S.G. Gocheva-Ilieva, A.V. Ivanov and I.E. Livieris. High Performance Machine Learning Models of Large Scale Air Pollution Data in Urban Area
.
Cybernetics and Information Technologies
, 2020.

Abstract - Preserving the air quality in urban areas is crucial for the health of the population as well as for the environment. The availability of large volumes of measurement data on the concentrations of air pollutants enables their analysis and modelling to establish trends and dependencies in order to forecast and prevent future pollution. This study proposes a new approach for modelling air pollutants data using the powerful machine learning method Random Forest (RF) and Auto-Regressive Integrated Moving Average (ARIMA) methodology. Initially, a RF model of the pollutant is built and analysed in relation to the meteorological variables. This model is then corrected through subsequent modelling of its residuals using the univariate ARIMA. The approach is demonstrated for hourly data on seven air pollutants (O3, NOx, NO, NO2, CO, SO2 and PM10) in the town of Dimitrovgrad, Bulgaria over 9 years and 3 months. Six meteorological and three time variables are used as predictors. High-performance models are obtained explaining the data with R2 = 90%-98%.