Annales Mathematicae et Informaticae
55 (2022) pp. 12–27
DOI: https://doi.org/10.33039/ami.2022.05.001
URL: https://ami.uni-eszterhazy.hu

Weather forecasting using DBSCAN clustering algorithm

Aida Chefrour

Computer Science Department, Mohamed Cherif Messaadia University,
Souk Ahras,41000, Algeria
LISCO Laboratory, Computer Science Department, Badji Mokhtar-Annaba University,
P.O. Box 12, Annaba, 23000, Algeria
aida.chefrour@univ-soukahras.dz

Abstract. The main objective of this study is the clustering of meteorological parameters and forecasting weather in the region of Annaba (Algeria) using clustering techniques. The proposed two-stage clustering approach is based on the first stage, on the proposition of ANN-DBSCAN, a combination of the DBSCAN algorithm and an Artificial Neural Network (ANN) for grouping the clusters. Internal indices of validation were used to compare and verify the correctness and efficiency of the results. Our experiments identified five groups, each of which was associated with the area's usual weather parameters. Our proposed incremental DBSCAN is employed in the second stage to determine the data pattern that can predict the future atmosphere. The natural molecules of the measured pollutants (nitrogen dioxide (NO2), ozone (O3), carbon dioxide (CO2), and sulfur dioxide (SO2)) are directly dependent on weather forecasting. The focus of this research is on a section of the Samasafia database. The proposed algorithm is used to determine the weather trend in that database. Advanced numerical analysis was applied to a few prediction tasks.

Keywords: Clustering, Forecasting weather, DBSCAN, Artificial Neural Network, SAMASAFIA

1. Introduction

Concern about decreasing air quality and its local and global consequences has increased significantly in recent years [10]. Rapid urbanization, population growth, and industrialization have resulted in alarming levels of air pollution. Scientific planning of analysis methodologies and pollution control are essential to prevent

Submitted: September 21, 2021
Accepted: May 9, 2022
Published online: May 11, 2022

continued declines in air quality. Within this framework, it is required to analyze and specify all pollution sources and their contributions to air quality; research the various elements that generate pollution; and develop instruments to minimize pollution by implementing control measures and alternative practices [18].

Algeria is a beautiful country with a rich and diversified geography. It does, however, have its own set of environmental challenges, as does every country in the world. This is particularly evident in highly industrialized and rapidly rising urban regions such as Annaba [8].

The city of Annaba is one of the most polluted cities in Algeria because of the existence of big industrial complexes such as the El Hadjar steel complex (Arcelor Mittal) and the fertilizer complex (Fertile). In addition, it is known for its dense roundabout traffic and overcrowding.

The main pollutants monitored in the Annaba air are chemicals such as ozone (O3), nitrogen oxide (Nox), carbon monoxide (Co2), and sulfur dioxide (SO2).

In our work, we have used the SAMASAFIA database, it will be described in section 2. Several research has been devoted to identifying the links between air pollution and meteorological variables exist in the literature. In the following, we outline the best known and most recent ones:

Using clustering algorithms, Khedairia and al [18] define meteorological conditions in the Annaba (Algeria) region. The Self-Organizing Maps (SOMs) and the well-known K-means clustering method are used in the suggested two-stage clustering strategy. To compare and validate the accuracy of the results, quantitative (using two kinds of validity indices) and qualitative criteria were used. Five classes emerged from the many experiments, all of which were related to typical weather circumstances in the area. The meteorological clusters obtained are then utilized to better understand the relationship between meteorology and air quality in the presence of seven pollutants. For modeling air contaminants and simulating their reactivity to meteorological parameters of interest, they used Artificial Neural Networks (ANNs), and more specifically, Multi-Layered Perceptron (MLP). This behavior is also examined using the correlation coefficient, where the results are displayed for comparison, and numerous relationships and conclusions are drawn.

Ghazi et al [6] report the construction of air pollution concentration prediction models for five major pollutants (O3, PM10, SO2, NOx, COx) utilizing two neurocomputing paradigms: Radial Basis Function and Elman Networks. As a result, each Artificial Neural Network (ANN) forecasts the concentrations of the five contaminants. These models were created to provide a 12-hour forecast for the Annaba region in northeast Algeria (north of Africa). The models are designed to predict air pollutant concentration at $t + 12$ hours after receiving measurements of air pollutant concentration and meteorological parameters (wind speed, temperature, and humidity) at time t. The performance of both ANN models is fully compared and assessed once anticipated pollutant concentrations are attained and the validity of each ANN model is verified. In light of the acquired results, the usage of one ANN network over another is justified.

