Genetic algorithm with a Bayesian approach for the detection of multiple points of change of time series of counting exceedances of specific thresholds.

Biviana Marcela Suárez Sierra (✉ bmsuarezs@eafit.edu.co)  
EAFIT University  
Arrigo Coen  
National Autonomous University of Mexico  
Carlos Alberto Taimal  
EAFIT University

Research Article

Keywords: Multiple Change-point Detection, Genetic Algorithms, Minimum Description Length, Non-homogeneous Poisson Processes, Maximum A Posteriori Estimation

Posted Date: February 13th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2566481/v1

License: ☑️ This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.
Genetic algorithm with a Bayesian approach for the detection of multiple points of change of time series of counting exceedances of specific thresholds.

Biviana Marcela Suárez Sierra · Arrigo Coen · Carlos Alberto Taimal

Abstract Although the applications of Non-Homogeneous Poisson Processes to model and study the threshold overshoots of interest in different time series of measurements have proven to provide good results, they needed to be complemented with an efficient and automatic diagnostic technique to establish the location of the change-points, which, when taken into account, make the estimated model fit poorly in regards of the information contained in the real model. For this reason, we propose a new method to solve the segmentation uncertainty of the time series of measurements, where the emission distribution of exceedances of a specific threshold is the focus of investigation. One of the great contributions of the present algorithm is that all the days that overflowed are candidates to be a change-point, so all the possible configurations of overflow days are the possible chromosomes, which will unite to have offspring. Under the heuristics of a genetic algorithm, the solution to the problem of finding such change points will be guaranteed to be non-local and the best possible one, reducing wasted machine time evaluating the least likely chromosomes to be a solution to the problem. The analytical evaluation technique will be by means of the Minimum Description Length (MDL) as the objective function, which is the joint posterior distribution function of the parameters of each regime and the change points that determines them and which account as well for the influence of the presence of said times. Thus, one
of the practical implications of the present work comes in terms of overcoming the need of modeling the time series of measurements, where the distributions of exceedances of certain thresholds, or where the counting of certain events involving abrupt changes, is the main objective. For this reason, the present proposal will be very useful in phenomena related to climate change, information security, epidemiology, to name a few.

Keywords Multiple Change-point Detection · Genetic Algorithms · Minimum Description Length · Non-homogeneous Poisson Processes · Maximum A Posteriori Estimation

1 Introduction

The study of exceedances of a specific threshold in the time series of environmental measurements [16],[18], data traffic measurements in computer security [10] or the measurement of the prevalence of tuberculosis in New York [3] in certain time periods are some examples of the practicality of a model that allows us to establish whether certain exceedance mitigation contingencies in the measurements have worked correctly in an observed period. That is, if there are marks or moments in the time series that indicate decrease, growth or steadiness in the rate of exceedance emission. These moments will be represented as abrupt changes, which in the model will be represented as change-points. One of the most commonly used methodologies for this purpose is based on Non-Homogeneous Poisson Processes, in which the change-points that determine the different regimes, which make up the entire time series, are estimated. In the same way, the parameters of the distribution of the excesses that compose each regime are estimated. For this, most of the time, the prior distribution of the change-points is considered to be the uniform distribution, with hyperparameters at the extremes of the interval where the researcher approximately observes the change-point. From there and from a Bayesian mechanism, the joint distribution of the change-points is updated with the rest of the parameters of each regime. The likelihood function of the observed overshootings and the aforementioned prior distribution are multiplied to obtain the posterior distribution. Using a Bayesian computational technique such as Markov Chain Monte Carlo (MCMC) on the posterior distribution, the parameters and change-points of interest are estimated.

One of the disadvantages that arise in the implementation of the aforementioned models is the model execution time, because the more change points are considered, the longer the time it will take to reach an optimum. But on the other hand, the greater the number of change points, the greater the adjustment of the estimated to the observed, allowing less loss of information. The latter can be seen by comparing the Deviance information criterion (DIC) of the models with zero, one, two or more change points in [18]. So we want a model that continues with the same goodness of fit, but automates the detection of change points, in short run times. Therefore, in the present work we
make a new proposal that unifies the principles of construction of the likelihood function from non-homogeneous Poisson processes (NHPP), as it is a counting problem, and on the other hand, Bayesian computational methods to estimate the parameters of each regime and its change points. After the joint posterior distribution obtained, it will be combined with information theory principles to penalize the integer and real parameters present in the segmentation of the time series. All the above under the heuristics of a genetic algorithm, will convert the penalized posterior distribution into the objective function. Hence, the problem of solving the segmentation uncertainty with the present proposal is reduced to finding the penalized maximum posterior estimator (MAP), which corresponds to a modification of the Minimum Description Length principle (MDL) and such we will refer to it as the Bayesian-MDL.

The cumulative mean function, which uniquely determines the NHPP, provides more information about the number of threshold exceedances up to a given time $t$. This mean function has an associated intensity function that will indicate the time trend in each of the segments and the speed of emission of the threshold overshoots of interest. The parameters of these functions will be estimated with the genetic algorithm, to which the starting conditions of the first generation of change-point configurations or chromosomes can be set. The latter will be responsible for generating the offspring with the best characteristics. To guarantee this, the mutation distribution, mating dynamics and reordering between chromosomes will be implemented and tested as in [11]. The execution time is also declared at the beginning of the algorithm and is determined by the number of generations. From each generation will come out the best of the children, which will be evaluated by the MDL, hence the chromosome with the lowest MDL among all the generations will be the best of all. In this type of algorithms, the execution time is compensated by the quality of the solution. Another advantage of this model is that once the cumulative mean function is known, the distribution of the number of overflows present at time $t$ can be explicitly established and from there different forecasting exercises can be performed.

It should also be established that as per the characterization exposed in [4] for the type of method, the here proposed procedure is of the off-line type as the whole dataset is considered at once in a retrospective manner and the conditions set for the genetic algorithm are evaluated to determine an optimal solution. At the end of the algorithm execution, it will return time segments that will be independent and identically distributed ($iid$) conditioned to the parameters of each segment. Assuming that the observations in each segment determined by the proposed algorithm are $iid$ may be wrong in the presence of correlation signals. However, [10] comments that in the presence of outliers, the assumption turns out to be robust, as in the case of [8] and the present work.
The presentation of this work is structured as follows. In the second section, a background review will be presented related to what has been developed in the path of NHPP as methods of adjustment of the univariate function of cumulative mean of the exceedances of a threshold of interest up to time \( t \) and its respective inferential analysis. In this same section the scheme of the genetic algorithm will be presented, displaying the start, pairing and stop conditions to guarantee that the solution is optimal. Section three will show the theoretical construction of the computational model for four different forms of the intensity function \( \lambda(t) \) proposed in the literature. In section four we will present the results with three sets of simulated data and one set of real data. From the simulated data, the performance of the algorithm will be shown to determine the number of change points, their location and the number of generations necessary to reach the optimal solution. With the real dataset, we will evaluate the ability of the algorithm to assess the impact of public policy on air pollution data in Bogotá by \( PM_{2.5} \) between the years 2018 and 2020. Finally, in section five we will establish the conclusions and future considerations of this work.

