Artificial Intelligence in Medicine 74 (2016) 44–52

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

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

Analysis of correlation between pediatric asthma exacerbation and exposure to pollutant mixtures with association rule mining

Giulia Toti[a,*], Ricardo Vilalta[a], Peggy Lindner[d], Barry Lefer[b,c], Charles Macias[e], Daniel Price[d]

[a] Department of Computer Science, University of Houston, Philip Guthrie Hoffman Hall, 3551 Cullen Blvd., Room 501, Houston, TX 77204-3010, USA
[b] Department of Earth and Atmospheric Sciences, University of Houston, Science & Research Building 1, 3507 Cullen Blvd, Room 312, Houston, TX 77204-5007, USA
[c] Now at: Earth Sciences Division, NASA Headquarters, 300 E St SW, Washington, DC 20546, USA
[d] Honors College, University of Houston, M.D Anderson Library, 4333 University Dr, Room 212, Houston, TX 77204-2001, USA
[e] Department of Pediatrics, Baylor College of Medicine/Texas Children's Hospital, One Baylor Plaza, Houston, TX 77030, USA

ARTICLE INFO

Article history:
Received 11 December 2015
Received in revised form
22 November 2016
Accepted 23 November 2016

Keyword:
Association rule mining
Rule redundancy
Risk assessment
Multiple exposures
Pediatric asthma
Outdoor pollution

ABSTRACT

Objectives: Traditional studies on effects of outdoor pollution on asthma have been criticized for questionable statistical validity and inefficacy in exploring the effects of multiple air pollutants, alone and in combination. Association rule mining (ARM), a method easily interpretable and suitable for the analysis of the effects of multiple exposures, could be of use, but the traditional interest metrics of support and confidence need to be substituted with metrics that focus on risk variations caused by different exposures.
Methods: We present an ARM-based methodology that produces rules associated with relevant odds ratios and limits the number of final rules even at very low support levels (0.5%), thanks to post-pruning criteria that limit rule redundancy and control for statistical significance. The methodology has been applied to a case-crossover study to explore the effects of multiple air pollutants on risk of asthma in pediatric subjects.
Results: We identified 27 rules with interesting odds ratio among more than 10,000 having the required support. The only rule including only one chemical is exposure to ozone on the previous day of the reported asthma attack (OR = 1.14). 26 combinatory rules highlight the limitations of air quality policies based on single pollutant thresholds and suggest that exposure to mixtures of chemicals is more harmful, with odds ratio as high as 1.54 (associated with the combination day0 SO_2, day0 NO, day0 NO_2, day1 PM).
Conclusions: The proposed method can be used to analyze risk variations caused by single and multiple exposures. The method is reliable and requires fewer assumptions on the data than parametric approaches. Rules including more than one pollutant highlight interactions that deserve further investigation, while helping to limit the search field.

1. Background and objectives

The adverse impact of air pollution on health is well established, and estimates of total mortality and of connections to individual disease suggest that even relatively low levels may impose substantial health burdens [1]. Exposure to pollutants has been linked to numerous health outcomes, including both the inception and triggering of asthma in children. Asthma is a chronic respiratory disease that is responsible for thousands of deaths every year in the U.S. and affects 1 in every 12 people [2]. Children are the most vulnerable to asthma because of their developing, narrower airways. They also inhale more air per pound of body weight than adults, which causes them to inhale, in proportion, a higher quantity of pollutants [3]. Available literature tends to agree on the existence of a positive correlation between asthma and outdoor pollutants, concerning both incidence and cases of exacerbation [4–7]. The assessment of the impact of the single chemicals present in the air still poses a challenge, due to the difficulty in controlling and separating the exposures [8]. The six criteria pollutants (carbon monoxide, lead, nitrogen oxides, ground-level ozone, particulate matter, and sulfur oxides) have been extensively monitored and individually regulated since the Clean Air Act of 1970. Currently, the US Environmental Protection Agency (EPA) issues air quality

* Corresponding author at: Honors College, M.D. Anderson Library, 4333 University Dr, Room 212, Houston, TX, 77204–2001, USA.
E-mail address: giulia.toti@kcl.ac.uk (G. Toti).

http://dx.doi.org/10.1016/j.artmed.2016.11.003
0933-3657/

warnings based on predicted high levels of any one of the pollutants [9]. Unfortunately, these warnings fail to account for potential combinatory effects of the chemicals, which may result in decreased effectiveness and unclear action plans for asthmatics.

There are multiple reasons for favoring the analysis of the effects of mixtures over single pollutants. First of all, because single chemicals have different effects on the respiratory and circulatory systems [5], it is reasonable to suspect that the action of a pollutant could be more harmful if another chemical has already weakened a sensitive part of the organism. This synergic action has already been observed in studies on combinatory effects of pollutants and aeroallergens [50,51]. Second, The chemicals in the atmosphere manifest high correlation between each other, with seasonal weather patterns, and with other seasonal or meteorological causes, such as humidity or pollen. An analysis with standard statistical approaches, such as logistic regression, would be inadequate to identify the effects of single pollutants on the risk of asthma exacerbation, because of well-known problematics in handling highly correlated variables. Furthermore, controlling for single pollutants, as for current EPA standard, may not be particularly useful if chemicals are always present in the atmosphere as a mixture. The environmental health community, including regulators, epidemiologists, and health practitioners, encourages the development of new paradigms to explore the diverse contributions of multiple air pollutants to health outcomes [10].

Relative strengths and weaknesses of different methodological approaches must be assessed for any new multi-pollutant paradigm to develop, although it is unlikely that a single approach will fit all the needs of the community of researchers. Several methods have been explored, and while concerns have been expressed about the effectiveness of traditional epidemiological approaches [11], interest is rising around techniques from the field of data mining and knowledge discovery. In [12] the authors provided a review of various statistical methods used to evaluate the effects of combinations of chemicals. Logistic regression, the gold standard in epidemiological studies, is described as struggling with assumptions of linearity with the logarithm of the odds, pollutants collinearity, difficulty of accounting for multiple interactions between pollutants, and measurement and sampling error [13–18]. Hierarchical Bayesian methods are cited as showing promise, since the approach allows for averaging over different models and retains the interpretability of traditional regression [19,20]. Other recent data mining approaches using sparse principal component analysis [21] and smoothing algorithms for the regressions may be able to provide an overall sense of which environmental variables are most responsible for asthma incidence and acute episodes. However, the existing approaches to the multi-pollutant problems remain poorly suited for isolating clinical and policy relevant effects of multiple pollutants [12]. Classification and Regression Trees (CART) are another viable option, and an attempt to identify and characterize harmful mixtures of exposures with CART has been proposed by Gass et al. [42]. The authors use a Poisson generalized linear model to iteratively determine the exposure with the smallest p-value, and select it as split node. The main problem with this approach lays in the hierarchical structure typical of a classification tree, which could potentially miss synergic actions between chemicals. CART would not recognize a dangerous mixture unless at least one of the included chemicals is identified to be harmful independently (nodes in a classification tree are selected in order, on bases such as changes in entropy in the resulting subgroups). The authors also state that this method is susceptible to collinearity, which could result in the selection of the exposure with the smallest measurement error, rather than the real causal exposure. The wrong selection of a node would affect all subsequent branches of the tree. Finally, approaches that allow for treating air pollution as a homogenized mixture, especially in the context of indoor exposures, have been developed [22], although skeptics can point to the growing risk of ecological fallacies, and of data massaging based on researcher bias, that emerge when the individual researchers are categorizing groupings of interest that rely on interpretation at the same time that they are preparing and analyzing the data.