Alioua et al [1], present the characterization of air pollution in the region of Annaba. The survey has been conducted using different complementary approaches. On one hand, results were recorded by the monitors operating in the air quality and control network in the region of Annaba (called Sama Safia), and on the other hand, results were provided by a bio-indicator, a lichen species called Xanthoria parietina. A relevant sampling strategy, space and time follow-up of measurements of certain physiological parameters (chlorophyll, proline, breathing), and the proportioning of NO2 have permitted us to characterize the impact of pollution resulting, on the one hand, from intense road traffic and, on the other hand, from the proximity of an iron and steel complex and a phosphate fertilizer complex. The results from the two monitoring techniques used, on one hand, the physico-chemical sensors and, on the other hand, bioindication, have shown a significant correlation not only between the analyzed pollutant (NO2) and the physiological parameters measured (chlorophyll, proline, respiration), but also between the bioindicator and the physical-chemical sensors. This work has allowed a better characterization of air pollution in this region.

The remainder of the paper is organized as follows: the study region and user data are introduced in the following section. The essential concepts of the DBSCAN clustering method, as well as how the suggested clustering strategy ANN-DBSCAN and validity indices are utilized to identify and compare clustering results, are discussed in Sections 3 and 4. In Section 5, we look at the correlations between air pollution and meteorological characteristics, as well as how we applied our novel approach to a portion of the SAMASAFIA database and analyzed the results. Section 6 concludes with a conclusion and recommendations for future research.

2. Studies area and dataset

2.1. Studied area

The city of Annaba is located in the east of Algeria (600 km from Algiers) between the latitudes of 36°30' Nord and 37°30' North and the longitudes of 07°20' East and 08°40' East, with 12 communes with a total area of 1411.98 km^2. It is situated on a wide plain separated on the northwest by a mountain range that gradually decreases in height towards the southwest, and on the east by the Mediterranean Sea.

These characteristics allow pollutants to accumulate and, as a result, their concentration to develop. I addition, the movement of polluted air is aided by sea and land breezes. Pollutants are transported out to sea by the land wind and then returned to the city by westerly winds along the Seraidi mountain. In the shape of a circle, the clouds stare down on the city. Because the pollutants are deposited slowly by gravity, pollution affects all three receivers (sea, land, and air). Depending on Annaba's industrial operations, contaminant air pollutants are spread variably. The industry in Annaba is both a source of growth and a source of environmental deterioration, with the majority of industrial sites (complexes) located near the city, such as the Asmidal phosphate and nitrogen fertilizer complex and the El Hadjar metal steel complex. These industrial operations are the primary source of particulate matter and sulfur oxides, whereas the transportation sector is the primary source of carbon monoxide, nitrogen, and lead emissions. Air pollution has risen due to an increase in the number of vehicles (a 5% annual increase in Algeria) and a lack of emission controls. In the open air, waste incineration (domestic, industrial, hospital, toxic) is also a source of pollution.

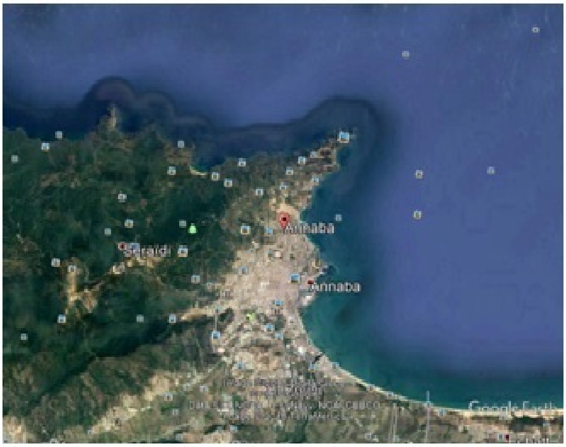

Figure 1. Geographic of Annaba (Google Earth).

