Comparison of Machine Learning Methods to Predict the Air Quality Impact of Wildfires

Gautam Agarwal, Jack Cai, Collin Frink, Brian Hu, Eliot Kim, Shreyans Saraogi
UW Madison ProjectX Deliverable 2, November 10, 2020

1 Explored and Final Methodology

Currently, an efficient model development and evaluation pipeline has been iterated on and finalized. This enables timely and important changes to the datasets and learning algorithms as necessary. For this submission, only preliminary results using data from 2019 for each algorithm have been finalized. Implementing these smaller models allowed for quicker important decision making regarding both the dataset (feature selection, spatial resolution, etc.) and model development (algorithm selection, initial hyperparameter tuning, etc.).

Preliminary results for the persistence, linear regression, boosted regression trees, support vector machine, artificial neural network, and recurrent neural network algorithms have been completed. Given that they have only been trained on a small subset of the total data, their current results are more of a proof of concept rather than actual usable results. One model we have yet to successfully implement is a convolutional neural network. For this model, the data has to be processed differently: instead of being flattened, it must have a 4D shape: (timestep, grid_x, grid_y, features). On this unflattened data, PCA cannot be performed to reduce the dimensionality of the data. Thus, training models will take significantly more time and computing resources for this data. Due to time constraints, this data has not been produced and processed and the convolutional neural network has not yet been built. However, given the format, complexity, and size of our data, a CNN has the potential to outperform the previously listed algorithms and remains a priority for the next iteration of models.

Another aspect of the preliminary results submitted in this deliverable that can be improved upon is the format of the air quality data. Currently, the data is taken from ground monitoring stations located across California. While this facilitates the easy generation of the output air quality labels, the locations of each station are relatively sparse and do not densely cover the region of interest. Instead of using these stations for air quality measurements, then, the next dataset iterations use satellite air quality data products that provide far greater spatial resolution and coverage. Also, the specific satellite dataset (MODIS Aerosol Product) which will be used allows prediction of air quality measures that have not been extensively covered in the literature, such as aerosol optical depth (AOD) as discussed in Deliverable 1. Thus, gridded satellite air quality data will improve the utility of the air quality output feature and enable the implementation of more complex learning models.

In regards to model implementation, a number of different learning algorithms, data structures, features, and evaluation methods have been explored. Many of these have been attempts to solve the difficulties of using supervised learning algorithms on time-series data. One such difficulty is the continuity of time-series data. Within our spatial region of interest, the fire season lasts roughly from June to October each year. In order to apply supervised learning to time-series data, a "sliding window" must be created by taking a set number of previous time steps from the original dataset as a feature, and using the next time step from the labels as the label for this "feature". This can be expanded to multi-step prediction by using multiple future time steps as the label as well. To mitigate this, a couple of different overarching model structures were considered. One approach would be to split up the different fire seasons, train separate models (for each algorithm) on a single fire season, and use a cross-validation framework to evaluate the best model. Another would be to use a weighted (either statistical or machine learning-inspired) combination of the predictions from each model. Both of these methods, however, are unlikely to fully utilize the information from the dataset. So, the chosen approach is to use data from the whole year for training and simply concatenate each year to create a continuous time series.

The main focus of this paper is to compare the effectiveness of predicting air quality of naive predictors, traditional supervised learning methods, and deep learning methods. Final model selection was inspired by existing literature, as each of the listed supervised algorithms have shown success in modeling wildfire-related data: linear regression, random forests, boosted regression trees, support vector machines, and artificial neural networks [5][10][11][12][14][15][18]. Statistical methods (naive persistence and ARIMA) were included to incorporate a baseline evaluation metric. For the deep learning models, a convolutional neural network and a recurrent neural network were chosen because they both particularly suit the nature of our data. Recurrent neural networks are inherently designed to work on time series data, while convolutional neural networks are able to find relationships between data points and their neighbors and are thus suited to both time-series and grid-based data. Other models that have been considered include clustering methods (K-means, mean-shift, K-nearest neighbor), VARMA, and gradient boosting. Ultimately, supervised learning models were chosen because as reference

2 Papers Consulted

The literature review process began with the goal of understanding the detrimental effects of larger and more frequent wildfires on anthropogenic climate change. Prior research has indicated that weather patterns are leading to more frequent wildfires, and these weather patterns are greatly influenced by human actions that directly impact climate change [4][6][13]. Further review indicated that the increasing trend of fire activities is strongly correlated with anthropogenic increases in temperature and vapor pressure deficit, which significantly enhanced fuel aridity, a major factor behind forest fires [7][9].

