Predicting Levels of Ambient Fine Particulate using a Neural Net

Rei Bertoldi, Akhil Ghanta, Sarah Gill, Rohen Shah

March 13, 2020

Abstract

Air pollution has become an increasingly salient public health concern, given its association with lung cancer, cardiovascular disease, respiratory disease, and metabolic disease. Globally, the distribution of $PM_{2.5}$ monitoring technologies are not equally distributed, contributing to $PM_{2.5}$ data scarcity. Insufficient data can substantially limit government capacities for mitigating $PM_{2.5}$ related mortality and public health risks. As a solution to this, we incorporate simple, globally accessible, ground level data such as temperature, geographical parameters such as latitude and longitude, and satellite observations into a deep learning architecture that predicts out-of-sample, ground level $PM_{2.5}$ concentrations. Predictions are made using a deep learning model for three cities within the continental United States with varying efficacy. The learning algorithm works best for areas with spatially homogeneous air quality and low climactic variability.

1 Background

A leading global environmental risk factor for mortality and disease burden is exposure to ambient fine particulate matter, $PM_{2.5}$, and is associated with global welfare costs in the trillions of dollars (Tagaris et al., 2007). The $PM_{2.5}$ pollutant is defined as fine, inhalable particles with diameters 2.5 micrometers or smaller. The biggest impact of particulate air pollution on public

health is understood to be from long-term exposure to $PM_{2.5}$, which increases the age-specific mortality risk, particularly from cardiovascular causes (W.H.O., 2006). Therefore, it is essential to monitor the level of $PM_{2.5}$ in order to manage and mitigate the impacts of air pollution on public health. Knowing the levels of $PM_{2.5}$ will not only help us assess which climate policies aimed at reducing it are working, but will also help identify areas that are dangerous if the levels exceed a certain threshold.

The high importance of monitoring $PM_{2.5}$ has led many regions around the world, particularly in developed countries, to establish ground and satellite monitoring stations. However, there are still many regions around the world where we are not able to directly monitor the level of $PM_{2.5}$ (Martin et al., 2019). Figure 1 illustrates the number of PM2.5 monitors per million people by country. According to the map, only 24 of 234 countries have more than 3 monitors per million. This is less than 9% of the world's population.

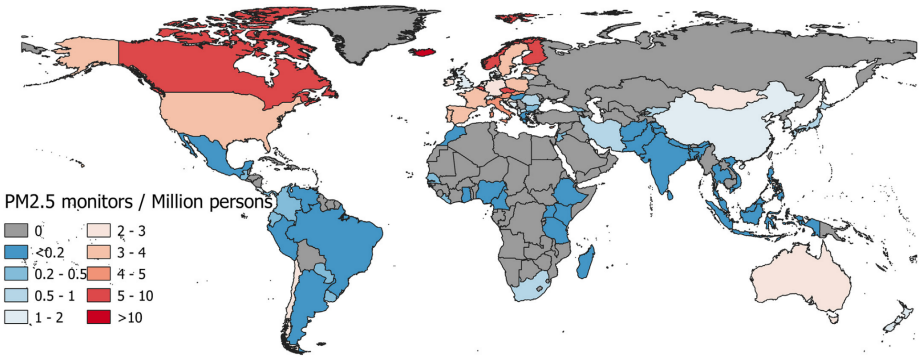


Figure 1: Number of $PM_{2.5}$ monitors per million inhabitants by country for any year between 2010-2016 (Martin et al. (2019)).

Much of the past literature has found that monitoring is difficult despite many attempts, and has focused on predicting $PM_{2.5}$ in a region *in the future* using the lagged level of $PM_{2.5}$ from previous years (Badura et al., 2018). Snider et al. (2015) recommends controlling for sun-photometer measurements of aerosol **Optical Depth** to improve the accuracy of satellite -based estimates of

$PM_{2.5}$.