2.2. Dataset and pre-processing

The dataset used in this study was collected from the SAMASAFIA network center (`www.samasafia.dz/journaux`) over a 24-hour period from 2003 to 2004. Nitric oxide (NO), carbon monoxide (CO), ozone (O3), particulate matter (PM10), nitrogen oxides (NOx), nitrogen dioxide (NO2), and sulfur dioxide (SO2) are among the pollutants that are regularly monitored in the air (SO2).The dataset includes Wind Speed, Temperature, and Relative Humidity.

Outliers are introduced by faulty measuring equipment operation or erroneous data collection and processing, and their detection is dependent on a number of criteria, such as the median value, the mode, and the mean,... Outliers are carefully scrutinized since they can skew the prediction model's calibration. Measurement instrument failures result in missing data. Because of the experimental nature of the measurement stations, this is typically caused by power outages and various analyzer problems (Samasafia, 2004). We must follow the same process as in [18], in which we estimate the model's parameters by analyzing observed data without accounting for missing data. Nevertheless, the results may be incorrect because considerable information is lost [7]. The missing data percentages for each year are: 2003 (08.31%), and 2004 (23.67%).

3. The proposed ANN-DBSCAN clustering algorithm

We will recall the principle of classic DBSCAN and Artificial Neural Network:

The DBSCAN algorithm was first proposed by [14], and it is based on the density-based cluster concept. The density of points can be used to identify clusters. Clusters are visible in regions with a high density of points, whereas clusters of noise or outliers are seen in regions with a low density of points. This technique is well suited to deal with huge datasets and noise, as well as identifying clusters of various sizes and forms.

The DBSCAN algorithm's main idea is that for each cluster point, the neighborhood of a specified radius must contain at least a certain number of points, i.e., the density in the neighborhood must surpass a certain threshold. Three input parameters (Eps and MinPts) are required for this algorithm [16]:

- k, the neighbour list size;
- Eps, the radius that delimitates the neighbourhood area of a point (Eps-neighbourhood);
- MinPts, the minimum number of points that must exist in the Eps-neighbourhood.

The clustering technique uses density relations between points (directly density-reachable, density-reachable, density-connected) to construct clusters after classifying the points in the dataset as core points, border points, and noise points.

- **Core Object:** object with at least MinPts objects within a radius 'Eps-neighborhood';
- **Border Object:** object that is on the border of a cluster of NEps(p): q belongs to $D \mid \text{dist}(p, q) \leq \text{Eps}$;
- **Directly Density-Reachable:** a point p is directly density-reachable from a point q w.r.t Eps, MinPts if p belongs to NEps(q) and $|\text{NEps}(q)| \geq \text{MinPts}$;
- **Density-Reachable:** a point p is density-reachable from a point q w.r.t Eps, MinPts if there is a chain of points $p_1, \ldots, p_n$, $p_1 = q$, $p_n = p$ such that p_{i+1} is directly density-reachable from p_i;
- **Density-Connected:** a point p is density-connected to a point q w.r.t Eps, MinPts if there is a point o such that both, p and q are density-reachable from o w.r.t Eps and MinPts.

The algorithm of DBSCAN is as follows [14]:

- Choose a point p at random;
- A cluster is formed if p is a core point;

- If p is a border point, DBSCAN visits the database's next point since no points are density-reachable from p;
- Repeat the process until you've processed all of the points.

Artificial Neural Networks (ANNs) have grown in popularity as a useful approach for modelling environmental systems in recent years. They've already been used to model algal development and transportation in rivers, anticipate salinity and ozone levels to predict air pollution and the functional aspects of ecosystems, and simulate the export of nutrients from river basins [9].

Now, let us recall the principle of ANN, which is a mathematical model that simulates the structure and functionalities of biological neural networks. The artificial neuron builds the basic block of every artificial neural network, that is, a simple mathematical model (function).

It is a set of neurons arranged and placed with each other under a structure of synaptic connections.

An ANN can be divided into three layers:

- **Input layer:** this layer is in charge of receiving information (data) from the outside environment. These inputs are normally normalized within the activation function's limit values;
- **Hidden, intermediate layers:** these layers are composed of neurons that are responsible for extracting data associated with the processor system being analyzed;
- **The output layer:** this layer, like the previous levels, is made up of neurons and is in charge of producing and displaying the final network outputs, which are the consequence of the processing done by the neurons in the previous layers.