### 2 Inferential Analysis of the Univariate Non-Homogeneous Poisson Process

Based on the construction of the likelihood from a non-homogeneous Poisson process discussed in [5], we have the following expression, from which we will start to construct the penalized MAP function. The developments that will be shown in the following sections will be based on the assumption that the true intensity function of the overshoots corresponds to some of the function in (1). Then the process that we have correspond to a NHPP with intensities varying with respect to \( t \),

\[
\begin{align*}
\lambda^{(W)}(t|\theta) &= (\alpha/\beta)(t/\beta)^{\alpha-1}, & \alpha, \beta > 0 \\
\lambda^{(MO)}(t|\theta) &= \frac{\beta}{t^\alpha}, & \alpha, \beta > 0 \\
\lambda^{(GO)}(t|\theta) &= \alpha \beta \exp(-\beta t), & \alpha, \beta > 0 \\
\lambda^{(GGO)}(t|\theta) &= \alpha \beta \gamma t^{-1} \exp(-\beta \gamma t), & \alpha, \beta, \gamma > 0 \\
\end{align*}
\]

(1)

Each letter in the superscript of the left-hand term corresponds to the initial letter of the name of the distribution used as intensity function as follows, Weibull \((W)\) [12,15], Musa-Okumoto \((MO)\) [13], Goel-Okumoto \((GO)\) and a generalization of the Goel-Okumoto model \((GGO)\) [9], for which the cumulative average function, \( m(t|\theta) \) is defined respectively by:

\[
\begin{align*}
m^{(W)}(t|\theta) &= (t/\beta)\alpha, & \alpha, \beta > 0 \\
m^{(MO)}(t|\theta) &= \beta \log \left(1 + \frac{t}{\beta}\right), & \alpha, \beta > 0 \\
m^{(GO)}(t|\theta) &= \alpha [1 - \exp(-\beta t)], & \alpha, \beta > 0 \\
m^{(GGO)}(t|\theta) &= \alpha [1 - \exp(-\beta \gamma t)], & \alpha, \beta, \gamma > 0. \\
\end{align*}
\]

(2)
In the first three cases, the vector of parameters is \( \theta = (\alpha, \beta) \) and for the last one, \( \theta = (\alpha, \beta, \gamma) \).

The way the rate functions of the process are formulated shows that the observations within the same segment are equally distributed and independent on the parameters of the emission intensity function of the observations in that segment. Hence, this information is taken into account in the Bayesian-MDL to correspond to the growth, decrease or steadiness of excess emissions in each interval of the partition determined by the change points. For example as exposed in [16], when considering change points in air pollution models, the Weibull rate function shows a better fit to the GGO, without and with change points.

The present model determines the number of change points and their specific location over the time series of interest. Since the data suggested to be studied with this type of model are exceedances of a specific threshold, the nature of the change points will represent a before and after abrupt variations in the rate of growth or decay of the exceedances in the cumulative mean function, which in turn will represent abrupt changes in the observations over the original time series.

2.1 Likelihood function with change points

A change point is defined as an instant where the structural pattern of a time series shifts [4,19]. We assumed the presence of \( J \) change-points, \( \{\tau_1, \tau_2, \ldots, \tau_J\} \) such that there are variations on the model parameters in between segments \( \tau_{j-1} < t < \tau_j, j = 0, 1, 2, \ldots, J + 1, j_0 = 1, j_{J+1} = T \). These changes can be attributed to environmental policies or legislations in a certain year, the suspension of some network station due to maintenance for a climate case, macroeconomic policies from an economic standpoint or the presence of a stimulus in a neuroscience context. On the other side, assuming the overshoots from a given threshold between two regimes determined by the presence of a change-point can be modeled after a NHPP, then the intensity functions have the form,

\[
\lambda(t|\theta) = \begin{cases} 
\lambda(t|\theta_1), & 0 \leq t < \tau_1, \\
\lambda(t|\theta_j), & \tau_{j-1} \leq t < \tau_j, j = 2, 3, \ldots, J, \\
\lambda(t|\theta_{J+1}), & \tau_J \leq t \leq T,
\end{cases}
\]  

(3)

where \( \theta_j \) is the vector of parameters between the change-points \( \tau_{j-1} \) and \( \tau_j \), for \( j = 2, \ldots, J \), and \( \theta_1 \) and \( \theta_{J+1} \) are the parameter vectors before and after the first and last change-points, respectively. With \( n \) observations, the functions for the means are (see, e.g., [16]),
\[ m(t|\theta) = \begin{cases} 
 m(t|\theta_1), & 0 \leq t < \tau_1, \\
 m(\tau_1|\theta_1) + m(t|\theta_2) - m(\tau_1|\theta_2), & \tau_1 \leq t < \tau_2, \\
 m(t|\theta_{j+1}) - m(\tau_j|\theta_{j+1}) + \sum_{i=2}^{j}[m(\tau_i|\theta_i) - m(\tau_{i-1}|\theta_i)], & \tau_j \leq t < \tau_{j+1}, j = 2, 3, \ldots, J, 
\end{cases} \]

where \( \tau_{J+1} = T \). That is, because \( m(t|\theta_1) \) represents the average number of exceedances of the standard, before the first change-point. \( m(\tau_1|\theta_1) + m(t|\theta_2) - m(\tau_1|\theta_2) \) is the average number of exceedances of the standard between the first change-point \( \tau_1 \) and the second one \( \tau_2 \), given that the vector of parameters \( \theta_2 \) is known, and so on.

Be \( D = d_1, \ldots, d_n \), where \( d_k \) (as in the case without change-points), is the time of occurrence of the \( k \) \( t \) event (the \( k \) \( t \) time the maximum level of the environmental standard is exceeded, for example), with \( k = 1, 2, \ldots, n \), the likelihood function is determined by the expression below where \( N_{\tau_i} \) represents the number of exceedances before the change-point \( \tau_i \), with \( i = 1, 2, \ldots, J \) (see [20, 2]).

\[
L(D|\phi) \propto \prod_{i=1}^{N_{\tau_1}} \lambda(d_i \mid \theta_1) e^{-m(\tau_1|\theta_1)} 
\times \prod_{j=2}^{J} \left[ \prod_{i=N_{\tau_{j-1}}+1}^{N_{\tau_j}} \lambda(d_i \mid \theta_j) e^{-[m(\tau_j|\theta_j) - m(\tau_{j-1}|\theta_j)]} \right] 
\times \prod_{i=N_{\tau_J}+1}^{n} \lambda(d_i \mid \theta_{J+1}) e^{-[m(T|\theta_{J+1}) - m(\tau_{J+1}|\theta_{J+1})]},
\]

Using the expression (5), we infer the parameters \( \phi = (\theta, \tau) \), with \( \theta = (\theta_1, \ldots, \theta_J) \) and \( \tau = (\tau_1, \ldots, \tau_J) \) using a Bayesian approach. This perspective consists of finding the relationship between the a priori distribution of the parameter \( \theta \), on whose intensity function \( \lambda(t|\theta) \) is dependent and the a posteriori distribution of the same, after taking into consideration the observed information \( D \). In [1], this method was applied to obtained results very close to the observed ones, hence the descriptive capacity of the model and the methodology used. In such work, the criteria used to select the model that best fits the data together with the graphic part was the MDL.
2.2 Detection of multiple change points using genetic algorithm

2.2.1 MDL framework

Since finding J change-points implies finding out J + 1 regimes for the time series or fitting J + 1 models with different parameters, statistical criteria has been used for such purpose in the available literature. Some include the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), Cross-Validation methods, and MDL-based methods. For problems involving regime shift detection, MDL methods usually provide superior empirical results. This superiority is probably due to the fact that both AIC and BIC apply the same penalty to all parameters, regardless of the nature of the parameter. On the other hand, MDL methods can adapt penalties to parameters depending on their nature be it continuous or discrete, bounded or not. In short, MDL defines the best fitting model as the one that enables the best compression of the data minimizing a penalized likelihood function. That is,

$$MDL = -\log_2(L_{opt}) + P. \quad (6)$$

Here $\log_2(L_{opt})$ is the required amount of information needed to store the fitted model, term taken from information theory. More details on this can be found in [6]. $L_{opt}$ is obtained from replacing the maximum likelihood estimator in the likelihood function (5). This will be explained in more detail in the next section.

