ISSN 1870-4069

Unsupervised Learning Algorithms are Able to Identify Relevant Patterns in the Pollution Data in Mexico City

Victor Lomas-Barrie[1], Tamara Alcántara[2], Sergio Mota[3], Antonio Neme[4]

[1] Universidad Nacional Autónoma de México,
Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas,
Mexico

[2] Universidad Nacional Autónoma de México,
Dirección General de Cómputo y de Tecnologías de Información y Comunicación,
Mexico

[3] Universidad Nacional Autónoma de México,
Postgraduation Program in Computer Science,
Mexico

[4] Universidad Nacional Autónoma de México,
Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas,
Mexico

{victor.lomas, antonio.neme}@iimas.unam.mx, talcantarac@unam.mx

Abstract. Air pollution is a major problem in almost every large city since the affections to human health are numerous, including damage to tissues and an increase in respiratory-related events. Many cities maintain a monitoring system in order to measure the level of several contaminants such as ozone and carbon monoxide that are particularly harmful to humans. By analyzing the temporal dynamics of those pollutants, authorities may decide to increase mobility constraints or activate contingency plans aiming to reduce the pollution levels. The Air quality authority in Mexico maintain a system of over 20 monitoring stations that serves the Metropolitan Area of Mexico City, covering an area of over $300km^2$, and sampling every hour the air for seven pollutants. Based on public data, we applied unsupervised learning algorithms, in particular anomaly detection algorithms, to unveil relevant patterns in data. An anomaly is an observation that does not resemble, under an unknown metric, the vast majority of instances within a dataset. By applying existing anomaly detection algorithms, we identified several observations of pollutant concentrations that differ from the rest of the observations. The existence of anomalies in the air pollution dataset indicates a qualitative change in the pollution dynamics over time, and the adequate identification of anomalies provide specialists with more information about those changes.

Keywords: Air pollution, monitoring system, contaminants, anomaly detection, pollution dynamics.

1 Introduction

The identification of elements that do not resemble the remaining objects from the same collection is a sign of intelligence [5]. An anomaly is an instance that, under certain unknown metric, do not resemble the rest of the elements in the same dataset. Detecting such anomalies is an open task, and several disciplines have dedicated considerable effort to try to solve it. Artificial intelligence has proposed some ideas aiming to identify anomalies by one path or another. In particular, the techniques defined as unsupervised learning have proven to be particularly relevant.

Unsupervised learning is a field in artificial intelligence aiming to learn from data. It is an open task since it is usually unclear what can be learnt from data, and how to fulfil this task is a prolific field. Several aspects can be learnt from data. A rather common aspect to learn is the separation in clusters. A different aspect to learn is whether an observation is anomalous with respect to the rest of the instances within a dataset [28].

The identification of such instances is an open task, and the techniques and approaches that aim to identify them is known as anomaly detection (AD) [21]. Given a dataset, a rather important question to ask is if all observations, instances, data, or any other synonymous term, were generated by the same mechanism. Whatever the process or structure under study, AD algorithms aim to identify a subset of observations that differ, under an usually unknown metric, to the rest of the elements. Anomaly detection aims to identify, within an unlabelled dataset, those instances or vectors that deviate from a common description found in the vast majority of vectors.

There are, in fact, two instances of anomaly detection. The first one is closely related to classification under unbalanced classes. In this scenario, each observation or vector is labelled as either common, normal, or any other synonym or as anomaly. The former is in general much more abundant than the latter, and thus, there is an unbalance in the classes. This scenario is of higher relevance, since in many applications of data science, it is not known before hand what instances constitute anomalies and which ones are common observations.

We are more interested in the second scenario for anomaly detection. In it, the vectors are not labelled and thus, the algorithm has to infer the class of the vectors, or assign an anomaly degree to them, based on undisclosed properties of the data.

Since the properties of the data that are to be taken into consideration for telling apart anomalies from common vectors are not unique, several alternatives exist. Some vectors can be identified as anomalies under certain assumptions, and not under a different set of premises. The working hypothesis is that observations that significantly differ from the common or usual observations are an indication of the presence of an additional mechanism that threads in the usual mechanism.

In this contribution, we face the problem of detecting anomalies in the air pollution levels in Mexico City from 2011 to May 2022. In this context, an anomaly corresponds to a set of measurements of different pollutants that do not resemble the vast majority of the observations.

Anomalies are relevant since they indicate that, besides the obvious errors from faulty equipment or human error, the observed system is affected by an additional mechanism. The dynamics of the atmosphere, although well understood, are far from being completely characterized.

When, in the context of urban pollution, an anomaly is present, it us suspected that changes in the variables that affect the density of pollutants have occurred. The occurrence of such changes is important in order to apply relevant decisions to diminish the use of vehicles and to reduce the activity in certain industrial sectors. The rest of this contribution goes as follows. In section 2 we briefly describe the problem we aim to understand, that is, the air pollution in Mexico City.

We briefly describe the impact in health of some of the measured pollutants. We also describe the monitoring system that allows the existence of massive data. In section 3 we describe the anomaly detection algorithms that are to be applied to the pollution data. We proceed to describe some of the main results in section 4, and we end by offering some conclusions and discussing what we think are some the most prominent aspects of this contribution in section 5.