3.1. Proposed algorithm

We have used neural networks in our algorithm because they encourage the principle of evolution and make it easy to estimate problems that are real and complex. The following are the steps of our proposed clustering algorithm, ANN-DBSCAN:

1. Apply the DBSCAN algorithm;
2. Create neural networks:

   For every cluster:

   - The centroid is the neuron of the first layer;
   - Other points from neurons constitute the second layer;
   - Every noise point forms a neural network of a single layer and a single neuron.

3. When a new data X arrives:

    - X is considered a neuron.

    For $n \leftarrow 1$ to m do (m is the number of neural networks):

    - Calculate the Euclidian distance d_m between this neuron and the neurons of n neural networks;
    - The neuron which has the minimal distance d_m between this neuron and the new neuron is called a winning neuron and the opposite is a losing neuron.

    We compared d_m with a threshold T:

    - if $d_m \leq T$ then the new neuron will be inserted into the cluster which has $d_m \leq T$ and the winning neuron will form the new centroid of this new cluster; otherwise, X will be considered as a noise point.

4. The most lost neurons will be deleted;

5. Apply the same algorithm if the new data is a cluster.

Remark: The threshold T is the minimal distance between all the centroids and the noise points.

4. Results and discussion

4.1. The cluster validity measures

Despite the fact that clustering methods strive to optimize a criterion, finding the global optimum is not guaranteed [19]. It is necessary to evaluate the quality and validity of results because there is no knowledge of priority in the process of clustering, there are no predefined classes, and there are no examples of what types of acceptable associations should be legitimate among the data. Two different approaches to validity indices are used for comparing the results: External criteria compare the resulting clustering to a ground-based truth available externally, either by previous research or by subjective knowledge from field experts [18]. Internal criteria use quantities and features available within the dataset. In this study, two internal validity indices were applied to determine the appropriate number of clusters within the data set.

4.1.1. Silhouette index (SC)

In contrast to the above-mentioned indices, the Silhouette value of a data object represents the degree of confidence in its clustering assignment [15]:

$$S_i = (b_i - a_i)/\max\{a_i, b_i\}'$$

where a_i is the average distance between point I and other points in the same cluster, and b_i is the average distance between itself and the "nearest" neighbouring cluster:

$$a_i = \frac{1}{|C_i|} \sum_{j \in C_i} d(i,j)$$
$$b_i = \min_{C_k, k \neq i} \sum_{j \in C_k} d(i,j)/|C_k| \tag{4.1}$$

Let us set as the Silhouette index of a clustering scheme, the average Silhouette score of its objects:

$$\mathrm{SC} = \frac{1}{N} \sum_{i=1}^{N} S_i.$$

The values of the index lie in $[-1; 1]$ with unity denoting a perfect assignment. Thus, in practice, values around 0.5 and higher are considered acceptable.

4.1.2. Davies–Bouldin index

The ratio of within-cluster scatter to between-cluster separation (see (4.1)) determines this index:

$$\mathrm{DBI} = \frac{1}{n} \sum_{i=1}^{n} \max_{i \neq j}((S_n(Q_i) + S_n(Q_j))/S(Q_i, Q_j)$$

where n is the number of clusters, S_n is the average distance between cluster centers, and $S(Q_i, Q_j)$ is the distance between cluster centers. As a result, if the clusters are compact and far apart, the ratio is low. As a result, for a decent cluster, the Davies–Bouldin index will have a tiny value [15].

We used the SAMASAFIA database, which is described in Section 2 [18] must deal with two types of data issues: missing data and outliers. Outliers are primarily caused by faulty measuring instrument operation or incorrect data collection and analysis methods. Outlier detection in our situation is based on median values, with the standard deviation of the meteorological factors taken into account. Outliers are carefully evaluated since they can generate bias in the prediction model's calibration. The failure of measurement instruments is the most common cause of missing data. Due to the experimental nature of the measurement stations, this is primarily caused by power outages and other faults in various analyzers (Samasafia, 2004). Missing data might throw off a statistical study by introducing systematic components of mistakes in parameter estimation in the prediction model.

Table 1 shows an example of an outlier detected in attributes: temperature, humidity, and wind speed that appeared on the second day of our database.

Table 2 shows examples of missing data for the different attributes: temperature, and humidity.

The results of clustering were obtained using DBSCAN to group 300 data lines of the database "Meteorological" by changing the input parameters: MinPts = 5

Table 1. Example of an outlier detected in attributes.