Because of the above, it is possible to make the natural connection between the likelihood and the MDL objective function by means of the penalty P (see [6]). The broad penalty methodology is summarized in three principles as stated by [11]. The first one is to penalize the real valued parameters by the number of observations. Say $k$ that are used to estimate it, then, the penalty will be $\log_2 k$. For this principle, it is important to take into consideration how the observations are arranged to calculate the parameter of interest because this arrangement will be reflected in the penalty.

The second principle involves the penalty of how many integer parameters, such as the number of change points $J$ and where they are located represented by $\tau_1, ..., \tau_J$ should be charged. This charging is calculated based on the value for each of them. For example, the quantity $J$, which is bounded by the total number of observations $T$ is charged an amount of $\log_2 T$. For each of the $\tau_j$ with $j = 1, ..., J$, we have that $\tau_j < \tau_{j+1}$, therefore the cost of its penalty will be $\frac{\log_2 \tau_{j+1}}{2}$ for $j = 2, ..., J$.

The last principle, mentioned in [11], is the additivity principle. It involves constructing $P$ based on the sum of all the partial penalties mentioned above. The more parameters the model has, the higher $P$ will be. However, if despite adding parameters, the expression $\log_2(L_{opt})$ does not grow larger than the
penalty $P$ of the extra parameters, the simpler model will be preferred. For the purposes of this paper, the following will be used as the penalty function $P_\tau(\theta)$ for a fixed change point configuration,

$$P_\tau(\theta) = R \sum_{j=1}^{J+1} \frac{\ln(\tau_1 - \tau_{i-1})}{2} + \ln(J) + \sum_{j=2}^{J} \ln(\tau_j), \quad (7)$$

where $R = 2, 3$ depending on whether $\theta = (\alpha, \beta)$ or $\theta = (\alpha, \beta, \gamma)$, i.e. if $\theta$ has one or two parameters. The first summand of the right-hand term of expression represents that each of the real-valued parameters $(\alpha_j, \beta_j, \gamma_j)$ will be penalized by $\frac{\ln(\tau_1 - \tau_{j-1})}{2}$ of the $j$-th regime to which they belong and since there are $J + 1$ regimes, the sum goes from 1 to $J + 1$. The second summand of the right-hand term is derived from the penalty of the number of points of change, and the last term comes from the sum of each of the penalties of each of the change points.

2.2.2 Genetic Algorithm Schema

As exposed in [11] the total possible cases to evaluate the MDL corresponds to $(\binom{T}{J})$, where $T$ is the number of observations in the time series and $J$ is the number of change points. However, this number of parametric configurations is a quantity that does not make a computationally efficient optimization algorithm that aims to choose the best of the parametric configurations that minimize the MDL. For this reason, we will use the genetic algorithm that, by natural selection criteria will establish the best of the parameters configurations that we will call chromosomes. Each chromosome will be labeled as $(J, \tau_1, ..., \tau_J)$, where the first component $J$ stands for the number of change points, located respectively at times $\tau_1, ..., \tau_J$, corresponding to the respective coordinates. The following is to establish how the genetic algorithm (GA) evaluates each of the chromosomes, while avoiding those with a low probability of being optimal.

Let us now see how a complete generation is produced from an initial one with a given size, although the size can also be a couple. For this purpose, suppose there are $k$ individuals or chromosomes in the initial generation set at random. Each of the $T$ observations in the time series is allowed to be a change-point, independent of all other change-points, with probability, for example as seen in [11], of 0.06. The number of change-points for each chromosome in the initial generation has a binomial distribution with parameters $T - 1$ and 0.06, respectively.

Two chromosomes are taken out of the initial generation, one mother and one father chromosome, to make a child of the next generation. This is done by a probabilistic combination of the parents. The principle of natural selection in this context will be performed by selecting a pair of chromosomes that best
optimize the expression (6) since this couple is considered to have the highest probability of having offspring. Therefore, the chromosomes are arranged from the most likely to the least likely to have children, and each individual of the same generation is assigned a ranking, say $S_i$, being the ranking of the $j$th individual, with $S_j = 1, \ldots, k$. If $S_j = k$ then $j$ is the individual that best fits the objective function (6). If $S_j = 1$ then $j$ is the individual that least well fits the objective function (6).

Once this ranking has been made for each of the chromosomes of the same generation, we proceed to establish the probability of selection using the following expression that simulates the principle of natural selection of the parents that will generate the next generation.

$$S_j \sum_{i=1}^{k} S_i$$

The chromosome that has the highest probability of being selected from the $k$ chromosomes, is chosen as the mother. Among the remaining $(k-1)$ chromosomes, the father is chosen under the same selection criteria as the mother. Suppose that the mother chromosome has $m$ change points, located in some $\tau_1, \ldots, \tau_m$, i.e. with the parameter configuration $(m, \tau_1, \ldots, \tau_m)$. Similarly, suppose that the father chromosome has the parameter configuration $(n, \delta_1, \ldots, \delta_n)$. A child of these parents can arise simply by joining the two chromosomes, i.e., the child chromosome will initially have the following configuration, $(m + n, \epsilon_1, \ldots, \epsilon_{m+n})$, where the $m + n$ change-points contain the mother’s $m$ and the father’s $n$ change-points.

After this, we remove the duplicated change-points from the child $(m + n, \epsilon_1, \ldots, \epsilon_{m+n})$. From this last configuration, we keep all or some change-points. For this, in [11] use the dynamics of flipping a coin for each change-point in the child configuration. If heads comes up, the change-point is left, otherwise it is removed. That is, a binomial distribution will be used with probability parameter $1/2$ and number of trials, the length of the configuration of the child minus duplicitities. All this with the aim that the offspring will keep traits of the parents, without being exact replicas.