In response to this context, our research team chose to investigate a variant of an association rule mining (ARM) algorithm. Compared to other methods, ARM has excellent interpretability, even for people who do not have data mining expertise. For this reason, ARM has found several applications in the medical domain, including research on chronic hepatitis, septic shock, heart disease, association deficit disorder, cancer prevention, and more [23–27].

Beside interpretability, other reasons make ARM a valid candidate for shedding more light on the difficult problem of correlation between asthma and outdoor pollution. First, the rules produced by this algorithm are capable of summarizing the impact of several factors in combination in a non-hierarchical fashion. In other words, the risk produced by a mixture of chemicals is not modeled as the effect of a pollutant of interest worsened by the simultaneous presence of other substances, but rather as the total effect of the exposures. Such combinatory effects include both joint effects between pollutants (two or more chemicals acting simultaneously on the body to increase the overall chance of an asthma attack) and interactions (synergic action of two or more chemicals that causes more harm than the combined exposure to the single elements). Second, our modified version of association rule mining retains all original risk factors and allows for the analysis of their possible combinations, shifting the work of establishing statistical significance toward understanding which ones of the proposed rules are meaningful (previous research on ARM has produced a considerable variety of useful metrics that establish statistical significance [43]). Finally, this approach limits unverifiable researcher presuppositions, such as hypothesis of cause-effect through time. This is an important improvement considering recent attacks on the possibility of meeting those conditions, even under the most carefully constructed randomized control trials. A mathematical method that does not rely on implicitly linear constructions of cause and effect should be preferred [28].

When applying well-known ARM algorithms, like *Apriori* [44], to extract new knowledge from data, one often has to face the limitations posed by the *support-confidence framework*. A detailed description of the meaning of support and confidence is offered in the Method section (2). For now, it is sufficient to know that this implies looking for frequent and strong associations in the data. In epidemiological studies, however, relevant associations may not be particularly strong or frequent (in a typical population, subjects with the condition of interest are outnumbered by healthy subjects). If we want to employ ARM to extract knowledge from data in this scenario, we must allow it to search for less frequent associations, and then use other criteria to identify relevant ones in what will possibly be a very long list.

In our studies, we developed a new methodology for rules selection from epidemiological data, based on traditional association rule mining. In this paper, we present the resulting configuration, comprising an *Apriori* ARM implementation [44,45], two criteria to eliminate non-statistically significant and redundant rules, and a training-testing strategy necessary to mine more interesting associations while limiting noise. We then discuss the results obtained by applying the proposed methodology on a large dataset on pediatric asthma cases and pollution levels in the greater Houston area. We find the results easy to interpret on a rule-by-rule basis, and the statistical techniques for building a set of rules into a model is appropriately insulated from problems of massaging categories to fit a model or unjustifiably smoothing underlying non-linear relations. Emerging patterns in the final set of rules could be further explored with other techniques, considerably reducing the

Fig. 1. Representation of dataset expansion following case-crossover study design. The assumption made is that a subject who visits the emergency department on a given day did not visit again 1 or 2 weeks before and after the event. A similar approach has been used by Raun et al. [30].

search space. Although this method was designed within the scope of a multipollutant exposure analysis, it is suitable for other epidemiological studies involving multiple variables and requiring a non-parametric approach.

A detailed description of the data and of the rule mining process is reported in the Methods section. In the Result section we report the list of relevant rules produced by the algorithm when applied to the asthma and pollution study. Rules include one or multiple exposures. In the Discussion section, we will present our conclusions on the effects of air pollution on asthma incidence and how the modified ARM helped in this discovery. Our findings are summarized in the Conclusions section.

2. Methods

2.1. Data

For our research, we used data related to pediatric emergency department (ED) visits attributed to asthma collected by the Texas Emergency Department Asthma Surveillance (TEDAS), a network of four hospitals in the Houston area, which shared the data related to 20,959 pediatric ED visits from 01/01/02 to 31/12/12. The only patients excluded from the records were those also diagnosed with cardiovascular or pulmonary disease. The features of the database include demographics, insurance status, primary care provider, diagnosis and severity assessment performed by the admitting staff. Since patients were not provided with a unique identifier, it is not possible to isolate returning individuals.

The database was expanded following the design of a case-crossover study to include controls (days when patients did not experience an asthma attack and/or did not visit the ED). In a case-crossover study it is simpler to select a large and well representative sample population, because it is not required to divide the subjects in cohorts or to find appropriately matched controls. The design also allows using routinely collected data for studies at individual levels [29]. We made the assumption that no subject experienced severe asthma 1 and 2 weeks before, or 1 and 2 weeks after the date of the ED visit (Fig. 1). Therefore, for every subject we have four control days. The database size after expansions includes 104,795 events. Raun et al. [30] used a similar approach in their research for the Houston emergency medical service. This technique is robust against demographic confounders such as age or gender. Unfortunately, it still cannot fully account for potentially correlated transient patterns, such as the presence of pollen, non measured co-pollutants such as $PM_{1.0}$, or weather.

Pollution levels were gathered from the Texas Commission on Environmental Quality (TCEQ) monitor network. Over the Houston area (TCEQ Region 12), a total of 103 monitors are available, including 6 for CO, 7 for SO_2, 21 for NO and NO_2, 42 for O_3, and 6 for $PM_{2.5}$ (particulate matter of diameter of 2.5 μm or smaller). The sensors of the TCEQ network record the level of pollutants every 5 min and at various locations over the city of Houston (Fig. 2). The raw data have been grouped according to date and hour and averaged into a single hourly value, in order to reduce noise due to possible measurement errors and sudden fluctuations in the chemical levels. In the next step, for every sensor and every pollutant, the maximum value recorded for each day has been annotated. The result of this operation was a complete database tracing the maximum level reached by every pollutant every day and at any available location during the desired period of time (01/01/02 to 31/12/12). A summary of the distribution of the six pollutants of interest over the Houston area from January 1st 2002 to December 31st 2012 is presented in Table 1. Fig. 3 shows instead the correlation between the pollutant distributions.