Hours	Temperature (C°)	Humidity (%)	Wind speed (m/s)
4	8.5	74	1.1
5	**37989,20833**	**37990,20833**	**37991,20833**
6	8.1	75	1.3
7	8.0	74	2.4

Table 2. Example of missing data detected in attributes.

Hours	Temperature (C°)	Humidity (%)	Wind speed (m/s)
9	9.4	79	5.0
10	**No data**	**No data**	**10**
11	**No data**	**No data**	**11**
12	8.5	85	3.7

and $\varepsilon = 5$, for each run are presented in Table 3, using the following criteria: result, ε, MinPts, number of data lines (NDL), number of clusters (NCL), and evaluation by Silhouette Index (SI) and Davies-Bouldin (DB).

Table 3. DBSCAN algorithm results.

Result	ε	MinPts	NDL	NCL	Evaluation SI	Evaluation DB
1	5	5	300	4	0.021645	0.874169
2	3	5	300	2	0.036589	0.325648
3	7	5	300	3	0.239746	0.345873
4	4.7	3	300	17	0.099475	0.187456

We have chosen partitioning number 1, which gives the best values of Davies Bouldin and Silhouette Index MinPts = 5 and $\varepsilon = 5$. We take the results of the DBSCAN algorithm (Table 3, line1), and we apply the first proposed incremental DBSCAN algorithm to group the 64 new data lines with the same input parameters MinPts = 5 and $\varepsilon = 5$. We get the following results (see Table 4):

Table 4. ANN-DBSCAN algorithm results with MinPts = 5 and $\varepsilon = 5$.

NDL	NCL	Evaluation SI	Evaluation DB
364	5	0.015846	0.896542

Note that the number of clusters in the proposed ANN-DBSCAN algorithm (=5) is great compared with the static DBSCAN algorithm result (=4). The optimal number of clusters proposed by this index is 5, according to the Davies–Bouldin index and the Silhouette index. Figure 2 shows the results of these indices.

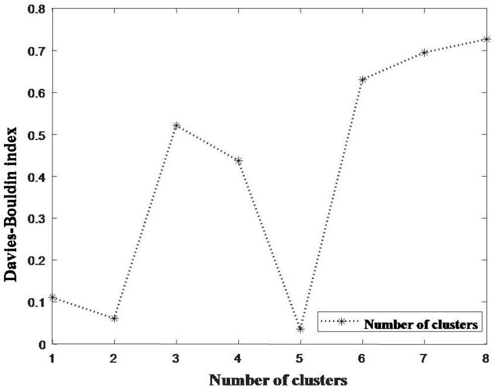

Figure 2. The Davies–Bouldin index minimum value indicates the optimal number of clusters as determined by the proposed ANN-DBSCAN.

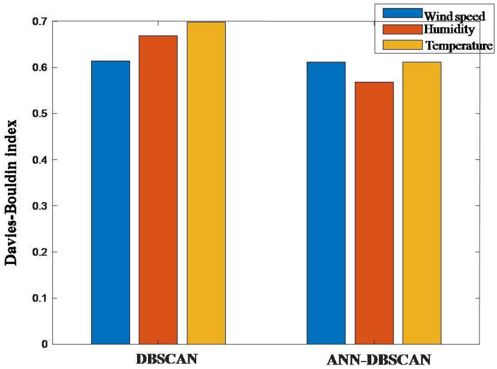

Figure 3. Davies–Bouldin index for the static algorithm and the ANN-DBSCAN algorithm.

From Figure 3, we see the following remarks: For the proposed ANN-DBSCAN algorithm, the Davies–Bouldin index values are low for the three attributes temperature, humidity and wind speed compared to the static DBSCAN algorithm. This explain that the clustering by ANN-DBSCAN is better than the static DBSCAN.

