Chemometrics and Intelligent Laboratory Systems 109 (2011) 217–227

Contents lists available at SciVerse ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

Evolutionary association rules for total ozone content modeling from satellite observations

M. Martínez-Ballesteros [a], S. Salcedo-Sanz [b,*], J.C. Riquelme [a], C. Casanova-Mateo [c], J.L. Camacho [d]

[a] Department Languages and Information Systems, Universidad de Sevilla, Seville, Spain
[b] Department of Signal Theory and Communications, Universidad de Alcalá, Madrid, Spain
[c] Department of Applied Physics, Universidad de Valladolid, Valladolid, Spain
[d] Meteorological State Agency of Spain (AEMET), Madrid, Spain

ARTICLE INFO

Article history:
Received 30 March 2011
Received in revised form 5 September 2011
Accepted 23 September 2011
Available online 1 October 2011

Keywords:
Association rules
Evolutionary algorithms
Total Ozone Content (TOC)
Satellite data

ABSTRACT

In this paper we propose an evolutionary method of association rules discovery (EQAR, Evolutionary Quantitative Association Rules) that extends a recently published algorithm by the authors and we describe its application to a problem of Total Ozone Content (TOC) modeling in the Iberian Peninsula. We use TOC data from the Total Ozone Mapping Spectrometer (TOMS) on board the NASA Nimbus-7 satellite measured at three locations (Lisbon, Madrid and Murcia) of the Iberian Peninsula. As prediction variables for the association rules we consider several meteorological variables, such as Outgoing Long-wave Radiation (OLR), Temperature at 50 hPa level, Tropopause height, and wind vertical velocity component at 200 hPa. We show that the best association rules obtained by EQAR are able to accurate modeling the TOC data in the three locations considered, providing results which agree to previous works in the literature.

1. Introduction

Modeling ozone series from satellite observations, past data and its relationship with meteorological variables is an important topic quite often tackled in the literature [1–11]. The interest in modeling ozone series started on the early 70's, when changes into the stratospheric ozone were claimed to be caused by catalytic reactions in the stratosphere that originated losses in the total amount of ozone [12,13]. More specifically, other studies on this topic focused on the role of chlorine [14] and the CFCs [15] in ozone losses at the stratosphere. Those theories were confirmed by the observation of a sharp decrease in the stratospheric ozone levels over Antarctica at the start of the southern spring season in the middle 80's over several polar bases of this continent [16]. A wide review on concepts and history of ozone depletion can be found in [17,18].

In recent years, ozone variation has been related to climate change, so ozone modeling has become an important indicator of deep changes in the atmosphere. That is why very different approaches can be found in the modeling of ozone series in recent bibliography. Specifically, a large amount of works dealing with Total Ozone Content (TOC) of the atmosphere have been published in the last few years, since it seems that variations in these TOC series are a more complete indicator of climate change than only stratospheric ozone series. Thus, there are important works devoted to comparison of different satellite and terrestrial measurements of TOC over different sites [19–21]. The influence of aerosols in total ozone measures is analyzed in [22], where ground and satellite measures are considered. Studies on rare events related to ozone content are studied as well in the literature [23], this includes cases located at the Iberian Peninsula [24]. Also, the modeling of TOC variability has been previously studied, treating different aspects such as its relationship with atmospheric circulation and dynamics or with greenhouse gasses [1,8,25,26]

In this paper, we present an analysis of TOC series modeling in the Iberian Peninsula using Association Rules (ARs) obtained by an evolutionary algorithm. The discovery of ARs is a non-supervised learning and descriptive tool, which explains or summarizes the data, i.e., ARs are used to explore the properties of the data, instead of predicting the class of new data [27]. The aim of ARs mining is discover the presence of pairs (attribute–value), which appear in a dataset with certain frequency, in order to obtain rules that show the existing relationships among the attributes. There exist many algorithms for obtaining ARs from a dataset, such as AIS [28], Apriori [29], and SETM [30]. However, many of these tools that work in continuous domains just discretize the attributes by using a specific strategy and deal with these attributes as if they were discrete, which may lead to poor results in real continuous datasets. Another important class of techniques for ARs discovery is based on evolutionary algorithms (EAs), which have been extensively used for the optimization and adjustment of models in data mining tasks. EAs are search algorithms which generate solutions for optimization problems using

* Corresponding author at: Department of Signal Theory and Communications, Universidad de Alcalá, 28871 Alcalá de Henares, Madrid, Spain. Tel.: +34 91 885 6731; fax: +34 91 885 6699.
E-mail address: sancho.salcedo@uah.es (S. Salcedo-Sanz).

0169-7439/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2011.09.011

techniques inspired by natural evolution [31]. They are implemented as a computer simulation in which a population of abstract representations (chromosomes) of candidate solutions (individuals) for an optimization problem evolves toward better solutions. EAs can be used to discover ARs, since they offer a set of advantages for knowledge extraction and specifically for rule induction processes. In this work, the evolutionary algorithm proposed in [32] has been extended and called EQAR (Evolutionary Quantitative Association Rules). EQAR is applied to the ARs extraction to explain TOC data. The new features added improve the AR mining task and result in the TOC modeling in the Iberian Peninsula. We show that the best rules obtained by the EQAR approach are able to accurately model the TOC data in the three locations considered, providing results which agree to previous works in the literature.

The structure of the rest of the paper is as follows: next section presents the available TOC data and input meteorological variables collected for this study, the description of measurement location and their characteristics. In this section we also detail the prediction variables used in the paper. Section 3 describes the main characteristics of the evolutionary algorithm used to obtain the associated rules. Section 4 presents the main results obtained using the associative rules obtained in the explanation of the TOC series in the three locations considered within the Iberian Peninsula. Section 6 closes the paper giving some final conclusions.

2. TOC data over the Iberian Peninsula and prediction variables

Monthly mean satellite measurements of TOMS (Total Ozone Mapping Spectrometer, on board the NASA Nimbus-7 satellite [33,34]) data for the period 1979–1993 have been used in this study. In addition, a group of several meteorological variables has been selected as input (prediction) variables. Specifically: tropopause height (hPa), *TP*, outgoing longwave radiation (Wm^{-2}), OLR, temperature at 50 hPa (K), t_{50} and air vertical velocity at 200 hPa (hPa/s), a_{200}. All these variables have been obtained with a spatial resolution of 2.5 degree latitude × 2.5 degree longitude from NCEP/NCAR reanalysis [35,36]. These four meteorological variables have been selected because all of them have a close relation to TOC concentration:

1. Temperature at 50 hPa (t_{50}): Many studies have shown that maps of total ozone and 50 hPa temperature look very similar, reflecting a very close coupling between them [8,37]. These studies highlight the fact that, as a rule of the thumb, a 10 Dobson Units (DU) change in total ozone corresponds to a 1 K change of 50 hPa temperature. Consequently, this meteorological variable should be correlated with TOC values.
2. Tropopause height (*TP*): The tropopause is a transition layer between the troposphere and the stratosphere. It is not uniformly thick, and it is not continuous from the equator to the poles. As well, tropopause separates the well-mixed ozone poor troposphere and the stratified ozone rich and well mixed stratosphere. This fact gives the key to use the tropopause as a proxy to analyze TOC values. According to [8], in a tropospheric high pressure system, sinking air in the troposphere leads to an adiabatic warming, causing tropopause and low stratosphere air to rise. As a consequence of these vertical movements, the lower stratosphere cools adiabatically and ozone-poor air moves up, decreasing total ozone. The opposite occurs in tropospheric low pressure systems. Thus, it can be said that high tropopause values are correlated with low total ozone and a low tropopause values with high total ozone [7].
   As has been shown in [38], the selected definition of the tropopause, thermal or dynamical, is not critical. Therefore we have decided to use the thermal one, following the standard criterion of the World Meteorological Organization (WMO) to define thermal tropopause: the lowest level at which the lapse rate decreases to $2\,K\cdot km^{-1}$ or less, provided also the average lapse rate between this level and all higher levels within 2 km does not exceed $2\,K\cdot km^{-1}$ [39]. To determine each thermal tropopause from NCAR reanalysis data, we have used the methodology proposed by [40] using European Centre for Medium-Range Weather Forecasts (ECMWF) reanalysis (ERA) data.
3. Vertical wind velocity (ω_{200}): As stated before, vertical movements through the tropopause bring ozone-poor air into the stratosphere, attenuating ozone-layer. Conversely, descending air from the upper layers of the stratosphere bring ozone-rich air into the ozone-layer, increasing the density of this layer. In [41,42] the authors have proposed a phenomenological model to explain this idea (another discussion about this model can be found in [43]). Thus, in order to deepen in the correlation between vertical movements and variations in total ozone, ω (total time derivative of the pressure -isobaric coordinate system-) can be used for this purpose. ω negative values will indicate ascending movements, whereas positive omega values will indicate descending movements.
4. Outgoing longwave radiation (OLR): Among other gasses, ozone is one of the most important absorbers in the atmosphere. The ozone molecule has a relatively strong rotation spectrum. The three fundamental ozone vibrational bands occur at wavelengths of 9.066, 14.27, and 9.597 μm, respectively. The very strong 9.597 μm and moderately strong 9.066 μm fundamentals combine to make the well-known 9.6 μm band of ozone [44]. Because this 9.6 μm band is a portion of the infrared region of the electromagnetic spectrum, a direct relationship exists between ozone and the OLR [45] and can be used to characterize TOC.

