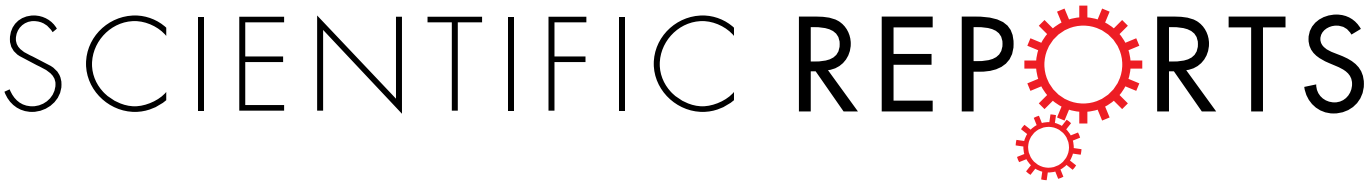

OPEN

Statistical analysis and estimation of the regional trend of aerosol size over the Arabian Gulf Region during 2002–2016

Received: 5 December 2017
Accepted: 8 June 2018
Published online: 22 June 2018

Alina Barbulescu[1], Yousef Nazzal[2] & Fares Howari[2]

In this article, we present the results of the regional estimation of the evolution of monthly mean aerosol size over the Arabian Gulf Region, based on the data collected during the period July 2002 – September 2016. The dataset used is complete, without missing values. Two methods are introduced for this purpose. The first one is based on the partition of the regional series in sub-series and the selection of the most representative one for fitting the regional trend. The second one is a version of the first method, combined with the k-means clustering algorithm. Comparison of their performances is also provided. The study proves that both methods give a very good estimation of the evolution of the aerosol size in the Arabian Gulf Region in the study period.

Aerosols are tiny particles suspended in the atmosphere which result from natural or anthropic sources. The natural aerosols are classified in: product of sea spray evaporation, mineral aerosol, volcanic aerosol, particles of biogenic origin, smokes from burning on land, sulphates[1].

Ginoux *et al.*[2] did an extensive study for attribution of anthropogenic and natural dust sources. Haywood *et al.*[3] indicate that the aerosols cause a strong radiative forcing of climate because of their efficient scattering of solar radiation. Since they are acting as cloud and ice concentration nuclei, the aerosols have a significant impact on the climate variation[4]. Their diameters vary from less than one nanometer to 100 μm, those of the natural aerosols being generally higher than those of the human-made ones[5]. Particles larger than about 1 μm are also called coarse particles.

One of the most abundant aerosols in the atmosphere is the dust. Desert dust particles, also called mineral aerosol, are soil particles suspended in the atmosphere in the area with easily erodible dry soils, strong winds and little vegetation[5]. They are composed of oxides (silica, iron oxides), quartz, feldspar, gypsum and hematite etc[1]. Their inhalation is dangerous for the human health because they can get deposited in the gas-exchange region of the lungs[6,7]. Therefore, their production and transport must be investigated.

Scientists found that dust particles enter the lower atmosphere primarily through a mechanism called saltation bombardment, which is strongly dependent on the meteorological conditions near the surface, as well as on the soil texture and particle size[8,9]. According to Astitha *et al.*[10], massive amounts of dust are lofted in deep atmospheric boundary layers over the hot deserts. Spyrou *et al.*[11] emphasized that extensive plumes can travel across multiple countries at high altitudes up to the middle troposphere, while other scientists emphasized the transport of the aerosol from Africa and Asia[12–14]. Levin *et al.*[15] analyzed the interaction mechanism between the mineral dust, sea-salt particles, and clouds, while Smoydzin *et al.*[4] analyzed the role of the meteorological conditions on the pollution over the Arabian Gulf. They stated that the dust is emitted as hydrophobic particles, relatively ineffective as cloud condensation nuclei, but during their transport in the atmosphere, and the interaction with gaseous and particulate air pollutants, their hygroscopicity increases, enhancing the efficiency of the dust removal through precipitation[16]. Alfaro and Gomes[17] proposed a model for mineral aerosol production by wind erosion by combining preexisting models of saltation and sandblasting processes that lead to mineral aerosol release in arid areas, while the impact of dust particles on the condensation nuclei, important in the precipitation formation and its spatial distribution has been analysed by Karydis *et al.*[18].