5. The influence of meteorological parameters on air pollution

5.1. Effects of air pollution data on the weather

- **CO2:** carbon dioxide is a significant heat-trapping (greenhouse) gas that is released by human activities like deforestation and fossil fuel combustion, as well as natural processes like respiration and volcanic eruptions [13]. During the past several hundred thousand years of glacial cycles, there has been a substantial correlation between temperature and carbon dioxide levels in the atmosphere. When the concentration of carbon dioxide in the atmosphere rises, so does the temperature. When the amount of carbon dioxide in the atmosphere decreases, the temperature decreases [12];

- **SO2:** sulfur dioxide has a strong stench and is colorless, thick, poisonous, and nonflammable. Temperatures and pressures are typical. Sulfur dioxide is a pollutant that causes respiratory issues, and it irritates the lungs in particular. Sulfur oxides are a primary contributor to acid rain production. Sulfur dioxide oxidizes to sulfur trioxide via many chemical processes. Sulfuric acid is formed when sulfur trioxide reacts with water vapor or droplets (H2SO4). Acid rain contains a variety of acids, including sulfuric acid [2];

- **NOx:** nitrogen dioxide has an unpleasant odor. Some nitrogen dioxide is created by plants, soil, and water, while some is formed naturally in the atmosphere by lightning. However, this method produces just around 1% of the total nitrogen dioxide contained in our cities' air. Nitrogen dioxide is a significant air contaminant because it leads to the creation of photochemical smog, which has serious health implications [17]. It is in charge of raising the temperature;

- **PM:** the most common pollutant that contributes to unhealthy days on the Air Pollution Index is Particulate Matter. Dust, smoke, fumes, mist, fog, aerosols, fly ash, and other pollutants are produced. Increased fine particle levels in the air have been related to health problems such as heart disease, lung problems, and lung cancer [11].

We have used the same database for forecasting weather, as is shown in Table 5. The influence of various air pollution molecules on the weather is seen in Table 6.

Figure 4 shows the average concentration levels of the pollutant CO. This graph represents the primary variances in pollution levels during the four seasons of 2004.

We have applied ANN-DBSCAN algorithm to the air pollutant on the data for the Spring season of 2014, we obtain the following results: Minpts = 5 and eps = 5.

Table 5. Hourly data.

Date	Temp	CO	NO2	O3	PM10	SO2
01/02/2004	14	1.56	82.12	37.6	158.46	26.41
02/02/2004	12	1.95	63.85	22.74	162.45	3.654
03/02/2004	20	1.75	122.64	29.45	228.96	5.96
04/02/2004	11	1.44	156.45	44.14	189.75	9.45
05/02/2004	16	0.96	74.28	30.96	154.85	9.88

Table 6. Parameters table.

Concentration	CO	NO2	O3	PM10	SO2
Low	Temp low	No Effect	No Effect	No Effect	No Effect
Normal	Temp Normal	Dry	No Effect	Dust	Dry
High	Temp High	Fog, dry	Humid high	Smog, dust, fog	Smog, dry
Extreme high	Temp Extreme High	Fog, dry	Humid high	Smog, dust, fog	Smog, dry

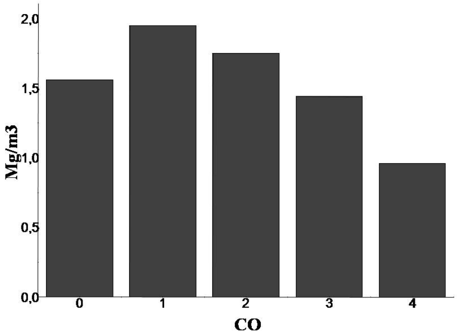

Figure 4. Pollutant concentrations for the four seasons of 2004 are averaged.

5.2. Evaluation measures for forecasting

A diagram identical to the one shown in figure is used to analyze the forecasts. The area “H” represents the intersection of the forecast and observed areas, or the “Hits” area; “M” represents the observed area that was missed by the forecast area, or the “Misses” area; and “F” represents the part of the forecast that did not overlap with an area of observed precipitation, or the “False Alarm” area, in this diagram [20].

We represent this situation by the “contingency table” as shown in Table 7. (Marginal of forecast: MF)

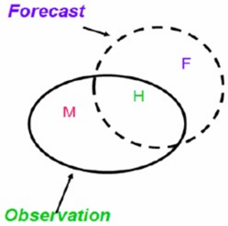

Figure 5. Diagram showing hits, misses, and false alarms for dichotomous forecast/observations.

Table 7. The contingency table for (yes-no) events.

