Machine Learning on Spark for the Optimal IDW-based Spatiotemporal Interpolation

Weitian Tong[1], Jason Franklin[1], Xiaolu Zhou[2], Lixin Li[1*], Gina Besenyi[3]

[1]Department of Computer Sciences,
[2]Department of Geology and Geography,
Georgia Southern University,
P.O. Box 7997, Statesboro, GA 30460, USA
Emails: {wtong; jf00936; xzhou; lli}@georgiasouthern.edu

[3]Clinical and Digital Health Sciences, CAHS,
Augusta University, Augusta, GA 30912, USA
Email: gbesenyi@augusta.edu

Abstract

In order to improve current spatiotemporal interpolation methods for public health applications (Li *et al.*, 2010), we combine the extension approach (Li and Revesz, 2004) and several machine learning methods, employ the efficient k-d tree structure to store data, and implement our method on Spark (Spark, 2016). The preliminary results demonstrate the excellent computation ability and amazing scalability of our method, which outperforms the previous work (Li *et al.*, 2014). Future research will continue exploring the current method to improve the interpolation accuracy and computation efficiency, as well as establishing associations between air pollution exposure and adverse health effects.

1. Introduction

To implement the spatiotemporal interpolation method, Li and Revesz (2004) proposed an *extension approach*, which reduces the spatiotemporal interpolation into a higher-dimensional spatial interpolation by treating time as an *asymmetric* dimension in space. Unfortunately, modern work on spatiotemporal interpolation (Pebesma, 2012; Gräler *et al.*, 2013; Losser *et al.*, 2014; Li *et al.*, 2014, *etc*) utilizes simplistic methods to scale the range of the time dimension. In recent work, Li *et al.* (2014) extended the inverse distance weighted (IDW) method (Shepard, 1968) to model the $PM_{2.5}$ exposure risk by scaling the time domain with a parameter c, which is a similar concept to the *spatiotemporal anisotropy parameter* (Gräler *et al.*, 2014).

The spatiotemporal IDW method was extended to estimate the $PM_{2.5}$ (particulate matter with a mean aerodynamic diameter less than or equal to 2.5 micrometers) exposure risk. The details of our method are omitted due to the page limit and will be elaborated in our future full paper. Our main contribution is to apply machine learning methods to efficiently learn the optimal model parameters. In particular, considering the expensive computation on the large data set, we implemented our new method with *Apache Spark* (Spark, 2016), which is a lightning-fast cluster computing framework and represents the avant-garde of big data processing tools.

[*] Correspondence Author

2. Methods

2.1 Data Sets

To demonstrate the efficacy and efficiency of our new method, we explored three daily $PM_{2.5}$ data sets for comparison with the results from Li *et al.* (2014). The first data set was air pollution data from AQS (Air Quality System), the EPA's repository of ambient air quality data, containing 146,125 $PM_{2.5}$ measurements collected at 955 monitoring sites on all 365 days of the year 2009 (Figure 1).

The second and third data sets contain centroid locations of 3109 counties and 207,630 census block groups in the contiguous U.S., respectively. Census block groups (the smallest geographical unit for which the Census Bureau publishes sample data) contain roughly 600~3000 people and are commonly used spatial units to explore population health variables (Iceland and Steinmetz, 2003, Krieger *et al*., 2002).

We will train our IDW-based spatiotemporal interpolation model to estimate the daily $PM_{2.5}$ concentration values in 2009 at the centroids of counties and census block groups (the second and third data sets) for the entire contiguous U.S., using the existing $PM_{2.5}$ measurements (the first data set) as the training set.