Each point of change in the child chromosome, can undergo a mutation; a schema taken from [11] is that one of the following three alternatives may happen. We start by generating, with some random mechanism, the numbers $-1$, 0 and, 1 with respective probabilities 0.4, 0.3, 0.4. If $-1$ comes out, the change point is subtracted by one unit; if 0 comes out, it stays at the time it is at, and if 1 comes out, the current change point is added by one unit. Again, duplicates are eliminated. With this last procedure, we have finished the construction of Child 1. Child 2 up to $k$ are generated in the same way as the previous one. New parents are selected if chromosomes are duplicated in the same generation with the previous parents.
The process of generation is repeated as many times as generations are to be obtained. In fact, one of the criteria for establishing the completion of the genetic algorithm is to fix the number of generations \( r \). Another approach could be to reach the solution that minimizes the objective function or one that does not improve after some given number of iterations or generations. Thus the objective function we used was 
\[
\ln P_\tau(\theta) - \ln f_\tau(D|\theta) - \ln f(\theta)
\]
and the optimization problem we intend to resolve through the means of the genetic algorithm takes the following form,

\[
\hat{\theta}_{BAYESIAN-MDL} = \underset{\theta,\tau}{\text{argmax}} \left( \ln P_\tau(\theta) - \ln f_\tau(D|\theta) - \ln f(\theta) \right) \tag{9}
\]

In other words, \( \hat{\theta}_{BAYESIAN-MDL} \) is the maximum argument of the objective function, in which case it will be the optimal solution to the problem of finding the best configuration of shift points and the respective parameters of the regimes they determine.

### 3 Methods and Models

We are particularly interested in the likelihood function of expression (9) in order to establish what we have called the corresponding Bayesian-MDL. Therefore, for any \( m(t, \theta) \) defined in expression (2) and its respective intensity function defined in (1), it follows that,

\[
L(D|\phi) \propto e^{-m(\tau_1|\theta_1)} \prod_{i=1}^{N_{\tau_1}} \lambda(d_i|\theta_1) \\
\times \prod_{j=2}^{J} \left( e^{-[m(\tau_j|\theta_j)-m(\tau_{j-1}|\theta_j)]} \prod_{i=N_{\tau_{j-1}}+1}^{N_{\tau_j}} \lambda(d_i|\theta_j) \right) \prod_{i=N_{\tau_J}+1}^{n} \lambda(d_i|\theta_{J+1})
\tag{10}
\]

since the product with respect to \( i \) only affects the functions \( \lambda(t, \theta) \) in each of the different \( J+1 \) regimes determined by the \( J \) change points in the vector \( \tau = (\tau_1, \tau_2, \ldots, \tau_J) \). Using that for \( J \) change points, \( \tau_{J+1} := T \) where \( T \) is the number of daily measurements, \( \tau_0 = 0, N_0 = 0 \) and \( m(0|\theta) = 0 \), it follows that the expression (10) reduces to,

\[
L(D|\phi) \propto \prod_{j=1}^{J+1} \left( e^{-[m(\tau_j|\theta_j)-m(\tau_{j-1}|\theta_j)]} \prod_{i=N_{\tau_{j-1}}+1}^{N_{\tau_j}} \lambda(d_i|\theta_j) \right) \tag{11}
\]

Taking the logarithm of (11) we have,
\[
\log L(D|\phi) = \left( \sum_{j=1}^{J+1} m(\tau_{j-1}|\theta_j) - m(\tau_j|\theta_j) \right) \\
+ \left( \sum_{j=1}^{J+1} \sum_{i=N_{\tau_{j-1}}+1}^{N_{\tau_j}} \log \lambda(d_i|\theta_j) \right) \\
= \sum_{j=1}^{J+1} \left( m(\tau_{j-1}|\theta_j) - m(\tau_j|\theta_j) + \sum_{i=N_{\tau_{j-1}}+1}^{N_{\tau_j}} \log \lambda(d_i|\theta_j) \right)
\]

(12)

Also, the Bayesian principle states that,

\[
\mathbb{P}(\theta|D) \propto L(D|\theta)f(\theta),
\]

(13)

where \( L \) is the likelihood function depending on the observations \( D \), given the parameter vector \( \theta \) and, \( f(\theta) \) the a priori function of the parameter vector \( \theta \). Taking logarithm for (13) we have that,

\[
\ln f(\theta|D) \propto \ln(L(D|\theta)) + \ln(f(\theta))
\]

(14)

From (14), we start by establishing the general form of the a priori joint function \( f(\theta) = f(\alpha, \beta, \tau_j) \) in the first three cases, and \( f(\theta) = f(\alpha, \beta, \gamma, \tau_j) \) in the last function \( m \) of the term (2).

3.1 A priori distributions

If we take \( \alpha \sim \text{Gamma}(\phi_{11}, \phi_{12}) \), then,

\[
f(\alpha) = \frac{\phi_{12}^{\phi_{12}}}{\Gamma(\phi_{12})} \alpha^{\phi_{12}-1} e^{-\phi_{12} \alpha}
\]

After applying logarithm we obtain,

\[
\log f(\alpha) = \log \left( \frac{\phi_{12}^{\phi_{12}}}{\Gamma(\phi_{12})} \alpha^{\phi_{12}-1} e^{-\phi_{12} \alpha} \right) \\
= \phi_{12} \log \phi_{11} - \log \Gamma(\phi_{12}) + (\phi_{12} - 1) \log \alpha - \phi_{11} \alpha \\
\propto (\phi_{12} - 1) \log \alpha - \phi_{11} \alpha
\]

(15)

Similarly for \( \beta \), if we take \( \beta \sim \text{Gamma}(\phi_{21}, \phi_{22}) \), then,

\[
\log f(\beta) \propto (\phi_{22} - 1) \log \beta - \phi_{21} \beta
\]

(16)
On the other hand, assuming every time in the series can be chosen as a change-point $\tau_j$ with the same probability, thus $\tau_j \sim \text{Uniform}(0, T)$, $j = 1, 2, \ldots, J$.

Then we have,

$$f(\tau_j) = \frac{1}{T}$$  \hspace{1cm} (17)

Taking logarithm we obtain,

$$\log f(\tau_j) = -\log(T)$$  \hspace{1cm} (18)

Rebuilding the joint function for $\theta = (\alpha, \beta, \tau_j)$ under the assumption of independence, we have,

$$\log f(\alpha, \beta, \tau_j) \propto (\phi_{12} - 1) \log \alpha - \phi_{11} \alpha + (\phi_{22} - 1) \log \beta - \phi_{21} \beta - \log(T)$$  \hspace{1cm} (19)

In the three-parameter model for the intensity function, we have that, under the assumption of independence and that all the parameters have an a priori gamma distribution; considering also, the distribution for the change-points, thus we have,

$$\log f(\alpha, \beta, \gamma, \tau_j) = -\alpha \phi_{11} + (\phi_{12} - 1) \log \alpha - \beta \phi_{21} + (\phi_{22} - 1) \log \beta - \gamma \phi_{31} + (\phi_{32} - 1) \log \gamma - \log(T).$$  \hspace{1cm} (20)

Up to this point, the second summand of the right-hand side of (14) has been obtained. Next, the first summand of the right-hand side of (14) will be derived, but this will be done depending on the intensity function of the non-homogeneous Poisson process that was established previously and whose cumulative mean functions are expressed in the four possibilities of (2).