[1]Ovidius University of Constanta, Constanta, Romania. [2]College of Natural and Health Sciences, Zayed University, Abu Dhabi, United Arab Emirates. Correspondence and requests for materials should be addressed to A.B. (email: emma.barbulescu@yahoo.com)

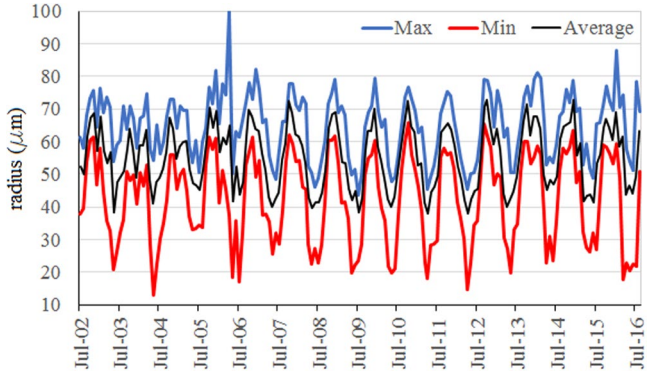

Figure 1. Maxima, minima and the average series during the study period.

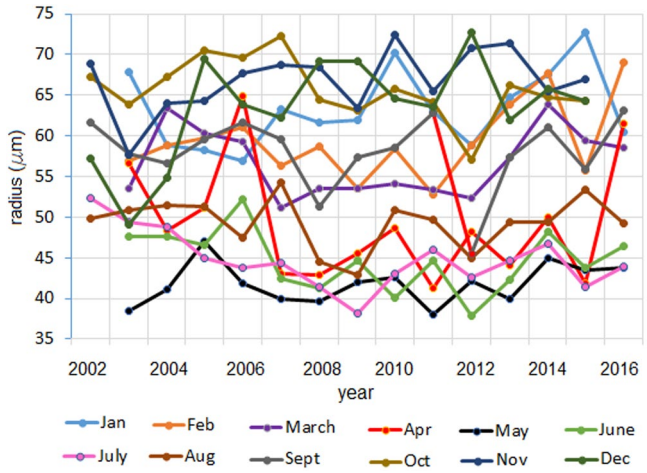

Figure 2. Regional monthly average series during the study period.

Namikas[19] investigated the causes of wind erosion in sand dunes. His research revealed that wind speed varies regionally, frequently occurring during certain periods and its interaction with the land-use disturbances may produce big quantities of atmospheric dust. It was shown that much of the desert dust mass transported in the atmosphere occurs during a few events. Part of them, occurring in South and Central Asia have been extensively studied by Chen *et al.*[12] and Huang *et al.*[20].

Recent research indicates a significant variability of the airborne desert dust concentrations that during the past decades in the Middle East, Africa, Central Asia and South America[12,21–23], and the augmentation of aerosol optical depth worldwide, especially in spring and summer, suggesting a relationship with the dust abundance[24]. Analyses of airborne desert dust and atmospheric concentrations of mineral dust particles extending from the Sahara, across the Arabian Peninsula and the Middle East, to South and Central Asia have been provided, as well[2,13,14,20].

Most article focused on classifying the aerosol types using the satellite data. The classification are based on the aerosol optical depth[25,26], Angstrom exponent and index of refraction[27–29], or fine mode fraction and the aerosol index[30].

The reviews on the impact of the dust size on climate and biogeochemistry, are focusing on the characterization of the size distributions of the aerosol particles. It was shown that the particles with sizes between 0.2–2 μm produce the largest shortwave radiative effect per unit mass[20,31]. As condensation nuclei, the number of particle above a given size is important[32]. Point of view of geochemistry, the amount of dust deposition is essential.

Given the importance of the study of aerosols dimensions, we notice only few articles treating this topic[20,33].

Therefore, we aim at analyzing the series of aerosols radius over the Arabian Gulf Region and to model the trend of the regional series. Two approaches are proposed, based on the partition of the regional series into subseries and the selection of the most representative one, used for building the regional trend.

In the following, we shall understand by *particular series*, a series recorded at a certain point and by the *regional series*, the multidimensional series containing all the particular series.

Results and Discussions

Figure 1 represents the maxima, minima and average regional series during the study period and Fig. 2 presents the monthly average of the series in the study period.

