Multiple changepoint detection on air pollution via genetic algorithms with bayesian-MDL on nonhomogeneous Poisson periods

Arrigo Coen
Faculty of Sciences, UNAM, México
coen@ciencias.unam.mx

Introduction

Particulate matter, also referred to as PM, is a mixture of solid particles and liquid droplets found in the air. Some particles, such as dust, dirt, soot or smoke, are large or dark enough to be seen with the naked eye. Others are so small they can only be detected using an electron microscope.

Because particulate matter is so small, it can be inhaled and cause serious health problems. Particles less than 10 micrometers (PM10) in diameter pose the greatest problems as they can get deep into peoples lungs and some may even get into the bloodstream.

One way to find out if a new environmental policy has a positive effect is to look for changes in the probabilistic behavior of the time series after the implementation of a new policy.

Figure 1: Comparison of Bogota sky with and without pollution

Model

It is assume the existence of changepoints $\tau = \{\tau_1, \ldots, \tau_m\}$, with $\tau_i < \tau_{i+1}$, such that the number of PM10 days that exceed the threshold between two subsequent points can be modeled as nonhomogeneous Poisson processes. The question is how to find the solution to:

$$\begin{aligned}\text{Bayesian}-\text{MDL} =& \arg\min_{\theta,\tau}\ \log P_{\theta,\tau} - \log f(\theta|D)\\ =& \arg\min_{\theta,\tau}\ \log P_{\theta,\tau} - \log f(D|\theta) - \log f(\theta)\end{aligned}$$

where $P_{\theta,\tau}$ is the penalty

$$\log P_{\theta,\tau} = \sum_{i=1}^{m+1} \frac{\log(\tau_i - \tau_{i-1})}{2} + \log(m) + \sum_{i=2}^{m} \log(\tau_i),$$

$f(\theta|D)$ is the posterior distribution, $f(D|\theta)$ is the likelihood, and $f(\theta)$ is the prior distribution, where D denotes the data. [1].

Data

Here is analyzed the time series of particulate matter exceedances from January 1, 2004 to August 31, 2015 of Bogota, Colombia. In particular, we analyzed the time series of exceedances of the 100 level.