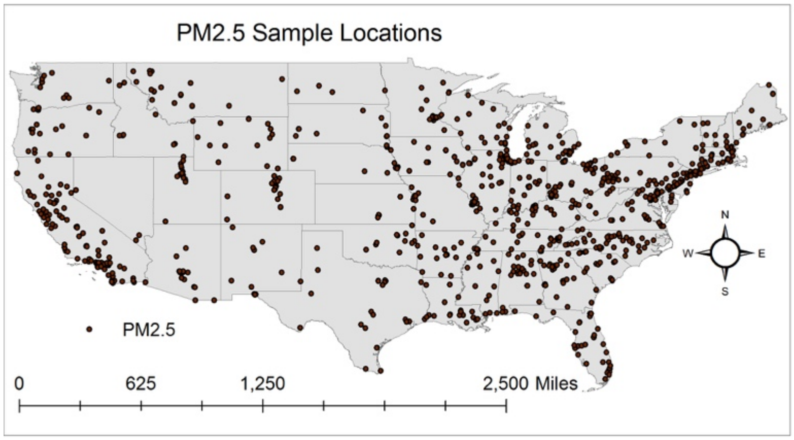

Figure 1. $PM_{2.5}$ Sample Locations

2.2 IDW-based Spatiotemporal Interpolation Method

To extend the spatial IDW method to interpolate the spatiotemporal data based on the *extension approach*, we developed the following formulas

$$w(x, y, ct) = \sum_{i=1}^{n} \lambda_i w_i, \qquad \lambda_i = \frac{(1/d_i)^p}{\sum_{k=1}^{N} (1/d_i)^p},$$

where $w(x, y, ct)$ represents the unknown value to be calculated at the un-sampled location (x, y) and time instance t, c is the spatiotemporal anisotropy parameter, p is the exponent that influences the weighting of w_i, and n is the number of nearest neighbors. It is essential to determine the optimal parameters c, p and n in order to estimate the daily $PM_{2.5}$ concentration values at unknown points.

2.3 Computationally Efficient Methods to Learn Optimal Parameters

Applying k-fold cross validation to the training set can discover the optimal parameters c, p and n. However, due to the volume of the data, this task is quite time-consuming. In order to improve the efficiency, we first employ the k-d tree data structure to quickly search nearest neighbors (Li *et al.*, 2014). Then we parallelize our algorithms in the Apache Spark. The Spark ecosystem (Figure 2) is a unified and powerful open source cluster processing engine, packaged with high-level libraries that support SQL queries, streaming, machine learning and graph processing. These libraries can be used to parallelize computationally expensive jobs while simultaneously easing the burden on the developer. Engineered from the bottom-up for performance, Spark can be up to 100 times faster than Hadoop MapReduce by allowing iterative, in-memory processing (Zaharia *et al*., 2014).

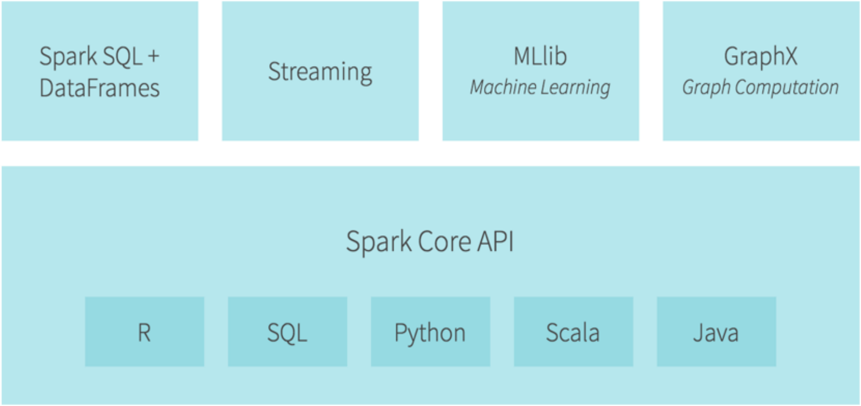

Figure 2. Spark Ecosystem (Spark, 2016)

3. Preliminary Results

A pilot version of the IDW-based spatiotemporal interpolation method on Spark was implemented using the same experiment settings from Li *et al.* (2014). Preliminary results demonstrate that our method and implementation is extremely fast compared to previous work.