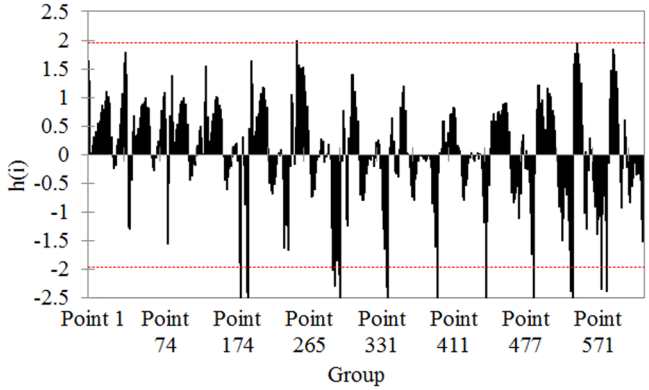

Figure 3. Results of Mandel's test.

We remark a seasonal variation of the series. At the beginning of the period the highest means have been recorded in October, November and December, but after 2010, we notice the highest monthly averages in November, December and January. The months when the smallest aerosols radii have been registered are May, June and July.

The analysis of the aerosols' radii at all the observation sites provide the following limits: the minimum is between 12.99 and 39.76 μm, the maximum is between 71.65 and 100 μm, the average is in the interval 51.21–56.79 μm, with the standard deviation in the interval 8.67–15.21. The series' skewness is between −0.41 and 0.48, 39% being in the interval [−0.05, 0.05], 31% greater than 0.05 and 33% less than −0.05, indicating that the majority of the series distributions are left or right skewed. The excess kurtosis is between −1.31 and 1.05: only 5 values are greater than zero, and approximately 90% around −1. Therefore the majority of the distributions of the dust aerosol size are platykurtic, only five of them being leptokurtic.

The normality hypothesis for the individual series has been rejected for all the series by the Shapiro-Wilk and Anderson-Darling tests. Jarque-Bera's test rejected the normality hypothesis for 95% of series. This result is in concordance with the finding related to the skewness and excess kurtosis.

After performing the Henze – Zirkler test, the null hypothesis was rejected as well. Therefore, the regional series does not follow a multivariate Gaussian distribution.

After applying the dcor t-test, the independence hypothesis has been rejected since the dcor coefficients are higher than 0.586. Since the p-value associated with the Kruskall-Wallis test is less than 0.0001, we can't accept the hypothesis that all the samples come from the same population. Therefore, the series are not independent, but they don't have the same distribution.

For emphasising the differences between the pairs of series, after the rejection of the null hypotheses by the Kruskall-Wallis test, the post-hoc Dunn's test was performed. Its results show that 82.2% of pairs of series don't come from the same population.

Mandel's test has been performed for detecting if there are outlying means among the series averages. It was found that there are only 18 outlying means of the study series. Figure 3 illustrates the result of this test. The dotted lines represent the critical values of the test (1.956 and −1.956), and the vertical lines, the values of h-statistic for the individual series. Most of these values are in the range from −1.25 to 1.25.

The Brown-Forsythe test rejected the homoscedasticity hypothesis. Therefore, we can conclude that the variances of the individual series are not homogenous.

For modeling the regional distribution of the aerosols' size, we used the following two algorithms.

Method I.

(1) Given the data series recorded at m different observation points, on n consecutive periods, build the matrix of the regional series, $Y = (y_{ji})_{\substack{j=\overline{1,n} \\ i=\overline{1,m}}}$, where y_{ji} is the series recorded at the moment j at the point i. So, a column of the matrix Y contains the data collected at a specific point, and a raw of the matrix contains all the values collected at a certain moment at all the sites.
For each $j = \overline{1,n}$, perform the steps (2)–(6).

(2) Compute the extreme values (maximum and minimum) on each row of the matrix (that is the regional extrema at each moment) and the amplitude, as the difference between the maximum and the minimum values.
For example, denoting respectively by $y_{j\,\max}$ and $y_{j\,\min}$ the maximum and minimum values at the moment j, the amplitude at the moment j will be defined by:

$$A_j = y_{j\,\max} - y_{j\,\min} \tag{1}$$

(3) Divide the intervals $[y_{j\,\min}, y_{j\,\max}]$ into a convenient number of sub-intervals, m_j, of length $L_j = A_j/m_j$, such as each sub-interval contains enough values.
Let us denote by I_{jl} the sub-interval l of the period j.