After gaining sufficient background knowledge about wildfire research, a thorough literature investigation about the application of machine learning to predict air quality impacts of wildfires was conducted. Here we have listed and briefly discussed a few of the most impactful papers.

Yao et al. utilizes satellite and terrain data to train random forests for predicting wildfire-induced smoke column height [15]. The datasets and data manipulation techniques were especially useful, because they used similar data products and had detailed explanations of their procedure. An interesting facet of this paper was their almost exclusive focus on fire-related variables and the omission of many meteorological factors.

Reid et. al. provides an example of comparing several machine learning models which take fire locations, air quality, land cover, and meteorological variables as inputs to predict PM2.5 concentrations [10]. This paper provided a solid foundation upon which to structure the project. The cross-validation procedure described in this paper was especially helpful in developing a method of evaluation.

We used the methodology in the Yeo-Johnson Power Transformations paper [16] to transform and normalize (equation number) was used for all quantitative features (list them here maybe) and labels to normalize each to a mean of 0 and variance of 1.

Modeling Financial Time Series with S-PLUS was primarily consulted to study and implement a vector autoregressive model, which is a multivariate variant of ARIMA that uses autoregressive terms but does not difference the data [17]. This model was determined to be the most appropriate for the AQS PM2.5 time series data.

After reviewing research papers, we implemented a RNN model using long-short-term memory(LSTM) nodes, since they have been shown to solve the short-term memory issues related to standard tanh recurrent nodes. Gated recurrent units were also considered and may be used and tested as an alternative to LSTM nodes. Currently, LSTMs are being used in the RNN [3][2].

Liang et al. utilized meteorological factors and trained a back propagation neural network and recurrent neural networks while incorporating time series and LSTM models to form predictive models with about 90.9% accuracy for predicting the scale of the forest fires [8]. This was a major driving force for use to choose to implement LSTMs in our research.

3 Difficulties

In this project, temperature, wind, humidity, and precipitation play an important role because they often lead to certain atmospheric conditions which might enable the spread of harmful particles in the air. We used the NARR dataset to extract these features, since it captures data every 3 hours. However, NARR satellites only capture data in 32 km by 32 km grid cells, which is a relatively low spatial resolution. This required interpolation of the NARR data to fit the custom grid.

There were several problems that had to be solved after the data was processed and formatted. Even after dimensionality reduction, the dataset was quite large and required significant computational resources. Furthermore, a clear pipeline from the data into training and testing the models was necessary to efficiently tune and optimize each model's hyperparameters. The key to solving these issues was Google Cloud. The AI Platform service on Google Cloud provides sufficient computing resources to train the specified models and also facilitates a standard process for all team members to run and tune models, once they are uploaded to the cloud.

Due to the substantial time needed to process the data and acquire the necessary resources to run the models, model runs have only been performed on 2019 data. This is problematic in terms of analyzing our results for several reasons: when splitting the 2019 data into training and testing sets, we had to split the typical fire season between the training and testing data, to ensure that both datasets included some portion of the fire season. A train-test split where the training data included most or very little of the fire season would result in poor generalization results. However, due to the split, the models were trained and tested on data of which a majority was not during the fire season, making the results somewhat distant from the goal of the paper.

