

Conference Proceedings Paper

Time Series Clustering to Estimate Particulate Matter Contributions from Deserts

Á. Gómez-Losada ^{1,*}, J. C. M. Pires ² and R. Pino-Mejías ¹

Published: 15 July 2016

¹ Departamento de Estadística e Investigación Operativa, Facultad de Matemáticas, Universidad de Sevilla, Avda. Reina Mercedes s/n, 41001 Sevilla, España

² LEPABE, Departamento de Engenharia Química, Faculdade de Engenharia, Universidade do Porto, Rua Dr. Roberto Frias s/n, 4200-465 Porto, Portugal; jcpires@fe.up.pt

* Correspondence: alvaro.gomez.losada@gmail.com

Abstract: Exploratory analysis of time series (TS) data is an important approach in experimental studies, with a large range of applications in many different fields, including air pollution studies. To identify structures in single (univariate) TS, main clustering analyses are based on general-purpose clustering algorithms (e.g., k-means, hierarchical clustering methods) and made the assumption that the samples (data) of a TS are independent, ignoring the correlations in consecutive sample values in time. This is specially the case of air pollutant studies based on monitoring data. Air pollutants TS can be studied using TS clustering techniques and as a result, pollution profiles or concentration regimes detected as well as the dependency structure between consecutive data is preserved. Once TS clustering applied over the TS data stream, a set of clusters group the data according to their similar concentration values, and therefore, different pollution profiles can be defined and their estimated range of concentration values. Hidden Markov Models (HMMs) are flexible general-purpose models for univariate and multivariate TS. The TS data are assumed to have a Markov property, and may be viewed as the results of a probabilistic walk along a fixed set of (no directly observable) states. This class of approach considers that each TS is generated by a mixture of underlying probability distributions, typically the Gaussian ones. In this study, HMMs were applied to cluster daily average particulate matter with aerodynamic diameter of 10 μm or less (PM_{10}) TS collected at background monitoring stations from the Iberian Peninsula and Canarian Archipelago (Spain). As a result, PM_{10} concentration regimes were studied and in particular, the contribution to PM_{10} ambient concentration levels from the regimes associated to transport of air masses from North Africa deserts was estimated. Regarding this last contribution, we later compared to those obtained using the monthly moving 40th percentile (P40) method over the same TS and no significant quantitative differences were detected. However, the results obtained with HMMs seem to correct the net load of PM_{10} given by the P40 method, and attributes less impact on areas suffering greater influence from African episodes. The method proposed in this work to estimate PM_{10} from deserts could improve the P40 method in two ways since it avoids: (i) the smoothed effect which is implicit in the P40 methods after applying a mobile procedure in the TS treatment; and (ii) the empirical approach based on a correlation analysis applied in order to select this particular percentile (40th). Moreover, the use of statistical replicative techniques (bootstrap) together with HMMs has let to obtain an interval confidence in the PM_{10} contribution estimates from North African deserts. This methodology may be used to estimate particulate matter contributions from any desert; however, a consensus among experts is required to give the regimes obtained with HMMs a definition.

Keywords: time series clustering; hidden Markov models; desert contributions; monthly moving 40th percentile method



© 2016 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons by Attribution (CC-BY) license (<http://creativecommons.org/licenses/by/4.0/>).