Table 1
Summary of the distribution of the six pollutants under analysis over the Houston area from January 1st 2002 to December 31st 2012.

	CO (ppb)	SO_2 (ppb)	NO (ppb)	NO_2 (ppb)	O_3 (ppb)	$PM_{2.5}$(μg/m^3)
1st quart.	297.61	1.36	2.11	9.21	33.70	14.07
Median	473.17	3.45	6.57	16.84	43.98	19.04
Mean	612.15	6.67	21.59	19.69	47.78	21.20
3rd quart.	760.12	7.95	21.33	27.70	58.53	25.79
St. Dev.	522.23	9.60	42.11	13.33	20.12	12.09

In order to understand the relationship between ED visits due to asthma attacks and air quality, each subject of the expanded patients dataset has been associated with pollutant levels recorded on the date of the visit (or control dates, for non-asthma events). The TEDAS database reports the location of the patient domicile (as zip code), so it is possible to associate the patient to the closest sensors. We do not know where the patient was at the moment of the event, but since the database is related to young subjects, we assumed a close distance to their home. We imposed a limit of 20 km between sensor and subject zip code centroid, and picked the closest sensor when more than one was available within that radius. This is a reasonable approximation and it is more reliable than trying to interpolate between sensors locations, because movements of chemicals in the atmosphere are influenced by a very high number of factors, such as wind and precipitation. The total number of subjects whose zip code is within the Houston city limits and with no missing pollutant data is 14,704, including 2973 cases and 11,731 controls. Because some sensors were activated later in time, no event antecedent to 27/02/06 had complete pollution records.

In the dose effect literature, delayed actions of chemicals on the human body have been observed [30–32]. For this reason, the pollutant levels recorded on the same day of the event have been augmented by the levels recorded from 1 to 4 days before, for a total of 30 possible exposures per subject. ARM treats the exposures as separate events, initially, and then allows for analyses of the statistical significance of combined exposures.

The strategy adopted for data collection and processing is one of the pillars of this project. A large patient dataset is available, with records collected over more than a decade. The study case-crossover design used to model the patients' dataset is the best suited for transient exposures and health outcomes. Additionally, the spatial resolution obtained using the large number of sensors of the TCEQ network gives us a better estimate of the actual subject exposure in comparison with studies that used fewer monitors or a citywide average [33–35].

2.2. Analysis with association rule mining

Association rule mining was first introduced in [37]. Let $I = \{i_1, i_2, \ldots, i_m\}$ be a generic set of items. A subset of items $X \subseteq I$ is called an *itemset*. Tuples of the form t, X, including a transaction identifier t and a set of items, are called *transactions*. All the transactions

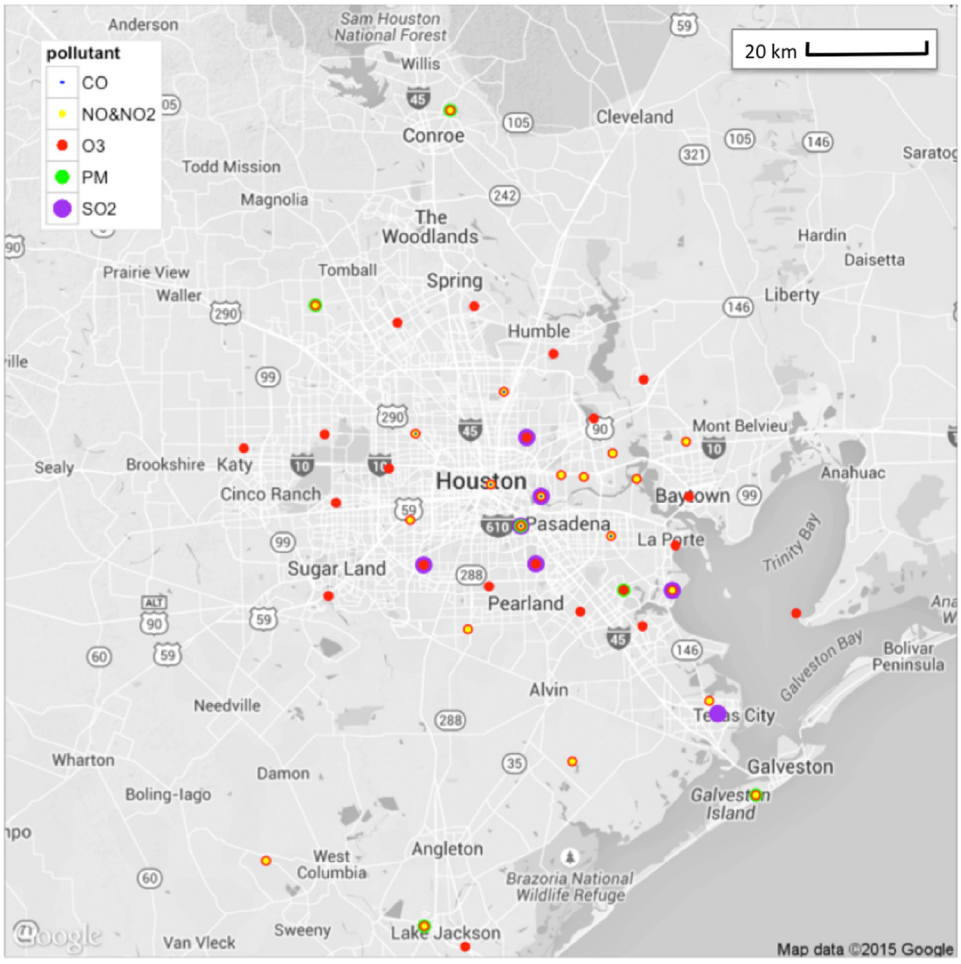

Fig. 2. Representation of distribution of sensors of the Texas Commission on Environmental Quality network over the Houston area. Some sensors are located beyond the map boundaries, and they are too far from any registered patient to be of interest, therefore were not included in this map.