A prominent study that estimated $PM_{2.5}$ levels in China found that controlling for **Season** is crucial to improving accuracy (Zhang and Cao, 2015). Other studies have shown that meteorological variables such as **Temperature** and **Precipitation** can often account for as much as 50% of the variance in levels of $PM_{2.5}$ (Tagaris et al., 2007). **Elevation** has also been shown to have a significant relationship with $PM_{2.5}$ (Hu et al., 2014), and this effect might even co-vary based on the season (Silcox et al., 2012).

In this paper, we aim to develop a model capable of predicting $PM_{2.5}$ for regions which are regionally and meteorologically dissimilar, using only readily available information. In the current study, five unrelated U.S. counties are used, four to train the neural net, and one to test.

2 Data

The data underlying our predictions came from three distinct sources - ground measured $PM_{2.5}$ concentration, satellite-derived MODIS aerosol optical depth and meteorological data (see Table 1). The study period was for 2010.

2.1 Study Area

The study area for this paper is at the county level for five counties in five unique states. These counties are Cook County, IL, Los Angeles County, CA, Denver County, CO, Hillsborough County, FL, and Anchorage County, AK.

- Cook County is located at 41.73° North and 87.69° West, with a population in 2010 of 5.1 million. Chicago's climate is typically continental with cold and snowy winters, warm summers, with frequent fluctuations in temperature, humidity, cloudiness and wind direction.

Data Type	Parameter	Abbreviation	Unit	Source
	Precipitation (month to date average)	mtd_prcp_normal	in	
	Precipitation (year to date average)	ytd_prcp_normal	in	
	Snow (month to date average)	mtd_snow_normal	in	
Meteorological	Snow (year to date average)	ytd_snow_normal	in	NOAA (National Centers for Environmental Information)
Stations	Temperature (daily average)	dly_tavg_normal	°F	
	Latitude	-	°	
	Longitude	-	°	
	Elevation	-	m	
	Temperature (daily diurnal average)	dly_dutr_normal	°F	
	Temperature (daily max average)	dly_tmax_normal	°F	
	Temperature (daily min average)	dly_tmin_normal	°F	
Air Quality Monitoring Station	PM 2.5	daily_mean_pm_2_5_concentration	μg/m^3 LC	EPA (United States Environmental Protection Agency)
Satellite Products	AOD 470nm		unitless	MODerate resolution Imaging Spectroradiometer (MODIS)
	AOD 550nm		unitless	

Table 1: Variables

- Los Angeles county is located at 34.05° North and 118.24° West, with a population in 2010 of 9.8 million. Los Angeles County's climate is classified as a Mediterranean climate, with year-round mild to hot, mostly dry weather, with hotter temperatures.

- Denver County is located at 39.73° North and 104.99° West, with a population in 2010 of 600,158. It features a semi-arid climate with mild summers and snowy winters, with very low humidity. The weather in the city is influenced by its proximity to the Rocky Mountains to the west.

- Hillsborough County is located at 27.99° North and 82.30° West, with a population in 2010 of 1.2 million. Hillsborough County has a humid subtropical climate closely bordering a tropical monsoon climate near the waterfront areas. Here, there are mild winters and hot, rainy summers.

- Anchorage County is located at 61.21° North and 149.90° West, with a population in 2010

State	Min Elevation	Max Elevation	Mean Elevation	Min Precipitation	Max Precipitation	Mean Precipitation	Min T_{avg}	Max T_{avg}	Mean T_{avg}	NOAA Station	PM2.5 Station
Cook County, IL	196.9	226.7	232.4	0.05	4.86	1.67	20.2	74.8	49.13	4	14
Los Angeles County, CA	4.3	932.7	276.17	0	12	0.77	16.7	83	60.7	21	13
Denver County, CO	1611.2	1611.2	1611.2	0.01	2.3	0.67	29.3	74.2	50.25	1	3
Hillsborough County, FL	5.8	33.5	19.65	0.05	8.62	2.15	60.1	83.3	72.91	2	3
Anchorage County, AK	40.2	688.8	188.72	0.01	9.27	1.25	14.4	60.5	35.87	5	3

Table 2: Summary Statistics

of 293,370. Anchorage has a subarctic climate, with short, cool, summers and cold, snowy winters.