		Observation			
Forecast	-	Yes	No	-	MF
	Yes	H: Hits	F False Alarm	H + F Yes forecast	
	No	M: Misses	N: Correct rejection	M + N ’No forecast’	
		H + M ’Yes observe’	F + N ’No observe’	Total forecast	
		Marginal of Observe			

Equitable Threat Score:

The Threat Score (TS) is a competence score for binary occurrences that necessitates the setting of a threshold. It’s calculated as the ratio of hits to the total of hits, false alarms, and misses. The Equitable Threat Score (ETS) is a modified Threat Score that subtracts the number of hits that would be predicted by chance alone from the total number of hits [20].

$$ETS = (H - CH)/(F + M - H - CH)$$

where $CH = (F \times M)/N$.

The number of random hits is CH, and the number of points in the verification domain is N. The ETS is the same as TS, but with a bias adjustment for random hits.

Bias score:

The number of times an event was predicted against the number of times it was observed is referred to as bias [20].

$$B = F/M$$

This score is above (below) 1 when the predicted precipitation rate is higher (lower) than the observed. $ETS = 1$ and $BIAS = 1$ are the results of a perfect forecast.

The number of verification grid-boxes containing observations affects the validity of these indexes.

Accuracy:

The ratio between the number of well-predicted (correctly classified) examples and the total number of examples. It is given by the following formula:

$$\text{Acc} = (F + H)(H + M)/(H + F + M + N)$$

5.3. Prediction evaluation

We evaluate the prediction by using the proposed ANN-DBSCAN algorithm. Using DBSCAN, we were able to determine the number of clusters for each air pollutant data set. Weather conditions are often the same inside a cluster. Now for the forecast, we have used the measures explained earlier (ETS, Bias and accuracy) for the purpose of the given date mentioned in [5]:

PWC: Predict weather conditions;
PT: Predict temp;
AT: Actual temp.

Table 8. The resultant table

Date	PWC	PT	AT	Hit	Miss
01/03/2014	Smog, fogs, Lightning, and cold	11-21	13	+	
02/03/2014	Humid, Smog	11-21	16	+	
03/03/2014	Humid, mist, dust	13-24	20	+	
04/03/2014	Smog, Lightning Thunder, and mist	09-17	11	+	
05/03/2014	Humid, fog	12-19	15	+	

These results are according to the cluster data and their nature, calculating the temperature range. The above table shows the comparison between the current temperature and the forecast temperature.

The forecasts will be evaluated using the Equitable Threat Score (ETS) and the Bias Score ($BIAS$), as well as their accuracy (acc):

$$\mathbf{ETS} = (364 - 311)/(5 + 968 - 364 - 311) = \mathbf{0.117}$$
$$\mathbf{Bias} = 398/968 = \mathbf{0.411}$$
$$\mathbf{Acc} = 364/398 \times 100 = \mathbf{91.45}\%$$

According to the use of the SAMASAFIA database and the result of the accuracy, we conclude that we have obtained the best score and the perfect forecast.

6. Conclusion and perspectives

In this work, two parts have been presented to depict and identify the Annaba region's meteorological day types; we have focused on the database Samasafia, and the following concluding statements can be made:

- We proposed the ANN-DBSCAN algorithm, combining the DBSCAN and Artificial Neural Network, based on density and proximity concepts;
- We evaluated our proposed ANN-DBSCAN algorithm versus static DBSCAN, using internal criteria: Davies Bouldin index and Silhouette index. There are five distinct clusters that have been found. Our results suggest that the proposed methodologies outperform the DBSCAN;
- Each cluster's meteorological parameters are simple to interpret based on previuos work. We have predicted weather using the ANN-DBSCAN clustering algorithm on the Samasafia database (Spring season);
- Verification of the Forecast Metrics like accuracy are calculated using their respective hit and miss times. Our results suggest that the proposed method produces more accurate results.

In the future:

- We will apply our proposed algorithm to the same aim as [18];
- We will use our proposed algorithm AMF IDBSCAN for the same objective as in this paper [4];
- We will utilize another incremental algorithm [3] for forecasting.

Acknowledgements. The author would like to thank the DGRSDT (General Directorate of Scientific Research and Technological Development) – MESRS (Ministry of Higher Education and Scientific Research), ALGERIA, for the financial support of LISCO Laboratory.

References

[1] A. Amel, M. Naila, M. Lyllia, T. Ali: *Caractérisation de la pollution par le NO2 à l'aide d'un couplage de technique biologique et physico-chimique dans la région d'Annaba (Algérie)*, Association pour la prévention de la pollution atmosphérique 69 (2008), pp. 2268–3798, doi: https://doi.org/10.4267/pollution-atmospherique.1367.