3.1.1 Weibull intensity rate (W)

After taking the expressions for the intensity function $\lambda^{(W)}(t|\theta)$ and the cumulative mean function $m^{(W)}(t|\theta)$ using (2) and (1) respectively, and replacing these in (12) we have,
\[
\log L(D|\phi) = \sum_{j=1}^{J+1} \left[ m(\tau_{j-1}|\theta_j) - m(\tau_j|\theta_j) + \sum_{i=N_{\tau_{j-1}}+1}^{N_{\tau_j}} \log \lambda(d_i|\theta_j) \right] \\
= \sum_{j=1}^{J+1} \left( \left( \frac{\tau_j}{\beta_j} \right)^{\alpha_j} - \left( \frac{\tau_{j-1}}{\beta_j} \right)^{\alpha_j} \right) \\
+ \sum_{i=N_{\tau_{j-1}}+1}^{N_{\tau_j}} \log \left( \frac{\alpha_j}{\beta_j} \left( \frac{d_i}{\beta_j} \right)^{\alpha_j-1} \right) \\
= \sum_{j=1}^{J+1} \left( \frac{\tau_j^{\alpha_j}}{\beta_j^{\alpha_j}} - \frac{\tau_{j-1}^{\alpha_j}}{\beta_j^{\alpha_j}} \right) + (N_{\tau_j} - N_{\tau_{j-1}}) (\log(\alpha_j) - \alpha_j \log(\beta_j)) \\
+(\alpha_j - 1) \sum_{i=N_{\tau_{j-1}}+1}^{N_{\tau_j}} \log(d_i) \tag{21}
\]

Substituting the expressions (7), (19), (21) in the objective function of the expression (9) we have that,

\[
\ln P_r(\theta) - \ln f_r(D|\theta) - \ln f_r(\theta) = 2 \sum_{i=1}^{J^+} \frac{\ln(\tau_i - \tau_{i-1})}{2} + \ln(J) + \sum_{i=2}^{J} \ln(\tau_i) \\
- \sum_{j=1}^{J+1} \left( \frac{\tau_j^{\alpha_j}}{\beta_j^{\alpha_j}} - \frac{\tau_{j-1}^{\alpha_j}}{\beta_j^{\alpha_j}} \right) \\
+(N_{\tau_j} - N_{\tau_{j-1}}) (\ln(\alpha_j) - \alpha_j \ln(\beta_j)) \\
+(\alpha_j - 1) \sum_{i=N_{\tau_{j-1}}+1}^{N_{\tau_j}} \ln(d_i) \right) \\
- \sum_{j=1}^{J^+} \left( (\phi_{12} - 1) \ln \alpha_j - \phi_{12} \alpha_j \right) \\
+(\phi_{22} - 1) \ln \beta_j - \phi_{21} \beta_j) + J \ln(T) \tag{22}
\]

3.1.2 Musa-Okamoto (MO)

Likewise, taking the expressions for the intensity function \(\lambda^{(MO)}(t|\theta)\) and the cumulative mean function \(m^{(MO)}(t|\theta)\) from (2) and (1) respectively, and replacing these values in (12) we have,
\[
\log L(D|\phi) = \sum_{j=1}^{J+1} \left( m(\tau_{j-1}|\theta_j) - m(\tau_j|\theta_j) + \sum_{i=N_{\tau_{j-1}}+1}^{N_{\tau_j}} \log \lambda(d_i|\theta_j) \right)
\]
\[
= \sum_{j=1}^{J+1} \left( \beta_j \log \left( \frac{\alpha_j + \tau_j - 1}{\alpha_j} \right) - \beta_j \log \left( \frac{\alpha_j + \tau_j}{\alpha_j} \right) \right)
+ \sum_{i=N_{\tau_j-1}+1}^{N_{\tau_j}} \log \left( \frac{\beta_j}{\alpha_j + d_i} \right)
\]
\[
= \sum_{j=1}^{J+1} \left( \beta_j \log \left( \frac{\alpha_j + \tau_j - 1}{\alpha_j} \right) - \beta_j \log \left( \frac{\alpha_j + \tau_j}{\alpha_j} \right) \right)
+ (N_{\tau_j} - N_{\tau_j-1}) \log(\beta_j) - \sum_{i=N_{\tau_j-1}+1}^{N_{\tau_j}} \log(\alpha_j + d_i)
\]
\[
= \sum_{j=1}^{J+1} \left( \beta_j \log (\alpha_j + \tau_j - 1) - \log (\alpha_j + \tau_j) \right)
+ (N_{\tau_j} - N_{\tau_j-1}) \log(\beta_j) - \sum_{i=N_{\tau_j-1}+1}^{N_{\tau_j}} \log(\alpha_j + d_i)
\]

Now, after replacing (7), (23), and (19) in (9) we have,
\[
P_{\tau}(\theta) - \ln f_{\tau}(D|\theta) - \ln f_{\tau}(\theta) = 2 \sum_{i=1}^{J+1} \frac{\ln(\tau_i - \tau_{i-1})}{2} + \ln(J) + \sum_{i=2}^{J} \ln(\tau_i)
- \sum_{j=1}^{J+1} \left( \beta_j \log (\alpha_j + \tau_j - 1) - \log (\alpha_j + \tau_j) \right)
+ (N_{\tau_j} - N_{\tau_j-1}) \log(\beta_j)
- \sum_{i=N_{\tau_j-1}+1}^{N_{\tau_j}} \log(\alpha_j + d_i)
- \sum_{j=1}^{J+1} ((\phi_{12} - 1) \ln \alpha_j - \phi_{11} \alpha_j + \phi_{22} - 1) \ln \beta_j - \phi_{21} \beta_j) + J \ln(T)
\]

3.1.3 Goel-Okumoto (GO)

As for the previous cases, we take the expressions \(\lambda^{(GO)}(t|\theta)\) and \(m^{(GO)}(t|\theta)\) from (2) and (1) respectively, and replace this values in (12). Then we have,
\log L(D|\phi) = \sum_{j=1}^{J+1} \left( m(\tau_{j-1}|\theta_j) - m(\tau_j|\theta_j) + \sum_{i=N_{\tau_j}+1}^{N_{\tau_j}} \log \lambda(d_i|\theta_j) \right) \\
= \sum_{j=1}^{J+1} \left( \alpha_j \left[ 1 - e^{-\beta_j \tau_{j-1}} \right] - \alpha_j \left[ 1 - e^{-\beta_j \tau_j} \right] \right) \\
+ \sum_{i=N_{\tau_j}+1}^{N_{\tau_j}} \log \left( \alpha \beta e^{-\beta d_i} \right) \\
= \sum_{j=1}^{J+1} \left( \alpha_j \left[ e^{-\beta_j \tau_j} - e^{-\beta_j \tau_{j-1}} \right] \right) \\
+ (N_{\tau_j} - N_{\tau_j-1}) \log(\alpha \beta) - \beta \sum_{i=N_{\tau_j-1}+1}^{N_{\tau_j}} d_i \right) \tag{25}

Replacing the expressions (7), (25), (19) in the objective function of the expression (9) we have that

\begin{align*}
\mathbb{P}_{\tau}(\theta) - \ln f_{\tau}(D|\theta) - \ln f_{\tau}(\theta) &= 2 \sum_{i=1}^{J+1} \frac{\ln(\tau_i - \tau_{i-1})}{2} + \ln(J) + \sum_{i=2}^{J} \ln(\tau_i) \\
&\quad - \sum_{j=1}^{J+1} \left( \alpha_j \left[ e^{-\beta_j \tau_j} - e^{-\beta_j \tau_{j-1}} \right] \right) \\
&\quad + (N_{\tau_j} - N_{\tau_j-1}) \ln(\alpha \beta) - \beta \sum_{i=N_{\tau_j-1}+1}^{N_{\tau_j}} d_i \right) \tag{26}
\end{align*}