The decision to utilize five different counties from five distinct states, was motivated by the input information we wanted to go into the model. The idea being, feeding climatic information from various, unique environments would enrich the model and increase it's predictive capacity. Using information from distinct climates, feeds the model with diversity of feature observations and increases the probability of accurately predicting out-of-sample environments. Table 2 presents various summary statistics across key variables for each county in our study area.

2.2 $PM_{2.5}$ Air Pollution Data

$PM_{2.5}$ data, as well as station latitude and longitude parameters, were retrieved from the United States Environmental Protection Agency (EPA) Outdoor Air Quality data portal. Data is collected via on-the-ground sensors and returns data summarized at the daily level, calculated on midnight to midnight basis in local time. From this source, we acquire our key, independent variable of interest - daily average $PM_{2.5}$ concentration.

2.3 Aerosol Optical Depth (AOD) Data

AOD is a measure of optical depth of ultraviolet wavelengths of light (470 nm and 550 nm) measured from low earth orbital satellites. When atmospheric aerosol concentration is high, the measured optical depth for a given line will be lowered. Thus, AOD provides a strong characterization of man-made particulate matter pollution. Daily AOD data came from the MODerate resolution Imaging Spectroradiometer (MODIS), with nearly global coverage available in two spatial resolutions of 10 and 3 km. The AOD is calculated based on variations in the Dark Target and Deep Blue Aerosol retrieval algorithm, over urban areas. The most recent year for which AOD data is available, 2010, was used. Retrieval of AOD data required usage of the Google Earth Engine API, a cloud-based geospatial analysis library that enables analysis of satellite data.

2.4 Meteorological Data

The climatic features in this study include daily average air temperature, diurnal temperature, maximum and minimum air temperature, elevation, snowfall, and precipitation. These features were retrieved from the National Center for Environmental Information (NOAA) U.S. Daily Climate Normals (1981-2010) database. This database provides quality checked, 30-year averages of meteorological data from on-the-ground sensors.

2.5 Created Variables

We also used the precipitation and snowfall variables with 1 and 2 days lag, so new columns were added to the database. As well, we added indicators for the day of the week, season and a dummy indicator for weekday.

2.6 Variable Choice

As we have seen, prior research indicates that key variables that are correlated with $PM_{2.5}$ include Season, Optical Depth, Elevation, Precipitation, and Temperature. Since we did not have a way to measure the traffic level, we use an indicator for being a Weekday as a proxy, which is true if more people are driving on a typical weekday than a typical weekend.

Additionally, Joharestani et al. (2019) conducted a study using deep learning to quantify feature importance for $PM_{2.5}$ prediction. In their results, they find that the feature that absorbs the most variation, is a lag variable on historical $PM_{2.5}$. As the purpose of our research is to predict $PM_{2.5}$ for out-of-sample sites, using historical $PM_{2.5}$ data is impractical.

Joharestani et al. (2019) also found wind speed to be a strong feature. However, for the counties analyzed in this paper, data on wind speed was either not made available by NOAA, or fully missing. Since one premise of this paper is predicting $PM_{2.5}$ using accessible, publicly available data sources, exclusion of wind speed due to the challenge in acquiring wind data is not a limitation but rather a feature.

Joharestani et al. (2019) rank features based on the median of feature importance using feature permutation impact on prediction performance of a well-trained deep neural network model. Their results rank our variables as minimum temperature (6), elevation (7), average temperature (8), maximum temperature (12), season (13), weekday (14), AOD (16), precipitation lag 2 (17), latitude (19), precipitation lag 1 (20), longitude (21), and precipitation (22).

3 Data Pre-processing and Matching

Merging was necessary because the data was obtained from different sources. Matching was necessary because the $PM_{2.5}$ and climatic variables did not come from equivalently located monitoring

stations.

3.1 Dropping Missing Variables

In the data cleaning process, whole monitoring sites were dropped as a result of missing values. These individual monitoring sites were either entirely missing data, or completely missing information for particular variables. When dropping missing observations, these sites were effectively dropped. This manner of missing data strongly indicates the presence of inactive or deficient monitoring sites in the data. We might be concerned about the nature of this missing data contributing to selection bias if sensor failure or inactivity is not at random. However, artificial neural networks are quite robust in the presence of external selection biases. ANNs are capable of learning these biases over the duration of their training time. In particular, each neuron in the network contains an extra weight parameter referred to as a bias. This term essentially acts as an extra "baseline" that will be visible in the predictions of non-classification problems.

