking feasibility

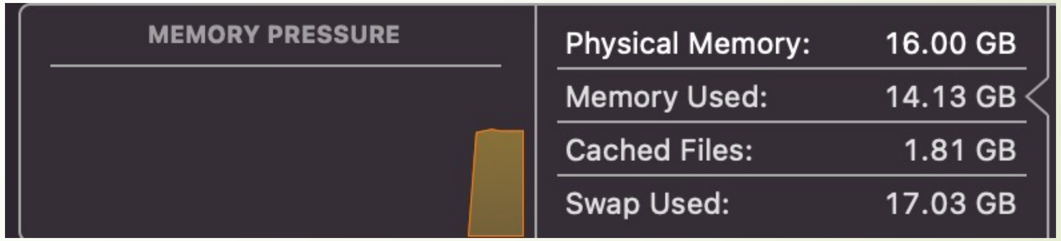
Gaussian process is feasible for smaller dataset but in case of large number of samples, it can lead to memory overhead.

For the Nitrogen dataset, shape of the covariance matrix would be (25000x25000) and each element is 4 byte.

Total memory = 25000*25000*4 = 2.5e9

= 2500GB

Not feasible for a normal computer .



Physical Memory: 16.00 GB
Memory Used: 14.13 GB
Cached Files: 1.81 GB
Swap Used: 17.03 GB

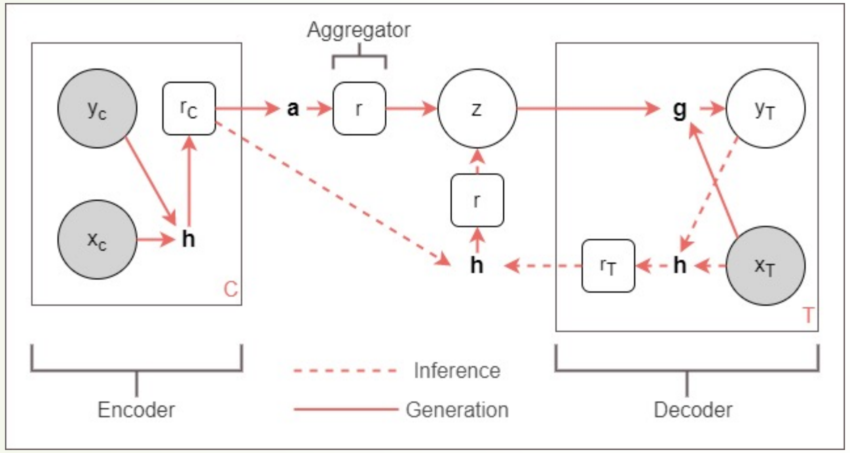
Tried the risk with a small sample, it went on for hours without any result😔

Deepmind, the saviour

- Google’s Deepmind has done amazing works in the field of machine learning.
- In 2018, They developed a neural network based gaussian process model known as Neural process.
- It is very much computationally feasible and can leverage the GPU for computations.
- Can also measure the non-linear relationship among the gaussian distributions in higher dimensional space.