Our study is focused in three locations of the Iberian Peninsula: Lisbon (38.70 N, 9.10 W), Madrid (40.40 N, 3.70 W) and Murcia (38.00 N, 1.10 W). The four meteorological variables have been calculated using a spatial grid covering the three locations. In addition, having into account the strong correlation between tropopause height and TOC, we have decided to calculate this meteorological variable with a customized grid for each of the three locations, i.e., *TP* variable is divided into four different variables depending on each location (the global *TP* for the Iberian Peninsula (TP_G) and the *TP* variable calculated with a grid centered at each location (TP_W, TP_C and TP_E for Lisbon, Madrid and Murcia, respectively)). Table 1 summarizes the grids used in this study.

3. Methods

In this section we introduce the main AR concepts necessary to follow the rest of the paper, and also the evolutionary algorithm proposed in this work in order to look for ARs.

3.1. Association rules

The massive use of computational processing techniques has revolutionized the scientific research due to the high volume of data

Table 1
Meteorological input variables and associated grid size. "IP" stands for Iberian Peninsula.

Area	Met. variable	Variable name	Grid coordinates
IP	50 hPa temperature	t_{50}	35–42.5 N, 12.5 W–5E
IP	Outgoing Longwave Radiation	OLR	35–42.5 N, 12.5 W–5E
IP	Omega at 200 hPa	ω_{200}	35–42.5 N, 12.5 W–0E
IP	Tropopause height	TP_G	35–42.5 N, 12.5 W–5E
Lisbon	Tropopause height	TP_W	35–42.5 N, 12.5 W–5 W
Madrid	Tropopause height	TP_C	35–42.5 N, 7.5 W–0 W
Murcia	Tropopause height	TP_E	35–42.5 N, 2.5 W–5E

which can be obtained. Data mining is one of the most used instrumental tool for discovering knowledge from transactions. In the field of data mining, the learning of ARs is a popular and well-known research method for discovering interesting relations among variables in large databases [29,46].

Formally, ARs were first defined by Agrawal et al. in [28] as follows. Let $I=\{i_1,i_2,\ldots,i_n\}$ be a set of n items and $D=\{t_1,t_2,\ldots,t_N\}$ a set of N transactions, where each t_j contains a subset of items. Thus, a rule can be defined as $X\Rightarrow Y$, where X,Y I and $X\cap Y=\varnothing$. Finally, X and Y are called antecedent (or left side of the rule) and consequent (or right side of the rule), respectively.

When the domain is continuous, the ARs are known as Quantitative Association Rules (QAR). In this context, let $F=\{F_1,\ldots,F_n\}$ be a set of features, with values in $\mathbb{R}$. Let A and C be two disjunct subsets of F, that is, $A\subset F$, $C\subset F$, and $A\cap C=\varnothing$. A QAR is a rule $X\Rightarrow Y$, in which features in A belong to the antecedent X, and features in C belong to the consequent Y, such that X and Y are formed by a conjunction of multiple boolean expressions of the form $F_i\in[v_1,v_2]$. The consequent Y is usually a single expression. In this proposal, QAR is used, since the domain variable (TOC) is a continuous one.

3.2. Quality measures for association rules

The following paragraphs detail the most popular quality measures used to evaluate an AR. Note that it is very important to have a measure of the quality of a given rule in order to select the best set of rules. In the ARs mining process, probability-based measures that evaluate the generality and reliability of ARs have been selected. In particular, the *support* measure is used to represent the generality of the rule and the *confidence*, the *lift* and the *leverage* are normally used to represent the reliability of the rule [47,48]. The formal definitions of these variables are the following:

- *Support(X)* [48]: The support of an itemset X is defined as the ratio of instances in the dataset that satisfy X. Usually, the support of X is named as the probability of X.

$$sup(X)=P(X)=\frac{n(X)}{N}. \tag{1}$$

  where $n(X)$ is the number of occurrences of the itemset X in the dataset, and N is the number of instances forming such dataset.
- *Support*$(X\Rightarrow Y)$ [48]: The support of the rule $X\Rightarrow Y$ is the percentage of instances in the dataset that satisfy X and Y simultaneously.

$$sup(X\Rightarrow Y)=P(Y\cap X)=\frac{n(XY)}{N}. \tag{2}$$

  where $n(XY)$ is the number of instances that satisfy the conditions for the antecedent X and consequent Y simultaneously.
- *Confidence*$(X\Rightarrow Y)$ [48]: The confidence is the probability that instances satisfying X, also satisfy Y. In other words, it is the support of the rule divided by the support of the antecedent.

$$conf(X\Rightarrow Y)=P(X|Y)=\frac{sup(X\Rightarrow Y)}{sup(X)} \tag{3}$$

- *Lift*$(X\Rightarrow Y)$ [49]: Lift or interest is defined as how many times more often X and Y are together in the dataset than expected, assuming that the presence of X and Y in instances is statistically independent. Lifts greater than one involve statistical dependence in simultaneous occurrence of X and Y, in other words, the rule provides successful information about X and Y occurring together in the dataset.

$$lift(X\Rightarrow Y)=\frac{sup(X\Rightarrow Y)}{sup(X)sup(Y)}=\frac{conf(X\Rightarrow Y)}{sup(Y)} \tag{4}$$

- *Leverage*$(X\Rightarrow Y)$ [50]: Leverage measures the proportion of additional cases covered by both X and Y above those expected if X and Y were independent of each other. Leverage takes values inside $[-1,1]$. Values equal or under value 0, indicate a strong independence between antecedent and consequent. On the other hand values near 1 are expected for an important association rule. Values above 0 are desirable. In addition, leverage is a lower bound for support, and therefore, optimizing only the leverage guarantees a certain minimum support (contrary to optimizing only the confidence or only the lift).

$$lev(X\Rightarrow Y)=sup(X\Rightarrow Y)-sup(X)sup(Y) \tag{5}$$

- *Accuracy*$(X\Rightarrow Y)$ [48]: Accuracy measures the degree of veracity thus, the degree of fit (matching) between the obtained values and the actual data. An accuracy of 100% means that the measured values are exactly the same as the given values. In the field of mining association rules, accuracy measures the sum of the percentage of instances in the dataset that satisfy the antecedent and the consequent and the percentage of instances in the dataset that do not satisfy neither the antecedent nor the consequent. Accuracy takes values inside [0, 1] and values near 1 are expected for a rule with high quality and veracity.

$$Acc(X\Rightarrow Y)=sup(X\Rightarrow Y)+sup\left(\neg X\Rightarrow\neg Y\right) \tag{6}$$

  where ¬ means negation, therefore $sup(\neg X\Rightarrow\neg Y)$ is the percentage of instances in the dataset that do not satisfy X and Y simultaneously.

In most cases, it is enough to focus on a combination of support, confidence, and either lift or leverage to obtain a good measure of the rule "quality". However, how good a rule is for modeling a dataset in terms of usefulness and actionability is a subjective concept, and depends on the particular domain and the business objectives.

For a better understanding of these quality measures, we give a small example, by using a dataset comprising eight instances and three features are shown in Table 2. Consider then two example rules, henceforth called Rule (7) and Rule (8), respectively:

$$F_1\in[32,35]\wedge F_2\in[179,188]\Rightarrow F_3\in[84,94] \tag{7}$$

$$F_1\in[32,35]\wedge F_2\in[179,188]\Rightarrow F_3\in[46,94] \tag{8}$$

In Rule (7), the support of the antecedent is 12.5%, since one instance, t_1, simultaneously satisfy that F_1 and F_2 belong to the intervals [32,35] and [179,188], respectively (one instance out of eight, $sup(X)=0.125$). As for the support of the consequent, $sup(Y)=0.375$ because instances t_1, t_3 and t_7 satisfy that $F_3\in[84,94]$. Regarding the confidence, only one instance t_1 satisfies all the three features (F_1 and F_2 in the antecedent, and F_3 in the consequent) appearing in the rule; in other words, $sup(X\Rightarrow Y)=0.125$. Therefore, $conf(X\Rightarrow Y)=0.125/0.125=1$, that is, the rule has a confidence of 100%. The lift

Table 2
Illustrative dataset.

Instance	F_1	F_2	F_3
t_1	35	183	88
t_2	42	154	47
t_3	37	186	93
t_4	30	199	112
t_5	33	173	83
t_6	24	178	75
t_7	63	177	91
t_8	22	167	60

is $lift(X \Rightarrow Y) = 0.125/(0.125 \cdot 0.375) = 2.66$, the leverage is $lev(X \Rightarrow Y) = 0.125 - (0.125 \cdot 0.375) = 0.078$ and the accuracy is $acc(X \Rightarrow Y) = 0.125 + 0.625 = 0.75$, since $sup(X \Rightarrow Y) = 0.125$, $sup(\neg X \Rightarrow \neg Y) = 0.625$, $sup(X) = 0.125$ and $sup(Y) = 0.375$, as discussed before.