[2] J. Bachmann, R. Damberg, J. Caldwell, C. Edwards, P. Koman: *Review of the national ambient air quality standards for particulate matter: Policy assessment of scientific and technical information. OAQPS staff paper. Final report*, tech. rep., Environmental Protection Agency, Research Triangle Park, NC (United States . . .), 1996.

[3] A. Chefrour: *Incremental supervised learning: algorithms and applications in pattern recognition*, Evolutionary Intelligence (2019), pp. 1–16, doi: https://doi.org/10.1007/s12065-019-00203-y.

[4] A. Chefrour, L. Souici-Meslati: *AMF-IDBSCAN: Incremental Density Based Clustering Algorithm Using Adaptive Median Filtering Technique*, Informatica 43.4 (2019), doi: https://doi.org/10.31449/inf.v43i4.2629.

[5] S. C. Chou, M. Justi da Silva: *Objective evaluation of Eta model precipitation forecasts over South America*, Revista Climanálise 1.1 (1999), pp. 1–17.

[6] S. Ghazi, T. K. Med: *Combination of artificial neural network models for air quality predictions for the region of Annaba, Algeria*, International Journal of Environmental Studies 69.1 (2012), pp. 79–89, doi: https://doi.org/10.1080/00207233.2012.644900.

[7] E. Hamdy, A.-G. Hala: *Estimation of air pollutant concentrations from meteorological parameters using artificial neural network*, Journal of Electrical engineering 57.1-2 (2006), pp. 105–110.

[8] A. Hamza, C. M. Reda, H. Ahmed: *Forecasting PM 10 in Algiers: Efficacy of multilayer perceptron networks*, Environmental Science and Pollution Research 23.2 (2016), pp. 1634–1641, doi: https://doi.org/10.1007/s11356-015-5406-6.

[9] M. Holger, D. Grame: *Neural network based modelling of environmental variables: a systematic approach*, Mathematical and Computer Modelling 33.6-7 (2001), pp. 669–682.

[10] T. M. Kamal, S. Najib: *Arab environment: Future challenges*, in: Proceedings of the 2008 Symposium on Interactive 3D Graphics and Games, Beirut: Arab Forum for Environment and Development, 2008, doi: https://doi.org/10.1145/1342250.1342261.

[11] F. J. Kelly, J. C. Fussell: *Air pollution and public health: emerging hazards and improved understanding of risk*, Environmental geochemistry and health 37.4 (2015), pp. 631–649, doi: https://doi.org/10.1007/s10653-015-9720-1.

[12] S. G. Kulkarni, H. M. Mehendale: *Carbon dioxide* (2005), doi: https://doi.org/10.1130/0091-7613.

[13] S. G. Laws: *Chemistry libretexts.*

[14] E. Martin, K. Hans-Peter, S. Jörg, X. Xiaowei, et al.: *A density-based algorithm for discovering clusters in large spatial databases with noise*, in: kdd, vol. 96, 34, ACM, 1996, pp. 226–231, doi: https://doi.org/10.1023/A:1009745219419.

[15] A. L. Mary, K. S. Kumar: *A density based dynamic data clustering algorithm based on incremental dataset* (2012), doi: https://doi.org/10.3844/jcssp.2012.656.664.

[16] A. Moreira, M. Y. Santos, S. Carneiro: *Density-based clustering algorithms–DBSCAN and SNN*, University of Minho-Portugal (2005), pp. 1–18.

[17] W. H. Organization: *Air quality guidelines: global update 2005: particulate matter, ozone, nitrogen dioxide, and sulfur dioxide*, World Health Organization, 2006.

[18] K. Soufiane, K. M. Tarek: *Impact of clustered meteorological parameters on air pollutants concentrations in the region of Annaba, Algeria*, Atmospheric research 113 (2012), pp. 89–101, doi: https://doi.org/10.1016/j.atmosres.2012.05.002.

[19] I. J. Turias, F. J. Gonzalez, M. L. Martín, P. L. Galindo: *A competitive neural network approach for meteorological situation clustering*, Atmospheric Environment 40.3 (2006), pp. 532–541, doi: https://doi.org/10.1016/j.atmosenv.2005.09.065.

[20] D. S. Wilks: *Statistical methods in the atmospheric sciences*, vol. 100, Academic press, 2011.