2 Air Pollution Monitoring in Mexico City

Air pollution has several consequences in human health. It can increase respiratory problems and damage tissues. [9, 15, 19, 24, 26]. Some of the suspended pollutants with the highest impact in human health are:

1. CO. When carbon monoxide is inhaled, it replaces the oxygen in the blood. CO causes damages in vital organs like the brain and heart.
2. NO. Nitric oxide causes irritation in the nose, throat and lungs. In high concentration, NO reduces the oxygen in blood causing headaches and fatigue. A longer exposure may cause pulmonary edema.
3. NO2. Breathing Nitrogen Dioxide can aggravate respiratory diseases and produce asthma.
4. NOX. As well as NO2, NOX can produce asthma and increase risk of respiratory diseases.
5. O3. Ozone produce throat irritation, chest pain, lung inflammation and asthma.
6. PM10. These small particles can infiltrate the lung tissue and get into bloodstream, provoking heart or lung disease.
7. PM2.5. Breathing PM2.5 can damage lung function causing asthma and heart disease.
8. SO2. Sulfure dioxide can cause inflammation of the throat and the lungs. Also can produce asthma.

The Mexico City Atmospheric Monitoring System (SIMAT) is composed by eight automatic equipment and seven manual equipment; and it is divided in four sub-systems:

1. Automatic Atmospheric Monitoring Network (RAMA).
2. Manual Atmospheric Monitoring Network (REDMA).
3. Meteorology and Solar Radiation Network (REDMET).
4. Atmospheric Deposit Network (REDDA).

In addition, a laboratory for the physicochemical analysis of samples (LAA) and a data processing and dissemination center (CICA) are also supporting SIMAT.

In this contribution, we rely on data generated by the RAMA system. SIMAT started operations in the year 2000, and in 2003 it incorporated the measurement of PM2.5 particles; and it is responsible for the permanent measurement of the main air pollutants in Mexico City and its metropolitan area, with more than 40 air quality monitoring stations.

The monitoring carried out in the metropolitan area of the Valley of Mexico covers the 16 delegations of Mexico City, as well as 12 suburban municipalities of the State of Mexico, which are: Acolman, Atizapán de Zaragoza, Chalco, Coacalco de Berriozábal, Ecatepec of Morelos, Naucalpan de Juárez, Nezahualcóyotl, Ocoyoacac, Tepotzotlán, Texcoco, Tlalnepantla de Baz and Tultitlán [6].

An atmospheric monitoring station consists of a stand that contains various equipment intended to measure the concentrations of one or more air pollutants and certain meteorological parameters. Manual stations, normally, after carrying out the sampling of contaminants, the sample is transferred to a laboratory for analysis. Automatic stations are those that are integrated with automatic and continuous measurement equipment. Each monitoring station is classified by its coverage area, following the U.S.

Environmental Protection Agency criteria (micro, local, neighborhood, city or regional), its location (urban or rural) and the predominant source of air contamination. The emissions inventories are defined by the predominant source of air contamination in the area where the monitoring station is located. The main emission sources in an urban development include generally industrial plants of all kinds, vehicles with diesel engines, internal combustion, power plants, incinerators, and heating equipment. The stations are classified into [12]:

1. Mobile or vehicular traffic, when the predominant source of emission is from roads, parking lots and/or vehicle service shops.
2. Area, when the predominant emission source is from services such as restaurants, dry cleaners, wineries, shopping malls, etc.
3. Biogenic, when the predominant emission source is related to streets unpaved, parks or empty lots.
4. Fixed, when the predominant emission source is from an industrial area.

The prediction of pollutants in several cities have been tackled by several artificial intelligence techniques. In [2], authors applied neural networks to detect changes in the ozone concentration in urban areas in Vilnius. Prediction of ozone in a large metropolitan area was performed via machine learning and statistical methods in [20]. A deep learning approach was applied in [3] with the objective of predict the concentration of several pollutants. In [18], neural networks were applied to detect temporal pollution patterns in a large metropolitan area.

3 Anomaly Detection Algorithms

An outlier is an instance or observation that falls off the range of the expected or usual data [10]. The term outlier is usually associated to observations that were obtained by a faulty process, such as errors in measurement, transmission, or human-caused mistakes.

In general, outliers tend to be discarded from datasets since they tend to affect performance metrics, and are considered errors. The term anomaly has been applied to refer to those instances that are different from the rest but that are not considered as errors. More modern on anomalies suggest that they may be an early indicator that some changes in the forces behind the observed phenomena are changing [16], or that a different mechanism is in play [21].

The identification of anomalies is an unsupervised learning task. What the algorithm has to learn is a function that tells apart expected or usual observations from the anomalies within the data. It is an open task since it is not clear neither what that function should be nor what parameters should take.

Traditional statistical techniques have proven valuable to detect outliers. Statistical approaches have offered a deep understanding of air pollution dynamics based on a detailed analysis of air quality data. For example, fig. 1 shows the result of applying two statistical approaches to identify outliers.

The first method is based on the Z-score, which constitutes a distance between the mean of the sample and the observations, weighted by standard deviations. This method identifies as outliers the observations that fall at the extremes in the range.

The second method is median absolute deviation (MAD). In MAD, if the difference between the observation and the mean of the sample is greater than a certain value, expressed in standard deviations, that observation is declared an outlier.

However, the use of statistical methods presents constraints. First, only observations below or above a certain threshold are identified as outliers, which clearly is insufficient to cope with the complexities of real-world phenomena. Second, when the number of dimensions increases, these techniques fall short of being reliable. In third place, the questions that can be answered based on this approach are limited.

From the same data, relying on unsupervised learning algorithms, a different set of questions can be answered. For example, we can ask What is the typical profile of the observations within a certain period for a large group of pollutant, or How different are two groups of observations in terms of their measured pollutant concentration.

In order to try to answer these last two questions, and some other relevant ones, we relied on four anomaly detection algorithms. These four algorithms are of different nature from each other. The four methods make different assumptions in order to compute a metric that is common in the vast majority of the observations, and that is not present in the anomaly set of instances.