Neural Process

Computational model



Graphical outline of NP

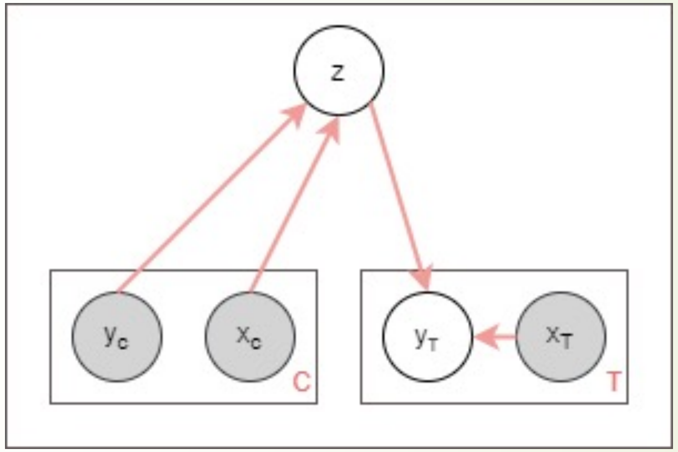
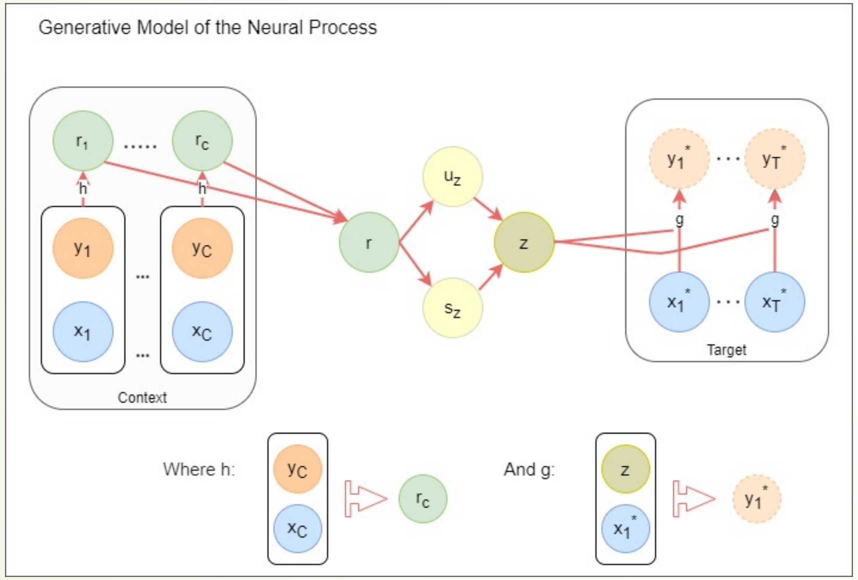


Image Source: Garnelo, M., Schwarz, J., Rosenbaum, D., Viola, F., Rezende, D. J., Eslami, S. M., & Teh, Y. W. (2018). Neural processes. arXiv preprint arXiv:1807.01622.

NP Computational Model

Generative model



Inference

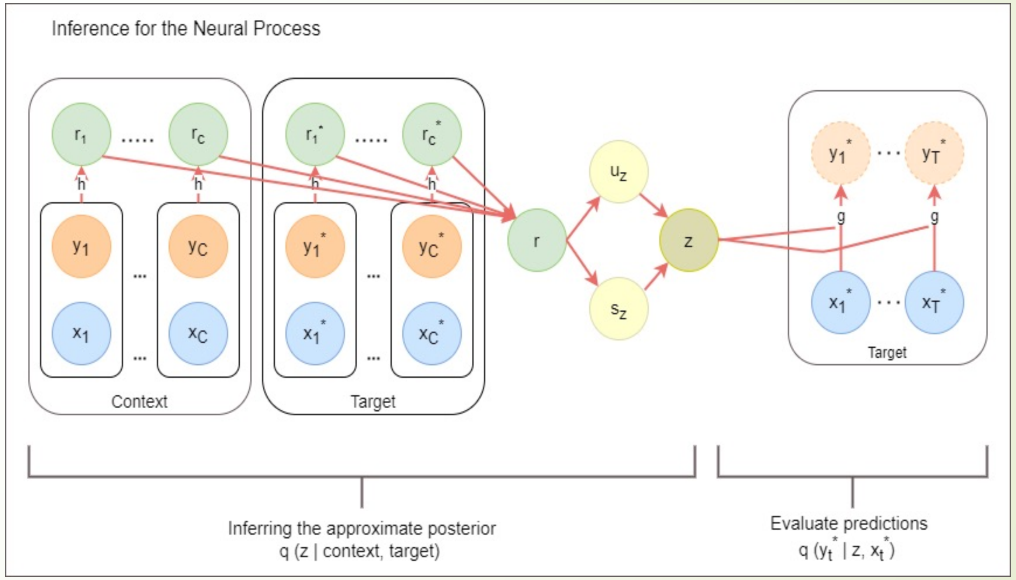


Image Source: Garnelo, M., Schwarz, J., Rosenbaum, D., Viola, F., Rezende, D. J., Eslami, S. M., & Teh, Y. W. (2018). Neural processes. arXiv preprint arXiv:1807.01622.