In Rule (8), the support of the antecedent is the same as in Rule (7), i.e. 12.5%, since one instance, t_1, simultaneously satisfy that F_1 and F_2 belong to the intervals [32,35] and [179,188]. However, the support of the consequent is $sup(Y) = 0.875$ because all instances except t_4 satisfy that $F_3 \in [46,94]$. The confidence in this rule is the same that in Rule (7), only one instance satisfies all the three features appearing in the rule, i.e., $sup(X \Rightarrow Y) = 0.125$. Therefore, the confidence value of this rule is also 100%. Regarding the lift or interest, $lift(X \Rightarrow Y) = 0.125/(0.125 \cdot 0.875) = 1.14$, the leverage is $lev(X \Rightarrow Y) = 0.125 - (0.125 \cdot 0.875) = 0.016$ and the accuracy is $acc(X \Rightarrow Y) = 0.125 + 0.125 = 0.25$, since $sup(X \Rightarrow Y) = 0.125$, $sup(\neg X \Rightarrow \neg Y) = 0.125$, $sup(X) = 0.125$ and $sup(Y) = 0.875$, as discussed before.

Note that confidence does not take into account the support of the rule consequent, because the confidence is the same in the two Rules (7) and (8). The lift of the rule should be considered to solve this drawback. Lift or interest measures the degree of dependence between the antecedent and the consequent. The lift of Rule (7) and Rule (8) is 2.66 and 1.14 respectively. Here, lift of Rule (7) is larger than the lift of Rule (8), which corresponds to our intuition that the first rule is more interesting than the second one. Regarding the values of accuracy and leverage are also higher for the Rule (7). Therefore, we can conclude that first rule has better quality, accuracy, interest and strong dependency between the antecedent and consequent than the second one even if they have the same confidence.

3.3. EQAR: an effective evolutionary algorithm for AR searching

As has been previously mentioned, EAs have been quite used to discover ARs, since they offer several advantages for knowledge extraction and specifically for rule induction processes. In [51] the authors proposed an EA to obtain numeric ARs, dividing the process in two phases. Another EA was used in [52] to obtain QAR where the confidence was optimized in the fitness function. In [53] a multi-objective pareto-based EA was presented in which the fitness function was composed by four different objectives. A study of three evolutionary ARs extraction methods was presented in [54] to show their effectiveness for mining ARs in quantitative datasets. Other EAs that use a weighted scheme for the fitness function which involved several evaluation measures of rules were presented in [55] and [32]. The main motivation of these works was to develop an algorithm able to find QAR over datasets with continuous attributes without a previous discretization in the process. In fact, in this paper, we use the basic scheme algorithm proposed in [32] and we extend this approach, henceforth called EQAR (Evolutionary Quantitative Association Rules), with new features in order to improve the ARs mining task. The results were obtained by EQAR in our problem of TOC modeling in the Iberian Peninsula.

EQAR follows the general scheme of the CHC binary-coded evolutionary algorithm proposed by Eshelman in 1991 [56]. The original CHC presents an elitist strategy for selecting the population that will make up the next generation and includes strong diversity in the evolutionary process through mechanisms of incest prevention and a specific operator of crossover called Half Uniform (HUX). Furthermore, the population is re-initialized when its diversity is poor. However EQAR adopts a more conservative re-initialization strategy and a less disruptive crossover operator than the HUX crossover procedure.

The search of the most appropriate intervals is carried out by means of EQAR and the intervals are adjusted to find ARs with high quality. Each individual constitutes a rule in the population. Each gene of an individual represents the limits of the intervals and the type of each attribute to indicate whether it belongs to the antecedent, consequent or not belonging to the rule. Thus, the representation of an individual consists in two data structures as shown in Fig. 1. The upper structure includes all the attributes of the database, where l_j is the lower limit of the range and u_j is the upper limit. The bottom structure indicates the membership of an attribute to the rule represented by an individual. The type of each attribute t_j, can have three values: 0 when the attribute does not belong to the rule, 1 if it belongs to the antecedent of the rule and 2 when it belongs to the consequent part.

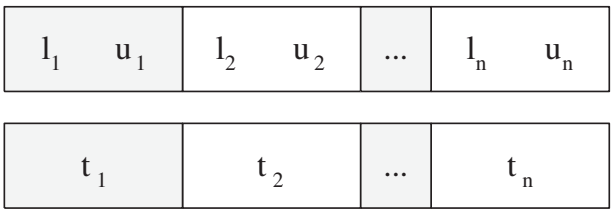

Fig. 1. Representation of an individual of the evolutionary algorithm's population.

An illustrative example of Rule (7) is depicted in Fig. 2. In particular, the rule $F_1 \in [32,35] \wedge F_2 \in [179,188] \Rightarrow F_3 \in [84,94]$ is represented. Note that attributes F_1 and F_2 appear in the antecedent and F_3 in the consequent. Therefore $t_1 = t_2 = 1$ and $t_3 = 2$.

The individuals of the population are subjected to an evolutionary process in which both crossover operator with incest prevention and re-initialization of the population are applied. At the end of this process, the fittest individual is designated as the best rule. Moreover, the fitness function has been provided with a set of parameters so that the user can drive the search process depending on the desired rules. The proposed algorithm is based on the Iterative Rule Learning (IRL) [57]. The punishment of the covered instances allows the subsequent rules found by EQAR to try to cover those instances that were still not covered. General scheme of the IRL is shown in Fig. 4.

In addition to the features of the algorithm described above, new ones have been added in order to improve the performance and the quality of the rules obtained in this specific problem of TOC analysis. The generation of the initial population for each evolutionary process has been modified to help the examples that are covered by a few rules, and also the fitness function has been expanded. These new functionalities of EQAR are detailed in the following subsections.

3.3.1. Generation of the initial population

The generation of the initial population is carried out at the beginning of each evolutionary process. It must be noted that the generation of the rules in EQAR is different to the algorithm proposed in [32], in which the process for generating the initial population was carried out in such a way that at least one randomly chosen sample or instance of the dataset was covered. However, in EQAR the samples of the dataset are not randomly selected but they are selected based on their level of hierarchy. The hierarchy is organized according to the number of rules which cover a sample. Thus, the records are

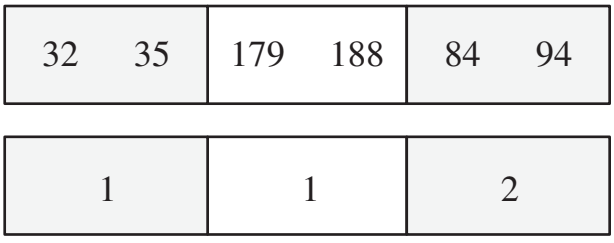

Fig. 2. Representation of Rule (6) example.

31.35 34.65	169.54 176.46	76.36 89.64
1	0	2

Fig. 3. Representation of an individual example of generation of initial population.

sorted by the number of rules that are covered and the samples covered by few rules have a higher priority.

A sample is selected according to the inverse of the number of rules which cover such sample. Intuitively, the process is similar to roulette selection method where the parents are selected depending on their fitness. In the roulette selection method, a sample is represented by a portion of roulette inversely proportional to the number of rules that cover such sample. Thus, the samples covered by a few rules have a greater portion of the roulette and, therefore, they will be more likely selected. In the first evolutionary process, all samples have the same probability to be selected.

This process for generating the initial population can be described by means of a pseudo-code, as follows.

(1) For all instances of the database the cumulative sum, *totalSum*, of the inverse of the number of rules that cover every instance is calculated.
(2) A random number *R* between 0 and *totalSum* is generated.
(3) For each instance of database, if the *totalSum* is greater or equal than *R*, then the current example is selected.

Constraints to generate individuals are given by the following settings:

- number of attributes that belong to rule represented by an individual.
- number of attributes in the antecedents and consequents.
- structure of the rule (attributes fixed or not fixed in consequent).

For a better understanding of the generation of initial population, we describe one example of generation of an individual following the Table 2. For each iteration one instance is selected based on their level of hierarchy. In this case, we have assumed that the algorithm is starting, that is, the first evolutionary iteration of the process, and all instances have the same probability to be selected.

Assuming that the instance t_5 of Table 2 is randomly selected, the values of each attribute are: $F_1 = 33$, $F_2 = 173$ and $F_3 = 83$. In order to generate an individual (one rule), we have to randomly select the number of attributes appearing in the rule and the type and interval of each attribute. We have supposed that the number of attributes chosen is 2 and the type for each attribute is 1 for F_1, 0 for F_2 and 2 for F_3. Then, we have to select a random number between 0 and a maximum amplitude (10%) for generating the intervals for each attribute. The value obtained is added and subtracted to the value corresponding for each attribute to the instance selected (t_5) in Table 2. For example: 5% for F_1, 2% for F_2 and 8% for F_3. Therefore, the intervals of each attribute are $[33 \pm (0.05 \cdot 33)]$ for F_1, $[173 \pm (0.02 \cdot 173)]$ for F_2 and $[83m(0.08 \cdot 83)]$.