(4) Attach to each sub-interval interval I_{jl} its frequency, f_{jl}, defined as the number of values from I_{jl}.
(5) Choose that interval I_{jl} whose frequency is maximum. Denote it by $I_{j\max}$ and by $f_{j\max}$ the corresponding frequency. If the highest frequency appears more than once, $I_{j\max}$ will be that interval whose average is the closest to the average of the entire period j. For example, suppose that the maximum frequency is $f_{j3} = f_{j5}$, the average of the values in I_{j3} is 24.5, the average of the values in I_{j5} is 29.5 and the average of the period j is 30. Then, we select $I_{j\max} = I_{j5}$ because 29.5 is closer to 30 than 24.5.
(6) Choose the representative value for the period j to be equal to the average of the values from the interval $I_{j\max}$, and denote it by $\overline{y}_{j\max}$.
(7) Build the trend series that fit the regional one using $(\overline{y}_{j\max})_{j=\overline{1,n}}$.
(8) Estimate the fitting quality, computing the mean absolute error and the mean standard error of each series i $(i = \overline{1,m})$[34].

The mean absolute error and the mean standard error of the series i (denoted respectively by MAE_i and MSE_i) are defined by:

$$\mathrm{MAE}_i = \frac{1}{n_i}\sum_{q=1}^{n_i}|x_{iq} - (x_{iq})_e|, \tag{2}$$

$$\mathrm{MSE}_i = \sqrt{\frac{1}{n_i}\sum_{q=1}^{n_i}\left[x_{iq} - (x_{iq})_e\right]^2}, \tag{3}$$

where n_i is the number of values of the series i, x_{iq} is the q^{th} value of the series i, $(x_{iq})_e$ is the value estimated by the model for the q^{th} value of the series i.

Lower MAE_i and MSE_i are, better the modeling quality is.

Method II. This algorithm is based on the previous one, but the selection of the interval with the maximum frequency is replaced by the selection of the cluster with the highest number of elements, after pruning the k-means clustering algorithm, for classifying the series in clusters (homogeneous and disjoint groups within which the patterns are similar). The number of clusters, k, is a priori specified and the algorithm k-means stores k centroids used to define clusters.

The clustering idea is finding groups (clusters) that minimise an error criterion, as, for example, Sum of Squared Error (SSE), which measures the total squared Euclidian distance of instances to their representative values[35].

Considering $Z = (z_1, z_2, \ldots, z_m)$, $z_i \in \boldsymbol{R}^n$, $i = \overline{1,m}$ being the given points, the k-means clustering algorithm has the following steps[32]:

(a) Choose the number of clusters, k.
(b) Initialize the cluster centroids $v_1, v_2, \ldots, vk \in \mathbf{R}^n$ randomly.
(c) Compute the distance between each data point and the cluster centers.
(d) Assign the data point to the cluster whose distance from the cluster center is the minimum of all distances to the cluster centers.
(e) Compute the new cluster center by

$$v_i = \frac{1}{c_i}\sum_{j=1}^{c_i} z_j\,, \tag{4}$$

where c_i is the number of the data points in i^{th} cluster.
(f) Restart from (b) till no data point will be reassigned to the cluster. Then stop.

For more information about other clustering algorithms, the reader could refer to Rokach and Maimoon[36], Everitt *et al.*[37], Xu and Wunsch[38].

In our case, $Z = X$ and the element z_i is the column i of the matrix X.

The stages or *Methods II* are the following:

(I) Similar to step (1) of the Method I.
(II) Choose the number of clusters, k, and perform the k-means clustering.
(III) Determine the cluster with the highest number of elements and build the matrix X_c with the columns of X that contain the values recorded at the observation points included in this cluster.
(IV) Choose the representative value for the period j to be equal to the average of the values from the j^{th} row of X_c. Denote it by $\overline{y}_{j\,C}$.
(V) Build the series that fit the regional one, using $(\overline{y}_{j\,C})_{j=\overline{1,n}}$, and estimate the fitting quality by computing MAE_i and MSE_i of each series $i = \overline{1,m}$.

The comparison of the methods' performances is done by computing the mean absolute error and mean standard error of each series and the overall the mean absolute error and mean standard error.

The overall mean absolute error is defined by:

Class	1	2	3	4	5	6
1	0	54.557	62.361	57.625	46.242	58.259
2	54.557	0	36.057	59.761	28.393	38.034
3	62.361	36.057	0	45.503	43.235	23.329
4	57.625	59.761	45.503	0	54.052	42.183
5	46.242	28.393	43.235	54.052	0	40.089
6	58.259	38.034	23.329	42.183	40.089	0

Table 1. Distances between the class centroids.