Working Principles

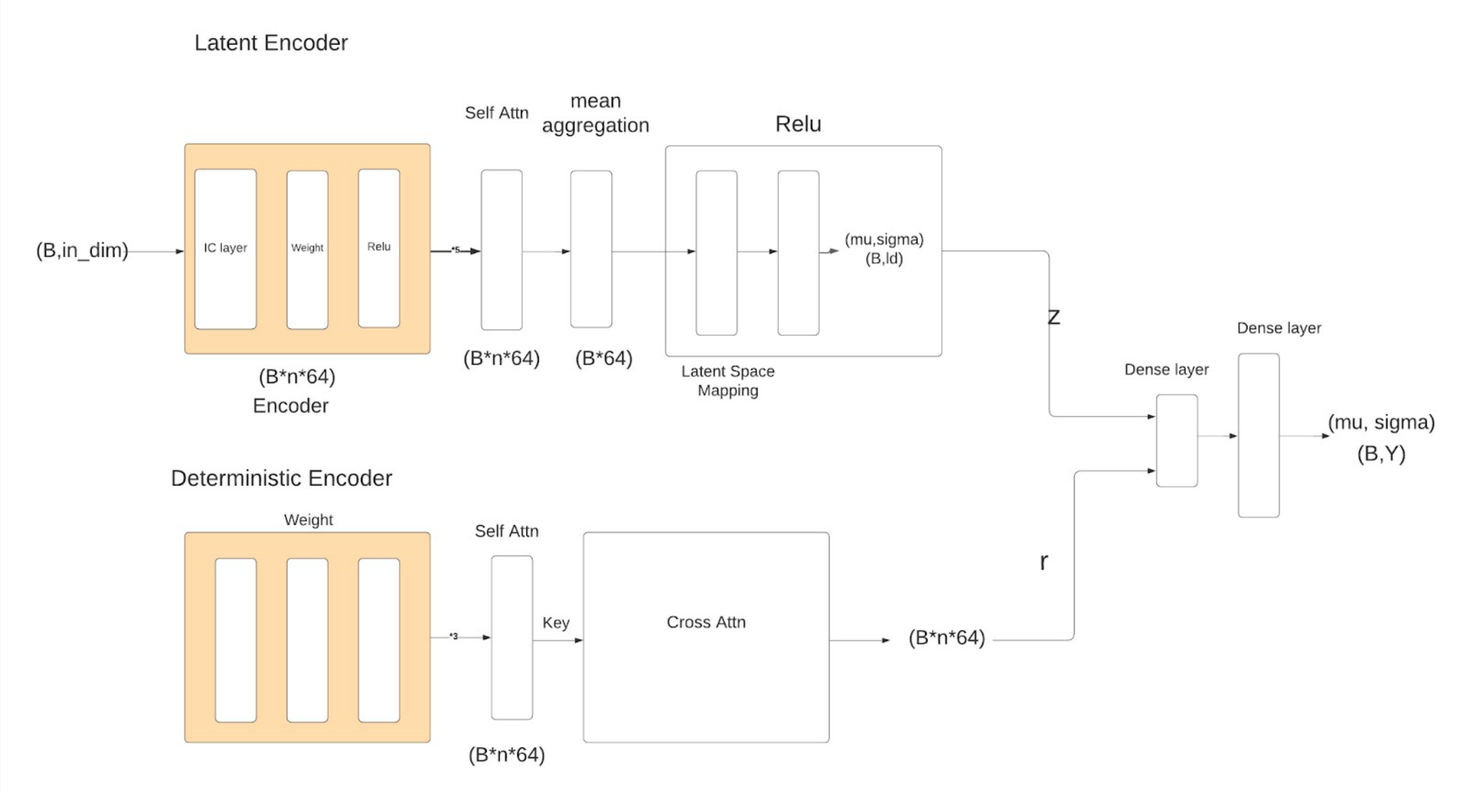
Generative model

- First, x_C and y_C are mapped to r_C through neural network h.
- r_C is aggregated to parametrize global latent variable z .
- To obtain y* , z is concatenated to x* and mapped through neural network g to obtain y*.