The individual generated is shown in Fig. 3 and the rule obtained is represented as follows:

$$F_1 \in [31.35, 34.65] \Rightarrow F_3 \in [76.36, 89.64] \quad (9)$$

3.3.2. Fitness functions proposed

The fitness of each individual in the evolutionary algorithm allows determining which are the best candidates to remain in subsequent generations. In order to make this decision, its calculation involves several measures that provide information about the rules. In this work, two fitness functions have been designed to maximize different objectives depending on the desired rules. Both are formed by the combination of different measures of association rules but their goals are different.

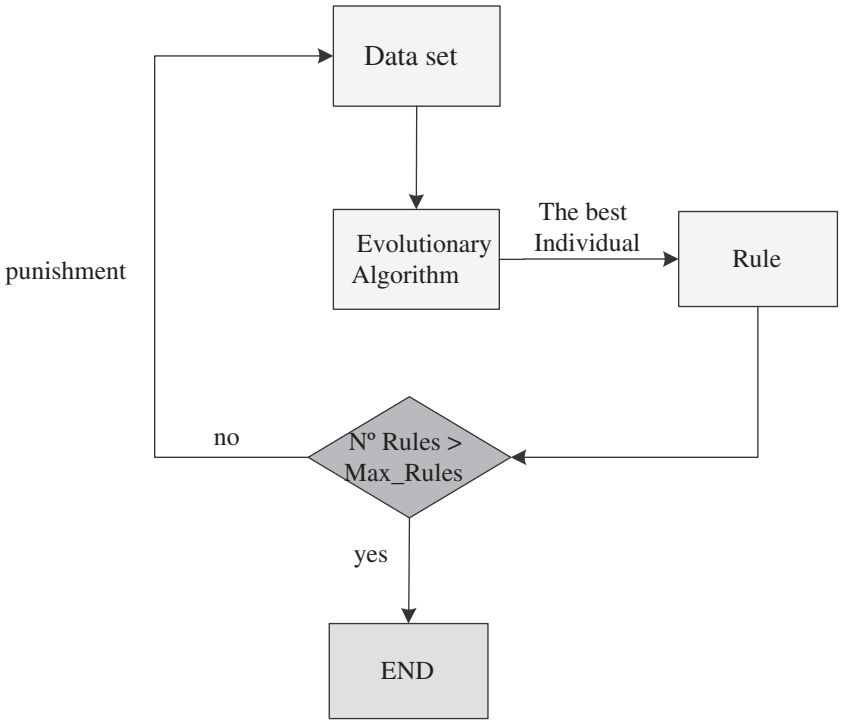

Fig. 4. Scheme of the Iterative Rule Learning algorithm.

Table 3
Ozone quartiles (DU) for each location.

Quartile	Lisbon	Madrid	Murcia
1°	[255.7, 291.2]	[253.9, 291.1]	[259.92, 293.03]
2°	[291.2, 326.7]	[291.1, 328.35]	[293.03, 326.15]
3°	[326.7, 362.2]	[328.35, 365.6]	[326.15, 359.26]
4°	[362.2, 397.7]	[365.6, 402.8]	[359.26, 392.38]

As first fitness function of guide the evolutionary search, we propose the following:

$$f(i) = w_s \cdot sup + w_c \cdot conf - w_r \cdot recov - w_a \cdot ampl \quad (10)$$

where *sup* is the support of the rule, *conf* is the confidence of the rule, *recov* is the number of recovered instances (it is used to indicate when a sample has already been covered by a previous rule, thus, rules covering different regions of search of space are preferred), *ampl* is the average size of intervals of the attributes belong to the rule and w_s, w_c, w_r and w_a are weights in order to drive the process of rules searching. Note that this function takes into consideration the support and the confidence of the rule. This function is used when QAR with high support and confidence is desired. High values of w_s imply that more samples from the database are covered and high values of w_c imply rules with greater reliability, that is, rules with fewer errors.

Nevertheless, only the support is usually not enough to calculate the fitness, because the algorithm would try to enlarge the amplitude of the intervals until the whole domain of each attribute would be completed to get a great support. For this reason, this fitness function includes a measure to limit the growth of the intervals during the evolutionary process. In addition, this function is able to find rules that cover different regions of the search space because it also includes a measure to negatively affect an instance that has already been covered by a previous rule.

However, this function is not entirely appropriate in some situations because the confidence has some drawbacks. Specifically, confidence does not take into account the support of the rule consequent hence it is not able to detect negative dependencies between items.

For this reason, a new fitness function has been proposed as alternative to the support and confidence measures. The second fitness function to be maximized used by EQAR is given by the following expression:

$$f(i) = w_i \cdot lift + w_l \cdot lev - w_r \cdot recov \quad (11)$$

where *lift* is the lift or interest of the rule and *lev* is the leverage of the rule and w_i, w_l and w_r are weights in order to drive the process of search of rules.

This function considers lift and leverage measures instead of support and confidence measures. This function is used when QARs with a high lift and high leverage are desired. High values of w_i ensure a degree of dependence between antecedent and consequent. The higher this value, the more likely that the existence of antecedent and consequent together in an instance is not just a random occurrence, but because there is some relationship or dependency between them. High values of w_l guarantee a certain minimum support. Thus, for leverage, values above 0 are desirable, whereas for lift, we want to see values greater than 1. Note that leverage and lift measure similar things, except that leverage measures the proportion of additional cases covered by both antecedent and consequent above those expected if antecedent and consequent were independent of each

Table 4
Association rules for TOC concentration at Lisbon. The "Code" of the rules describes the location and TOC concentration, i.e., L_{m1} stands for Lisbon, medium TOC rule 1, L_{m2} stands for Lisbon, medium TOC rule 2, and L_{h1} stands for Lisbon, high TOC rule 1, and so on.

Code	Rules	TOC (DU)	Scores						Fitness
			Sup(%)	Conf(%)	Ampl(%)	Lift	Lev	Acc(%)	
L_{m1}	TP_G[12.5, 12.1] & t_{50}[215.9, 216.9]	[332.4350.5]	5.2	100	10.0	6.0	0.04	88.5	Eq. (11)
L_{m2}	TP_W[11.5, 11.2] & t_{50}[214.6, 215.7] & *OLR*[237.5, 256.6]	[353.2367.4]	3.5	100	9.2	14.4	0.03	96.6	Eq. (11)
L_{h1}	TP_G[11.4, 10.6] & t_{50}[212.7, 216.8] & *OLR*[224.4, 239.8]	[349.0387.8]	10.4	81.8	21.3	4.3	0.08	89.1	Eq. (10)
L_{h2}	t_{50}[215.1, 217.3] & *OLR*[231.5, 257.1] & ω_{200}[−5, 8]	[349.6,387.8]	10.4	62.1	22.0	3.4	0.07	85.8	Eq. (10)
L_{v1}	t_{50}[215.6, 217.7] & *OLR*[231.5, 245.4]	[369.5,397.7]	4.0	87.5	14.8	12.6	0.04	96.5	Eq. (11)

Table 5
Association rules for TOC concentration at Madrid. The "Code" of the rules describes the location and TOC concentration, i.e., M_{m1} stands for Madrid, medium TOC rule 1, M_{m2} stands for Madrid, medium TOC rule 2, M_{h1} stands for Madrid, high TOC rule 1, and so on.

Code	Rules	TOC (DU)	Scores						Fitness
			Sup(%)	Conf(%)	Ampl(%)	Lift	Lev	Acc(%)	
M_{m1}	TP_G[14.1, 12.5]	[285.6,327.8]	23.1	97.6	37.1	1.9	0.11	71.1	Eq. (10)
M_{m2}	TP_G[12.4, 11.9] & *OLR*[262.9, 270.3] & t_{50}[215.4, 217.2]	[329,347.5]	5.8	100	11.2	5.8	0.05	88.4	Eq. (11)
M_{h1}	TP_G[12.0, 11.3] & t_{50}[215.1, 216.4]	[334.1, 361]	5.8	100	16.1	4.6	0.05	83.8	Eq. (11)
M_{h2}	TP_G[11.5, 10.8] & t_{50}[214.4, 216.1]	[356.3,392.5]	6.9	100	20.6	6.2	0.06	90.8	Eq. (11)
M_{v1}	TP_G[11.2, 10.6] & *OLR*[224.4, 245.4] & t_{50}[214.5, 216.8]	[368.1,402.8]	6.4	91.7	16.9	11.3	0.06	97.7	Eq. (11)

Table 6
Association rules for TOC concentration at Murcia. The "Code" of the rules describes the location and TOC concentration, i.e., U_{m1} stands for Murcia, medium TOC rule 1, U_{m2} stands for Murcia, medium TOC rule 2, U_{h1} stands for Murcia, high TOC rule 1, and so on.