3.1.4 Generalized Goel-Okumoto (GGO)

Finally and once again, we take \(\lambda^{(GGO)}(t|\theta)\) and \(m^{(GGO)}(t|\theta)\) from (2) and (1) respectively, and replace these values in (12) and then,
\[
\log L(D|\phi) = \sum_{j=1}^{J+1} \left( m(\tau_{j-1} | \theta_j) - m(\tau_j | \theta_j) + \sum_{i=N_{\tau_j-1}}^{N_{\tau_j}} \log \lambda(d_i | \theta_j) \right) \\
= \sum_{j=1}^{J+1} \left( \alpha_j \left( 1 - e^{-\beta_j \tau_j^{\gamma_j}} \right) - \alpha_j \left( 1 - e^{-\beta_j \tau_j^{\gamma_j-1}} \right) \right) \\
+ \sum_{i=N_{\tau_j-1}+1}^{N_{\tau_j}} \ln \left( \alpha_j \beta_j \gamma_j d_i^{\gamma_j-1} e^{-\beta_j \tau_j^{\gamma_j}} \right) \\
= \sum_{j=1}^{J+1} \left( \frac{N_{\tau_j} \tau_j^{\gamma_j}}{\tau_j^{\gamma_j-1}} - \frac{N_{\tau_j} \tau_j^{\gamma_j-1}}{\tau_j^{\gamma_j-1}} \right) \ln(\alpha_j \beta_j \gamma_j) \\
+ (\gamma_j - 1) \left( \sum_{i=N_{\tau_j-1}+1}^{N_{\tau_j}} \ln d_i \right) - \beta_j \left( \sum_{i=N_{\tau_j-1}+1}^{N_{\tau_j}} d_i^{\gamma_j} \right) \\
\tag{27}
\]

Substituting (7), (27), and (20) in (9) we have,

\[
P_{\tau}(\theta) - \ln f_{\tau}(D|\theta) - \ln f_{\tau}(\theta) = 3 \sum_{i=1}^{J+1} \frac{\ln(\tau_i - \tau_{i-1})}{2} + \ln(J) + \sum_{i=1}^{J} \ln(\tau_i) \\
= \sum_{j=1}^{J+1} \left( \alpha_j \left( e^{-\beta_j \tau_j^{\gamma_j}} - e^{-\beta_j \tau_j^{\gamma_j-1}} \right) \right) \\
+ (N_{\tau_j} - N_{\tau_j-1}) \ln(\alpha_j \beta_j \gamma_j) \\
+ (\gamma_j - 1) \left( \sum_{i=N_{\tau_j-1}+1}^{N_{\tau_j}} \ln d_i \right) \\
- \beta_j \left( \sum_{i=N_{\tau_j-1}+1}^{N_{\tau_j}} d_i^{\gamma_j} \right) \\
- \sum_{j=1}^{J+1} \left( \alpha_j \phi_{11} + (\phi_{12} - 1) \ln(\alpha_j) \\
- \beta_j \phi_{21} + (\phi_{22} - 1) \ln(\beta_j) \\
- \gamma_j \phi_{31} + (\phi_{32} - 1) \ln(\gamma_j) \right) + J \ln(T)
\]

Therefore, the expressions (22), (24), (26) and (28) are the objective functions that by minimizing the Bayesian MDL is obtained for each of the different functions \(\lambda(t|\theta)\) of the expression (1), respectively.
Each of the members of the same generation will have a Bayesian-MDL, of which the smallest is chosen. This is done for all the generations. At the end, we will have as many Bayesian-MDLs as generations, and the minimum corresponding to the solution sought in the problem of determining the points of change of the time series of intervals is chosen.

4 Results and discussion

In this section, we proceed to assess the performance of the algorithm to detect multiple change-points. For this purpose we consider two datasets; the first one being simulated observations and the second one consistent of records of particulate matter of less than 2.5 microns of diameter ($PM_{2.5}$) in the city of Bogotá, Colombia collected during the period 2018-2020 on a daily basis.

For the implemented experiments, only the Weibull mean cumulative function (23) was considered with optimal values for the parameters $\alpha$ and $\beta$, 0.1 and 0.5 respectively, which were estimated via the optim function of the statistical software R [14]; on the other hand, for $\alpha$ and $\beta$ we used as a priori distributions $Gamma(\phi_{i1}, \phi_{i2})$, $i = 1, 2$ as we previously defined in section (3.1.1) and the optimal values for the hyperparameters were found to be $\phi_{11} = 1$, $\phi_{12} = 2$, $\phi_{21} = 3$ and $\phi_{22} = 1.2$ estimated using Markov-Chain Monte Carlo (MCMC) methods such that, the objective function takes the form of (21).

We start by analyzing the simulated data under three settings which will be described soon in a detailed manner and then we proceed with the real data. For this task we used the statistical software R; the scripts and datasets can be shared on reader’s request.

4.1 Simulation study

To assess the performance of the algorithm on simulated data, we proceed under a similar scheme as the one proposed by [11]. Three different settings were considered for the number of change-points $J$, 1, 2 and 3 and their locations, $\tau_1, \tau_2, \ldots, \tau_J$ were selected in a convenient manner to illustrate, such that these are presented in table (1).

Taking into account that the length of the $PM_{2.5}$ series for Bogotá during the period 2018-2020, was 1096, such number of observations were simulated from a log-normal distribution with scale parameter $\mu \in \mathbb{R}$ and of shape $\sigma > 0$ (see expression (28)). The data was approximated to this distribution according to the results obtained through the library fitdistrplus [7] in R and the ones published by [17].
\[ f(x) = \begin{cases} \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(\frac{-(\ln(x) - \mu)^2}{2\sigma^2}\right), & x > 0 \\ 0, & x < 0 \end{cases} \] (28)

Now, let’s remember that \( J \) change-points split the time series into \( J + 1 \) sub-series or regimes, such that for each \( J \), every regime was generated by incrementally varying the scale parameter \( \mu \) in 0.5 units while the parameter \( \sigma \) was held constant and equal to 0.32. Thus, the values of \( \mu \) and \( \sigma \) used to generate the \( J + 1 \) regimes, \( J \in \{1, 2, 3\} \) are presented in table (2) while the series graphical behavior for the three settings of \( \tau_1, \tau_2, \ldots, \tau_J \) can be appreciated in figure (1) such that the vertical dashed lines represent the change-points.

On the other hand, again to illustrate, we defined as threshold for possible exceedances the arithmetic mean of the 1096 simulations, \( \bar{X} = \frac{1}{1096} \sum_{t=1}^{1096} X_t \).

For each of the three settings, the number of change-points were estimated, the optimal Bayesian-MDL and the cumulative mean function, \( m(t|\theta) \) and for every run of the genetic algorithm, 50 generations with 50 individuals each were used; the proportion of the times used to generate the initial population was of 6% and the mutation probability was 3% as in [11]. These results are presented and analyzed in the following section.