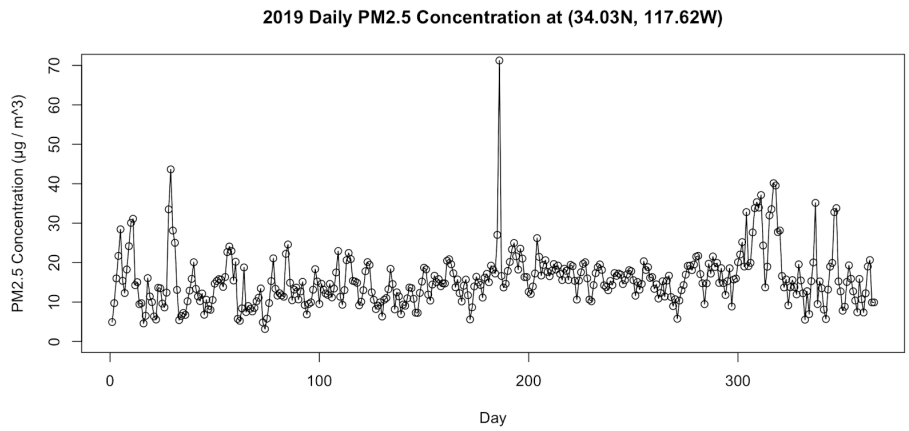

Figure 1: Air Quality at $(34.0^oN, 117.6^oW)$

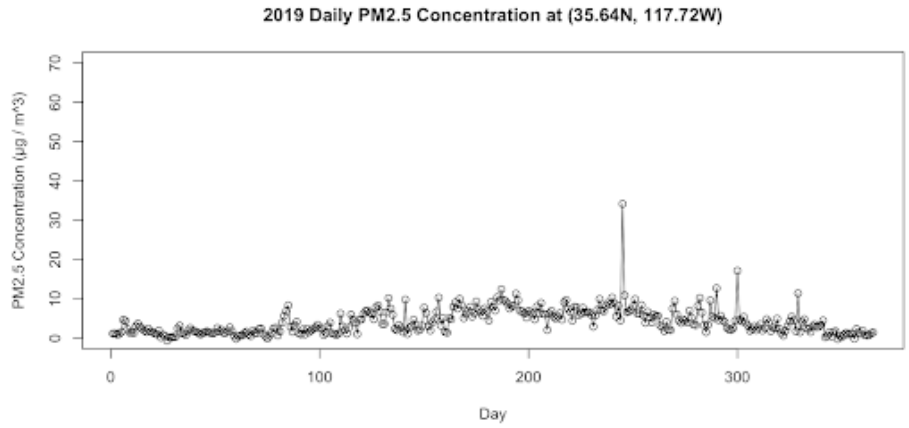

Figure 2: Air Quality at $(39.76^oN, 121.84^oW)$

4 Interesting Findings

4.1 MODIS Fire Product

The extraction of latitude and longitude coordinates from the original HDF-EOS format was difficult. Thus, the fire data was obtained from annua country-based summaries in a CSV format, which were stored in a data archive maintained by Fire Information for Resource Management System (FIRMS). The CSV data is coherent with the HDF data but does not contain all the information contained in the HDF files. Another interesting finding was that

4.2 EPA AQS PM2.5 Dataset

Time series plots of daily 2019 PM2.5 concentrations were plotted for several individual monitoring stations using R and the TSA package [1]. One key finding was that monitoring stations have varying features in their time series plots. The 2019 daily PM2.5 concentration plot for the monitoring station at (34.0o N, 117.6o W), which is in a highly urban area of the Los Angeles metropolitan area, is shown in Figure 1.

4.3 NARR Dataset

The NARR data published by the University Corporation for Atmospheric Research (UCAR) is accessible in multiple formats. An interesting finding in the 2019 NARR dataset was that for three full days, all the data values for all the features were labeled as 0. One possible explanation for this is that satellite data was not

available for the three days. NARR data also carried a wide range of information on atmospheric conditions at various pressure levels across the entire North America. This allowed us to access

4.4 MODIS Land Cover Dataset

The MODIS Land Cover grid exactly matches the custom grid defined in Section III in the research paper. Thus, no re-gridding was necessary.

References

[1] Chan, Kung-Sik, and Brian Ripley. "Time Series Analysis." Package 'TSA', CRAN, 11 Sept. 2020, cran.r-project.org/web/packages/TSA/TSA.pdf.

[2] Chung, Junyoung & Gulcehre, Caglar & Cho, KyungHyun & Bengio, Y.. (2014). Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling.

[3] Hochreiter, Sepp & Schmidhuber, Jürgen. (1997). Long Short-term Memory. Neural computation. 9. 1735-80. 10.1162/neco.1997.9.8.1735.

