Genetic algorithm based technique for defining threshold regression models

J. C. M. Pires
F. G. Martins
M. C. M. Alvim-Ferraz
M. C. Pereira

Follow this and additional works at: https://scholarsarchive.byu.edu/iemssconference


This Event is brought to you for free and open access by the Civil and Environmental Engineering at BYU ScholarsArchive. It has been accepted for inclusion in International Congress on Environmental Modelling and Software by an authorized administrator of BYU ScholarsArchive. For more information, please contact scholarsarchive@byu.edu, ellen_amatangelo@byu.edu.
Genetic algorithm based technique for defining threshold regression models

J.C.M. Pires, F.G. Martins, M.C.M. Alvim-Ferraz and M.C. Pereira
LEPAE, Departamento de Engenharia Química, Faculdade de Engenharia, Universidade do Porto, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal
(dce05005/fgm/aferraz/mcsp@fe.up.pt)

Abstract: This study proposes a new technique based on genetic algorithms to define threshold regression models (TR-GA). The threshold regression assumes that the behaviour of the dependent variable changes when it enters in a different regime. The change from one regime to another depends of a specific value (threshold value) of an explanatory variable (threshold variable). In this study, the threshold regression models were composed by two linear equations. The application of genetic algorithms allows evaluating, at the same time: (i) the threshold variable; (ii) the threshold value; and (iii) the statistically significant regression parameters in each regime. The aim of this study was to evaluate the performance of TR-GA models in the prediction of next day hourly average ozone ($O_3|d+1$) concentrations. The considered predictors were hourly average concentrations of carbon monoxide (CO), nitrogen oxide (NO), nitrogen dioxide ($NO_2$) and ozone ($O_3$) and meteorological data (temperature - T, solar radiation - SR, relative humidity - RH and wind speed - WS). The studies were performed in the period from May 2004 to July 2004. Different TR-GA models were obtained, corresponding to different threshold variables and threshold values. Considering both training and test periods and comparing with multiple linear regression (MLR) approach, better performance indexes were achieved in four of these models. The model that presented the best results showed that the $O_3|d+1$ change its behaviour at the temperature of 23 ºC. For temperatures below 23 ºC, $O_3|d+1$ depended on CO, NO, $NO_2$, SR, WS and $O_3$, for higher temperatures, it depended on CO, NO, T, WS and $O_3$, while for MLR, the most important variables were NO, $NO_2$, SR, RH, WS and $O_3$.

Keywords: Threshold regression models; Genetic algorithms; Multiple linear regression; Ozone concentrations prediction.

1. INTRODUCTION

Ozone ($O_3$) is a photochemical oxidant found in different layers in the atmosphere. In the stratosphere, $O_3$ plays an important role for the protection of the human health, absorbing much harmful ultraviolet radiation. However, in the troposphere this reactive gas has negative impacts on human health, climate, vegetation and materials [Alvim-Ferraz et al., 2006; Chan et al., 1998]. Concerning human health, $O_3$ is responsible for the irritations in the respiratory system [Kley et al., 1999; Leeuw, 2000]. It is considered a greenhouse gas, being also associated to the increase of the temperature [Bytnerowicz et al., 2006]. Additionally, it is considered toxic to plants, reducing the crop yields [Alvim-Ferraz et al., 2006] and as a strong oxidant, it contributes for the increase of the corrosion of the materials [Leeuw, 2000].

$O_3$ is the result of the combination of three basic processes: (i) photochemical production by the interaction of hydrocarbons and nitrogen oxides (emitted by gasoline vapours, fossil fuel power plants, refineries, and other industries) under the action of suitable ambient...
meteorological conditions [Guerra et al., 2004; Strand and Hov, 1996; Zolghadri et al., 2004]; (ii) tropospheric/stratospheric exchange that causes the transport of stratospheric air, rich in ozone, into the troposphere [Dueñas et al., 2002]; and (iii) horizontal transport due to the wind that transports ozone produced in other regions.

Many studies were done aiming the prediction of O$_3$ concentrations [Abdul-Wahab et al., 2005; Sousa et al., 2006 and 2007; Schlink et al., 2003]. Linear (multiple linear and principal component regressions) and nonlinear (neural networks, time series and generalised additive models) models were applied and no significant difference was found between their performances. In this study, O$_3$ concentrations were predicted using a threshold regression model composed by two linear equations. These models are not only applied to economy [Fouquau et al., 2007; Hansen, 1999; Lee and Chang, 2007], but also in biology [Gutzwiller and Barrow, 2003], medicine [Whitmore and Su, 2007] and environment [Cakmak et al., 1999]. The threshold regression assumes that the behaviour of the dependent variable changes when it enters in a different regime [Terui and Dijk, 2002].