<table>
<thead>
<tr>
<th>Number of Change-points</th>
<th>Locations</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( \tau_1 = 825 )</td>
</tr>
<tr>
<td>2</td>
<td>( \tau_1 = 365, \tau_2 = 730 )</td>
</tr>
<tr>
<td>3</td>
<td>( \tau_1 = 548, \tau_2 = 823, \tau_3 = 973 )</td>
</tr>
</tbody>
</table>

Table 1 Different Change-points simulations considered

<table>
<thead>
<tr>
<th>Number of Change-points</th>
<th>Number of regimes</th>
<th>Distribution for each regime</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>( \text{log-normal}(\mu = 3.5, \sigma = 0.32), ) ( \text{log-normal}(\mu = 4.0, \sigma = 0.32) )</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>( \text{log-normal}(\mu = 3.5, \sigma = 0.32), ) ( \text{log-normal}(\mu = 4.0, \sigma = 0.32), ) ( \text{log-normal}(\mu = 4.5, \sigma = 0.32) )</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>( \text{log-normal}(\mu = 3.5, \sigma = 0.32), ) ( \text{log-normal}(\mu = 4.0, \sigma = 0.32), ) ( \text{log-normal}(\mu = 4.5, \sigma = 0.32), ) ( \text{log-normal}(\mu = 5.0, \sigma = 0.32) )</td>
</tr>
</tbody>
</table>

Table 2 Settings for the time series regimes for different number of change-points
Genetic algorithm with a Bayesian approach.

4.1.1 First Setting: One Change-point

We start by considering the first setting or the presence of one change-point which was located at time 825, i.e., $\tau_1 = 825$. The genetic algorithm was implemented such that the mean cumulative function was fitted, the Bayesian-MDL values were computed for each generation and as well the optimal number of change-points and their location. The results can be found in figure (2).

Starting by the upper left panel of (2) in a clockwise manner, the first plot indicates how the algorithm succeeds in capturing the presence of the
change-point for the mean cumulative function $m(t|\theta)$; before the change-point $\tau_1 = 825$ the values taken by the function grow in a steady smooth manner until the presence of the instant where the break or rupture is experienced. After such break $m(t|\theta)$ take larger values which is consistent with the behavior of the simulated data; it is also noteworthy how the confidence interval at 95% (upper and lower solid blue lines) contain both the estimation of the mean cumulative function and its real values.

On the other hand, according to the next plot (upper right panel) the minimum reached for the Bayesian-MDL was found around the 40th generation taking a value of 797.27; such generation is marked explicitly in the lower right panel by the vertical dashed line while the optimal number of change-points detected was $J = 3$ as per the horizontal dashed line.

Despite the real value of $J = 1$ being different to the estimated one, in the last panel (lower left) the bar associated with the more frequent change-point can be found in a neighborhood of $\tau_1 = 825$, furthermore, taking as a reference such real value (vertical solid line) we have that the estimations are in average, two units over the optimum.

Now, the figure (3) shows using the solid horizontal line, the sample mean associated to each regime defined by the estimated change-points. Thus the algorithm accomplishes for the whole series capturing the trend of the values under the presence of said points even while there are differences regarding the estimated $J$ and its real value.
4.1.2 Second Setting: Two Change-points

![Graphs showing confidence intervals, AG rate with priori Gamma-Gamma, repeated change points, and best chromosomes with 4 change points.]

Fig. 4 Genetic algorithm results - 2 change-points, third setting

![Graph showing mean fitting for regimes.]

Fig. 5 Second setting - Mean fitting for the regimes

Now we have the second setting or the presence of two change-points, \( \tau_1 = 365 \) and \( \tau_2 = 730 \) with the results of the implementation of the genetic
algorithm presented in the figure (4). We start by analyzing the fitting of the mean cumulative function (upper left panel) and move in a clockwise manner.

In the first plot the fit of $m(t|\theta)$ is presented (middle solid red line) such that its values overlap with the real ones (middle solid black line) while the upper and lower lines (in blue) correspond to the 95% confidence intervals for the estimate of the mean cumulative function. Again, as in the previous case such estimates grows in a piecewise manner following the behavior of the real values with breaks defined by the change-points.

Thus, the first regime for $m(t|\theta)$ is defined for the times $0 < t < 365$ approximately such that from the first time the function grows in a steady monotonous manner before experiencing local maximum around time 365; then a similar growth is experienced from around the time 365 until the second change-point $\tau_2$. Then we find the next regime which goes from $\tau_2 = 549$ until the third change-point $\tau_3 = 823$ and finally we have the fourth regime which goes from the previously mentioned instant until the end of the time window.

Next we have the evolution of the Bayesian-MDL in terms of each generation of the genetic algorithm (upper right panel). Here the algorithm detects as the optimum for said function a value of 726.48; such value is reached according to the vertical dashed line in the third plot (lower right panel) around the 42nd generation.

The genetic algorithm estimated $J = 4$ according to the horizontal dashed line in the third plot, being its real value $J = 2$; nevertheless as can be seen in the last plot (lower left panel) the detected change-points cluster around the real values of $\tau_1$ and $\tau_2$. Again, using those values as a reference (vertical solid red lines), in the first case the estimations are apart from the real optimum in average, two units, while in the second one, these overlap with the real change-point or can be found apart in just one unit.

On the other hand, in figure (5) through the solid horizontal line the sample mean is again represented for each of the found regimes after estimating the change-points; as in the previous case we appreciate how the algorithm allows for a proper fitting of the observed data and the variations of the statistic of interest under the presence of said times despite the estimation of $J$ which was superior to the real value.

4.1.3 Third Setting: Three Change-points

Finally we have now the case of three change-points such they are located at $\tau_1 = 169, \tau_2 = 413$ and $\tau_3 = 1027$ and the results of applying the genetic algorithm are presented in figure (6).
Beginning with the analysis of the mean cumulative function (upper left panel), the algorithm captures the behavior of the real values of $m(t|\theta)$ such that for all the path of the series the estimates of the function (middle red line) overlap with the real values and the confidence interval at 95% (upper and lower blue lines) captures the fitted values and the real ones. As in the other cases there are breaks in the mean cumulative function, associated to the first two change-points, $\tau_1 = 169$ and $\tau_2 = 413$ such that the other rupture
can barely be seen.

Now we analyze how does the Bayesian-MDL evolves according to the generations of the genetic algorithm (upper right panel) such that its behavior seems to be approximately constant for all its path with some decreases such that the optimal value was found to be 584.13. Said optimum was reached in generation 20 as noted by the vertical dashed line in the third plot (lower right panel); on the other hand, the optimal number of change-points was found to be \( J = 3 \) according to the horizontal dashed line in the third panel, the real number of change-points.

Similarly as in the other two cases, the estimated change-points cluster around the real ones as per the last plot of figure (6) (lower left panel) and as per the values around the vertical solid lines which represent the real optimum; as in the other cases the more frequent values are far from the real optimum in average in just two units or one, this for the first two change-points while for the third one the algorithm captures such value with difficulty which is consequent with the behavior for the mean cumulative function. Finally, in figure (7) we have the fitting of the sample mean for all the considered regimes; the trend associated to the change-points it is not only captured by the algorithm but also it can be appreciated how the found ruptures are in a neighborhood close to the real change-points which is consistent with the results in the last plot of figure (6).