Additionally, AOD data points were missing for several days out of the year due for various reasons including satellite calibration errors, weather systems, and space traffic. The data for these days were dropped from our training dataset. This is valid since our predictive model does not account for time as a variable in determining the concentration of $PM_{2.5}$. Rather, it only requires a set of meteorological parameters coupled with AOD to account for human polluting factors such as car exhaust, industrial emissions, and other man-made particulate sources.

3.2 Matching

First, each $PM_{2.5}$ monitoring station was matched to it's closest NOAA station, by calculating and minimizing distance between each $PM_{2.5}$ and NOAA station. After assigning each $PM_{2.5}$ site with its appropriate NOAA counterpart, the climatic information was merged on site name and

date. Of the available 33 NOAA sites across the five counties in our study area, 17 were retained and used in our model.

The decision to match based on minimizing distance, rather than using a buffer and joining by intersections, was to assure the greatest accuracy that we could achieve with the data we had. Though, using intersections rather than one-to-one matching may have resulted in more data, this level of accuracy, that we value, would not be assured for the additional data.

Fortunately, $PM_{2.5}$ readings over time are very highly correlated between monitoring sites within a single county. For instance, in Cook County, IL, the least correlated two monitoring sites have a Pearson correlation of 0.765. On average, we find that monitoring sites are highly correlated, with the mean correlation of all pairwise correlations between sites being 0.907. This point estimate represents the average correlation of an individual site's with all other sites. Table 3 provides point mean and minimum correlation estimates for each county in our study area.

This high level of between-site correlation gives us confidence that pairing $PM_{2.5}$ monitoring sites with the nearest weather monitoring site will generate reliable results. Both weather and $PM_{2.5}$ appear to be relatively invariant within a single county, on any given day. Thus, even though some $PM_{2.5}$ monitoring sights are far away from the nearest weather monitoring site, weather recorded at that distant (within county) site likely has the same impact on $PM_{2.5}$ recorded at the distant $PM_{2.5}$ sight as on a closer one.

It is important to note that the correlation is notably lower in Los Angeles, CA, with a minimum correlation of 0.057 and mean correlation of 0.566. We will discuss the implications of this on our prediction in the Results section.

County	Min	Mean
Cook County, IL	0.765	0.818
Los Angeles County, CA	0.057	0.566
Denver County, CO	0.913	0.959
Hillsborough County, FL	0.802	0.903
Anchorage County, AK	0.057	0.566

Table 3: Correlations

4 Learning Methodology

To produce our predictive model, we used a feedforward neural network. This is a subset of artificial neural networks (ANNs) in which layers do not form closed cycles and instead feed directly into the next subsequent layer. The network architecture consisted of two hidden layers of sixty four nodes each and an input layer with sixteen features. The number of nodes within the hidden layers was selected to be the number of input features multiplied by a factor of four. In order to secure the ability of the network to generalize the number of nodes has to be kept as low as possible. If the network has a large excess of nodes, it becomes a memory bank that can recall the training set to perfection, but does not perform well on samples that was not part of the training set. At this number of nodes, the cross-validation loss is kept minimal. Similar reasons were used to select the number of hidden layers. Moreover, two hidden layers have been shown to model an arbitrary function (Hornik (1991)). The network is fully connected and utilizes the rectified linear unit (ReLU), expressed as $a(z) = \max(0, z)$, as an activation function. Our network was trained

to minimize the mean squared error (MSE) using gradient descent. Many standard techniques in deep learning utilize the activation functions $a(z) = \tanh z$ and $a(z) = (1 + e^{-z})^{-1}$. Although these functions are effective in characterizing nonlinear trends in training data, they are significantly slower in converging to a global minimum than the ReLU. As seen in Figure 2, ANNs with ReLUs train several times faster than their equivalents with tanh units.