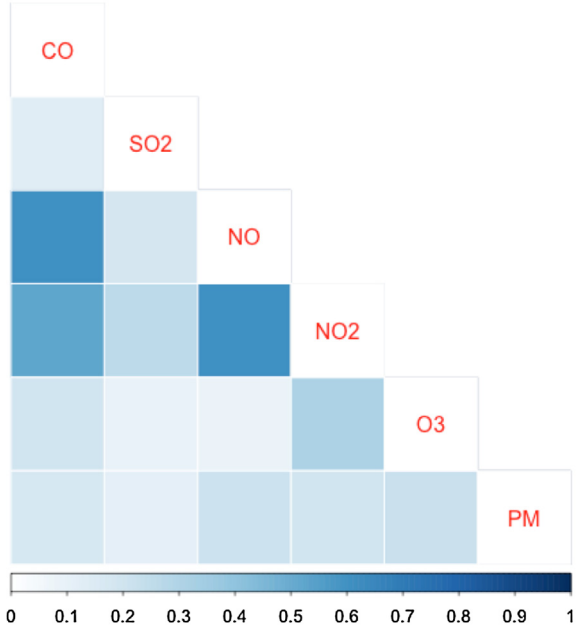

Fig. 3. Colormap of Pearson correlation coefficients between daily pollutant distributions. Darker values indicate high correlation, while light values indicate no correlation (independence).

available in the database under investigation are represented in form of a binary matrix **D**. The rows of **D** represent all the transactions of the database, while its columns represent the set of all possible items. $\mathbf{D}_{i,j}$ is true (1) if the item j is present in the ith transaction, and false otherwise. The transaction database **D** is searched for rules of the form $X \rightarrow Y$, where X and Y are itemsets of cardinality $1 \leq k \leq m$. The rule indicates that when the set of items X is present, it is also likely to find the set of items Y in the same transaction. Most algorithms for itemset mining search for rules with high *support* and *confidence*. The support of a rule indicates how frequently it appears in **D**, while its confidence measures the probability of finding Y when X is present in a transaction.

In this application, we are interested in rules of the form $X \rightarrow$ case, where $X \subseteq E$, $E = \{e1, e2, \ldots, em\}$ is the set of all possible exposures and case indicates a subject going to the ER because of an asthma attack. Since **D** can only include binary variables, it was necessary to transform the pollutant levels (numerical) into binary variables representing exposure to a particular pollutant. We defined an exposure in relation to monitor readings above the third quartile of the distribution of each pollutant. The threshold values used to bin each pollutant are reported in Table 2. This decision mirrors the approach currently used by the EPA, which uses single pollutant levels to issues air quality warnings. However, we decided not to use the threshold values indicated by the EPA because they were established to account for a range of health effects (particularly mortality rates), and less influenced by the more recent literature on asthma attacks. Note that the third

Table 2
Thresholds above which a subject is considered to have been exposed to a particular pollutant, compared with most recent EPA standards for 1-h average value regulation (with the exception of PM2.5, for which only a 24-h average limit has been established). NO is not currently regulated [36].

Pollutant	Exposure threshold	EPA standard
CO	696 ppb	35,000 ppb
SO_2	7 ppb	75 ppb
NO	37 ppb	–
NO_2	34 ppb	100 ppb
O_3	55 ppb	120 ppb
$PM_{2.5}$	24 μg/m^3	35 μg/m^3 (24-h avg)

quartiles are different from those listed in Table 1, because the database used for rule mining is a particular subset of all the available pollution records (as stated in Section 2.1, it was only possible to include records from 27/02/06 on, and pollution levels show a decreasing trend over the years).

With the sole exception of CO, all the third quartiles are sufficiently close to the range of suspected health effects on the human body. CO pollution has decreased dramatically because of the catalytic converter and the reported threshold of 696 ppb is considerably lower than EPA air quality standards. We are not confident in the precision of CO measurements at this level, given that the monitors were designed to give meaningful results around much higher concentrations [48]. Because we could not guarantee the necessary precision of CO, or its plausible pathway to health effects, those measurements have not been included in further analysis.

Using ARM for an epidemiological study is equivalent to evaluating up to $2^m - 1$ contingency tables, where m is the number of exposures under analysis (-1 accounts for the combination where no variable is selected as exposure, which would result in an indefinite OR). This is the first advantage in using ARM for studies of this type: the algorithm produces a complete analysis of all possible associations in the available database. In our study, we examine 6 pollutants with 0 to 4 days of lag (30 total variables, or single exposures), for a total of 1,073,741,823 possible contingency tables. Naturally, it would be impossible to evaluate by hand the odds ratio (OR) associated with every contingency table. It should be noted that 2^{30} is an upper bound to the number of contingency tables that the algorithm needs to evaluate, but it may not be necessary to evaluate some combinations if their support is too low (this is determined during the mining process).

When a rule includes more than one exposure, non-exposed subjects are considered to be those for whom *none* of the selected pollutants were in the top quartile for the specified day at that nearest monitor. This is an alternative strategy to the one commonly used, where the non-exposed group included every subject who was not exposed to *all* of the risk factors under analysis. We decided to use this approach because it allows always comparing homogenous groups. Furthermore, it preserves independence between the analysis of multiple and single exposures. Note that this decision only influences the computation of the odds ratio, its confidence interval (CI) and its p-value, but not other metrics that are not influenced by the definition of non-exposed population, such as rule support. A schematic representation of how the itemsets are selected from the database and used to compute an odds ratio is visible in Fig. 4.

We set low values for minimum support and confidence (0.5% and 0%): this is necessary because itemsets associated with interesting changes in the odds of having the health outcome under investigation may not appear in the database very frequently. A basic *Apriori* algorithm working with so low values of minimum support and confidence is likely to output an incredibly high number of rules. To preserve interpretability and only focus on rules with interesting odds ratios, we defined some post-processing pruning criteria:

- A rule is pruned if its 95% CI for the OR crosses the value of 1 (with a tolerance of 0.03), implying that it we cannot be confident that it shows an effect.
- A rule is pruned if it is redundant, that is, one or more of the exposures considered in the rule could be removed without significant changes in the associated odds ratio. To assure significant OR difference, we require no overlapping of the 95% CI of the rule with all of its parents, as proposed by [39]. A representation of how the criterion categorizes redundant rules is visible in Fig. 5.

The database has been mined in the R environment using the *Apriori* function from the package *arules,* available on the CRAN website [44,45]. Ad-hoc functions have been designed to compute odds ratio [38], non-redundancy based on [39], and chi-squared statistics [40]. Table 3 presents a summary of the equations associated with each criterion used for mining and post-pruning.

