Contribution ID: **11** Type: **Oral presentation**

Graph machine learning for improved imputation of missing tropospheric ozone data

Monday, 6 March 2023 12:00 (20 minutes)

Gaps in the measurement series of atmospheric pollutants can impede the reliable assessment of their impacts and trends. Data imputation methods to close the gaps in the observation series range from simple linear interpolation to machine learning. We propose a new method for missing data imputation of the air pollutant tropospheric ozone by using the graph machine learning algorithm 'correct and smooth'. This algorithm uses auxiliary data that characterize the measurement location and, in addition, ozone observations at neighboring sites. Specifically, we apply this method to the missing data of a preliminary dataset from 278 stations of the year 2011 of the German Environment Agency (Umweltbundesamt - UBA) monitoring network. These data exhibit three distinct, frequently occurring gap patterns: shorter gaps in the range of hours, longer gaps of up to several months in length, and gaps occurring at multiple stations at once. We apply correct and smooth as a post-processing algorithm after imputing the missing data with different statistical and machine learning methods.

For short gaps of up to five hours, linear interpolation is most accurate with R^2 values of 0.91 - 0.97, RMSEs of 2.43 - 4.44 ppb, and indexes of agreement of 0.98 - 0.99. Longer gaps at single stations are most effectively imputed by a random forest in connection with correct and smooth, with R^2 values of 0.86 - 0.87, RMSEs of 5.64 - 6.18 ppb, and an index of agreement of 0.96. This case exhibits strong improvement through the correct and smooth algorithm, as the RMSEs decreased by 0.57 - 0.76 ppb compared to the random forest alone. For longer gaps at multiple stations, the correct and smooth algorithm improved the random forest RMSE by 0.07 ppb, despite a lack of data in the neighborhood of the missing values. Based on these results, we suggest applying a hybrid of linear interpolation and graph machine learning for the imputation of tropospheric ozone time series.

ML method

Other

Main air pollutant of interest

Tropospheric ozone and precursors

Primary author: BETANCOURT, Clara (Forschungszentrum Jülich)

Co-authors: Ms LI, Cathy W. Y. (Max-Planck Institut für Meteorologie); Mr KLEINERT, Felix (Forschungszentrum Jülich); Mr SCHULTZ, Martin G. (Forschungszentrum Jülich)

Presenter: BETANCOURT, Clara (Forschungszentrum Jülich)

Track Classification: Machine learning applications