3.1 Local Outlier Factor

Several families of anomaly detection algorithms have been created in more than two decades of active research. In particular, those focused on the analysis of nearest neighbors are of particular relevance, since the relative size of the neighborhood are a free parameter and thus, a wide sensitive analysis can be conducted.

Local Outlier Factor, o *LOF* [4], or LOF, is one of the best-known anomaly detection algorithms that take into account the surroundings of each vector in order to compute an anomaly index. Here, a vector v is characterized in terms of its k nearest neighbors.

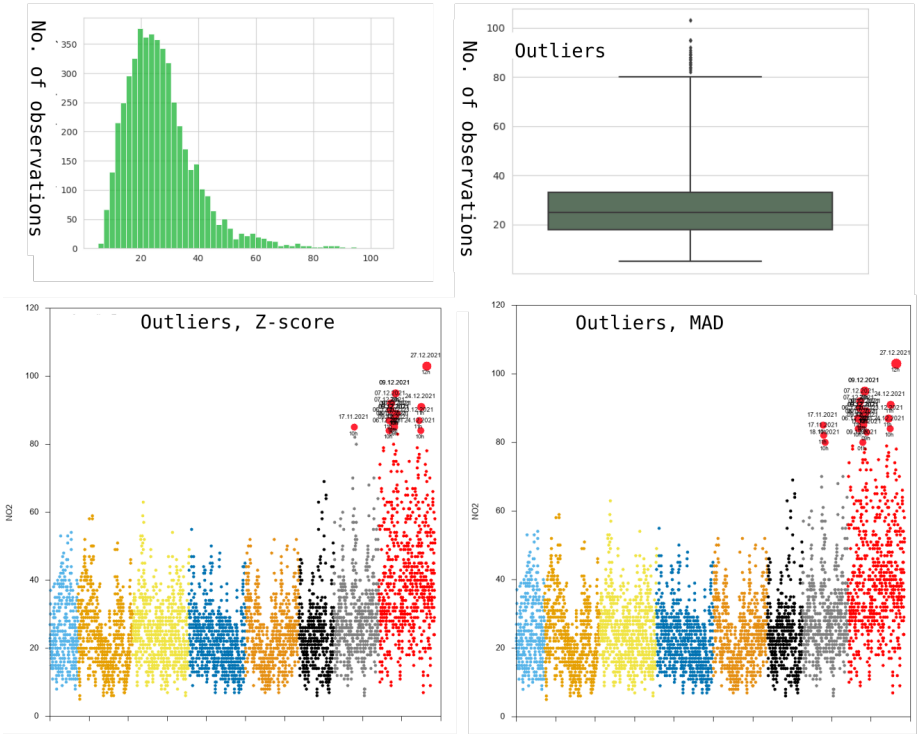

Fig. 1. Identification of outliers based on statistical tools. Top left: Histogram of O_3 concentration at the *Iztapalapa* station for several years. Top right: Boxplot of the same information. Bottom: Time series of the concentration of the pollutant per day. The days that constitute outliers are always in the extremes of the range of values for O_3.

Each of those k neighbors is in turn characterized in terms of its nearest k neighbors. Once the characterizations are concluded, the descriptions obtained from v are compared to those obtained from its k neighbors.

Technically, a vector v is described by a k-distance. k-distance(v) is the distance from v to the $k-$nearest neighbor. The set of neighbors within reach of v based on k-distance(v) is denoted as $N_k(v)$. The reachibility distance from a second vector w and v is given by reachability$-$distance$_k(v,w) = \max($k-distance$(w), d(v,w))$, where d is a distance function. All $k-$neighbours of w will be characterized by the same reachability distance. It should be noted that the reachability distance may be greater than the actual distance. The benefit of this substitution is that it offers more stability for certain distributions.

From the reachability distance, vector v is further described by its local reachability density, defined as:

$$lrd_k(v) = 1/\frac{(\sum_{w \in N_k(v)} \text{reachability} - \text{distance}_k(v,w))}{|N_k(v)|}, \tag{1}$$

$lrd_k(v)$ is a measure of the reachability of vector v from its neighbors. In particular, it is the expected value over all the elements in $N_k(v)$, that is, its $k-$neighbors. From this quantity, the *local outlier factor* or lof is computed:

$$LOF_k(v) = \frac{\sum_{w \in N_k(v)} lrd_k(w)}{|N_k(v)| \times lrd_k(v)}, \tag{2}$$

when $LOF_k(v) > 1$, the local density of v compared to that of its neighbors $N_k(v)$ is lower. On the other hand, if $LOF_k(v) < 1$, it means that vector v presents a higher density of vectors. The former case defines v as an outlier, whereas the latter defines it as an inlier. In this contribution we will refer to both cases as anomalies. The more distant from 1 , the higher the anomaly level.

The control parameter k allows for an increase of the neighborhood and thus. In the extreme case, when k equals the number of elements in the dataset, leads to a global comparison. There is not, however, a formal criteria to identify the correct value of k. As in any other anomaly detection algorithm, if the criteria, in this case defined by the neighborhood size changes, the outcome can also change. This leads to instabilities, but is a problem not tracked in this contribution.

3.2 Isolation Forests

In a high-dimensional feature space, the relative isolation or concentration of a vector offers a path for comparison. Instead of relying on concepts of distance, which are well-known to affect high-dimensional data, the algorithm of isolation forests (*IF*) aims to quantify the anomaly level of each vector based on the effort of isolating it via random decision boundaries [14].

