EDA and a Tailored Data Imputation Algorithm for Daily Ozone Concentrations

Gualán, R. a, b, Saquicela, V. a, Tran-Thanh, L. b

aDepartment of Computer Science, University of Cuenca, Cuenca, Ecuador
bSchool of Electronics and Computer Science, University of Southampton, Southampton, SO17 1BJ, United Kingdom

Abstract

Air pollution is a critical environmental problem with detrimental effects on human health that is affecting all regions in the world, especially to low-income cities, where critical levels have been reached. Air pollution has a direct role in public health, climate change, and worldwide economy. Effective actions to mitigate air pollution, e.g. research and decision making, require of the availability of high resolution observations. This has motivated the emergence of new low-cost sensor technologies, which have the potential to provide high resolution data thanks to their accessible prices. However, since low-cost sensors are built with relatively low-cost materials, they tend to be unreliable. That is, measurements from low-cost sensors are prone to errors, gaps, bias and noise. All these problems need to be solved before the data can be used to support research or decision making. In this paper, we address the problem of data imputation on a daily air pollution data set with relatively small gaps. Our main contributions are: (1) an air pollution data set composed by several air pollution concentrations including criteria gases and thirteen meteorological covariates; and (2) a custom algorithm for data imputation of daily ozone concentrations based on a trend surface and a Gaussian Process. Data Visualization techniques were extensively used along this work, as they are useful tools for understanding the multi-dimensionality of point-referenced sensor data. © 2019, Springer Nature Switzerland AG.
The Air Data project from the U.S. Environmental Protection Agency (EPA) collects air quality and weather measurements from more than 4000 outdoor monitors across the United States, Puerto Rico, and the Virgin Islands [22]. The data provided cover a time period from 1980 to 2017 in hourly, daily and annual aggregation. For this study, we downloaded the pre-generated daily files for the year 2016. The available files are grouped in two categories: (1) Criteria Gases such as Ozone, Sulfur dioxide (SO2), Carbon monoxide (CO), and Nitrogen dioxide (NO2); and (2) Particulates, such as PM2.5 (particles with a diameter of 2.5 μm or less), PM10 (particles with a diameter of 10 μm or less) [22].

Motivated by the general challenge of studying spatio-temporal modeling of air pollution concentrations measured from cheap sensor networks, we decided to assemble a real-life data set made up of ozone concentrations taken from the Air Data project from the U.S. Environmental Protection Agency (EPA) plus weather covariates taken from the NCEP/NCAR Reanalysis Project (NNRP). The biggest problem with this data set, was the missing values in the target variable (ozone). This problem was addressed by means of a tailored algorithm inspired by the hierarchical models. Throughout this study we extensively applied several Exploratory Data Analysis techniques and multivariate analysis to understand the data and the relationships between the features.

References (26)

<table>
<thead>
<tr>
<th></th>
<th>Author(s)</th>
<th>Title</th>
<th>Cited Times</th>
<th>Publisher</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Bakar, K.S., Sahu, S.K.</td>
<td>spTimer: Spatio-temporal bayesian modeling using R</td>
<td>39</td>
<td>Open Access</td>
</tr>
<tr>
<td>4</td>
<td>Cameletti, M., Lindgren, F., Simpson, D., Rue, H.</td>
<td><em>Spatio-temporal modeling of particulate matter concentration through the SPDE approach</em></td>
<td>124</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Campozano, L., Sánchez, E., Avilés, A., Samaniego, E.</td>
<td>Evaluation of infilling methods for time series of daily precipitation and temperature: The case of the ecuadorian andes</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Gelfand, A.E.</td>
<td>Hierarchical modeling for spatial data problems</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Gräler, B., Pebesma, E., Heuvelink, G.</td>
<td><em>Spatio-temporal interpolation using gstat</em></td>
<td>90</td>
<td></td>
</tr>
</tbody>
</table>
10 Hasenfratz, D., Saukh, O., Sturzenegger, S., Thiele, L. 
Participatory air pollution monitoring using smartphones 

11 Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., (...), Joseph, D. 
The NCEP/NCAR 40-year reanalysis project 
http://ams.allenpress.com 
View at Publisher

12 Mukhopadhyay, S., Sahu, S.K. 
A Bayesian spatiotemporal model to estimate long-term exposure to outdoor air pollution at coarser administrative geographies in England and Wales (Open Access) 
http://www.ingenta.com/journals/browse/bpl/rssa 
doi: 10.1111/rssa.12299 
View at Publisher

13 Pirani, M., Gulliver, J., Fuller, G.W., Blangiardo, M. 
Bayesian spatiotemporal modelling for the assessment of short-term exposure to particle pollution in urban areas (Open Access) 
doi: 10.1038/jes.2013.85 
View at Publisher

https://www.r-project.org/

15 (2012) *Matrix Factorization as Data Imputation / S3l* 
http://s3l.stanford.edu/blog/?p=66

16 Kumar, S.S., Bakar, K.S. 
Hierarchical Bayesian auto-regressive models for large space time data with applications to ozone concentration modelling 
doi: 10.1002/asmb.1952 
View at Publisher

17 Sahu, S.K., Gelfand, A.E., Holland, D.M. 
High-resolution space-time ozone modeling for assessing trends 
doi: 10.1198/016214507000000031 
View at Publisher
18. Samworth, R.J.
Optimal weighted nearest neighbour classifiers (Open Access)
http://projecteuclid.org/DPubS/Repository/1.0/Disseminate?
handle=euclid.aos/1359987536&view=body&content-type=pdfview_1
doi: 10.1214/12-AOS1049
View at Publisher

19. Seo, J., Youn, D., Kim, J.Y., Lee, H.
Extensive spatiotemporal analyses of surface ozone and related meteorological variables in South Korea for the period 1999-2010 (Open Access)
http://www.atmos-chem-phys.net/volumes_and_issues.html
doi: 10.5194/acp-14-6395-2014
View at Publisher

The changing paradigm of air pollution monitoring
doi: 10.1021/es4022602
View at Publisher

Representing situational knowledge acquired from sensor data for atmospheric phenomena
www.elsevier.com/inca/publications/store/4/2/2/9/2/1
doi: 10.1016/j.envsoft.2014.04.006
View at Publisher

22. Us, E.P.A.
(2016) Air Data Basic Information | Air Data: Air Quality Data Collected at Outdoor Monitors across the US | US EPA
https://www.epa.gov/outdoor-air-quality-data/air-data-basic-information

23. Wen, H., Xiao, Z., Markham, A., Trigoni, N.
Accuracy Estimation for Sensor Systems
doi: 10.1109/TMC.2014.2352262
View at Publisher

24. WHO Global Urban Ambient Air Pollution Database. Cited 119 times.
http://www.who.int/phe/healthtopics/outdoorair/databases/cities/en/
Spatio-temporal modeling of particulate air pollution in the conterminous United States using geographic and meteorological predictors

http://www.ehjournal.net/home/

View at Publisher

Imputation methods for filling missing data in urban air pollution data for Malaysia

http://uac.incd.ro/Art/v9n2a04.pdf

Gualán, R.; Department of Computer Science, University of Cuenca, Cuenca, Ecuador;
email: ronald.gualan@ucuenca.edu.ec
© Copyright 2018 Elsevier B.V., All rights reserved.