The change from one regime to another depends of a specific value (threshold value) of an explanatory variable (threshold variable) [Fouquau et al., 2007]. The dependent variable is given by:

$$
y_i = \begin{cases} 
\hat{\alpha}_0 + \sum_{i=1}^{k} \hat{\alpha}_i x_i + \epsilon_1, & \text{if } x_d \leq r \\
\hat{\beta}_0 + \sum_{i=1}^{k} \hat{\beta}_i x_i + \epsilon_2, & \text{if } x_d > r 
\end{cases}$$

where $x_i$ (i=1,…,k) are the explanatory variables, $\hat{\alpha}_i$ and $\hat{\beta}_i$ (i=0,…,k) are the regression parameters, $\epsilon_1$ and $\epsilon_2$ are the errors associated with the regressions, $r$ is the threshold value and $x_d$ is the threshold variable (one of the explanatory variables) that determines the division of the original data in two parts [Terui and Dijk, 2002]. Multiple linear regression (MLR) was then applied to each part of the data and the regression parameters were determined through the minimization of the sum of square errors (SSE) [Pires et al., 2008]. In both regression equations, only the statistically significant regression parameters should be considered. The statistical significance of the regression parameters was evaluated through the calculation of their confidence interval. Thus, the parameter $\hat{\nu}$ is statistically significant if [Hayter et al., 2006]:

$$
|\hat{\nu}| > \frac{t_{\alpha/2}}{\sqrt{\frac{SSE}{n-k-1}}} \sqrt{\sum_{i=1}^{k} Sxx_i}
$$

where $t$ is the Student $t$ distribution, $n$ is the number of points, $k$ is the number of parameters, $\alpha$ is the significance level, $\hat{\sigma}$ is the standard deviation given by $\sqrt{\frac{SSE}{n-k-1}}$ and $Sxx_i$ is the sum of squares related to $x_i$ given by $\sum_{j=1}^{n} (x_{i,j} - \bar{x}_i)^2$.

Genetic algorithms (GA) were applied to the threshold regression model (TR-GA model), aiming to optimize the values of $r$ (threshold value) and $d$ (index of threshold variable) with the constraint of all regression parameters must be statistically significant. The main objective of this study was to evaluate the performance of TR-GA models in the prediction of the next day hourly average ozone (O$_3|d+1$) concentrations.

2. GENETIC ALGORITHMS AND TR-GA PROCEDURE

GA is a search algorithm based on the mechanics of natural selection and population genetics [Goldberg, 1989; Holland, 1975]. This method starts with a set of individuals (the population) chosen randomly. These individuals (also called chromosomes) have genes that represent a solution of a given problem. GA generates a sequence of populations (the generations) by applying the genetic operators (selection, crossover and mutation) to the individuals. GA presents the following advantages: (i) optimization with continuous or
discrete variables; (ii) derivative function not necessary; (iii) dealing with a large number of variables; (iv) optimization of variables with extremely complex cost surfaces; and (v) providing a list of optimal solutions (not just a single solution). Even that, GA is not the best method to solve all the problems. For example, traditional methods quickly find the solution of a well-behave convex analytical function with few variables, while GA is still evaluating the initial population [Haupt and Haupt, 2004]. The optimizer should select the best method to solve the problem that has in hands. In this study, GA was selected due to the different type of parameters to optimize and the complexity of the constraints (ensure that all regression parameters are statistically significant).

The population size is the number of individual chromosomes that is presented in a population. A large number of chromosomes increases the population diversity, but it also increases the computation time due to the fitness evaluation step. Goldberg [1989] reported that the population size selected by many GA researchers usually ranges from 30 to 200. In this study, the population size was fixed to 100 chromosomes. Preliminary simulations showed that for this population size the number of generations should be high to achieve convergence. Thus, the number of generations was 500. Figure 1 shows the codification of chromosomes. Each chromosome was divided in four sub-strings that correspond to: (i) the value of \(d\); (ii) the value of \(r\); (iii) the explanatory variables used in the first regression (1 – consider the correspondent explanatory variable); and (iv) the explanatory variables used in the second regression (for \(x > r\)).