To further improve the quality of the output rules, we took precautions against the chance of overfitting the information carried by the entire database of 14704 entries. To protect us from the possibility of extracting from the database erroneous information produced by factors such as noise and statistical variance, we relied on a training/testing strategy. The database was randomly split into training and testing sets (5000 and 9704 entries, respectively). The training set was used to compute the binning thresholds and to

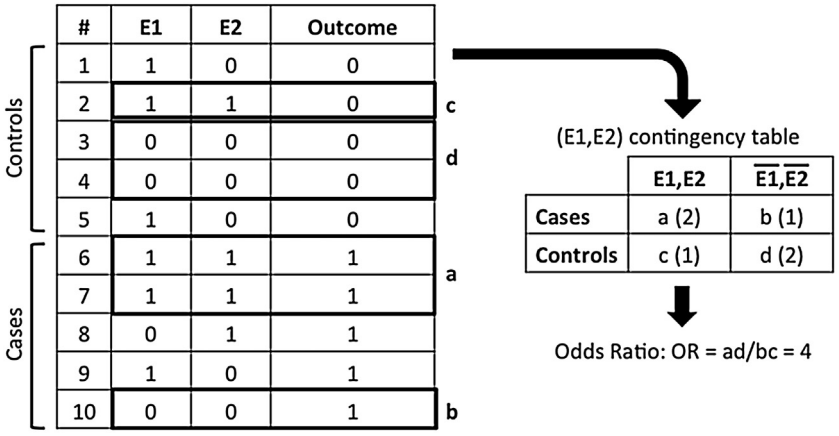

Fig. 4. Selection of cases and controls from the database in case of rules including multiple exposures ({E1, E2} → {Outcome}). Subjects partially exposed are not included in the computation of the odds ratio.

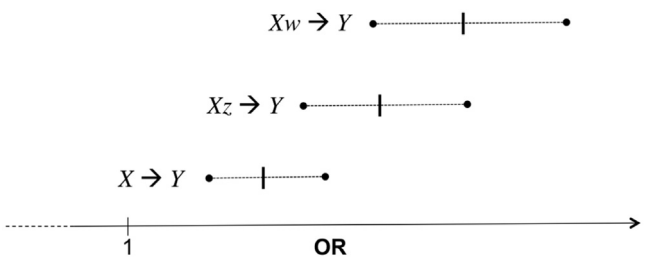

Fig. 5. Schematic representation of odds ratio confidence interval of different rules. Rule X → Y is the parent. By adding other exposures to the parent rule, we obtain the new rules Xz → Y and Xw → Y. Because the confidence interval of Xw → Y does not overlap with the parent rule, the rule is deemed different from the parent and carrying new relevant information, while Xz → Y is not substantially different and should be pruned.

Table 3
Summary of the parameters used for mining and post-pruning rules in the modified association rule mining algorithm, for rules of the form $X \rightarrow Y$.

Parameter	Equation
Support (*supp*)	$P(X \wedge Y) = supp(X \wedge Y)$
Confidence (*conf*)	$P(Y\|X) = supp\left(\frac{X \wedge Y}{X}\right)$
Odds Ratio (OR)	$\frac{supp(X \wedge Y) \cdot supp(\neg X \wedge \neg Y)}{supp(X \wedge \neg Y) \cdot supp(\neg X \wedge Y)}$
Redundancy	$95\%CI_{X \rightarrow Y} \cap 95\%CI_{X \wedge x \rightarrow Y} = \emptyset$ where $x \in I and x \cap X = \emptyset$

produce sub-training sets of different sizes. Particularly, we randomly sampled 10 sets for each size considered: 10, 20, 50, 100, 200, 500, 1000, 2000 and 5000 entries. In total, 90 different sets were generated from the original training set of 5000 entries, using a sampling with replacement strategy. Rules were mined in each of the training sets. The output of each set was then tested on the 9704 entries forming the testing set. If all the criteria listed above were respected also in the testing phase, the rule was approved and included in the final results, which will be discussed in the next section. Approved rules output from different training sets appear only once in the final list.

3. Results

After the testing phase, the algorithm reported 27 rules that fit the criteria of minimum support, significant OR interval and non-redundancy. Using the False Discovery Rate (FDR) controlling procedure proposed by Benjamini and Hochberg [46] on the 27 resulting hypothesis, we verified that the total FDR was less than 13%. Table 4 reports the 10 rules found more often across different training sets and those with the highest support. The tag "day " before a pollutant indicates how many days before the ED visit the value was recorded.

The support of the 27 rules varies from 0.54% to 5.82%, which means that every rule was computed from a 2×2 table including at least 52 and at most 564 exposed cases. The rule with the highest support is {day1 O_3} → {case}, which is also the only rule including only one pollutant exposure. The one with the lowest support is {day0 NO, day1 NO, day1 NO_2, day0 PM, day1 PM} → {case}. Naturally, rules including more exposures tend to have smaller support. The rule length varies from 1 to 5 risk factors in combination, with an average length of 2.81.

The highest OR (1.54, 95% CI 1.14–2.08) is associated with the rule {day0 SO_2, day0 NO, day0 NO_2, day1 PM} → {case}. Some exposures tend to appear across rules more often than others, as visible in Fig. 6.

Finally, we checked for correlation of pollutants within the 27 rules. Some results were expected, because of the correlation between pollutants previously shown in Fig. 3. For example, we found that NO and NO_2 appear together in many rules. Other occurrences are more surprising. For example, we found that every rule including day0 NO always includes also day1 PM (100%), and many rules including day0 NO_2 also include day0 PM (75%).

4. Discussion

The rules produced by this study indicate that a correlation between pediatric asthma and outdoor air pollution exists, when exposure is defined as proximity to chemical levels above the third quartile of their distribution. The rule {day1 O_3} → {case} supports previous results in the literature [4,30,34], thus reinforcing the hypothesis of dangerousness associated with exposure to high levels of ozone. Ozone levels higher than 54.72 ppb have been associated with increased risk of asthma exacerbation in children. The threshold used to identify high ozone levels is below the threshold currently employed by the EPA (120 ppb for the 1-h average exposure, although their primary metric for compliance is 70 ppb for an 8-h average [47]).

The algorithm did not find a significant correlation between exposure to NO_2 alone and odds of asthma exacerbation, which differs from some previous findings [4,30,35,41]. However, NO_2 appears to cause an increase in risk when associated with other pollutants. Analogous behavior was found for NO and PM.