Inference

- Inference is carried out in variational Inference Framework (Amortized VI) .
- instead of optimizing p(z) directly, a parametric function is introduced which maps from observation space to approximate posterior distribution.
- Since the exact posterior distribution $p(y_{1:n} | x_{1:n})$ is not known to us, we approximate the parametric distribution q(z |.) over variational parameters of latent variable z. Its evidence lower bound(ELBO) is given by
- **ELBO =** $E_{q(z|C,T|)}\left[\sum_{t=1}^{T} \log p(y_t^* \,||\, x_t^*) + \log \frac{q(z|C)}{q(z|C,T)}\right]$

Proposed Architecture



Results

Result for Nitrogen(N) on test set

	MAE	RMSE	R^2
Before Augmentation	0.09	0.14	0.37
After Augmentation	0.04	0.08	0.78

😳

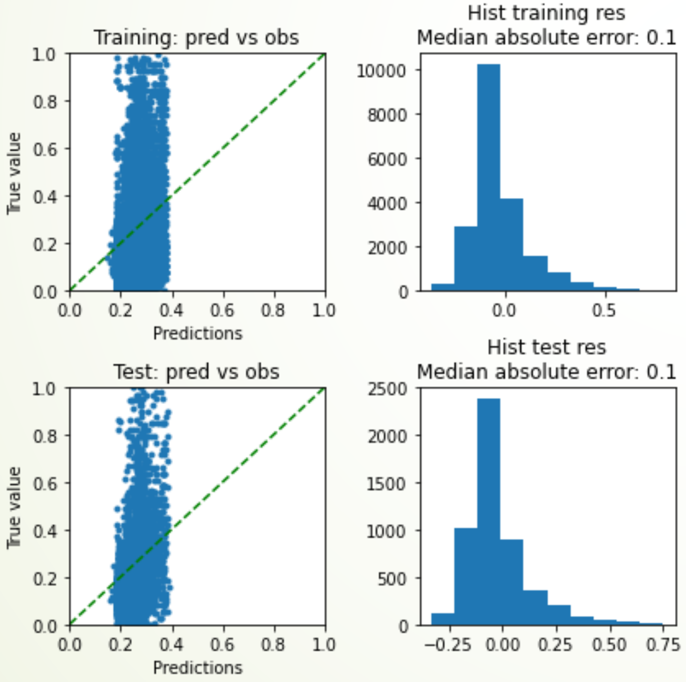
Results for other nutrients after augmentation on test set

Nutrients	MAE	RMSE	R^2
Potassium (k)	0.05	0.07	0.68
Phosphorous (P)	0.03	0.06	0.83
Organic compound	0.05	0.08	0.66

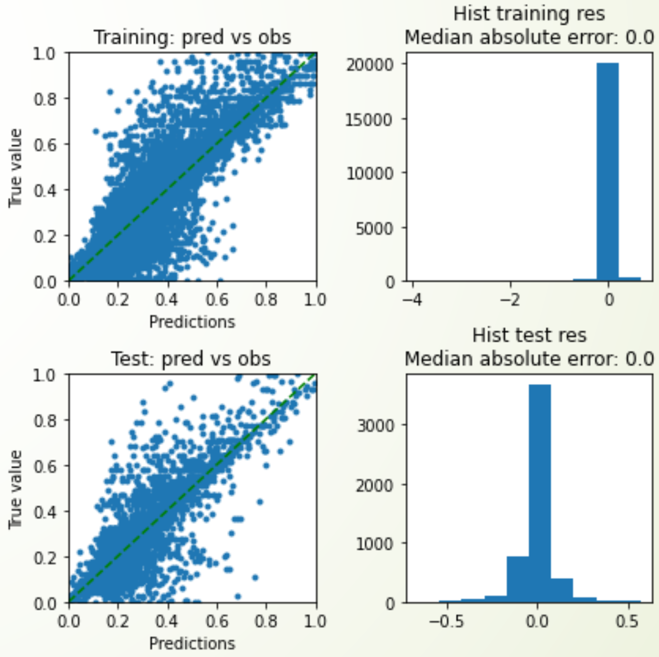
Residual Plots for NP

Residual plot for Nitrogen (N)