![Figure 1. Codification of chromosomes.](image)

The selection operator determines which chromosomes are used to generate the next population based on their fitness in the current generation (survival of the fittest). The fitness function measures the performance of the individual with respect to the particular search problem. The fitness function was defined as:

\[
\text{arg min } f = \sqrt{\frac{SSE_1 \times 10^{p_1} + SSE_2 \times 10^{p_2}}{nl}}
\]

where \(ip\) is the number of statistically insignificant regression parameters and \(nl\) is the number of the training points. The indexes 1 and 2 correspond to the first and second regressions, respectively. In many selection methods, the best solutions can be not selected to reproduce. Therefore, these solutions can be lost after the application of crossover and mutation. To avoid this situation, the best elements were copied to the next generation (elitism). However, this procedure decreases the population diversity in the next generations. To reduce this effect, all the chromosomes in the current generation had equal probability to be chosen by crossover and mutation procedures.

The crossover operator consists in exchanging genetic material (binary substrings) of two parents (two chromosomes of the current generation), creating two new individuals. High crossover rates increase the population diversity, promoting the mixing of chromosomes [Siriwardene and Perera, 2006]. The used crossover rate was 0.7.

The mutation operator consists in modifying the chromosomes at random. In bit string representation, the mutation is done by changing 0 to 1 and vice versa in one or more bits. High mutation rates increase the probability of destruction of the best chromosomes [Siriwardene and Perera, 2006]. The used mutation rate was 0.1.

Figure 2 shows how GA is applied for defining threshold regression models. First, the initial population is randomly created. Then, for each individual in the population, the values of \(r\) and \(d\) must be calculated to divide the initial data in two parts. Applying MLR to each part (taking into account only the explanatory variables selected by the chromosome), the
regression parameters are determined and also their statistical significance. After, the individual fitness is evaluated. Finally, the genetic operators are applied to create new individuals in the next generation. This procedure stops when the maximum number of generations is achieved.

Figure 2. Procedure to apply GA for defining threshold regression models.

3. DATA

Tropospheric $O_3$ is a secondary pollutant, predominantly formed by photochemical reactions involving other air pollutants, under suitable meteorological conditions. Thus, the MLR and TR-GA models should present relationships between $O_3$ concentrations and environmental and meteorological variables.

The environmental data, hourly average concentrations of carbon monoxide (CO), nitrogen oxide (NO), nitrogen dioxide (NO$_2$) and $O_3$, were collected in an urban site (Antas) with traffic influences situated in Oporto, Northern Portugal (Figure 3). This site belongs to the air quality monitoring network of Oporto Metropolitan Area that is managed by the Regional Commission of Coordination and Development of Northern Portugal (Comissão de Coordenação e Desenvolvimento Regional do Norte), under the responsibility of the Ministry of the Environment. The meteorological data, temperature (T), solar radiation (SR), relative humidity (RH) and wind speed (WS), were recorded on the left edge of Douro River, at an altitude of 90 m approximately.

The analysed period was from May to July 2004. It was divided in the training (1 May 2004 to 15 July 2004) and test (16 to 31 July 2004) periods.

4. RESULTS AND DISCUSSION

Different TR-GA models were obtained corresponding to different threshold variables. As MLR was the model selected for each regression in TR-GA models, it was considered the basis for comparison for the achieved models. For all models, a t-test (with $\alpha=0.05$) was performed to evaluate the statistical significance of the regression parameters. Table 1 presents the statistically significant regression parameters for TR-GA (1 to 6) and MLR models and corresponding root mean squared error (RMSE) values in the training period. The regression parameters $\hat{\nu}_i$ ($i=1$ to 8) correspond to CO, NO, NO$_2$, T, SR, RH, WS and
$O_3$, respectively. As all regression parameters were considered statistically significant, the fitness value (calculated in GA procedure) corresponded to the RMSE value in the training data. Therefore, all TR-GA models presented slightly better performances than MLR approach in the given period.

In almost TR-GA models, the MLR parameters were very similar (when validated at same time) to the parameters of the first regression of TR-GA models. Thus, the improvement of the achieved models was expected in the prediction of $O_3$ concentrations corresponding to their second regression.