The idea of *IF* is based on exploration of points based on binary trees. In a $N-$dimensional space, an hyperplane of dimension $N - 1$ is needed to create two non-overlapping regions (see fig. 2). For each vector, *IF* randomly selects the dimensions (axis) to create a boundary, and it decides the location of the boundary selecting at random a cut point within the available range.

If the vector of interest is the only within the newly formed region, then the vector is isolated and the number of decision or cuts is linked to the vector. If the vector of interest is not alone in the region, then the algorithm focuses its efforts in that specific region and recursively tries to isolate that vector.

Since *IF* asks binary questions (Is the vector isolated or not?), a binary tree is generated. Graph theory tells us that the number of questions (decisions) needed to identify a node within a binary tree is given by $C(N) = 2 \times H(N-1) - \frac{2(N-1)}{N}$, where N is the number of points in the dataset [22]. Based on $C(N)$, it is possible to compute an anomaly score. If the number of expected trees (decisions) that was needed to isolate vector v is $E(h(v))$, the anomaly level is given by:

$$s(v, N) = 2^{\frac{E(h(v))}{C(N)}}. \tag{3}$$

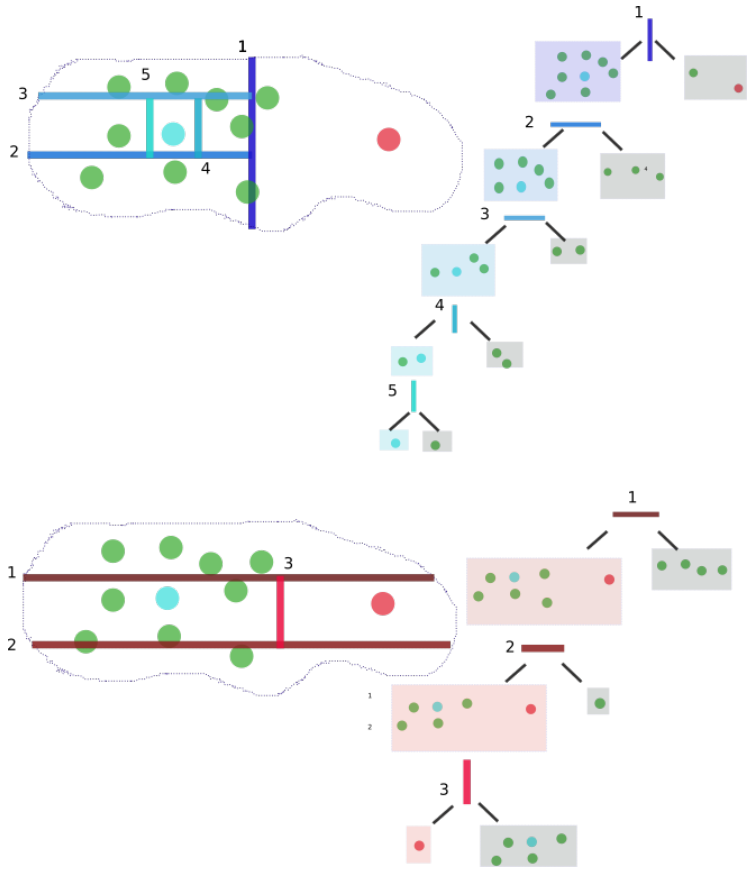

Fig. 2. The difficulty associated to isolating a vector based on random isolation trees is a measure of its anomaly level. The easier a vector is isolated from the rest, the higher its level of anomaly. A vector is isolated when no other vector is contained within the isolated region. The blue vector is harder to isolate than the red one. The process is repeated several times in order to attain a robust measure. The expected number of decisions is taken into account to assign an anomaly level to each vector. Two iterations are shown in the figure.

The closer to 1 is $s(v, N)$, the higher the anomaly level of vector v. The approach followed by *IF* is rather useful since it does not rely on distances, which can be a problem in high-dimensional.

3.3 Support Vector Machines

A support vector machine (*SVM*) can be thought of as an algorithm that maps data into a particularly relevant high-dimensional space. In this mapping space, vectors from two different classes tend to be placed in different regions so that an hyperplane can tell apart the label of the mapped vectors.

SVM constitute an instance of classifiers that map data to a different space so that a linear function can decide the class of the studied vectors [7]. In particular, the hyperplane is placed so that the distance from it to the closest vectors of each class, the support vectors, is maximized. *SVM* map data to the high-dimensional space via a kernel function. This kernel takes as arguments the dot product of the description of each vector. Based on a nice mathematical property derived from Mercer's theorem, the computationally demanding projection to that new space is not explicitly performed.

This *kernel trick* allows the generation of ultra high-dimensional (or infinite-dimensional) spaces in which the decision function can easily classify vectors. The mathematical details of the method, though powerful and highly interesting, are not required its application as anomaly detectors. What is required is the particular method known as one-class support vector machine [25].

In one-class *SVM*, the algorithm is trained with instances of the usual or expected class. The binary function computed by the trained *SVM* will return the same value for all vectors in the training set, which are assumed to belong to the same class, which is the usual or expected class. That value, by convention, is 1. When the trained *SVM* is presented with a vector that does not belong to the same class, that is, which constitutes an anomaly, the decision function returns a -1.

SVM have been successfully applied as anomaly detectors in several contexts. In particular, anomaly detection in time series has proven to be a relevant tool [27]. In [17], SVM are applied to detect anomalies in data obtained by hundreds of sensors in a petroleum facility. The performance of *SVM* is these and many other cases is outstanding.

3.4 Autoencoders

Deep architectures have been successfully applied in several classification tasks [8]. For the anomaly detection problem, in which there is no ground truth about the nature of the observations, an interesting approach comes from the application of autoencoders (*AE*) to detect anomalies in unlabeled data. Trained *AEs* aim to recover the input data at the output layer. The architecture of this type of networks consists of three blocks [11].