Code	Rules	TOC (DU)	Scores						Fitness
			Sup(%)	Conf(%)	Ampl(%)	Lift	Lev	Acc(%)	
U_{m1}	TP_E[11.5, 10.7] and t_{50}[212.5, 213.4] and *OLR*[211.5, 243.6]	[341.9,359.7]	4.6	100	15.7	6.9	0.04	90.1	Eq. (11)
U_{m2}	TP_G[11.8, 11.4] and TP_E[12.1, 10.6] and *OLR*[250.6, 256.6]	[343.0,354.5]	3.5	100	11.0	9.6	0.03	93.1	Eq. (11)
U_{h1}	TP_G[11.2, 10.6] and t_{50}[214.6, 216.5] and *OLR*[214.9, 217.7]	[343.7,387.5]	8.7	100	22.6	4.7	0.07	87.4	Eq. (11)
U_{v1}	TP_E[11.2, 10.8] and t_{50}[214.5, 217.7]	[359.3,392.4]	8.1	100	19.7	7.2	0.07	94.2	Eq. (11)

Table 7
TOC variation obtained with different association rules for different meteorological variables at Lisbon.

Met. variable	Ratio	Rule code				
		L_{m1}	L_{m2}	L_{h1}	L_{h2}	L_{v1}
With TP_G	DU/km	22.5	5.9	19.6	13	17.1
With TP_W	DU/km	15.3	13.9	14.2	8.5	18
With t_{50}	DU/K	6.4	6.6	6.2	6.4	4.3

Table 8
TOC variation obtained with different association rules for different meteorological variables at Madrid.

Met. variable	Ratio	Rule code				
		M_{m1}	M_{m2}	M_{h1}	M_{h2}	M_{v1}
With TP_G	DU/km	13.7	13.8	23.0	21.4	10.8
With TP_C	DU/km	13.4	9.3	22.2	21.4	11.1
With t_{50}	DU/K	9.0	9.1	5.7	6.6	6.0

Table 9
TOC variation obtained with different association rules for different meteorological variables at Murcia.

Met. variable	Ratio	Rule code			
		U_{m1}	U_{m2}	U_{h1}	U_{v1}
With TP_G	DU/km	9.6	8.7	14	14.7
With TP_E	DU/km	9.3	25.1	13.7	18.7
With t_{50}	DU/K	9.4	6.6	6.9	6.3

other. Leverage is also included because lift is susceptible to noise in small databases. Rare itemsets with low probability that per chance occur a few times (or only once) together can produce enormous lift values. In this function, the amplitude of intervals is not included because leverage is inversely proportional to the size of the intervals. If leverage is maximized, we ensure that the intervals of attributes do not extend to the whole domain. This function also includes a measure to negatively affect an instance that has already been covered by a previous rule in order to find rules that cover different regions of the search space.

In conclusion, the first fitness function corresponding to Eq. (10) should be used when rules covering many examples with a high degree of reliability are desired without interesting the degree of dependence between antecedent and consequent of the rule. High confidence and high support could imply interdependence between antecedent and consequent. While the second fitness function corresponding to Eq. (11) should be used when a rule with a high degree of dependence between antecedent and consequent is desired regardless of the number of instances covered by the rule. High lift and high leverage could imply low support.

4. Experimental results

In order to apply the methodology describe above, we have divided TOC values of each location (Lisbon, Madrid and Murcia) into four equal groups or quartiles, each representing a fourth of each TOC data set, i.e., first quartile [0%, 25%], second [26%,50%], third [51%, 75%] and fourth [76%, 100%]. Following this idea, ozone quartiles for each location can be calculated for the period of study considered in this work (1979–1993), as shown in Table 3. Once TOC values have been divided into the four quartiles, the following criteria to set different Ozone concentrations have been used:

- Medium ozone concentration: Rules which ozone values belong to third quartile or a lower quartile.
- High ozone concentration: Rules which ozone values belong to third and fourth quartiles.
- Very high ozone concentration: Rules which ozone values belong only to fourth quartile.

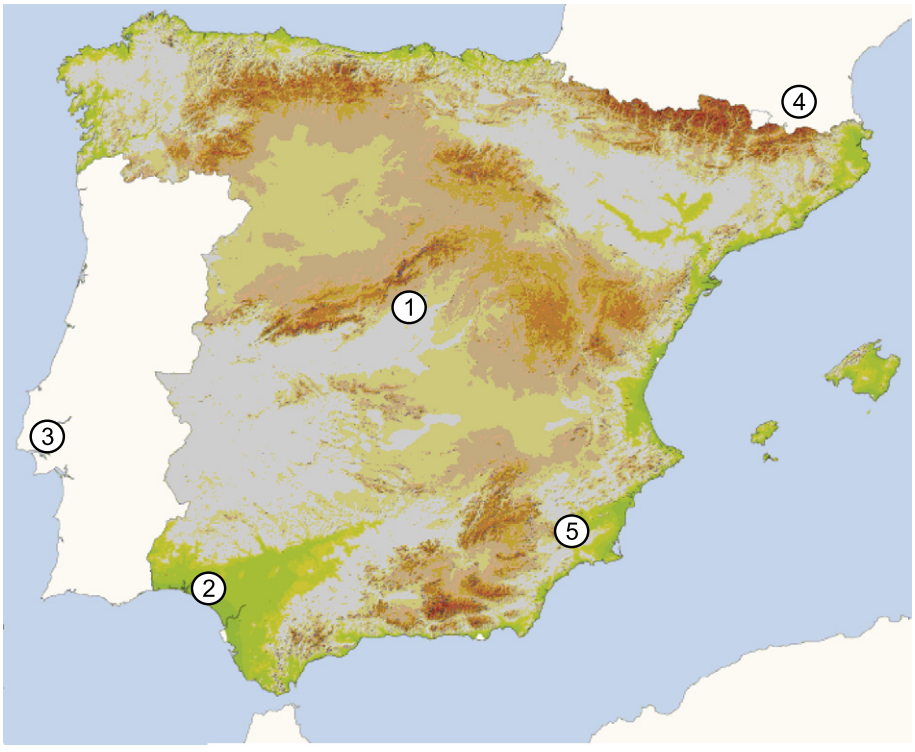

Fig. 5. Location of the different O_3 observing stations considered in the validation process of the results: 1. Madrid, 2. Arenosillo, 3. Lisbon, 4. Montlouis and 5. Murcia.

Table 10
Accuracy of association rules for TOC concentration trained in Lisbon data and tested in Lisbon, Murcia, Madrid, Arenosillo and Montlouis data separately.

Rule code Lisbon	Accuracy (%)				
	Lisbon	Murcia	Madrid	Arenosillo	Montlouis
L_{m1}	88.4	86.7	87.9	87.3	83.8
L_{m2}	96.5	90.2	90.8	91.3	90.2
L_{h1}	89.0	88.4	86.1	88.4	80.9
L_{h2}	82.1	81.5	78.0	81.5	76.3
L_{v1}	95.4	96.5	94.8	95.4	89.0

Table 11
Accuracy of association rules for TOC concentration trained in Murcia and tested in Lisbon, Murcia, Madrid, Arenosillo and Montlouis data separately.

Rule code Murcia	Accuracy (%)				
	Lisbon	Murcia	Madrid	Arenosillo	Montlouis
U_{m1}	85.5	90.2	86.7	87.3	83.2
U_{m2}	88.4	93.1	91.9	91.3	89.0
U_{h1}	83.8	87.3	83.2	89.6	73.4
U_{v1}	89.0	94.2	89.6	92.5	85.0

Thus it is possible to calculate association rules for each location and TOC range (medium, high and very high) and the considered input meteorological variables (given in Table 1).

Applying the EQAR described in Section 3.3, association rules for the TOC[1] and the considered input meteorological variables have been calculated. EQAR has been executed five times for the two fitness function represented by the Eqs. (10) and (11) considering different TOC concentration (medium, high and very high) in each dataset. The best QAR obtained, thus, the rules with support greater than 3% and accuracy greater than 70% have been examined by the group of expert authors in meteorological data. Tables 4 to 6 show the results obtained for the three locations, where we have displayed the rules selected by the expert authors from the best ones according to their meteorological relevance. The column *Scores* describes the values obtained for the different interestingness measures used to qualify the QAR (support, confidence, amplitude, lift, leverage and accuracy). The column *Fitness* indicates the number of equation used as fitness function that has been optimized to obtain each QAR respectively.

It can be shown that most of the QARs provide in these tables were obtained by the second fitness function (Eq. (11)) which shows that the enhancement carried out in EQAR adding a new fitness function to evaluate the individuals in the population provides better and more relevant rules in terms of TOC concentration. Therefore the results obtained by the second fitness function improve the results obtained by the first fitness function (Eq. (10)). This enhancement is due to the first fitness function that only optimizes confidence and support, while the second fitness function optimizes the interest of the rules and the degree of dependence among the attributes belonging to the antecedent and the consequent (TOC concentration in this paper).

It can be appreciated that the scores of the quality measures of the QARs are very good in terms of the confidence and accuracy. Most of them reaches values very close to 100% also the lift and leverage values are greater than 1 and 0 respectively, therefore, the rules obtained present have high accuracy, reliability, and strong dependence among the attributes belonging to the antecedent and the consequent.

