

Comparison of several linear statistical models to predict tropospheric ozone concentrations

José C. M. Pires, Fernando G. Martins,
Maria C. M. Alvim-Ferraz, and Maria C. Pereira

University of Porto, Portugal

Abstract

This study aims to evaluate the performance of five linear statistical models in the prediction of the next day hourly average ozone (O_3) concentrations. The selected models are: (i) multiple linear regression (MLR); (ii) principal component regression (PCR); (iii) independent component regression (ICR); (iv) quantile regression (QR); and (v) partial least squares regression (PLSR). As far as it is known, no study was presented comparing the performance of these five linear models for predicting tropospheric ozone concentrations. Moreover, it is the first time that ICR is applied with this aim. The considered ozone predictors are meteorological data (hourly averages of temperature, relative humidity and wind speed) and environmental data (hourly average concentrations of sulphur dioxide, carbon monoxide, nitrogen oxide, nitrogen dioxide and ozone) of the previous day collected in an urban site with traffic influences. The analysed period was May and June 2003.

The QR model, which tries to model the entire distribution of the O_3 concentrations, presents better performance in the training step, because it tries to model the entire distribution of the O_3 concentrations. However, it presents worst predictions in the test step. This means that a new procedure should be found better than the one applied (k nearest neighbours algorithm) that can be able to estimate the percentiles of the output variable in the test dataset with more precision. From the five statistical models tested in this study, the PLSR model presents the best predictions of tropospheric ozone concentrations.

Keywords

Air pollution, Tropospheric ozone, Statistical models, Concentration level prediction.