Central object	1 (Point 57)	2 (Point 88)	3 (Point 160)	4 (Point 242)	5 (Point 283)	6 (Point 296)
1 (Point 57)	0	62.988	74.500	72.723	55.341	67.226
2 (Point 88)	62.988	0	49.429	69.936	37.787	49.918
3 (Point 160)	74.500	49.429	0	60.791	55.932	34.831
4 (Point 242)	72.723	69.936	60.791	0	63.364	53.695
5 (Point 283)	55.341	37.787	55.932	63.364	0	49.296
6 (Point 296)	67.226	49.918	34.831	53.695	49.296	0

Table 2. Distances between the central objects.

Class	1	2	3	4	5	6
Objects' number	19	105	111	50	50	52
Within-class variance	2935.890	821.017	870.805	1950.387	1547.098	781.011
Minimum distance to centroid	36.693	17.443	19.743	24.940	19.993	17.046
Average distance to centroid	50.604	27.306	28.408	41.550	36.373	26.409
Maximum distance to centroid	95.749	61.838	59.983	77.899	82.159	51.011

Table 3. Results of the k-means clustering algorithm by class.

$$\mathrm{MAE} = \frac{1}{\sum_{i=1}^{m} n_i} \sum_{i=1}^{m} (n_i \cdot \mathrm{MAE}_i), \tag{5}$$

and the overall mean standard error is given as:

$$\mathrm{MSE} = \sqrt{\frac{1}{\sum_{i=1}^{m} n_i} \sum_{i=1}^{m} (n_i \cdot \mathrm{MSE}_i)}, \tag{6}$$

where MAE is the overall mean absolute error, MSE is the overall mean standard error, MAE_i is the mean absolute error of the series i, MSE_i is the mean standard error of the series i, n_i is the number of values in the series i.

Lower MAE (or MSE) is, better the modeling result is.

The modeling techniques have been applied for monthly average series collected at the study sites for 171 month. The data series and the computation done by applying Methods I and II can be found in the Supplementary Dataset 1 and Supplementary Tables S1 and S2.

For Method I, $n = 387$, $m = 171$. We chose $m_j = 6$, for all $j = \overline{1, 171}$ at stage (3), meaning that after detecting the range of the values in a given period, this interval was divided into 6 subintervals. Method II was applied using a number of clusters $k = 6$, for comparison reasons.

After performing the k - means algorithm we got the results presented in Tables 1–3.

We remark that the within-class variances, minimum, maximum and average distances to centroids are close to each other for the classes 2 and 3. The classes have different numbers of elements, the highest one being in the third one that will be selected for modeling with Method II.

The regional series obtained by applying the methods previously described are presented in Fig. 4, where Series I is obtained by Method I and Series II is obtained by Method II.

The corresponding errors (MAD and MSE) are presented in Figs 5 and 6. They were obtained based on the data provided in the Supplementary Tables S1 and S2. The values from the Supplementary Tables S1 and S2 can be synthetized in Table 4, where the minimum and maximum modeling error are presented.

Mean absolute deviations and mean standard error of the individual series (MAD_i, MSE_i) vary in larger limits when using Method II, than when using Method I. Comparing the MADs (respectively MSEs) from Supplementary Tables S1, one can see that the first method better performed for 223 series (respectively 218 series). This means that MADs (respectively MSEs) were smaller in 57.62% (respectively 56.33%) cases when

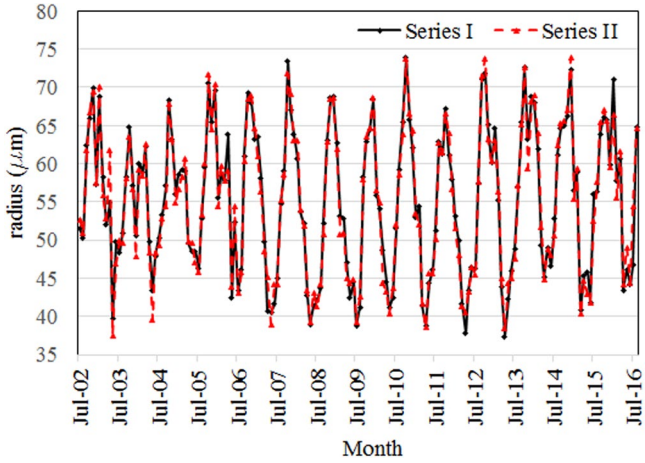

Figure 4. Models obtained by Method I (Series I) and Method II (Series II).