[4] Flannigan, M.D., Logan, K.A., Amiro, B.D. et al. Future Area Burned in Canada. Climatic Change 72, 1–16 (2005). https://doi.org/10.1007/s10584-005-5935-y

[5] Jain, Piyush et al. "A Review of Machine Learning Applications in Wildfire Science and Management." Environmental Reviews (2020): n. pag. Crossref. Web.

[6] Gillett, N. P., Weaver, A. J., Zwiers, F. W., and Flannigan, M. D. (2004), Detecting the effect of climate change on Canadian forest fires, Geophys. Res. Lett., 31, L18211, doi:10.1029/2004GL020876.

[7] Levin, Noam & Heimowitz, Aliza. (2012). Mapping spatial and temporal patterns of Mediterranean wildfires from MODIS. Remote Sensing of Environment. 126. 12-26. 10.1016/j.rse.2012.08.003.

[8] Liang, Hao & Zhang, Meng & Wang, Hailan. (2019). A Neural Network Model for Wildfire Scale Prediction Using Meteorological Factors. IEEE Access. 7. 1-1. 10.1109/ACCESS.2019.2957837.

[9] Medina, Susan & Vicente, Rubén & Nieto-Taladriz, María & Aparicio, Nieves & Chairi, Fadia & Vergara Diaz, Omar & Araus, Jose. (2019). The Plant-Transpiration Response to Vapor Pressure Deficit (VPD) in Durum Wheat Is Associated With Differential Yield Performance and Specific Expression of Genes Involved in Primary Metabolism and Water Transport. Frontiers in Plant Science. 9. 1994. 10.3389/fpls.2018.01994.

[10] Reid, Colleen & Jerrett, Michael & Petersen, Maya & Pfister, Gabriele & Morefield, Philip & Tager, Ira & Raffuse, Sean & Balmes, John. (2015). Spatiotemporal Prediction of Fine Particulate Matter During the 2008 Northern California Wildfires Using Machine Learning. Environmental Science & Technology. 49. 10.1021/es505846r.

[11] Shi, H., Jiang, Z., Zhao, B., Li, Z., Chen, Y., Gu, Y., et al. (2019). Modeling study of the air quality impact of record-breaking Southern California wildfires in December 2017. Journal of Geophysical Research: Atmospheres, 124, 6554– 6570. https://doi.org/10.1029/2019JD030472

[12] Watson, Gregory & Telesca, Donatello & Reid, Colleen & Pfister, Gabriele & Jerrett, Michael. (2019). Machine learning models accurately model ozone exposure during wildfire events. Environmental Pollution. 254. 10.1016/j.envpol.2019.06.088.

[13] Williams, A. P., Abatzoglou, J. T., Gershunov, A., Guzman-Morales, J., Bishop, D. A., Balch, J. K., & Lettenmaier, D. P. (2019). Observed impacts of anthropogenic climate change on wildfire in California. Earth's Future, 7, 892– 910. https://doi.org/10.1029/2019EF001210

[14] Yao, Jiayun Angela & Brauer, Michael & Raffuse, Sean & Henderson, Sarah. (2018). Machine Learning Approach To Estimate Hourly Exposure to Fine Particulate Matter for Urban, Rural, and Remote Populations during Wildfire Seasons. Environmental Science & Technology. 52. 10.1021/acs.est.8b01921.

[15] Yao, Jiayun Angela & Raffuse, Sean & Brauer, Michael & Williamson, Grant & Bowman, David & Johnston, Fay & Henderson, Sarah. (2018). Predicting the minimum height of forest fire smoke within the atmosphere using machine learning and data from the CALIPSO satellite. Remote Sensing of Environment. 206. 98-106. 10.1016/j.rse.2017.12.027.

[16] Yeo, In-Kwon & Johnson, Richard. (2000). A new family of power transformations to improve normality or symmetry. Biometrika. 87. 10.1093/biomet/87.4.954.

[17] Zivot, Eric, and Jiahui Wang. Modeling Financial Time Series with S-PLUS®. Springer New York, 2006.

[18] Zou, Y.; O'Neill, S.M.; Larkin, N.K.; Alvarado, E.C.; Solomon, R.; Mass, C.; Liu, Y.; Odman, M.T.; Shen, H. Machine Learning-Based Integration of High-Resolution Wildfire Smoke Simulations and Observations for Regional Health Impact Assessment. Int. J. Environ. Res. Public Health 2019, 16, 2137.