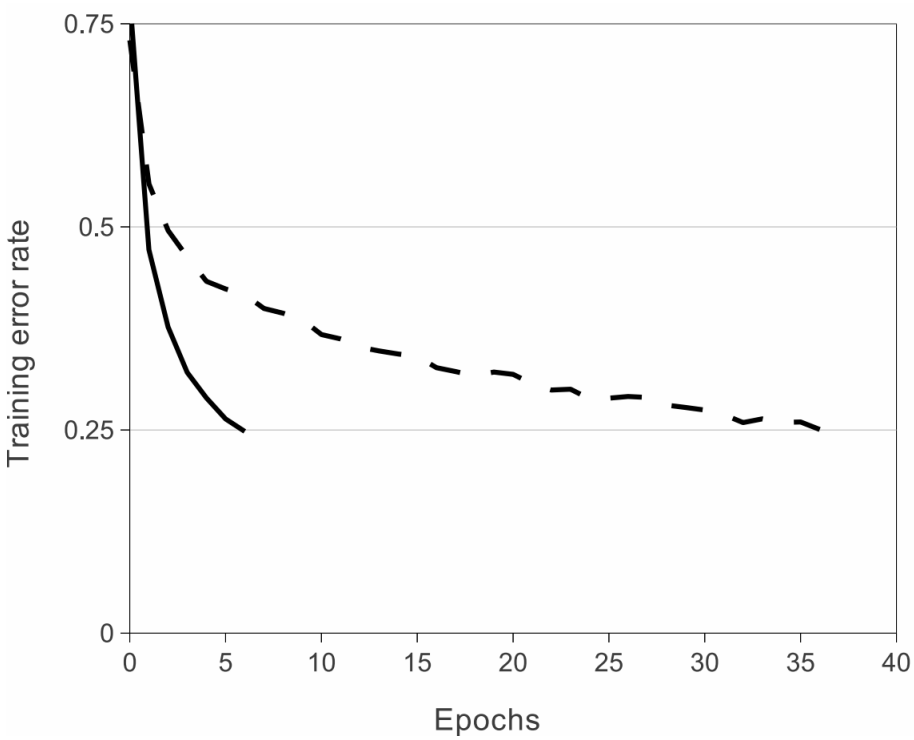


Figure 2: Training history for a four-layer convolutional neural network with ReLUs (solid line) reaches a 25% training error rate approximately six times faster than an equivalent network with tanh neurons (dashed line). (Khrizhevsky (2017))

Assembly of the neural network was done in Python, specifically using Keras, an API within TensorFlow used for building and training deep-learning models. The training set consisted of the data in Table 1 for four of the aforementioned city centers. The data for the fifth city was then used as a testing set to validate the efficacy of our trained model. The testing city was varied between Los Angeles, Tampa Bay, and Denver due to the diversity of their climactic regions. In each

scenario, the network was trained until the change in MSE between a number of epochs p, known as the patience of the training algorithm, was less than a set value δ. The value for δ was chosen through trial and error to allow for adequate training without over-fitting. It was simultaneously substantiated to make out of sample predictions through the use of three-fold cross validation in Keras. The number of epochs the network was trained for in each case is outlined in Table 4.

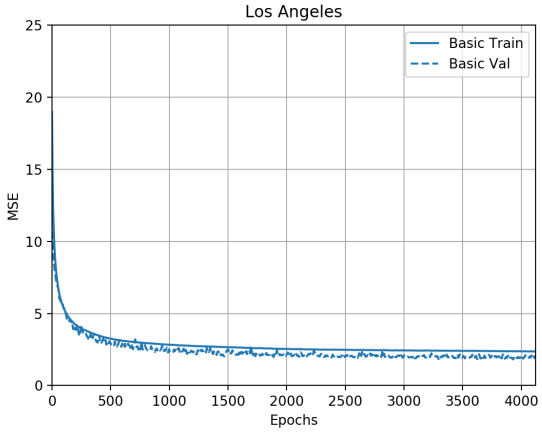
Test City	Epochs	Final MSE
Los Angeles County, CA	4056	2.468
Denver County, CO	5890	5.329
Hillsborough County, FL	4123	4.892