For test period, the regression equations obtained in the training step were applied to predict the $O_{3d+1}$ concentrations. The performance of the models was evaluated through the calculation of the commonly used statistical indexes: mean bias error (MBE), mean absolute error (MAE), RMSE, Pearson correlation coefficient ($R$) and index of agreement ($d_2$) [Chaloulakou et al., 2003; Gardner and Dorling, 2000; Sousa et al., 2006]. Table 2 shows the performance indexes presented by TR-GA and MLR models. MBE was always positive, showing that, in average, the predicted ozone concentrations were overestimated. The MAE, RMSE (absolute error measures), $R$ and $d_2$ give a global idea of the difference between the observed and modelled values. Thus, slightly better model predictions were

![Figure 3. Location of the air quality monitoring site (Antas).](image-url)
Table 1. Statistically significant regression parameters for TR-GA (1 to 6) and MLR models and correspondent RMSE value in the training data

<table>
<thead>
<tr>
<th>Model</th>
<th>$\hat{\nu}_0$</th>
<th>$\hat{\nu}_1$</th>
<th>$\hat{\nu}_2$</th>
<th>$\hat{\nu}_3$</th>
<th>$\hat{\nu}_4$</th>
<th>$\hat{\nu}_5$</th>
<th>$\hat{\nu}_6$</th>
<th>$\hat{\nu}_8$</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>TR-GA1</td>
<td>42.1</td>
<td>1.7</td>
<td>-4.1</td>
<td>8.5</td>
<td>3.6</td>
<td>3.2</td>
<td>17.8</td>
<td></td>
<td>19.9</td>
</tr>
<tr>
<td></td>
<td>22.6</td>
<td>-8.3</td>
<td>8.5</td>
<td>22.4</td>
<td>5.3</td>
<td>9.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TR-GA2</td>
<td>43.5</td>
<td>2.1</td>
<td>4.5</td>
<td>2.4</td>
<td>4.5</td>
<td>15.1</td>
<td></td>
<td></td>
<td>20.0</td>
</tr>
<tr>
<td></td>
<td>26.8</td>
<td>-5.4</td>
<td>14.5</td>
<td>3.7</td>
<td>13.7</td>
<td>17.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TR-GA3</td>
<td>41.2</td>
<td>-3.9</td>
<td>7.3</td>
<td>6.1</td>
<td>5.4</td>
<td>12.5</td>
<td></td>
<td></td>
<td>20.0</td>
</tr>
<tr>
<td></td>
<td>77.2</td>
<td>16.4</td>
<td>10.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TR-GA4</td>
<td>43.7</td>
<td>2.6</td>
<td>-5.1</td>
<td>8.2</td>
<td>3.5</td>
<td>-5.3</td>
<td>3.0</td>
<td>14.2</td>
<td>20.1</td>
</tr>
<tr>
<td></td>
<td>40.2</td>
<td>-6.9</td>
<td>6.9</td>
<td>5.1</td>
<td>8.4</td>
<td>6.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TR-GA5</td>
<td>12.5</td>
<td>8.4</td>
<td>5.3</td>
<td>-15.6</td>
<td>4.6</td>
<td>10.0</td>
<td></td>
<td></td>
<td>20.1</td>
</tr>
<tr>
<td></td>
<td>43.3</td>
<td>1.7</td>
<td>-3.8</td>
<td>7.0</td>
<td>3.3</td>
<td>-3.9</td>
<td>3.4</td>
<td>14.9</td>
<td></td>
</tr>
<tr>
<td>TR-GA6</td>
<td>57.5</td>
<td>23.9</td>
<td>5.4</td>
<td>4.9</td>
<td>2.2</td>
<td>-4.0</td>
<td>4.3</td>
<td>15.7</td>
<td>20.1</td>
</tr>
<tr>
<td></td>
<td>44.7</td>
<td>-2.8</td>
<td>4.3</td>
<td>3.1</td>
<td>-2.2</td>
<td>3.8</td>
<td>14.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MLR</td>
<td>42.6</td>
<td>-2.2</td>
<td>5.3</td>
<td>3.9</td>
<td>-3.7</td>
<td>3.5</td>
<td>13.9</td>
<td></td>
<td>20.7</td>
</tr>
</tbody>
</table>

obtained in four TR-GA models when compared to MLR. The TR-GA1 model presented the best results in both training and test periods. Figure 1 shows the codification of the correspondent chromosome. This model assumed that the $O_3|d+1$ behaviour changed at the temperature of 23 ºC. For temperatures below 23 ºC, $O_3|d+1$ depended on CO, NO, NO$_2$, SR, WS and $O_3$, while for higher values, it depended on CO, NO, T, WS and $O_3$.