The first one is the encoder. In this stage, the usually high-dimensionality of the input space is reduced. This stage constitutes a case of dimensionality reduction, in particular, a non-linear one. The encoder maps input data to a latent space, which constitutes the second block. The latent space is in general of a lower dimension that the input space. It is in this latent space that instances that are anomalies are revealed, since the usual or expected vectors tend to be clustered together, whereas the anomalies tend to form a different cluster [23]. The third block is the decoder, that tries to reconstruct the original or input data from its low-dimensional representation in the latent space.

In an autoencoder, the number of neurons in the input and output layers is the same. In particular, we built an *AE* with two hidden layers, each defined by three neurons. There are several paths to compute anomaly levels in an *AE*. The one we relied on is based on the expected distances in the latent space. By computing a histogram of the expected distances, a decision can be made concerning the cutoff for the discrimination of anomalies and regular or expected vectors. Those vectors with a large distance, compared to the distance shown by the majority, are identified as anomalies.

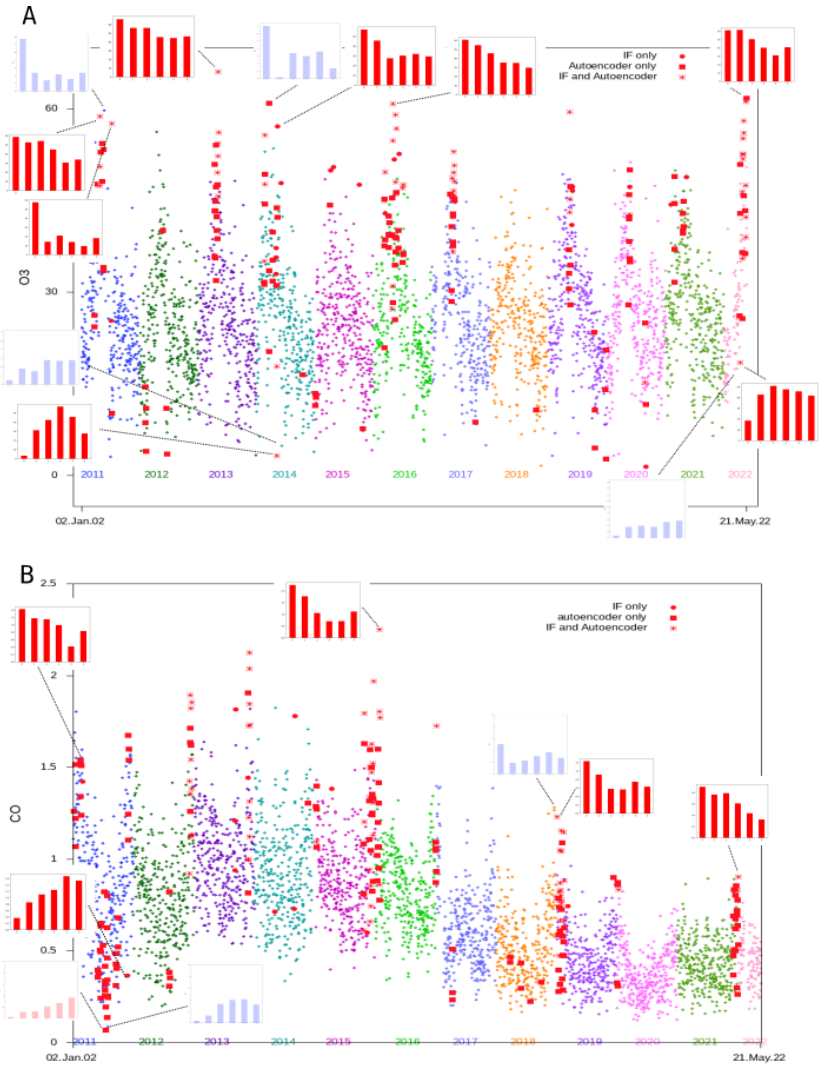

Fig. 3. Anomalies detected in the time series of the average daily concentration of O_3 (A) and CO (B) recorded at the *Tlalnepantla* station. The days detected as anomalies by *IF* only are shown as red filled circles, the days identified as anomalies by the autoencoder only are shown as red squares, and the days identified as anomalies by both methods are shown as stars.

4 Results

From the massive dataset of pollutants, several relevant questions can be answered by relying on machine learning, specifically, in unsupervised learning. The first and most obvious one within this contribution is that of the existence of anomalies. We present in this section some of the results of applying anomaly detection algorithms to the large dataset generated by SIMAT.

Our analysis was conducted focusing on monitoring stations separately in order to disregard the spatial dynamics of air pollution. For each station, we followed two paths of analysis. In the first one, a time series was constructed for each of the monitored pollutants. Instead of detecting changes in consecutive observations, we applied a different approach in order to detect more relevant changes.

For this, a sliding window of size $k = 6$ was applied to the time series in order to embed that point of k coordinates as a point into a $k = 6-$ dimensional space. This embedding is a rather common procedure in anomaly detection of time series [1]. This approach is able to detect relevant patterns in data. Once the time series is embedded as described, the anomaly detection algorithms are applied in the $k-$ dimensional embedding space.

Fig. 3 shows the time series of CO and O_3 for *Tlalnepantla* station from January 1st, 2011 to May 30, 2022. The days detected as anomalies by *IF* or by *AE* are indicated accordingly. It is also shown some of the anomalies as well as some of the expected or usual days. Some days are identified as anomalies by both methods, some others by only one of them, and the majority are not identified as anomalies.