It is interesting to observe that all the considered variables form part of the association rules obtained, with some interesting peculiarities: variable ω_{200} is used to explain high and very high TOC concentration in Lisbon, but it does not appear in Murcia nor Madrid. Also in the Murcia case the confidence score is always 100. Note also that, in general, the confidence score is better for Madrid and Murcia than for Lisbon.

In order to discuss the physical correctness of the obtained association rules, we will do a comparison of these rules with results in previous studies. Several previous works have shown the quasi-linear relation that exists between TOC and the meteorological variables tropopause height and temperature at 50 hPa [8,37]. Note that this quasi-linear relationship cannot be found for OLR and ω200. Thus, in order to analyze how good association rules obtained are, ratios DU/km and DU/K have been calculated to be compared against the results obtained by other authors using similar data sources. In Tables 7–9, ratios (in absolute value) for the different tropopause heights (DU/km) and the temperature at 50 hPa (t_{50}) (DU/K) are showed. In each table we show values of TOC variation for two *TP* variables (the global (TP_G) and the *TP* variable calculated with a grid centered in the point (TP_W, TP_C and TP_E) for Lisbon, Madrid and Murcia, respectively).

In general, values for the four tropopause heights considered and temperature at 50 hPa (t_{50}) agree with results in different previous studies. In the case of the tropopause height in [7,37] it is shown that TOC values change approximately between 8 and 25 DU per 1 km increase in tropopause height. Note that the results obtained with the proposed association rules also follow these range of TOC

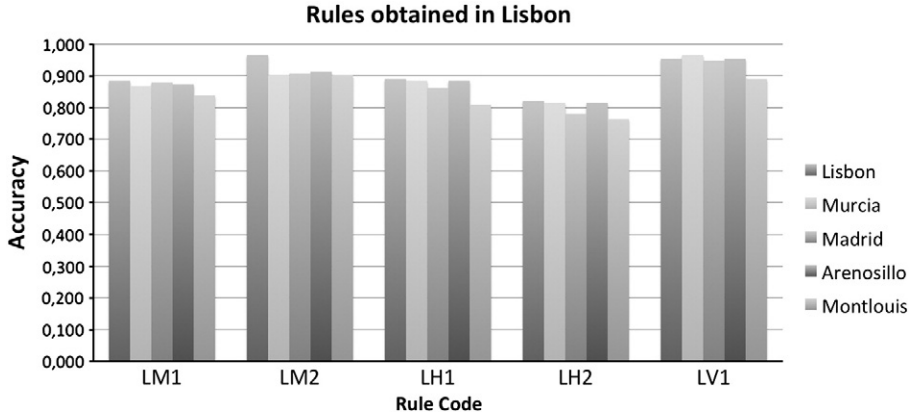

Fig. 6. Accuracy of association rules for TOC concentration trained in Lisbon data and tested in Lisbon, Murcia, Madrid, Arenosillo and Montlouis data separately.

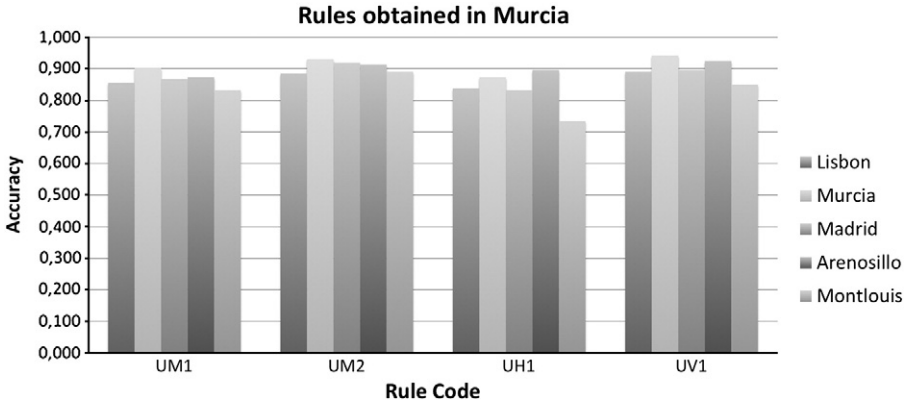

Fig. 7. Accuracy of association rules for TOC concentration trained in Murcia data and tested in Lisbon, Murcia, Madrid, Arenosillo and Montlouis data separately.

Table 12
Accuracy of association rules for TOC concentration trained in Madrid data and tested in Lisbon, Murcia, Madrid, Arenosillo and Montlouis data separately.

Rule code Madrid	Accuracy (%)				
	Lisbon	Murcia	Madrid	Arenosillo	Montlouis
M_{m1}	65.9	67.6	71.1	64.2	72.3
M_{m2}	80.9	81.5	83.8	85.0	81.5
M_{h1}	86.7	85.0	88.4	86.7	86.7
M_{h2}	90.2	91.9	90.8	90.2	87.9
M_{v1}	94.2	96.0	97.1	95.4	93.6

variation against all the TP variables considered at each location, and for all the rules (medium, high and very high TOC) considered. The only result out of this range in our rules is for Lisbon, medium TOC concentration (rule L_{m2}), in which a value of 5.9 DU/km is found for TOC variation against variable TP_G. For the case of TOC variation with variable t_{50} the variation ratio obtained in other works such as [8,37] is that 10 DU change in TOC corresponds to a roughly 1 K change of t_{50}. However, in other studies it is shown that these values can be quite affected by atmospheric variability, El Niño Southern Oscillation (ENSO), Quasi-Bienial Oscillation (QBO) [58], and values of TOC variation with t_{50} of 6, 12 or even 16 DU/K can be found at mid latitudes. Our results show that these thumb rule of 10 DU/K is very well fulfilled in the TOC variation against t_{50} in Madrid (medium TOC concentration) and in Murcia, mainly in the rule U_{m1}. The rest of the cases are not far away from these rule, showing a TOC variation with t_{50} between 6 and 7 DU/K which also agrees with values obtained in other works.

5. Validation of the obtained results

This section describes the tests carried out to validate the results obtained by EQAR in the previous section. In order to confirm that our model has no risk of over-fitting, the rules obtained by EQAR have been tested with six different datasets evaluating the accuracy of the rules. First, the rules obtained for each considered location (Lisbon, Madrid and Murcia) have been tested in the datasets corresponding to five locations (Lisbon, Madrid, Murcia, Arenosillo and Montlouis) separately (Fig. 5 shows these locations, the 3 previously considered and Montlouis and Arenosillo, newly added for this validation study). In addition, the rules have been tested in a dataset containing the TOC values of Arenosillo and Montlouis which consist of 346 instances in total.

Table 10 and Fig. 6 describe the accuracy values corresponding to the rules obtained in the dataset of Lisbon as training data for each level of TOC concentration (medium, high and very high) in the five locations as test data separately. Similarly, Table 11 and Fig. 7 show the accuracy values obtained for the rules belonging to the dataset of Murcia. Finally, Table 12 and Fig. 8 indicate the accuracy values obtained for the rules corresponding to the dataset of Madrid. It can be observed that the accuracy obtained for each rule discovered with the training datasets (Lisbon, Murcia and Madrid) are quite similar when they are applied to other dataset used as test data (Lisbon, Murcia, Madrid, Arenosillo and Montlouis). As can be seen, even in some cases there is greater value of accuracy on the test

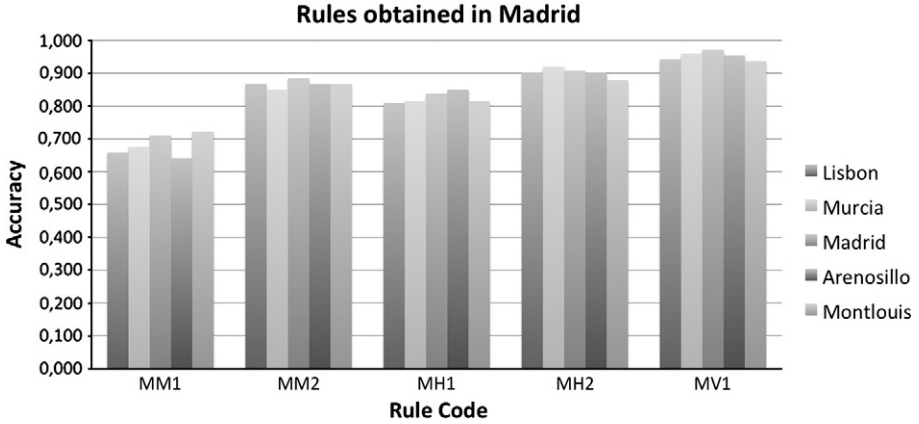

Fig. 8. Accuracy of association rules for TOC concentration trained in Madrid and tested in Lisbon, Murcia, Madrid, Arenosillo and Montlouis data separately.

Table 13
Accuracy of association rules trained in Lisbon, Murcia and Madrid and tested in a dataset containing all TOC concentration of Arenosillo and Montlouis.