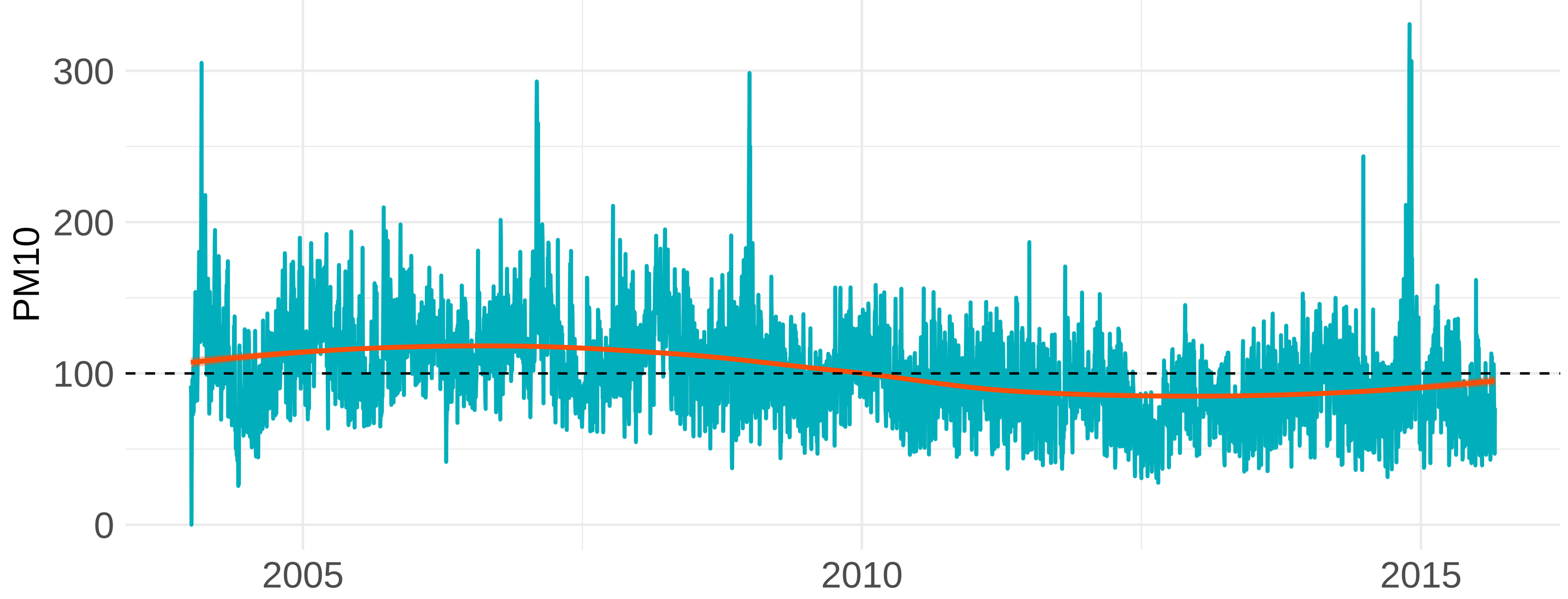

Figure 2: Time series of PM10 (green) with smoothed line (red) and the 100 threshold (black)

Nonhomogeneous Poisson Periods

For the i-period we assume the likelihood

$$f_i(D|\theta) \propto e^{-[\Lambda(\tau_i|\theta_i) - \Lambda(\tau_{i-1}|\theta_i)]} \prod_{j=N_{\tau_{i-1}}+1}^{N_{\tau_i}} \lambda(d_j|\theta_i), \qquad i = 1, \ldots, m+1,$$

where d_i are the exceedances days, λ is the rate function and Λ is the mean function of the nonhomogeneous Poisson processes, under the assumption that λ is nonnegative with $\Lambda(t) = \int_0^t \lambda(s)ds$. In particular, to model the data we use the next mean functions [2]:

$$m(t|\theta) = \beta \log(1 + t/\alpha) \qquad \text{Musa} - \text{Okumoto}$$

$$m(t|\theta) = -\log\left(1 - \left[1 - e^{-(t/\sigma)^\alpha}\right]^\beta\right) \qquad \text{Exponentiated} - \text{Weibull}$$

$$m(t|\theta) = \alpha\left(1 - e^{-\beta t^\sigma}\right) \qquad \text{Generalized Goel} - \text{Okumoto}$$

Genetic Algorithm

To find the solution to the minimization problem we apply a genetic algorithm to search the best number and position of changepoints. For each fixed period (τ_{i-1}, τ_i) we find the MAP parameters.