An anomaly in this context is a sequence of six consecutive hours or days, depending on the case, that, in the $k = 6$-dimensional space, does not resemble certain characterizations that are common along the vast majority of the observations. For the IF algorithm, this means, for instance, that the anomalies are rather isolated from the rest of the points in the embedding space since it was easier to isolate it than expected.

For the case of LOF, this means that the sequences detected as anomalies are characterized by neighborhoods that are rather different, in terms of proximity and density, than the neighborhoods of the majority of the vectors. Consecutive observations ,as those observed in 3-A upper left, may differ in nature, that is, one might be an anomaly, and the next one may be an expected observation. Once again, we remind the reader that the algorithm works in the embedding space, not in the time series itself.

In the second approach of anomaly detection, each station is characterized in terms of the concentration of six pollutants: $CO, O_3, NO, NO_2, PM10, SO_2$. That is, for a given station and hour, a point in the six-dimensional space of pollutants is generated. In this approach, anomaly detection algorithms are applied to the points in this six-dimensional space.

Although some sensors suffered occasional problems affecting the records, this does not affect the anomaly detection scheme, since we are not interested in consecutive hours, as is the case for time series analysis. The anomaly detection scheme is applied in the feature space defined by the concentration of the six mentioned pollutants.

Fig. 4 shows the 65,798 recorded hours from January, 2nd 2011 to 5th May 2022 at the *Tlalnepantla* monitoring station. Each hour is linked to a point in the six-dimensional space of pollutant concentration.

It is in that space that the four algorithms are invoked. In 4-A, it is shown, in the y-axis, the ratio of the average pollutant concentration at the corresponding hour and the distance, in the six-dimensional space, from that observation to the next available one. It is also indicated whether the observations at a certain hour were detected as anomalies by one of the four anomaly detection algorithms.

The numbers on top of A indicate the number of instances, per year, that were detected as anomalies by the specified algorithm (label at the right end). The number of usual or expected (non-anomalies) observations is also indicated. The number of observations that were detected as anomalies by one, two, or three anomaly detection algorithms is also displayed.

From fig. 4 it is already available some information that could not be obtained by traditional statistical approaches. As a preamble, 5,822 out of the 65,798 hours were identified as anomalies by at least one algorithm. From the algorithms included in this contribution, SVM is the more stringent one. Only 245 observations were detected as anomalies, and in several years, no anomalies were identified by this method.

Interestingly, though, is that the years 2021 and 2022 are the ones with the highest number of identified anomalies by SVM. This may indicate a change in the dynamics behind the sources of pollution. Indeed, this time frame corresponds to mobility constraints imposed by the government in order to reduce social contact as a policy to reduce *SARS-COV2* contagions. The remaining three methods do not present this change in the number of detected anomalies.

As was already stated in the Introduction, different anomaly detection algorithms make different assumptions in order to identify peculiar or dissonant observations. Fig. 4-B shows a comparison, based on visualization of different categories (a kind of Venn-diagram for several sets) [13], of the intersection among the four anomaly detection algorithms and the non-anomalies in the data. In blue, it is shown that the majority of observations were not detected as anomalies. 8.8% of the observations define the set of anomalies, identified by at least one of the methods.

The autoencoder (AE) is the most sensitive one, as observed by the high number of anomalies detected by it (4,707). *LOF* is the second most sensitive algorithm, as it records 1,181 anomalies. However, the observations detected as anomalies by these two algorithms is rather low, 68 exclusively detected by those two, plus 2 more anomalies detected also by *SVM*. The methods with the highest overlap were *IF* and the autoencoder, with 277 common observations.

In fig. 4-C, it can be seen the expected (typical) observation of the six pollutants detected as usual, or anomalies accordingly to one of the four described algorithms. It is clear the difference between the usual observations (blue) and the anomalies (red). In 4-D, it is shown the distribution of the number of anomalies in the specified hour of the day. The hour with the highest number of anomalies is at 8:00. The reasons of this are still under deeply research, but there is evidence that at this time, the changes in temperature and mobility are rather important factors.

Interestingly, no observation were detected as an anomaly by the four methods. Only 64 four observations were detected as anomalies by at three methods, and the only of such anomalies for 2022 is shown in fig. 4-E. This observation corresponds to March, 22nd. at 9:00.

The points that are anomalous indicate that at a certain hour, the concentration of the six pollutants was rather different to the concentration observed in the majority of points in the six-dimensional space.