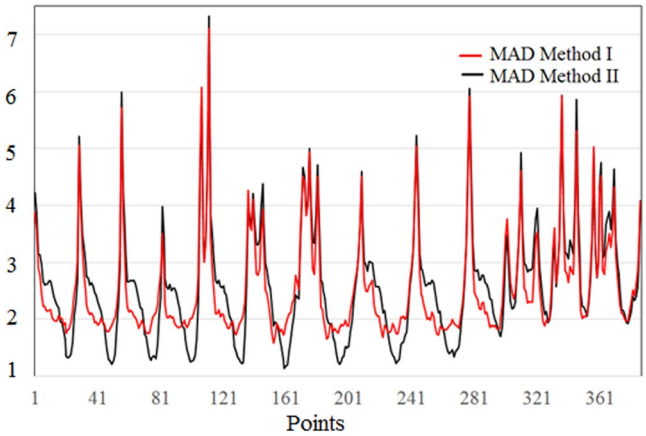

Figure 5. Mean absolute error (MAD) in the models.

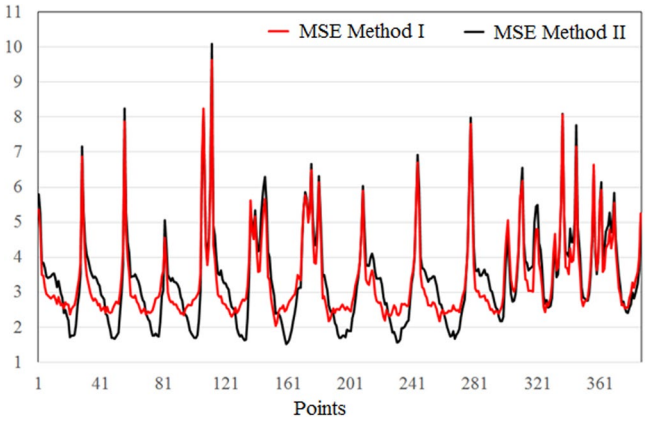

Figure 6. Mean standard error in the models.

Method I was used. Overall MADs for the models were respectively 2.536 and 2.607. Overall MSEs for the models were respectively 3.547 and 3.639. Therefore, the overall mean absolute deviations and mean standard errors are comparable for both methods.

The study series were not highly inhomogeneous. Only some outlying means were detected and the standard deviations were between 8.69 and 15.25. This could be a reason for which the goodness of fit of the methods are similar.

Method	MAD_i		MAD	MSE_i		MSE
	min	max		min	max	
I	1.584	7.111	2.536	2.040	9.629	3.547
II	1.137	7.730	2.607	1.514	10.083	3.659

Table 4. Comparison of modeling errors.

Figure 7. Observation area - inside the rectangle (https://www.google.com/mymaps/, 2018).

The Method I gave better results due to the procedure of selection of the values of the individual series that participate in fitting the regional one. While in the second method, all the values that participate to this process are taken from the same series, belonging to the cluster with the highest number of member, in Method I, the values taken into account at each moment can belong to different series. For example, the representative values for the time $t = 2$ could be in the subinterval 3 and those for the time $t = 4$ could be in the subinterval 6.

Conclusions

In this article, we presented two methods for estimating the evolution of the regional time series of the aerosols dimensions. Both methods are easy to use, the advantage of the second one being that the k-mean algorithm is implemented in many software. In both methods, the number of subintervals, respectively clusters must be specified from the beginning for the selection of the series that will finally participate in fitting the regional series. The fitting results are not significantly different (in terms of overall MADs and MSEs), but the first method gave a better estimation of the individual series values (223 and 218 series, respectively). For highly inhomogeneous series, we recommend the use of the first method, knowing that the k-mean algorithm is sensitive to the outliers' existence. Therefore statistical tests for the mean homogeneity and homoskedasticity, as well as that for the outliers' existence must be performed before deciding the modeling method.

