March 9, 2018

Partner Report - Milestone 2

Keyan Halperin, Christopher Hase, Justin S. Lee, Casey Meehan
Harvard IACS

Dear Christine and Ben:

As it is now halfway through the Spring 2018 semester, we want to give you a brief update on our work in the IACS Capstone pollution prediction project. We have included a short summary of work done, the current state of our experiments, and an overview of our path going forward for the rest of the semester.

In our last report to you delivered on February 20, 2018, we explained our predicted scope of work for the project, with an expected timeline and deliverables. By the current milestone, we expressed our intention to obtain satellite data for areas of the United States without sensors, run existing models on Harvard's Odyssey computing cluster, and prototype imputation methods and prediction models.

We were made aware in previous correspondence, as well as in discussions with Qian Di, that obtaining complete, relevant satellite data for the entirety of the continental United States would require a significant amount of time. Therefore, we have been focusing on predictive models for existing satellite locations and will consider geospatial interpolation after we have considered the former problem.

Additionally, since our last report, we have (1) performed an exploratory data analysis on the sensor data we were given, and explored new models/sources of data, (2) tested various imputation methods, and (3) began setting up our compute resources on Odyssey and started thinking about the project from a software product perspective. Below, we go into more detail on our work in each of these areas.

I: Current Stage

Per our discussions, we are focusing on advancing PM2.5 prediction models. To this end, we are focusing on 3 primary facets.

1. EDA and testing new prediction models: We have examined the model implemented in Qian Di's original paper, and have spoken with Ben about some new attempts. To add to this work, we are testing a variety of other prediction models, such as RNN's and additional convolutional layers to see if we can capture additional time and space information. Additionally, we are experimenting with including non-satellite data (e.g. U.S. census data obtained courtesy of Pavlos Protopapas, the course instructor) to see if it helps increase prediction accuracy.

2. Data imputation: There is a high degree of missingness in the satellite and sensor data on a daily level. In a random 0.5% sample we took of the data, nearly 80% of the response variable were missing. As such, we are testing a variety of imputation schemes to make the data more robust, such as random forest imputation.

3. Compute Resources/Making a useful modular software framework: We have set up accounts on Odyssey and have familiarized ourselves with how to run compute-intensive jobs on the server. Ultimately, we want to make our work into a useful tool for your team. We are trying to design our code in a way that is flexible and extensible for you. Ideally, our models will be able to incorporate any new data you wish to use, and allow your team to experiment with the its key parameters.

We have attached our mid-semester presentation with some more technical detail for reference. Please let us know if you have any questions.

II. Planned Next Steps

Going forward, we plan on continuing to explore and implement more imputation methods and prediction models. With respect to imputation methods, we want try to improve upon our iterative random forest methodology and will consider methods such as multiple imputation chained equations (MICE). In terms of models, recurrent and convolutional neural nets may be able to capture complex temporal relationships for each sensor. Also, convolutional neural nets may be able to capture complex relationships between nearby sensors.

Additionally, with the help of Odyssey/AWS, we plan on implementing these procedures on the full satellite/sensor data set. Our preliminary results have indicated that our model performance should increase significantly with the inclusion of additional data. There are some concerns about the potential computational limitations of some methods, but we expect to have a much better understanding of what is computationally feasible after some preliminary trial runs.

We hope that this provides a satisfactory overview of our progress in the project. If you have any questions, concerns, or suggestions, please feel free to reach out anytime. Thank you.

Sincerely,

Keyan, Chris, Justin, and Casey