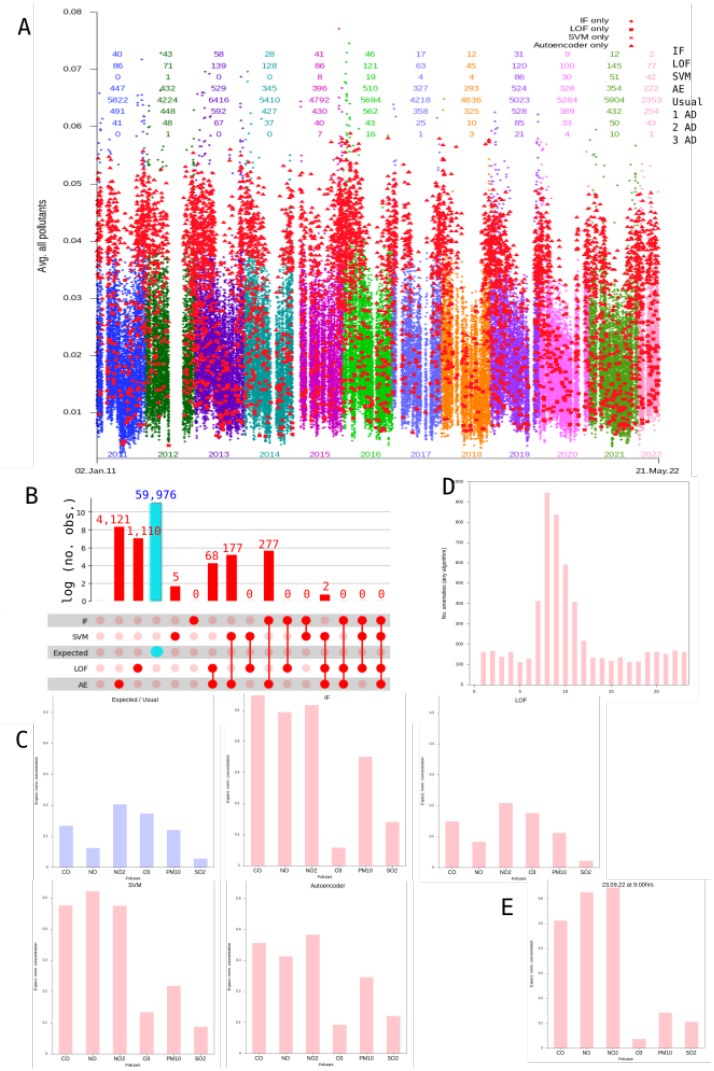

Fig. 4. Anomalies focused on hourly observations at *Tlalnepantla* station, from 02.01.2011 to 21.05.2022. A. Hourly measure of six pollutants. In the y-axis, it is shown the ratio of the average concentration of pollutants at the specified hour and the difference to the set of measurements in the next available hour. It is indicated whether a particular set of observations was identified as an anomaly by any of the four algorithms. It is also shown the number of observations detected by one, two or three algorithms (1AD, 2AD, 3AD). B. UpSet visualization of the intersections among the four anomaly detection algorithms. C. The distribution of anomalies along the 24 hours. The hour with more anomalies in this station was at 8:00. D. The expected concentration (normalized) of the pollutants for each of five cases: usual (expected) observations, detected as anomalies by: *IF, LOF, SVM, AE*. E. The only observation detected as anomaly by three methods during 2022.

5 Discussion and Conclusions

The identification of certain observations that do not resemble the rest of the observations in a dataset is a peculiar, and rather interesting case, of pattern recognition in particular, and of artificial intelligence in general. Although some researchers consider anomaly detection a special case of classification, we stick to a different perspective, in which both tasks are inherently different.

Classification relies on the existence of an assigned label or class to each vector, whereas in anomaly detection, the algorithm has to infer the label for each observation. The label may be either usual or anomalous observation. The second approach for anomaly detection is more complex since the metric to compare observations is unknown and has to be learnt from the existing data. Besides, the criteria to decide whether an observation constitutes an anomaly or not is not unique.

In this contribution, we applied existing anomaly detection algorithms to air pollution data in the Metropolitan Area of Mexico City. The main goal behind our work was to identify non-trivial observations, that is, groups of data from different pollutant sensors, that are rather different to the majority of observations. Those anomalous observations denote special atmospheric circumstances that may indicate transitory changes in the mechanisms and variables that affect the dynamics such as wind, temperature, changes in mobility, among others.

For the case of one station, that of *Tlalnepantla*, the anomaly detection algorithms identified some relevant patterns. For instance, the average concentration of six pollutants of the anomalies detected by the four methods present a wide range. Since in anomaly detection there is no ground truth, it is relevant to capture several possible profiles for some of the possible anomalies.

In particular, we applied Isolation Forests, Local Outlier Factor, support vector machines and autoencoders to the data collected by the Air Quality Authority of The Metropolitan Area of Mexico City. The existing data includes hourly observations of over thirty stations and covering seven different pollutants. We focused our efforts in a subset of the dataset in order to communicate the relevance of applying anomaly detection algorithms to air quality data. To our knowledge, this has not previously been conducted.

Artificial Intelligence tools provide insight into complex phenomena by detecting patterns that otherwise could not be elucidated. In this sense, this contribution describes the use of an instance of unsupervised learning to a complex phenomena, that of air pollution in large metropolitan areas.

Our main conclusion is that the nature of patterns that can be detected by the use of relevant tools is of a subtle nature, and this patterns provide more information to better understand, in this case, the dynamics of air pollution in Mexico City. Several paths are open for future work. For instance, several other anomaly detection algorithms can be applied to the same pollution dataset.

In a more insightful perspective, atmospheric attributes such as pressure, temperature, and humidity may be included in the analysis to gain a broader perspective of the dynamics of pollutants.

Acknowledgments. A.N thanks PAPIIT for the partial support of this research, with grant number IA103921.

