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

Journal
International Conference on GIScience Short Paper Proceedings, 1(1)

Authors
Tong, Weitian
Franklin, Jason
Zhou, Xiaolu
et al.

Publication Date
2016

DOI
10.21433/B3114dw721gn

Peer reviewed

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

To improve current spatiotemporal interpolation methods for public health applications (Li *et al.*, 2010), we combine the extension approach (Li and Revesz, 2004) with machine learning methods, employ the efficient k-d tree structure to store data, and implement our method on Apache Spark (Spark, 2016). Preliminary results demonstrate the computational power of our method, which outperforms the previous work in terms of speed and generates comparable results in terms of accuracy (Li *et al*., 2014). Future research will continue exploring this method to improve the interpolation accuracy and efficiency, with the long term objective of 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 resolves 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).

In applying the *extension approach* to the spatial IDW method to interpolate the spatiotemporal data, we arrived at the following formulae

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

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. Applying k-fold cross validation (k-CV) to the training set can discover the optimal parameters c, p and n for this data set in order to estimate the daily $PM_{2.5}$ concentration values at unknown points. Building upon this work, our method parallelizes the implementation of the original IDW algorithm using

* Correspondence Author

Apache Spark (Spark, 2016) (Figure 1), which is a lightning-fast cluster computing framework and represents the avant-garde of big data processing tools.

In short, our parallelized IDW process broadcasts structured data (the k-d tree) to worker nodes for distributed nearest-neighbor queries and, thus, rapid estimation of pollution levels at unmeasured locations. Naturally, the accuracy of the IDW method for pollution level estimation depends on certain model parameters. Previous work (Li *et al*., 2014) might search a few dozen parameterizations because of limited computational resources. Our system can search tens of thousands of parameterizations in a manageable amount of time. We attempt to learn the best model parameters with brute force, and Apache Spark allows for the application of tremendous force.

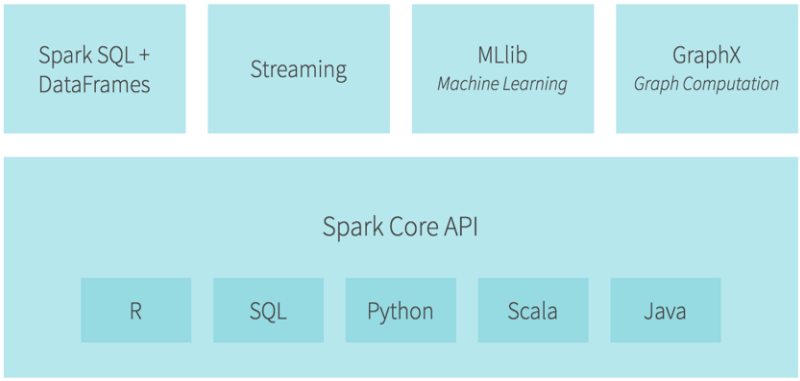

Figure 1. Spark Ecosystem (Spark, 2016)

2. 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 the EPA's AQS (Air Quality System) which provided 146,125 $PM_{2.5}$ measurements collected at 955 monitoring sites on all 365 days of the year 2009 (Figure 2).

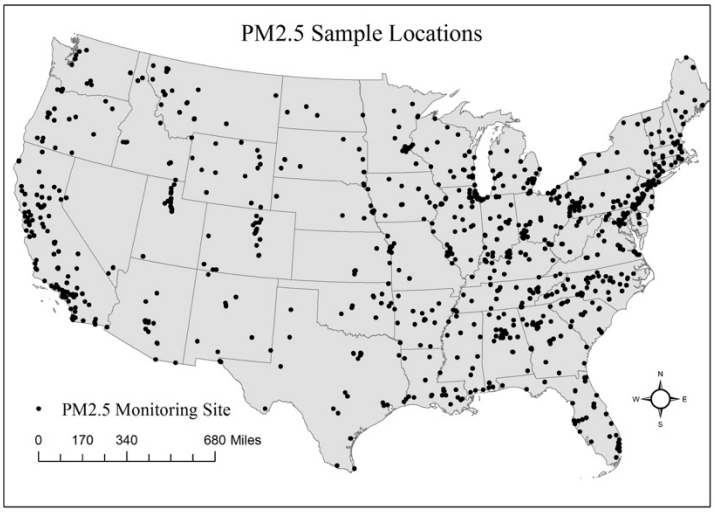

Figure 2. $PM_{2.5}$ Sample Locations

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 train our IDW-based model to estimate the daily $PM_{2.5}$ concentration values in 2009 at the centroid locations (the second and third data sets) for the contiguous U.S., using the existing $PM_{2.5}$ measurements (the first data set) as the training set.

3. Preliminary Results

Two pilot experiments using our general approach have been built. Preliminary results demonstrate that our method and implementation is extremely fast compared to previous work while achieving better prediction accuracy. Since our current method follows the work by Li *et al.* (2014), the main contributions are the larger learning ranges for the parameters in the model and the employment of the cutting-edge technique – Spark. The experiment details are shown as follows. The summaries also include runtimes on Spark with the equivalent time it would have taken in a sequential procedure (a statistic tracked by Spark).

Experiment 1: We exactly follow the work by Li *et al.* (2014), where the *spatiotemporal anisotropy parameter* c was fixed as 0.1086 and 45 parameter configurations were selected for inspection. The learning task in our system only took 2.3 minutes on Spark (70 minutes in sequential time). 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. For the prediction stage, where the daily $PM_{2.5}$ concentration values at the centroids of counties and census block groups are estimated, our implementation only took about 8% of Li *et al.* (2014)'s record.

Experiment 2: Instead of fixing the *spatiotemporal anisotropy parameter* c, we search for the optimal value. The parameters considered here include c, n, and p. Furthermore, each parameter configuration was run across three 10-CV partitions with the resulting error statistics averaged (to reduce the effect that using a particular partition might have). This set up amounted to 16,848 configurations that were tested in 144.6 hours (4,694.3 hours in sequential time). As expected, the prediction accuracy, measured by MARE, is increased from 1.2058 to 0.3791. This result is actually better than the current best accuracy under the 10-CV, which is 0.3866 and was produced by Li et al. (2012)'s shape-function-based method.

We are 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 four directions: (1) scanning a wider parameterization space and further optimizing search methods for the parameter configurations, (2) exploring alternate machine learning methods such as *leave-one-out cross validation* and *random forest*, (3) attempting other spatiotemporal methods such as *shape function* and *Kriging* based methods, and (4) analyzing 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, et al., 2012. Estimating Population Exposure to Fine Particulate Matter in the Conterminous U.S. using Shape Function-based Spatiotemporal Interpolation Method: A County Level Analysis. *GSTF: International Journal on Computing*, 1:24–30.

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.