Rule code	Accuracy (%)	Rule code	Accuracy (%)	Rule code	Accuracy (%)
Medium		High		Very high	
L_{m1}	85.5	L_{h1}	84.7	L_{v1}	92.2
L_{m2}	90.8	L_{h2}	78.9		
U_{m1}	85.3	U_{h1}	81.5	U_{v1}	88.7
U_{m1}	90.2				
M_{m1}	68.2	M_{h1}	83.2	M_{v1}	94.5
M_{m2}	86.7	M_{h2}	89		

data with respect to training data. Therefore the accuracy obtained is stable, since there are no distinct differences among the datasets, which indicates that there is not over-fitting among the rules learned and datasets used as training data.

Accuracy values of QAR tested in a dataset containing the TOC concentration of Arenosillo and Montlouis have been displayed in Table 13. Rules for each location used as training data have been separated by level of TOC concentration and accuracy values are shown graphically in Figs. 9 to 11. Fig. 9 represents the rules discovered in Lisbon, Murcia and Madrid for medium TOC concentration and their accuracy values obtained in the test dataset. Similarly, Fig. 10 describes the rules discovered in Lisbon, Murcia and Madrid for high TOC concentration and their accuracy values obtained in the test dataset. Finally, Fig. 11 shows the rules discovered in Lisbon, Murcia and Madrid for very high TOC concentration and their accuracy values obtained in the test dataset. These tests prove that the rules obtained separately for a particular location are valid for locations analyzed together. The results show accuracy rates above 80% except in one case (M_{m1}), and over 90% in many cases.

After this validation study, we can conclude that there is no over-fitting among rules obtained and the dataset used as training data and we can confirm that the EQAR approach has been really good in terms of the quality of the QAR found because the accuracy values are very high (exceeding 80%) and are very similar in all datasets used as test data.

6. Conclusions

In this paper we have described the application of the EQAR algorithm (Evolutionary Quantitative Association Rules) to a problem Total Ozone Content (TOC) modeling in the Iberian Peninsula. Different improvements in the initial population generation and fitness function have been incorporated to EQAR in order to improve its performance in this problem of TOC modeling. Experimental results have been carried out in TOC data from the Total Ozone Mapping Spectrometer (TOMS) measuring at three locations (Lisbon, Madrid and Murcia) of the Iberian Peninsula. As prediction variables for the association rules we have considered several meteorological variables, such as Outgoing Long-wave Radiation (OLR), Temperature at 50 hPa level, Tropopause height, and wind vertical velocity component at 200 hPa. The results obtained with the EQAR approach have been really good in terms of the quality of the association rules found. Also, the analysis of these rules agrees with the results obtained in other works dealing with TOC modeling, so we can conclude that the use of association rules in TOC modeling could be an interesting analysis method for the future in this and similar problems.

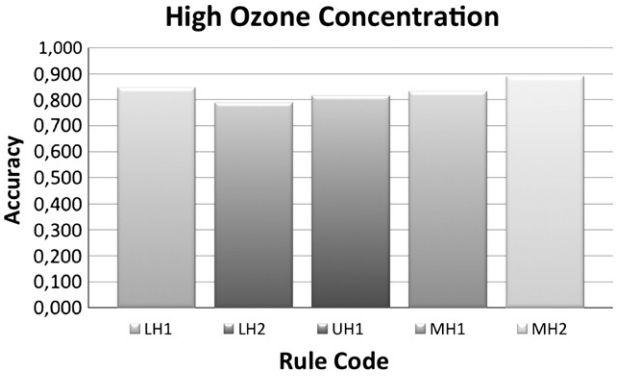

Fig. 10. Accuracy of association rules for high TOC concentration trained in Lisbon, Murcia and Madrid and tested in a dataset containing all location (Lisbon, Murcia. Madrid, Arenosillo and Montlouis).

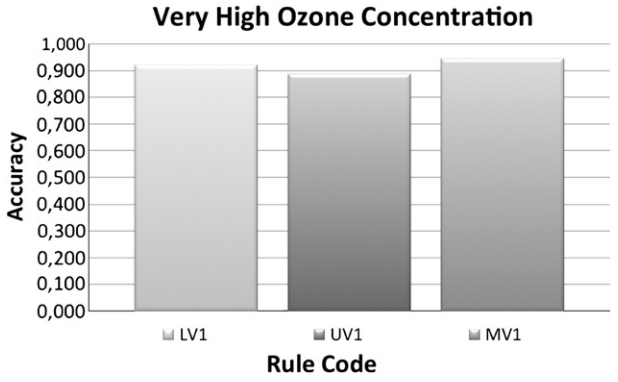

Fig. 11. Accuracy of association rules for very high TOC concentration trained in Lisbon, Murcia and Madrid and tested in a dataset containing all location (Lisbon, Murcia. Madrid, Arenosillo and Montlouis).

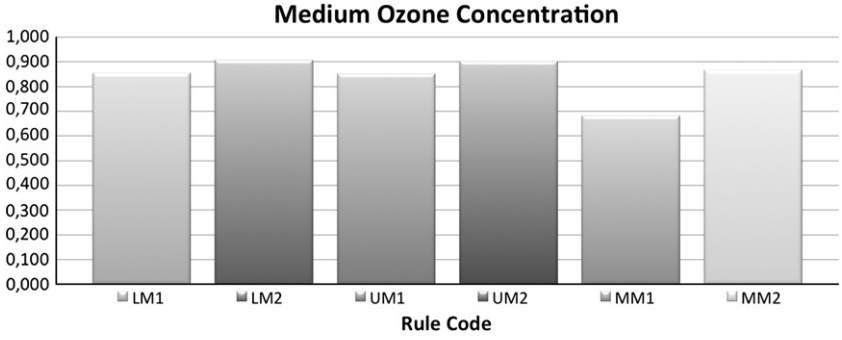

Fig. 9. Accuracy of association rules for medium TOC concentration trained in Lisbon, Murcia and Madrid and tested in a dataset containing all location (Lisbon, Murcia. Madrid, Arenosillo and Montlouis).

References

[1] S. Bronnimann, J. Luterbacher, C. Schmutz, H. Wanner, J. Staehelin, Variability of total ozone at Arosa, Switzerland, since 1931 related to atmospheric circulation indices, Geophysical Research Letters 27 (15) (2000) 2213–2216.

[2] B. Massart, O.M. Kvalheim, L. Stige, R. Aasheim, Ozone forecasting from meteorological variables: part I. Predictive models by moving window and partial least squares regression, Chemometrics and Intelligent Laboratory Systems 42 (1–2) (1998) 179–190.

[3] B. Massart, O.M. Kvalheim, L. Stige, R. Aasheim, Ozone forecasting from meteorological variables: part II. Daily maximum ground-level ozone concentration from local weather forecasts, Chemometrics and Intelligent Laboratory Systems 42 (1–2) (1998) 191–197.

[4] X. Jin, J. Li, C.C. Schmidt, T.J. Schmit, J. Li, Retrieval of total column ozone from imagers onboard geostationary satellites, IEEE Transactions on Geoscience and Remote Sensing 46 (2) (2008) 479–488.

[5] M. Palacios, F. Kirchner, A. Martilli, A. Clappier, F. Martín, M.E. Rodríguez, Summer ozone episodes in the Greater Madrid area. Analyzing the ozone response to abatement strategies by modelling, Atmospheric Environment 36 (2002) 5323–5333.

[6] J.W. Krzyscin, J.L. Borkowski, Variability of the total ozone trend over Europe for the period 1950–2004 derived from reconstructed data, Atmospheric Chemistry and Physics 8 (2008) 2847–2857.

[7] W. Steinbrecht, U. Köhler, K.P. Hoinka, Correlation between tropopause height and total ozone: implication for long-term trends, Journal of Geophysical Research 103 (1998) 19 183–19 192.

[8] W. Steinbrecht, B. Hassler, H. Claude, P. Winkler, R.S. Stolarski, Global distribution of total ozone and lower stratospheric temperature variations, Atmospheric Chemistry and Physics 3 (2003) 1421–1438.

[9] M.A. Barrero, J.O. Grimalt, L. Cantón, Prediction of daily ozone concentration maxima in the urban atmosphere, Chemometrics and Intelligent Laboratory Systems 80 (1) (2006) 67–76.

[10] G. Christakos, A. Kolovos, M.L. Serre, F. Vukovich, Total ozone mapping by integrating databases from remote sensing instruments and empirical models, IEEE Transactions on Geoscience and Remote Sensing 42 (5) (2004) 991–1008.

[11] M. Felipe-Sotelo, L. Gustems, I. Hernández, M. Terrado, R. Tauler, Investigation of geographical and temporal distribution of tropospheric ozone in Catalonia (North-East Spain) during the period 2000–2004 using multivariate data analysis methods, Atmospheric Environment 40 (2004) 7421–7436.

[12] P.J. Crutzen, The influence of nitrogen oxide on the atmospheric ozone content, Quarterly Journal of the Royal Meteorological Society 96 (1970) 320–327.

[13] P.J. Crutzen, Ozone production rates in an oxygen–hydrogen nitrogen oxide atmosphere, Journal of Geophysical Research 76 (1971) 7311–7327.

[14] R.S. Stolarski, R.J. Cicerone, Stratospheric chlorine: a possible sink for ozone, Canadian Journal of Chemistry 52 (1974) 1610–1615.