4.2 Real Data Analysis

Here we applied on the \( PM_{2.5} \) series for Bogotá in the period 2018 - 2020, in an exhaustive manner, the genetic algorithm with the already described settings at the beginning of this section and using as well as an objective function (21). As threshold, the one established by the environmental norm for Colombia was used which specifies that presence of this polluting agent can not overpass \( 37 \mu g/m^3 \). The series can be observed in figure (8).

Then, the optimal chromosome was found to be \( (8, 400, 408, 445, 488, 627, 654, 661, 798) \), such that there are \( J = 8 \) change-points. The horizontal solid line in the third plot (lower right panel) of figure (8) indicates the sample mean of every regime such that the start of each one of these represents the presence of a change-point, except for the first one and the last times. It can be observed as well, that the first 4 change-points are very close and they overlap. The table (3) shows the date corresponding to each change-point.

These surpasses occurred during the rainy season in Bogotá. If the dates in the table above are compared with the measurements in figure (8), the seasons
Genetic algorithm with a Bayesian approach.

Fig. 8 $PM_{2.5}$ Measurements in Bogota, January 1 2018 - December 31 2020

<table>
<thead>
<tr>
<th>Change Point</th>
<th>Day of the week</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>400</td>
<td>Monday</td>
<td>February 4, 2019</td>
</tr>
<tr>
<td>408</td>
<td>Tuesday</td>
<td>February 12, 2019</td>
</tr>
<tr>
<td>445</td>
<td>Thursday</td>
<td>March 21, 2019</td>
</tr>
<tr>
<td>488</td>
<td>Friday</td>
<td>May 3, 2019</td>
</tr>
<tr>
<td>627</td>
<td>Thursday</td>
<td>September 19, 2019</td>
</tr>
<tr>
<td>654</td>
<td>Wednesday</td>
<td>October 16, 2019</td>
</tr>
<tr>
<td>661</td>
<td>Wednesday</td>
<td>October 23, 2019</td>
</tr>
<tr>
<td>798</td>
<td>Sunday</td>
<td>March 8, 2020</td>
</tr>
</tbody>
</table>

Table 3 Points of change with their respective dates

with the most consistent peaks over time can be captured.

As can be seen in the other plots in figure (9), the one in the upper left panel shows the adjustment of the cumulative mean function $m(t|\theta)$ (black line) by means of the point averages, obtained from the vectors of parameters better qualified by the MDL (red line), and their respective 95% confidence intervals (blue lines).

Also, the corresponding optimal Bayesian-MDL was 778.44 as per the upper right panel which shows the evaluation of said metric for each one of the best chromosomes of each of the 45 generations such that the minimum is reached around the 45th one (vertical dashed line in the lower right panel).

Finally, in the lower left panel, the histogram shows the days in the time series that exceed the threshold, which appear most frequently in the first 50
chromosomes of the 50 generations analyzed.

![Graphs showing confidence intervals, repeated change points, AG rate W and priori Gamma-Gamma, and the best chromosomes with change points.](image)

**Fig. 9** Results for the Optimal Chromosome

The graph in figure (10) shows that before day 400, i.e., Monday, February 4, the rate of exceedances of the 37 $\mu g/m^3$ threshold had been decreasing sharply, but after it, the highest emission rate for eight consecutive days was recorded. This high average rate is around 1.004, as shown in (Table4) for the second regime. In the third regime, it decreased to an average rate of 0.9468; in the fourth, it jumps to 0.6748, and it is in the fifth regime that it achieves the lowest drop, 0.2090, before the intensity function rises again to levels above the 0.2 threshold overshoots per unit of time. This rate present in the fifth regime goes from May 3 to September 19, 2019. Thus, it is evident that the fifth regime occurred before the COVID-19 pandemic lockdown was declared in Colombia. Therefore, this may be the result of a public policy aimed at reducing the emissions of $PM_{2.5}$.

As part of the 2010-2020 ten-year plan for air pollution control, the use of emission control systems in cargo transport vehicles and motorcycles, and as well as the integrated public transport system (SITP for its acronym in Spanish) policies were implemented. The later includes the replacement of old buses with internal combustion engine with electric or hybrid buses. In addition to the above, a few days before the fifth regime, resolution 383 (see RES19 (2019)) was issued, which declared a yellow alert for particulate matter exceedances. Considering the first regime represented in figure (10), the rapid deceleration in the emission of threshold exceedances can also be seen as a consequence of this resolution. Such is the case of restrictions on the use of transportation and the mobility sector, in addition to those aimed at the
Fig. 10 Rate function for PM$_{2.5}$ by Days

operations of industries that use combustion processes associated mainly with the burning of biomass and the use of fossil or liquid fuels.

<table>
<thead>
<tr>
<th>Regime</th>
<th>Min</th>
<th>Mean</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.1894</td>
<td>0.2379</td>
<td>0.6555</td>
</tr>
<tr>
<td>2</td>
<td>1.002</td>
<td>1.004</td>
<td>1.007</td>
</tr>
<tr>
<td>3</td>
<td>0.9363</td>
<td>0.9468</td>
<td>0.9577</td>
</tr>
<tr>
<td>4</td>
<td>0.6657</td>
<td>0.6748</td>
<td>0.6841</td>
</tr>
<tr>
<td>5</td>
<td>0.1991</td>
<td>0.2090</td>
<td>0.2201</td>
</tr>
<tr>
<td>6</td>
<td>0.4801</td>
<td>0.4832</td>
<td>0.4864</td>
</tr>
<tr>
<td>7</td>
<td>0.5765</td>
<td>0.5773</td>
<td>0.5781</td>
</tr>
<tr>
<td>8</td>
<td>0.4761</td>
<td>0.4896</td>
<td>0.5043</td>
</tr>
<tr>
<td>9</td>
<td>0.1744</td>
<td>0.1849</td>
<td>0.1971</td>
</tr>
</tbody>
</table>

Table 4 Minimum, Maximum and Mean for each Regime

5 Conclusions and future work

A solution has been presented to determine change points in counting time series, particularly when they exceed an environmental standard of interest. So, the need for specifying a likelihood function was overpassed by modeling
the times of the series where existed exceedances through a non-homogeneous Poisson process and this specification allowed for better estimations of the number of change-points with a lower variability and a better estimation of parameters of the underlying generating process.

A family of such objective functions was derived using different definitions for the mean cumulative function of the non-homogeneous Poisson processes but it is yet to be seen the influence of using one or the other on the estimation of the change-points and their location.

The search method for the change-points was a genetic algorithm similar to the one exposed by [11], nevertheless, there is still room for improvement for it in terms of different operators available at literature and as well the use of other families of such algorithms.

The detection of change-points using a genetic algorithm and a Bayesian-MDL as selection criterion yielded results that agree with the public policies implemented in Bogotá, Colombia, both regarding the contamination alerts issued by the monitoring network, as well as the mobility restrictions due to the quarantine caused by the SARS-CoV-2 virus pandemic.

The good performance of the algorithm is partly due to the fact that in [1] and [2], the cumulative means function of the contamination exceedances was adjusted, but by then there was no computational technique for the automatic detection of change-points.

We hope that in the near future these methods will prove to be a useful tool for the government agencies in charge of measuring the effectiveness of the actions taken to reduce air pollution, and thus reduce its impacts both in the environment and the health of the inhabitants.

**Conflict of interest**

The authors declare that they have no conflict of interest.

**References**