Table 2. Performance indexes of the TR-GA and MLR models in the test period

<table>
<thead>
<tr>
<th>Model</th>
<th>MBE</th>
<th>MAE</th>
<th>RMSE</th>
<th>R</th>
<th>$d_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>TR-GA1</td>
<td>0.31</td>
<td>15.42</td>
<td>19.66</td>
<td>0.74</td>
<td>0.83</td>
</tr>
<tr>
<td>TR-GA2</td>
<td>1.61</td>
<td>15.63</td>
<td>19.77</td>
<td>0.74</td>
<td>0.83</td>
</tr>
<tr>
<td>TR-GA3</td>
<td>0.81</td>
<td>15.92</td>
<td>20.68</td>
<td>0.71</td>
<td>0.81</td>
</tr>
<tr>
<td>TR-GA4</td>
<td>0.14</td>
<td>16.01</td>
<td>20.94</td>
<td>0.70</td>
<td>0.80</td>
</tr>
<tr>
<td>TR-GA5</td>
<td>0.43</td>
<td>15.51</td>
<td>19.98</td>
<td>0.73</td>
<td>0.81</td>
</tr>
<tr>
<td>TR-GA6</td>
<td>1.66</td>
<td>15.45</td>
<td>19.80</td>
<td>0.74</td>
<td>0.83</td>
</tr>
<tr>
<td>MLR</td>
<td>0.25</td>
<td>15.59</td>
<td>20.29</td>
<td>0.73</td>
<td>0.81</td>
</tr>
</tbody>
</table>

The main differences of the regressions in TR-GA1 model were the incorporation of NO$_2$ concentrations and SR as important variables in O$_3$ formation for temperatures below 23 ºC, being T only important for temperatures above this threshold value. These observations could be explained by the relative influence of volatile organic compounds (VOC) and nitrogen oxides (NO$_x$) in O$_3$ formation. Seinfeld and Pandis [1998] presented the complex chemical reactions involved in this system and showed that there is a competition between VOC and NO$_x$ for the hydroxyl radical (OH), very important in the O$_3$ formation. At high VOC to NO$_x$ ratio, OH reacts with VOCs; otherwise, the NO$_x$ reaction predominates. In general, increasing VOC concentrations means the appearance of more ozone. The effect of
NO$_x$ concentration increase depends on the VOC to NO$_x$ ratio (positive correlation for high ratios and negative correlation for low ratios) [Seinfeld and Pandis, 1998]. The temperature has a great influence on O$_3$ formation. High temperatures favour VOC to NO$_x$ ratio because VOC concentrations increase more with temperature than NO$_x$ concentrations. Accordingly, the results showed that temperature increase lead to higher O$_3$ concentrations (positive correlation between O$_3$ concentrations and temperature was observed in the second regression of TR-GA1 model). Simultaneously, as expected, a positive correlation between NO and O$_3$ concentrations was also observed. At lower temperatures, VOC to NO$_x$ ratio decrease and O$_3$ concentrations are greatly dependent on NO$_x$ concentrations. As the correspondent chemical reactions are catalysed by solar radiation [Seinfeld and Pandis, 1998], this variable was considered statistically significant in the first regression of TR-GA1 model, presenting a positive correlation with O$_3$ concentrations. Furthermore, as expected for lower VOC to NO$_x$ ratio, a negative correlation between NO and O$_3$ concentrations was observed.

5. CONCLUSIONS

GA was applied to define threshold regression models for prediction of the next day hourly average ozone (O$_{3|d+1}$) concentrations. These models assume that the dependent variable changes its behaviour when an explanatory variable takes a specific value. Applying the procedure presented in this study, different TR-GA models were obtained corresponding to different threshold variables and threshold values. In both training and test periods, four of these models presented slightly better results than MLR approach. Additionally, the best model showed that O$_{3|d+1}$ changed its behaviour at the temperature of 23 ºC. For temperatures below that value, O$_{3|d+1}$ depended of CO, NO, NO$_2$, SR, WS and O$_3$, while for higher temperatures, it depended of CO, NO, T, WS and O$_3$.

ACKNOWLEDGEMENTS

Authors are grateful to Comissão de Coordenação da Direcção Regional-Norte and to Instituto Geofísico da Universidade do Porto, for kindly providing the air quality and meteorological data. This work was supported by Fundação para a Ciência e Tecnologia (FCT). J.C.M. Pires also thanks the FCT for the fellowship SFRD/BD/23302/2005.

REFERENCES

Alvim-Ferraz, M.C.M., S.I.V. Sousa, M.C. Pereira, and F.G. Martins, Contribution of anthropogenic pollutants to the increase of tropospheric ozone levels in Oporto Metropolitan Area, Portugal since the 19th century, Environmental Pollution, 140(3), 516-524, 2006.