For comparison with the proposed ARM method, we fitted a logistic model using the forward stepwise search implemented in R [49]. The first major difference is that no interaction emerges from the analysis done with the logistic regression, because this model cannot find or recommend automatically combinatory effects between pollutants. The researcher would have to add them to the model if their existence is suspected. The final model included 5 out of the 30 possible exposures (day2 PM, day4 NO, and day0 PM, O_3 and SO_2). Of these 5, only 3 were associated with a p-value less than 0.05 (day2 PM, day4 NO, and day0 PM). The logistic model parameters and associated odds ratio are reported in Table 5. The model identifies a few mild interactions between some exposures and asthma events. However, the high degree of collinearity between the data could have an adverse effect on the coefficient estimates.

Table 4
On the left, set of 10 rules with highest frequency across training sets. On the right, top 10 rules with highest support.

Rule	Exposures	OR	Freq.	Rule	Exposures	OR	Supp
1	day1 O3	1.14 (1.02–1.27)	8	1	day1 O3	1.14 (1.02–1.27)	0.06
2	day0 O3, day0 PM	1.20 (1.02–1.41)	5	2	day0 O3, day4 O3	1.21 (1.03–1.42)	0.02
3	day3 NO, day0 NO_2, day2 NO_2	1.34 (1.05–1.70)	3	3	day0 O3, day0 PM	1.20 (1.02–1.41)	0.02
4	day0 O3, day4 O3	1.21 (1.03–1.73)	3	4	day0 O3, day1 PM	1.22 (1.03–1.44)	0.02
5	day0 NO_2, day2 O3, day0 PM	1.33 (1.00–1.65)	3	5	day0 SO_2, day0 O3	1.23 (1.03–1.46)	0.02
6	day1 NO_2, day2 O3, day0 PM	1.29 (1.03–1.61)	3	6	day3 NO_2, day1 PM	1.30 (1.08–1.57)	0.02
7	day0 SO_2, day0 O3	1.23 (1.03–1.46)	2	7	day1 NO, day4 O3	1.25 (1.02–1.54)	0.01
8	day0 O3, day1 PM	1.22 (1.21–2.18)	2	8	day4 SO_2, day0 PM	1.26 (1.02–1.55)	0.01
9	day3 NO, day4 NO, day1 NO_2	1.34 (1.02–1.75)	2	9	day3 NO, day4 NO, day2 NO_2	1.28 (1.03–1.59)	0.01
10	day1 SO_2, day3 NO_2, day2 O3	1.36 (1.01–1.81)	2	10	day2 O3, day0 PM, day2 PM	1.27 (1.02–1.58)	0.01

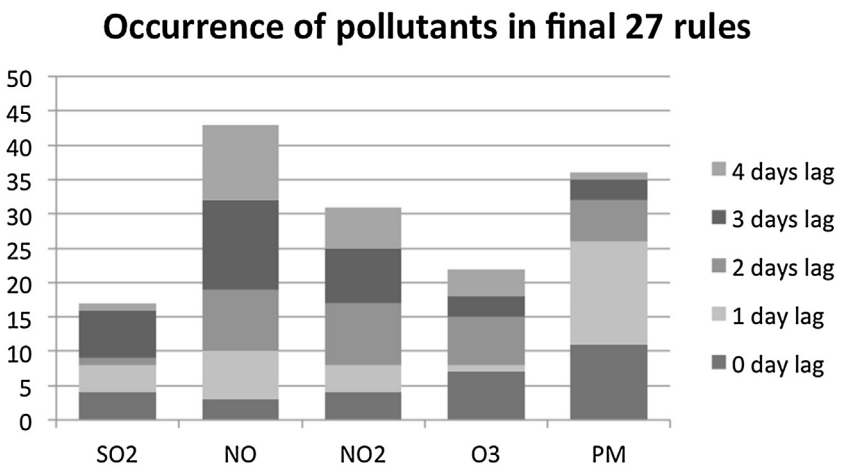

Fig. 6. Frequency with which different pollutants at different day lags appear in the final group of 27 rules.

Table 5
Coefficients of the logistic regression model fitted on the dataset using the forward stepwise search implemented in R. Odds ratio and 95% confidence interval associated with increments of 10 units are also reported. The symbol (*) marks p-values < 0.05.

Exposure	Coefficient	Std. Error	*p*-value	OR (10 units increment)
Day0 O_3	0.002	0.001	0.066	1.02 (1.00–1.04)
Day0 PM	0.006	0.002	0.009*	1.06 (1.01–1.11)
Day0 SO_2	−0.004	0.002	0.084	0.96 (0.92–1.00)
Day2 PM	0.006	0.074	0.003*	1.07 (1.02–1.11)
Day4 NO	0.001	0.002	0.000*	1.01 (1.01–1.02)

The post-pruning criteria designed to filter the output of the basic *Apriori* algorithm are effective in limiting the final number of associations when low values of support and confidence are used. In Fig. 7 it is possible to see how the number of rules found in the training sets is reduced at each step of the algorithm, thanks to the different filters implemented.

The training/testing strategy involving training sets of multiple sizes was necessary to find all the final 27 rules. By using the training set at the fixed size of 5000 entries, only a subset of the final 27 rules would be found. Rules of interest have been found using sets of different sizes, from 5000 to 20 entries. The only sets that did not yield results were the smallest ones, containing only 10 subjects. We believe this strategy to be an acceptable compromise between an inclusive analysis of the possible rules included in the training set and the risk of overfitting.

The rules found by the proposed ARM method point to hidden correlations in the data and can be grounds for further analysis. Like many other tools for data mining, the algorithm is capable of identifying interesting patterns within a multitude of combinations, but the output still requires interpretation by an expert in the domain. The method requires minimal data preprocessing and no human intervention in selecting the combination of exposures to be tested. It is particularly suitable when multiple interactions between risk factors are suspected and need to be investigated. The combinatory rules produced by the ARM method represent possible chemical interactions and it would have been challenging to identify them using interaction terms in a logistic model, particularly when more than 2 exposures are involved. The interactions

