Geospatial Big Data Competition

November 17, 2020

1 Timeline

Competition opens Monday, Nov 23rd , 2020 @ 10:00 am and closes on Dec 9th, 2020 @ 11:59 pm (KSA time)

2 Registration

Please fill the registration form and submit.

3 Motivation

Spatial datasets have received much attention in the recent past. With the explosion of spatial data coming from different sources such as sensors, maps, climate informatics, smartphones, and others, accurate and timely analysis of these big data has become a necessity. Various tools from different fields are used to handle and manage spatial data. They provide a better understanding of the nature of the data and the underlying data model.

In Geostatistics, the Maximum Likelihood Estimation (MLE) function is a common tool used to model the spatial data. It provides a set of parameters to describe the model. This set of parameters can be used to predict missing values on given locations. With recent trends in the uptake of machine learning techniques applied to different data structures,

we believe that they may have a role in improving modeling of the spatial data in different forms.

The goal of this competition is to provide a platform to benchmark modeling techniques for Geospatial Big Data with traditional as well as non-traditional methods, including but not limited to Machine learning, statistics, earth science, etc., and analyze spatial data in a uniform manner to guarantee a fair comparison and assessment. With the aid of the ExaGeoStat software (https://github.com/ecrc/exageostat), we have generated a set of synthetic Geospatial datasets for this competition.

4 Datasets

This competition focuses on two parts (i.e., competition I, and competition II). Each will be treated independently. Based on data sizes, competition I includes small-size datasets (100K), and competition II includes large-size datasets (1M). Teams have the choice to participate in competition I and/or competition II.

All datasets cover univariate variable indexed at a two-dimensional space. Using ExaGeoStat, we have generated datasets with different properties, 18 for small datasets (100K) and 18 for large datasets (1M). Each dataset has been randomly divided into 90% training data and 10% testing data. The participant should train his/her model using the training dataset and predict missing readings in the testing locations (i.e., testing dataset).

5 Assessment

The Mean Column-wise Root Mean Squared Error (MCRMSE) is used to assess prediction accuracy. The MCRMSE is simply the average of the Root Mean Square Error (RMSE) computed for each of the D datasets,

$$MCRMSE = \frac{1}{D}\sum_{d=1}^{D} RMSE_d$$

The RMSE for a given dataset d is,

$$RMSE_d = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{Z}_i - Z_i)^2}$$

where n is the number of testing locations, and $\hat{Z}(\boldsymbol{s}_i)$ and $Z(\boldsymbol{s}_i)$ are respectively the predicted true realization values at the prediction location $\boldsymbol{s}_i$ in the testing dataset.

The assessment will separately be carried out for each competition. Teams should submit their RMSE for each of the datasets using a single CSV file with 10k (100K) rows and 18 columns where each column contains the predictions for one of the 18 testing datasets. This CSV file should be uploaded to the Kaggle submission form to able to compute the MCRMSE for each competition. An example of the submitted CSV file can be found here. The final rank for each competition will be assigned by sorting the MCRMSE values in ascending order.

6 Submission deadline

Kaggle will stop accepting submissions on December 9th at 11:59 pm (KSA time).

7 Pre-requisites

Registrants are expected to have at least an intermediate level of understanding and expertise in Machine Learning/Deep Learning or Statistical methods.

8 Rules

- All the teams from different backgrounds are able to participate if they can provide the sought output. The two competitions will have separate rankings.
- We will use Kaggle leaderboards to rank the submissions. Details will be provided to registrants in communications to follow the registration process.

- Datasets are confidential and it is only available through this competition, the participants are not allowed to distribute without our permission.
- It is recommended that you submit results for all the datasets of a competition. Not submitting a result would result in auto-filling with 0s and will affect your ranking for that competition.
- The top 3 teams will be requested to submit their solution, the code along with information about the compute hardware with the competition hosts. By submitting the registration form, you will be agreeing to provide your code if you rank top on the leaderboard at the end of the competitions.
- Methodology and results will be included in a public report on the competition. By submitting the registration form, you will be agreeing to the inclusion of the methodology and name of participants in such a report.

9 Compute Infrastructure

We will be providing a finite amount of computing resources on one of our HPC clusters with GPU support. You are free to use your own computing infrastructure for your solution or work on the one we allocate (fairness of use will apply). Some training on accessing and using compute infrastructure and software resources will also be provided.

10 Contact

If you have any question about this competition, you can contact us at kaustcompspatml@gmail.com.

11 Rewards

The winners of each competition will be awarded a prize during the closing ceremony of the event. We will be publishing the outcome of the competition in a report which will also include the methods used for the top 3 ranks.