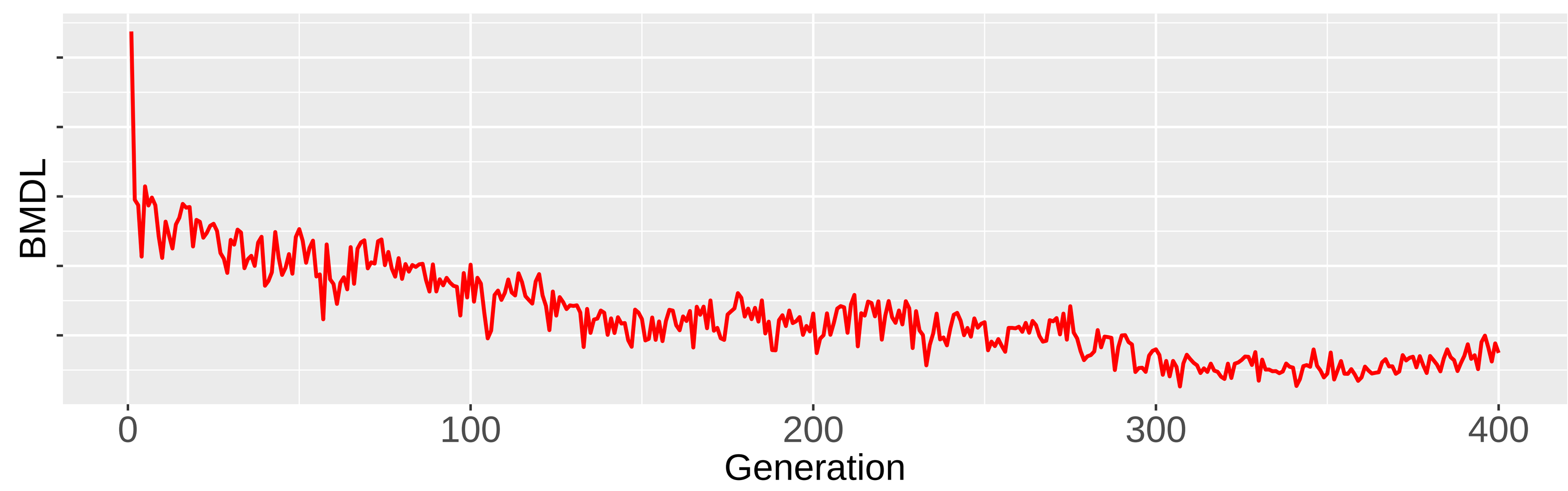

Figure 3: Evolution of genetic algorithm values

Under a loop we apply the steps: selection, crossover, mutation and acceptance. All these steps depend on the values of BMDL.

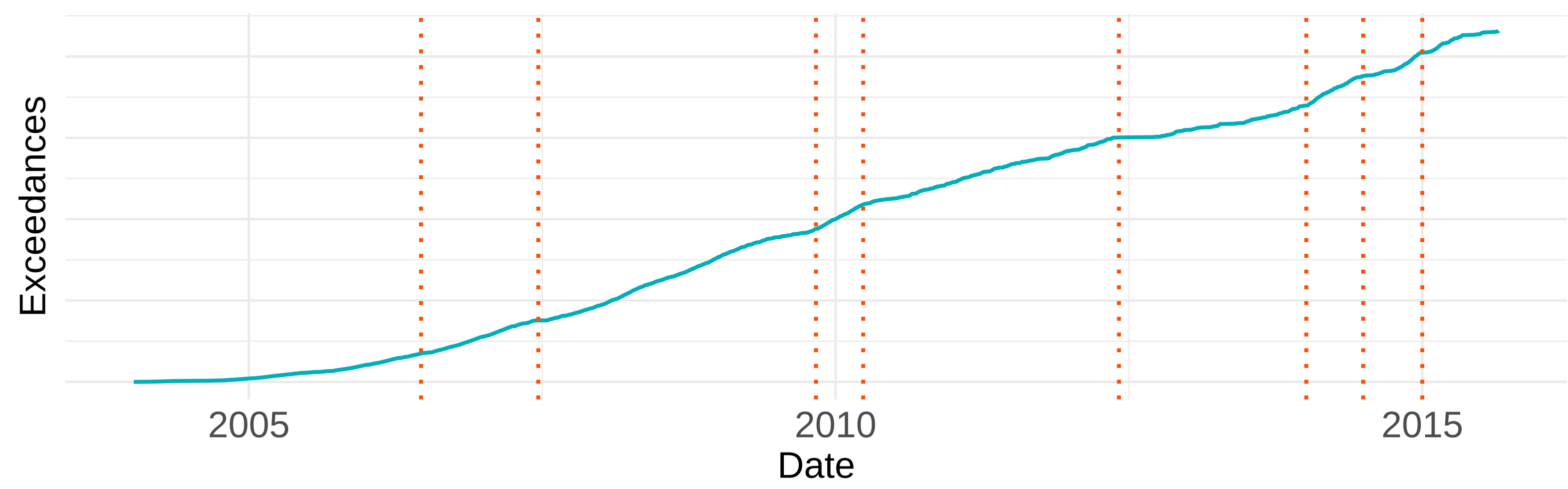

Figure 4: Exceedances data (green) and changepoints (red) of Exponentiated-Weibull model

Conclusions

- The changepoint detection model presented here correctly detects the environmental policies implemented on Bogota
- For the Bogota's data, the Exponential-Weibull model fits better
- The model could be applied to find the impact of environmental policies of other countries

Acknowledgements

The work of the author was supported by DGAPA Posdoctoral program

References

[1] Shanghong Li, Robert Lund, Shanghong Li, and Robert Lund. Multiple Changepoint Detection via Genetic Algorithms. *Journal of Climate*, 25(2):674–686, jan 2012.

[2] Eliane Regina Rodrigues and Jorge Alberto Achcar. *Applications of Discrete-time Markov Chains and Poisson Processes to Air Pollution Modeling and Studies*. SpringerBriefs in Mathematics. Springer New York, New York, NY, 2013.