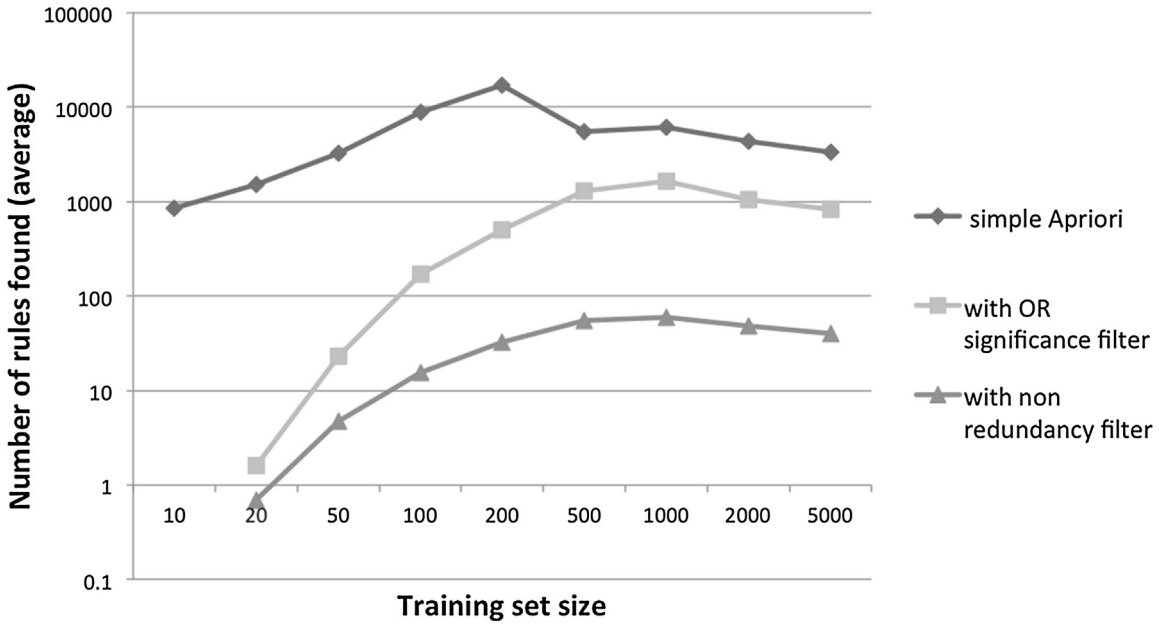

Fig. 7. Average number of rules found at each iteration using training sets of different size. When basic Apriori search is used, thousands of associations are reported. The other lines of the chart represent the effect of adding additional filters (in sequence). When all filters are used, less than 100 rules need to be verified on the testing set.

found using the modified ARM can be further investigated using other methods, but initial identification is much simpler.

5. Conclusions

We proposed an alternative method for analysis of odds ratio based on an *Apriori* ARM algorithm, two criteria to filter non-statistically significant and redundant rules, and a training-testing strategy to reinforce an inclusive analysis while limiting noise. Using this framework, it is possible to find interesting associations in terms of odds ratio even if their support is low. We demonstrated how the algorithm could be used in the analysis of asthma risk correlated to outdoor air quality. The addition of associations between asthma cases and chemicals in the form of *if – then* rules increases interpretability and makes the output of the proposed method accessible for inspection and validation by experts in the domain. This method could easily be extended to other areas of epidemiology where a high number of exposures are investigated and a non-parametric approach may help limit the assumptions necessary to model the data.

Our results confirm the already suspected dangers of exposure to ozone for asthmatic subjects under the age of 18. Additionally, combinatory rules have been found that suggest that the exposure to combinations of chemicals is even more harmful.

The associations presented in this paper constitute fertile ground for future studies, especially regarding the increased risk caused by simultaneous exposure to multiple pollutants. Our goal is to use these results toward the creation of a better alert system, which should not be limited to single pollutant thresholds. In the future, these rules will be tested for capability of predicting days with highly hazardous levels of pollution, in comparison with traditional statistic approaches such as logistic regression and other methods from the field of machine learning (i.e. decision trees). We also plan on developing a method to determine the optimal thresholds to convert the pollution levels in binary variables indicating exposure. In some scenarios, the user already knows the optimal thresholds to use (i.e. they have been previously established in the literature). In other cases, like the study described here, there is large uncertainty on the levels of exposure harmful for human subjects. Improving the algorithm so it can determine the optimal threshold to use based on the data under analysis would be extremely helpful, especially when threshold values are used in legislation to ensure and protect the quality of outdoor air.

Financial interests declaration

No financial interest.

Acknowledgments

We would like to acknowledge the financial support of the Data Analytics in Student Hands Program at the Honors College of the University of Houston. We would also like to acknowledge the reviewers for their valuable input during the publication process.