Table 4: Epochs

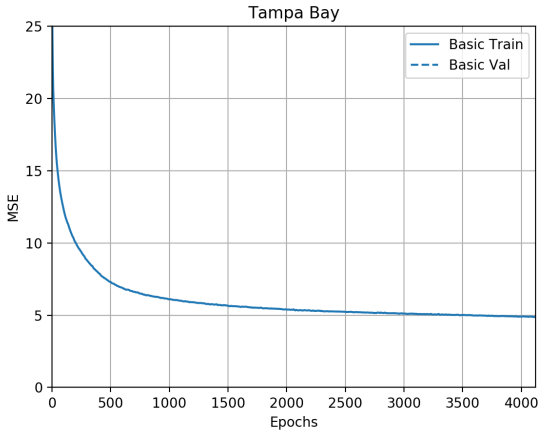
The learning rate was set at 0.001 and chosen at a low enough level to prevent overshooting the global minimum. Since we chose the MSE loss-function, we are guaranteed to converge due to its convexity so long as we do not train too quickly. The training histories for each of the scenarios are shown in Figure 3.

5 Results

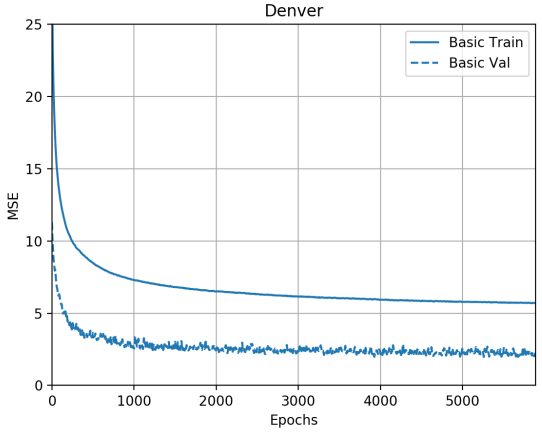
The results of our training are presented in Figure 4. The results shown are plots of the predicted values of $PM_{2.5}$ for the input data provided by the testing city, plotted against the true values of $PM_{2.5}$. The solid blue line indicates $y = x$, therefore the closer the plotted values are to the line, the stronger our predictive model.



(a)

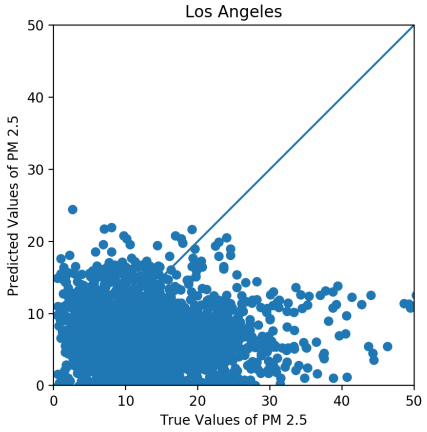


(b)

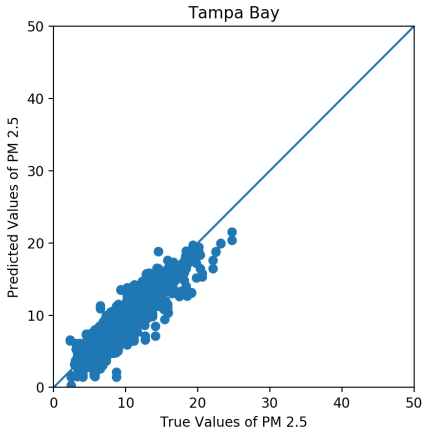


(c)

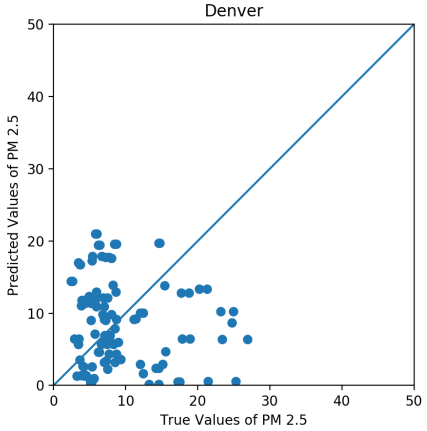
Figure 3: Training results, solid lines denote MSE, dashed lines denote cross-validation loss.