[15] M.J. Molina, F.S. Rowland, Stratospheric sink for chlorofluoromethanes: chlorine atom catalyzed destruction of ozone, Nature 249 (1974) 810–812.

[16] J.C. Farman, B.G. Gardiner, J.D. Shanklin, Large losses of total ozone in Antarctica reveal seasonal ClOx/NOx interaction, Nature 315 (1985) 207–210.

[17] S. Solomon, Stratospheric ozone depletion: a review of concepts and history, Reviews of Geophysics 37 (3) (1999) 275–316.

[18] United Nations Environment Programme, Environmental Effects Assessment Panel, Environmental effects of ozone depletion and its interactions with climate change: progress report 2005, Photochemical & Photobiological Sciences 5 (13) (2006).

[19] K. Bramstedt, J. Gleason, D. Loyola, W. Thomas, A. Bracher, M. Weber, J.P. Burrows, Comparison of total ozone from the satellite instruments GOME and TOMS with measurements from the Dobson network 1996–2000, Atmospheric Chemistry and Physics 3 (2003) 1409–1419.

[20] A.A. Silva, A quarter century of TOMS total column ozone measurements over Brazil, Journal of Atmospheric and Solar-Terrestrial Physics 69 (12) (2007) 1447–1458.

[21] V. Savastiouk, C.T. McElroy, Brewer spectrophotometer total ozone measurements made during the 1998 middle atmosphere nitrogen trend assessment (MANTRA) Campaign, Atmosphere-Ocean 43 (4) (2005) 315–324.

[22] K.M. Latha, K.V. Badarinath, Impact of aerosols on total columnar ozone measurements a case study using satellite and ground-based instruments, Atmospheric Research 66 (4) (2003) 307–313.

[23] Z. Xiangdong, Z. Xiuji, T. Jie, Q. Yu, C. Chuenyu, A meteorological analysis on a low tropospheric ozone event over Xining, North Western China on 26–27 July 1996, Atmospheric Environment 38 (2) (2004) 261–271.

[24] A. Pérez, I. Aguirre de Cárcer, F. Jaque, Low ozone event at Madrid in November 1996, Journal of Atmospheric and Solar-Terrestrial Physics 64 (3) (2002) 283–289.

[25] C. Appenzeller, A.K. Weiss, J. Staehelin, North Atlantic oscillation modulates total ozone winter trends, Geophysical Research Letters 27 (8) (2000) 1131–1134.

[26] T.G. Shepherd, A.I. Jonsson, On the attribution of stratospheric ozone and temperature changes to changes in ozone-depleting substances and well-mixed greenhouse gases, Atmospheric Chemistry and Physics 8 (2008) 1435–1444.

[27] J. Han, M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, 2006.

[28] R. Agrawal, T. Imielinski, A. Swami, Mining association rules between sets of items in large databases, Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, 1993, pp. 207–216.

[29] R. Agrawal, R. Srikant, Fast algorithms for mining association rules in large databases, Proceedings of the International Conference on Very Large Databases, 1994, pp. 478–499.

[30] M. Houtsma, A. Swami, Set-Oriented Mining for Association Rules in Relational Databases, IEEE Computer Society, 1995, pp. 25–33.

[31] A.E. Eiben, J.E. Smith, Introduction to Evolutionary Computing, Springer-Verlag, 2003.

[32] M. Martínez-Ballesteros, F. Martínez-Álvarez, A. Troncoso, J.C. Riquelme, Mining quantitative association rules based on evoluationary computation and its application to atmospheric pollution, Integrated Computer-Aided Engineering 17 (3) (2010) 227–242.

[33] R.D. McPeters, P.K. Bhartia, A.J. Krueger, J.R. Herman, B.M. Schlesinger, C.G. Wellemeyer, C.J. Seftor, G. Jaross, S.L. Taylor, T. Swissler, O. Torres, G. Labow, W. Byerly, R.P. Cebula, Nimbus-7 Total Ozone Mapping Spectrometer (TOMS) Data Products User's Guide, NASA Reference Publication, 1996.

[34] M. Anton, J.M. Vilaplana, M. Kroon, A. Serrano, M. Parias, M.L. Cancillo, B.A. de la Morena, The empirically corrected EP-TOMS total ozone data against Brewer measurements at El Arenosillo (Southwestern Spain), IEEE Transactions on Geoscience and Remote Sensing 48 (7) (2010) 3039–3045.

[35] E. Kalnay, M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Woollen, Y. Zhu, A. Leetmaa, R. Reynolds, M. Chelliah, W. Ebisuzaki, W. Higgins, J. Janowiak, K.C. Mo, C. Ropelewski, J. Wang, R. Jenne, D. Joseph, The NCEP/NCAR reanalysis 40-year project, Bulletin of the American Meteorological Society 77 (1996) 437–471.

[36] B. Liebmann, C.A. Smith, Description of a complete (Interpolated) outgoing longwave radiation dataset, Bulletin of the American Meteorological Society 77 (1996) 1275–1277.

[37] C. Varotsos, C. Cartalis, A. Vlamakis, C. Tzanis, I. Keramitsoglou, The long-term coupling between column ozone and tropopause properties, Journal of Climate 17 (2004) 3843–3854.

[38] K. Hoinka, H. Claude, U. Kohler, On the correlation between tropopause pressure and ozone above central Europe, Geophysical Research Letters 23 (1996) 1753–1756.

[39] World Meteorological Organization, Meteorology — a three dimensional science, World Meteorological Organization Bulletin 6 (1957) 134–138.

[40] G. Zangl, K. Hoinka, The tropopause in the polar regions, Journal of Climate 14 (14) (2001) 3117–3139.

[41] A. Rabbe, S.H. Larse, Ozone variations in the northern-hemisphere due to dynamic processes in the atmosphere, Journal of Atmospheric and Solar-Terrestrial Physics 54 (9) (1992) 1107–1112.

[42] A. Rabbe, S.H. Larse, 'On the low ozone values over Scandianvia during the winter of 1991–1992', Journal of Atmospheric and Solar-Terrestrial Physics 57 (4) (1995) 367–373.

[43] K. Henriksen, V. Roldugin, Total ozone variations and meteorological processes, Atmospheric Ozone Dynamics, in: Costas Varotsos (Ed.), Series I: Global Environmental Change, vol. 53, Springer Verlag, 1997.

[44] P.W. Menzel, Applications with meteorological satellites, World Meteorological Organization (WMO), Technical Document No. 1078, 2001.

[45] V. Williams, R. Toumi, The correlation between tropical total ozone and outgoing long-wave radiation, Quarterly Journal of the Royal Meteorological Society 127 (2001) 989–1003.

[46] B. Alatas, E. Akin, An efficient genetic algorithm for automated mining of both positive and negative quantitative association rules, Soft Computing 10 (3) (2006) 230–237.

[47] O. Berzal, I. Blanco, D. Sánchez, M. Vila, Ios press measuring the accuracy and interest of association rules: a new framework, , 2001.

[48] L. Geng, H. Hamilton, Interestingness measures for data mining: a survey, ACM Computing Surveys 38 (3) (2006) 9.

[49] S. Brin, R. Motwani, C. Silverstein, Beyond market baskets: generalizing association rules to correlations, Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data, vol. 26, no. 2, 1997, pp. 265–276.

[50] G. Piatetsky-Shapiro, Discovery, analysis and presentation of strong rules, Knowledge Discovery in Databases, 1991, pp. 229–248.

[51] J. Mata, J.L. Álvarez, J.C. Riquelme, Discovering numeric association rules via evolutionary algorithm, Lecture Notes in Artificial Intelligence 2336 (2002) 40–51.

[52] X. Yan, C. Zhang, S. Zhang, Genetic algorithm-based strategy for identifying association rules without specifying actual minimum support, Expert Systems with Applications: An International Journal 36 (2) (2009) 3066–3076.

[53] B. Alatas, E. Akin, A. Karci, MODENAR: multi-objective differential evolution algorithm for mining numeric association rules, Applied Soft Computing 8 (1) (2008) 646–656.

[54] J. Alcala-Fdez, N. Flugy-Pape, A. Bonarini, F. Herrera, Analysis of the effectiveness of the genetic algorithms based on extraction of association rules, Fundamenta Informaticae 98 (1) (2010) 1001–1014.

[55] M. Martínez-Ballesteros, F. Martínez-Álvarez, A. Troncoso, J. Riquelme, Quantitative association rules applied to climatological time series forecasting, Intelligent Data Engineering and Automated Learning — IDEAL 2009, ser. Lecture Notes in Computer Science, vol. 5788, 2009, pp. 284–291.

[56] L. Eshelman, The CHC Adaptative Search Algorithm: How to Have Safe Search when Engaging in Nontraditional Genetic Recombination, Morgan Kaufmann, 1991.

[57] G. Venturini, SIA: a supervised inductive algorithm with genetic search for learning attribute based concepts, Proceedings of the European Conference on Machine Learning, 1993, pp. 280–296.

[58] W.J. Randel, J.B. Cobb, Coherent variation of monthly mean total ozone and lower stratospheric temperature, Journal of Geophysical Research 99 (1994) 5433–5447.