Before Augmentation

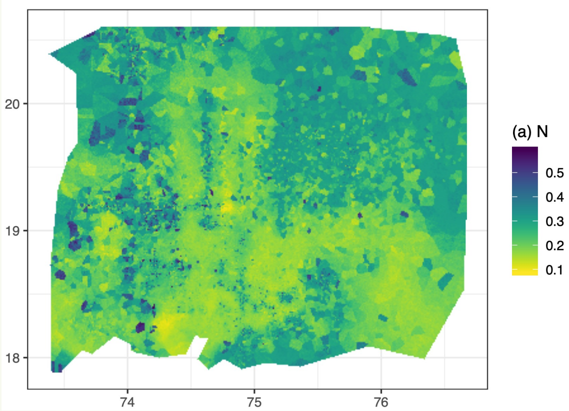


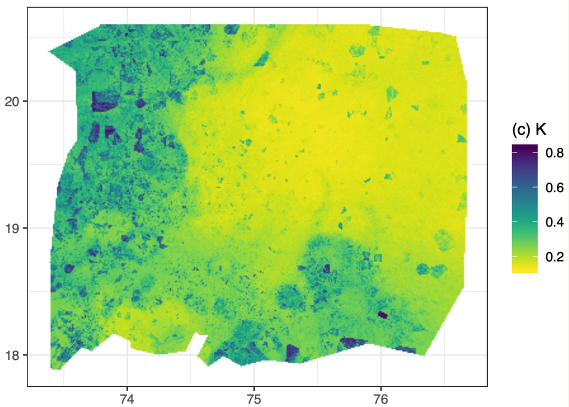
After Augmentation

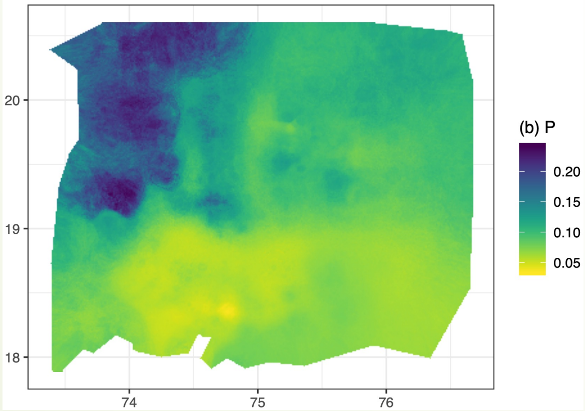


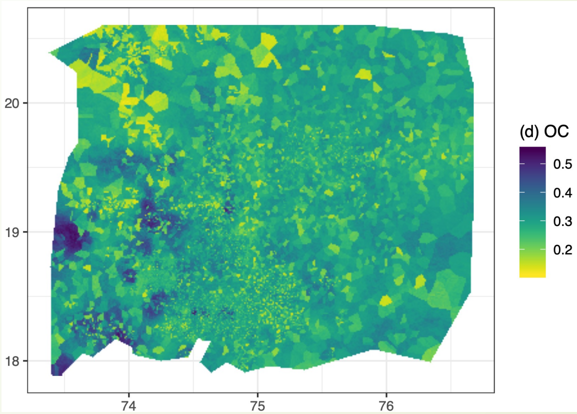
Final Spatial Map (N, K, P, OC)

Map plotted in R (by overlaying the shapefile for the region of study)









References

- Garnelo, M., Schwarz, J., Rosenbaum, D., Viola, F., Rezende, D. J., Eslami, S. M., & Teh, Y. W. (2018). Neural processes. arXiv preprint arXiv:1807.01622.
- Garnelo, M., Rosenbaum, D., Maddison, C., Ramalho, T., Saxton, D., Shanahan, M., ... & Eslami, S. A. (2018, July). Conditional neural processes. In International Conference on Machine Learning (pp. 1704-1713). PMLR.
- Oliver, M. A., & Webster, R. (1990). Kriging: a method of interpolation for geographical information systems. International Journal of Geographical Information System, 4(3), 313-332.
- Fu, W., Tunney, H., & Zhang, C. (2010). Spatial variation of soil nutrients in a dairy farm and its implications for site-specific fertilizer application. Soil and Tillage Research, 106(2), 185-193.
- Das, K., Mandal, S., & Thakur, M. High Resolution Spatial Mapping of Soil Nutrients Using K-Nearest Neighbor Based CNN Approach. In IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium (pp. 1102-1105). IEEE.
- https://www.youtube.com/watch?v=4vGiHC35j9s&t=2702s

Thank You