(a)



(b)



(c)

Figure 4: Training results.

As seen in the plots, Tampa Bay and Denver $PM_{2.5}$ concentrations are predicted fairly well, while Los Angeles is very poor. Upon more careful analysis, it was noted that Los Angeles has an extremely high city-wide variance of air quality index. This indicates that $PM_{2.5}$ concentration is strongly varied depending on spatial location within the city. AOD measurements, however, do not possess resolution small enough to ameliorate these variances, thus resulting in non-ideal predictions. Tampa bay, however, is a city center with homogeneous population density and little climactic variance supplemented by sparse road networks and a lack of industrial areas. Thus, the variance of air quality across the city center is significantly lower than in Los Angeles, therefore leading to better predictions.

6 Conclusion

Disparate data accessibility on $PM_{2.5}$ concentrations, impedes the implementation of informed, effective air quality management programs within countries with limited monitoring technologies. We have found that climatic variables, which are often much more readily accessible, can help inform out-of-sample $PM_{2.5}$ concentrations. Though, there are certainly limitations to our work and opportunities for improvement - such as disaggregating to finer spatial resolutions - we hope that the motivation of our paper will at least help advance issue salience surrounding the unequal distribution of technological resources globally, and can energize and empower the use of machine learning techniques to help bridge the divide in global information access.

A GitHub Link

Link to source code

References

Badura, M., Batog, P., Drzeniecka-Osiadacz, A., and Modzel, P. (2018). Evaluation of low-cost sensors for ambient pm2. 5 monitoring. *Journal of Sensors*, 2018.

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4:251–257.

Hu, X., Waller, L. A., Lyapustin, A., Wang, Y., Al-Hamdan, M. Z., Crosson, W. L., Estes Jr, M. G., Estes, S. M., Quattrochi, D. A., Puttaswamy, S. J., et al. (2014). Estimating ground-level pm2. 5 concentrations in the southeastern united states using maiac aod retrievals and a two-stage model. *Remote Sensing of Environment*, 140:220–232.

Joharestani, M., Cao, C., Ni, X., Bashir, B., and Talebiesfandarani, S. (2019). Pm2. 5 prediction based on random forest, xgboost, and deep learning using multisource remote sensing data. *Atmosphere*, 10(7):373.

Khrizhevsky, A. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60.

Martin, R. V., Brauer, M., Van Donkelaar, A., Shaddick, G., Narain, U., and Dey, S. (2019). No one knows which city has the highest concentration of fine particulate matter. *Atmospheric Environment: X*, 3:100040.

Silcox, G. D., Kelly, K. E., Crosman, E. T., Whiteman, C. D., and Allen, B. L. (2012). Wintertime pm2. 5 concentrations during persistent, multi-day cold-air pools in a mountain valley. *Atmospheric environment*, 46:17–24.

Snider, G., Weagle, C., Martin, R., Van Donkelaar, A., Conrad, K., Cunningham, D., Gordon, C.,

Zwicker, M., Akoshile, C., Artaxo, P., et al. (2015). Spartan: a global network to evaluate and enhance satellite-based estimates of ground-level particulate matter for global health applications.

Tagaris, E., Manomaiphiboon, K., Liao, K.-J., Leung, L. R., Woo, J.-H., He, S., Amar, P., and Russell, A. G. (2007). Impacts of global climate change and emissions on regional ozone and fine particulate matter concentrations over the united states. *Journal of Geophysical Research: Atmospheres*, 112(D14).

W.H.O. (2006). *Air quality guidelines: global update 2005: particulate matter, ozone, nitrogen dioxide, and sulfur dioxide*. World Health Organization.

Zhang, Y.-L. and Cao, F. (2015). Fine particulate matter (pm 2.5) in china at a city level. *Scientific reports*, 5:14884.