References

1. Aguayo, L., Barreto, G.A.: Novelty Detection in Time Series Using Self-Organizing Neural Networks: A Comprehensive Evaluation. Neural Processing Letters, vol. 47, no. 2, pp. 717–744 (2018). DOI: 10.1007/s11063-017-9679-2.
2. Bekesiene, S., Meidute-Kavaliauskiene, I., Vasiliauskiene, V.: Accurate Prediction of Concentration Changes in Ozone as an Air Pollutant by Multiple Linear Regression and Artificial Neural Networks. Mathematics, vol. 9, no. 356, pp. 1–21 (2021). DOI: 10.3390/math9040356.
3. Bekkar, A., Hssina, B., Douzi, S., Douzi, K.: Air-Pollution Prediction in Smart City, Deep Learning Approach. Journal of Big Data, vol. 8, no. 161, pp. 1–21 (2021). DOI: 10.1186/s40537-021-00548-1.
4. Breunig, M., Kriegel, H.P., Ng, R.T., Sander, J.: LOF: Identifying Density-Based Local Outliers. In: Proceedings of the ACM SIGMOD International Conference on Management of Data. vol. 29, pp. 93–104 (2000). DOI: 10.1145/335191.335388.
5. Buzsaki, G.: The Brain from Inside out. Oxford University Press, (2019). DOI: 10.1093/oso/9780190905385.001.0001.
6. Comision Ambiental de la Megalopolis: ¿Como se monitorea la calidad del aire en la ZMVM? Gobierno de México (2018)
7. Cristianini, N., Ricci, E.: Support Vector Machines. Encyclopedia of Algorithms, pp. 928–932 (2008). DOI: 10.1007/978-0-387-30162-4_415.
8. Emmert-Streib, F., Yang, Z., Feng, H., Tripathi, S., Dehmer, M.: An Introductory Review of Deep Learning for Prediction Models with Big Data. Frontiers in Artificial Intelligence, vol. 3, pp. 1–23 (2020). DOI: 10.3389/frai.2020.00004.
9. Harper, A., Croft-Baker, J.: Carbon Monoxide Poisoning: Undetected by Both Patients and Their Doctors. Age and Ageing, vol. 33, no. 2, pp. 105–109 (2004). DOI: 10.1093/ageing/afh038.
10. Hawkins, D.: Identification of Outliers. Monographs on Statistics and Applied Probability, (1980). DOI: 10.1007/978-94-015-3994-4.
11. Hinton, G.E., Salakhutdinov, R.R.: Reducing the Dimensionality of Data with Neural Networks. Science, vol. 313, no. 5786, pp. 504–507 (2006). DOI: 10.1126/science.1127647.
12. Instituto Nacional de Ecología y Cambio Climático: Primer catálogo de estaciones de monitoreo atmosférico en México, Instituto Nacional de Ecología (2013)
13. Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., Pfister, H.: UpSet: Visualization of Intersecting Sets. IEEE Transactions on Visualization and Computer Graphics, vol. 20, no. 12, pp. 1983–1992 (2014). DOI: 10.1109/TVCG.2014.2346248.
14. Liu, F.T., Ting, K.M., Zhou, Z.H.: Isolation Forests. In: Eighth IEEE International Conference on Data Mining, pp. 413–422 (2008). DOI: 10.1109/ICDM.2008.17.
15. Luttrell, W.E.: Nitrogen Dioxide. Journal of Chemical Health and Safety, vol. 21, no. 2, pp. 28–30 (2014). DOI: 10.1016/j.jchas.2014.01.008.
16. Markou, M., Singh, S.: Novelty Detection: A Review-Part 1, Statistical Approaches. Signal Processing, vol. 83, no. 12, pp. 2481–2497 (2003). DOI: 10.1016/j.sigpro.2003.07.018.
17. Martí, L., Sanchez-Pi, N., Molina, J.M., Bicharra-García, A.C.: Anomaly Detection Based on Sensor Data in Petroleum Industry Applications. Sensors, vol. 15, no. 2, pp. 2774–2797 (2015). DOI: 10.3390/s150202774.

18. Neme, A., Hernández, L.: Visualizing Patterns in the Air Quality in Mexico City with Self-Organizing Maps. In: Advances in Self-Organizing Maps - 8th International Workshop, vol. 6731, pp. 318–327 (2011). DOI: 10.1007/978-3-642-21566-7_32.
19. NPS: National Park Services Stats (2014)
20. Oufdou, H., Bellanger, L., Bergam, A., Khomsi, K.: Forecasting Daily of Surface Ozone Concentration in the Grand Casablanca Region Using Parametric and Nonparametric Statistical Models. Atmosphere, vol. 12, no. 6, pp. 1–19 (2021). DOI: 10.3390/atmos12060666.
21. Pimentel, M., Clifton, D., Clifton, L., Tarassenko, L.: A Review on Novelty Detection. Signal Processing, vol. 99, pp. 215–249 (2014). DOI: 10.1016/j.sigpro.2013.12.026.
22. Preiss, B.: Data Structures and Algorithms with Object-Oriented Design Patterns in C++. vol. 1 (1999)
23. Sakurada, M., Yairi, T.: Anomaly Detection Using Autoencoders with Nonlinear Dimensionality Reduction. In: Proceedings of the MLSDA 2nd Workshop on Machine Learning for Sensory Data Analysis, pp. 4–11 (2014). DOI: 10.1145/2689746.2689747.
24. Salladay, S.: Right to Know. Unexpcted Diagnosis . Nursing, vol. 27, no. 11, pp. 22–24 (1997). DOI: 10.1097/00152193-199711000-00013.
25. Schölkopf, B., Williamson, R., Smola, A., Shawe-Taylor, J., Platt, J.: Support Vector Method for Novelty Detection. Advances in Neural Information Processing Systems, pp. 582–588 (2000)
26. United States Environmental Protection Agency: Particulate Matter (PM) Basics (2021)
27. Yokkampon, U., Chumkamon, S., Mowshowitz, A., Fujisawa, R., Hayashi, E.: Anomaly Detection Using Support Vector Machines for Time Series Data. Journal of Robotics, Networking and Artificial Life, vol. 8, no. 1, pp. 41–46 (2021). DOI: 10.2991/jrnal.k.210521.010.
28. Zimek, A., Schubert, E., Kriegel, P.: A Survey on Unsupervised Outlier Detection in High-Dimensional Numerical Data. Statistical Analysis and Data Mining: The ASA Data Science Journal, vol. 5, no. 5, pp. 363–387 (2012). DOI: 10.1002/sam.11161.