References

[1] WHO, http://www.who.int/mediacentre/news/releases/2014/air-pollution/en/; [Accessed 27 October 2016].
[2] CDC, Asthma in the US Vital Signs May, 2011.
[3] Bateson TF, Schwartz J. Children's response to air pollutants. J Toxicol Environ Health—A 2008;71(3):238–43.
[4] Gasana J, Dillikar D, Mendy A, Forno E, Ramos Vieira E. Motor vehicle air pollution and asthma in children: a meta-analysis. Environ Res 2012;117:36–45.
[5] Guarnieri M, Balmes JR. Outdoor air pollution and asthma. Lancet 2014;383(9928):1581–92.
[6] Patel MM, Quinn JJ, Jung KH, Hoepner D, Diaz D, Perzanowski M, et al. Traffic density and stationary sources of air pollution associated with wheeze asthma, and immunoglobulin E from birth to age 5 years among New York City children. Environ Res 2011;111(8):1222–9.
[7] Sram RJ, Binkova B, Dostal M, Merkerova-Dostalova M, Libalova H, Milcova A, et al. Health impact of air pollution to children. Int J Hyg Environ Health 2013;216(5):533–40.
[8] Wendt JK, Symanski E, Stock TH, Chan W, Du XL. Association of short-term increases in ambient air pollution and timing of initial asthma diagnosis among medicaid-enrolled children in a metropolitan area. Environ Res 2014;131:50–8.
[9] EPA. http://www.airnow.gov; [Accessed 11 October 2015].
[10] Johns D, Stanek L, Walker K, Benromdhane S, Hubbell B, Ross M, et al. Practical advancement of multipollutant scientific and risk assessment approaches for ambient air pollution. Environ Health Perspect 2012;120:1238–42.
[11] Koop G, Tole L. Measuring the health effects of air pollution: to what extent can we really say that people are dying from bad air? J Environ Econ Manage 2004;47(1):30–54.
[12] Billionnet C, Sherrill D, Annesi-Maesano I. Estimating the health effects of exposure to multi-pollutant mixture. Ann Epidemiol 2012;22(2):126–41.
[13] Mauderly JL, Burnett RT, Castillejos M, Ozkaynak H, Samet JM, Stieb DM, et al. Is the air pollution health research community prepared to support a multipollutant air quality management framework? Inhal Toxicol 2010;22(Suppl. 1):1–19.
[14] Erbas B, Chang JH, Dharmage S, Ong EK, Hyndman R, Newbigin E, et al. Do levels of airborne grass pollen influence asthma hospital admissions? Clin Exp Allergy 2007;37(11):1641–7.
[15] Billionnet C, Gay E, Kirchner S, Leynaert B, Annesi-Maesano I. Quantitative assessments of indoor air pollution and respiratory health in a population-based sample of French dwellings. Environ Res 2011;111(3):425–34.
[16] Mauderly JL, Samet JM. Is there evidence for synergy among air pollutants in causing health effects? Environ Health Perspect 2009;117(1):1–6.
[17] Dominici F, Peng RD, Barr CD, Bell ML. Protecting human health from air pollution: shifting from a single-pollutant to a multipollutant approach. Epidemiology 2010;21(2):187–94.
[18] Greenland S. The effect of misclassification in the presence of covariates. Am J Epidemiol 1980;112(4):564–9.
[19] Thomas DC, Witte JS, Greenland S. Dissecting effects of complex mixtures: who's afraid of informative priors? Epidemiology 2007;18(2):186–90.
[20] Thomas DC, Jerrett M, Kuenzli N, Louis TA, Dominici F, Zeger S, et al. Bayesian model averaging in time-series studies of air pollution and mortality. J Toxicol Environ Health A 2007;70(3–4):311–5.
[21] Chang TS, Gangnon RE, Page CD, Buckingham WR, Tandias A, Cowan KJ, et al. Sparse modeling of spatial environmental variables associated with asthma. J Biomed Inform 2015;53:320–9.
[22] Roberts S, Martin MA. Investigating the mixture of air pollutants associated with adverse health outcomes. Atmos Environ 2006;40(5):984–91.
[23] Nahar J, Tickle KS, Ali AB, Chen YP. Significant cancer prevention factor extraction: an association rule discovery approach. J Med Syst 2011;35(3):353–67.
[24] Ohsaki M, Sato Y, Yokoi H, Yamaguchi T. A rule discovery support sys- tem for sequential medical data in the case study of a chronic hepatitis dataset. Int'l Workshop on Active Mining, IEEE Int'l Conf. on Data Mining 2002:154–65.
[25] Ordonez C, Ezquerra N, Santana CA. Constraining and summarizing association rules in medical data. Knowl Inf Syst 2006;9(3):259–83.
[26] Tai YM, Chiu HW. Comorbidity study of ADHD: applying association rule mining (ARM) to national health insurance database of Taiwan. Int J Med Inf 2009;78(12):75–83.
[27] Crespo J, Maojo V, Martin F. A frequent patterns tree approach for rule generation with categorical septic shock patient data. In: Medical Data Analysis. Berlin Heidelberg: Springer; 2001.
[28] Cartwright N. Nature's Capacities and Their Measurements. Clarendon Press; 1994.
[29] Jaakkola JJ. Case-crossover design in air pollution epidemiology. Eur Respir J Suppl 2003;40:81s–5s.
[30] Raun LH, Ensor KB, Persse D. Using community level strategies to reduce asthma attacks triggered by outdoor air pollution: a case crossover analysis. Environ Health 2014;13(58).
[31] Peters A, Dockery DW, Muller JE, Mittleman MA. Increased particulate air pollution and the triggering of myocardial infarction. Circulation 2001;103(23):2810–5.
[32] Pope CA, Dockery DW. Acute health effects of PM10 pollution on symptomatic and asymptomatic children. Am Rev Respir Dis 1992;145(5):1123–8.
[33] Ito K, Thurston GD, Silverman RA. Characterization of PM2.5 gaseous pollutants, and meteorological interactions in the context of time-series health effects models. J Expo Sci Environ Epidemiol 2007;17(Suppl. 2):45–60.
[34] McConnell R, Berhane K, Gilliland F, London SJ, Islam T, Gauderman WJ, et al. Asthma in exercising children exposed to ozone: a cohort study. Lancet 2002;359(9304):386–91.
[35] Schildcrout JS, Sheppard L, Lumley T, Slaughter JC, Koenig JQ, Shapiro GG. Ambient air pollution and asthma exacerbations in children: an eight-city analysis. Am J Epidemiol 2006;164(6):505–17.
[36] EPA. http://www3.epa.gov/airquality/; [Accessed 10 March 2016].

[37] Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases. In: Proceedings of the ACM SIGMOD International Cconference on Management of Data. 1993. p. 207–16.
[38] Tan P-N, Kumar V, Srivastava J. 'Selecting the right objective measure for association analysis. Inf Syst 2004;29(4):293–313.
[39] Li J, Liu J, Toivonen H, Satou K, Sun Y, Sun B. Discovering statistically non-redundant subgroups. Knowl—Based Syst 2014;67:315–27.
[40] Liu B, Hsu W, Ma Y. Pruning and summarizing the discovered associations. Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining 1999:125–34.
[41] McConnell R, Islam T, Shankardass K, Jerrett M, Lurmann F, Gilliland F, et al. 'Childhood incident asthma and traffic-related air pollution at home and school,'. Environ Health Perspect 2010;118(7):1021–6.
[42] Gass K, Klein M, Chang HH, Flanders WD, Strickland M. Classification and regression trees for epidemiologic research: an air pollution example. Environ Health 2014;13(17).
[43] Ohsaki M, Abe H, Tsumoto S, Yokoi H, Yamaguchi T. Evaluation of rule interestingness measures in medical knowledge discovery in databases. Artif Intell Med 2007;41(3):177–96.
[44] Agrawal R, Srikant R. Fast algorithms for mining association rules in large databases. Proceedings of the 20th International Conference on Very Large Data Bases 1994:487–99.
[45] Hahsler M, Gruen B, Hornik K. arules – A computational environment for mining association rules and frequent item sets. J Stat Softw 2005;14(15):1–25.
[46] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Se B (Methodol) 1995;57(1):289–300.
[47] EPA, https.//www.epa.gov/ozone-pollution/table-historical-ozone-national-ambient-air-quality-standards-naaqs; [Accessed 1 April 2016].
[48] Lefer Barry, Rappenglück Bernhard, Flynn James, Haman Christine. Photochemical and meteorological relationships during the Texas-II radical and aerosol measurement project (TRAMP). Atmos Environ 2010;44(33):4005–13.
[49] R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2016 https://www.R-project.org/.
[50] Kehrl HR, Peden DB, Ball B, Folinsbee LJ, Horstman D. Increased specific airway reactivity of persons with mild allergic asthma after 7.6 h of exposure to 0.16 ppm ozone. J Allergy Clin Immunol 1999;104:1198–204.
[51] Nel AE, Diaz-Sanchez D, Ng D, Hiura T, Saxon A. Enhancement of allergic inflammation by the interaction between diesel exhaust particles and the immune system. J Allergy Clin Immunol 1998;102:539–54.