Experiment 1: We built the k-d tree and attempted to learn optimal parameters for the model. This task only took 2.3 minutes on Spark while the same task executed sequentially on like hardware would need 70 minutes. Since Li *et al.* (2014) did not provide time consumptions for this learning process, we are not able to compare our result with their outcome.

Experiment 2: Li *et al.* (2014) did offer the time consumption of estimating the daily $PM_{2.5}$ concentration values at the centroids of counties and census block groups. In our implementation, the total time consumption of this interpolation stage only takes about 8% of Li *et al.* (2014)'s record. That is, our method is much faster. This is very inspiring as it demonstrates excellent scalability and shows promise as a more efficient and practical method for public health applications.

We are very confident that our experiments will efficiently learn the optimal parameters, and thus improve the estimation accuracy of the interpolation model, helping us to definitively establish more accurate associations between air pollution exposure and adverse health effects.

4. Future Work

Future research will extend our machine learning approach on Spark in the following three directions: (1) other machine learning methods such as *random forest*, (2) other spatiotemporal methods such as *shape function* and *Kriging* based methods, and (3) other data sets such as real-time hourly air pollution data from the AirNow government website service that provides hourly updates of pollution measurements data from sites across North America.

Acknowledgements

We would like to thank Brandon Kimmons, Director of Computational Research Technical Support at Georgia Southern University, for helping us set up Spark. Franklin, Tong and Zhou were supported in part by funds from the Office of the Vice President for Research & Economic Development at Georgia Southern University.

References

Gräler B, Rehr M, Gerharz LE and Pebesma E, 2013. Spatio-temporal analysis and interpolation of PM10 measurements in Europe for 2009. *ETC/ACM Technical Paper*.

Iceland, J, and Steinmetz, E, 2003. The effects of using census block groups instead of census tracts when examining residential housing patterns. *Bureau of the Census*.

Krieger, N, Chen, JT, Waterman, PD, Soobader, MJ, Subramanian, SV, and Carson, R, 2002. Geocoding and monitoring of US socioeconomic inequalities in mortality and cancer incidence: does the choice of area-based measure and geographic level matter? The Public Health Disparities Geocoding Project. *American journal of epidemiology*, 156(5):471--482.

Li L and Revesz, P, 2004. Interpolation methods for spatiotemporal geographic data. *Computers, Environment and Urban Systems*, 28:201–227.

Li L, Zhang, X and Piltner, R, 2010. An application of the shape function based spatiotemporal interpolation method on ozone and population exposure in the contiguous U.S. *Journal of Environmental Informatics*, 12:120–128.

Li L, Losser T, Yorke C and Piltner R, 2014. Fast Inverse Distance Weighting-based spatiotemporal interpolation: a web-based application of interpolating daily fine particulate matter $PM_{2.5}$ in the Contiguous U.S. using parallel programming and k-d tree. *International Journal of Environmental Research and Public Health*, 11(9): 9101-9141.

Losser L, Li L and Piltner R, 2014. A spatiotemporal interpolation method using radial basis functions for geospatiotemporal big data. *In Proceeding of the 5th International Conference on Computing for Geospatial Research and Application*, Washington DC, USA, 17-24.

Pebesma E, 2012. Spacetime: spatio-temporal data in R. *Journal of Statistical Software*, 51(7):1–30.

Shepard D, 1968. A two-dimensional interpolation function for irregularly spaced data. *In Proceedings of the 23nd National Conference ACM*, 517-524.

Spark, 2016. https://databricks.com/spark/about.

Zaharia M, Chowdhury M, Franklin MJ, Shenker S and Stoica I, 2010. Spark: Cluster Computing with Working Sets. *In Proceedings of the 2nd USENIX conference on Hot topics in cloud computing (HotCloud'10) ACM*, 7 pages.