Both methods could be successfully used to estimate the regional evolution of the pollutants' series when some particular series in the study area present missing values, but the neighboring series are complete. In the future works we shall study the sensitivity of both methods to the selection of the clusters' number. We shall compare the methods' goodness of fit for regional series of data (as precipitation, temperature, anthropic aerosol) before and after the seasonality removal, as well. We also aim at implementing both methods in a friendly - user software.

Methodology

Data series. Data used are monthly series containing the aerosol particle radius record from July 2002 till September 2016, over the Gulf Region (Fig. 7), retrieved by MODIS[39]. There were 1850 observation points. For modeling the regional evolution of the dimension of aerosol particles we selected 387 complete series (without missing data), each of them containing 171 values. The coordinates of these points can be found in Supplementary Dataset.

To determine the type of aerosols, we compared the aerosols dimensions and the optical depth (AOT) downloaded from MODIS site[39] for our series, with the results from the literature[25–29]. It resulted that the type of aerosol is mostly dust because AOT was larger than 0.3 (majority around 0.43). The aerosols' dimensions are in the range of the coarse aerosols (as per the data from the Supplementary Dataset). Due to lack of space and since our main purpose is the modeling, we shall not insist on other characteristics of the study aerosols. We intend to dedicate another article to an extensive study of their properties.

Statistical analysis. Before modeling, statistical analyses of the series have been performed, at the significance level $\alpha = 0.05$. For all the tests, when the p-value associated is lower than the significance level, one should reject the null hypothesis H_0, and accept the alternative hypothesis H_a.

In what follows we shall call *individual series* the series recorded a certain observation point and the *regional series* that one recorded at all the sites. Therefore, an individual series will be represented by a column vector and the regional series as a matrix formed by all the columns containing the individual series.

The statistical tests have been conducted using the R software.

For testing the normality hypothesis against the non-normality for the individual series, the Anderson-Darling[40], Jarque-Bera[41] and Shapiro-Wilk[42] tests were performed. They are implemented in 'fBasics' package in R[43].

The test statistic for the Anderson-Darling test is defined as:

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)[\ln p_{(i)} + \ln(1 - p_{(n-i+10)})], \tag{7}$$

where

$$p_{(i)} = \Phi([x_{(i)} - \overline{x}]/s), \tag{8}$$

Φ is the cumulative distribution function of the standard normal distribution, $\overline{x}$ is the mean and s is the standard deviation of elements in the sample,$\{x_i\}_{i=\overline{1,n}}$.

If the p-value associated to the test is less than the significance level, the normality hypothesis can be rejected.

The Jarque-Bera test statistic is defined as:

$$JB = n\left(\frac{k_3^2}{6} + \frac{k_4^2}{24}\right), \tag{9}$$

where n is the sample volum, k_3 is the sample skewness and k_4 is the sample excess kurtosis, defined as:

$$k_3 = \frac{1}{ns^3}\sum_{i=1}^{n}(x_i - \overline{x})^3, \tag{10}$$

$$k_4 = \frac{1}{ns^4}\sum_{i=1}^{n}(x_i - \overline{x})^4 - 3. \tag{11}$$

For large samples, the JB statistics is compared to a chi-squared distribution with 2 degrees of freedom ($\chi^2(2)$). The normality hypothesis is rejected if the test statistic is greater than $\chi^2(2)$.

The statistic of the Shapiro-Wilk test is defined as:

$$W = \frac{(\sum_{i=1}^{n} w_i x'_i)^2}{\sum_{i=1}^{n}(x'_i - \overline{x})^2} \tag{12}$$

where n is the sample volume, $x_1, x_2, \ldots, x_n$ are the original data, $x'_1, x'_2, \ldots, x'_n$ are the ordered data, $\overline{x}$ is the sample mean of the data, and the constants w_i are given by:

$$(w_1, w_2, \ldots, w_n) = \frac{M^T V^{-1}}{(M^T V^{-1} V^{-1} M^T)^{1/2}}, \tag{13}$$

where $M = (m_1, m_2, \ldots, m_n)^T$ (the transposed vector $(m_1, m_2, \ldots, m_n)$) formed by the expected values of the order statistics of independent and identically distributed random variables sampled from the standard normal distribution and V is the covariance matrix of those order statistics.

Small values of W indicate non – normality.

All software provide the p-values corresponding to this test statistics. If the p-value is less than the significance level, the normality hypothesis can be rejected.

For testing the multivariate normality (the normality of the regional series) against the non-normality hypothesis, we used the multivariate Henze - Zirkler test[44], implemented in the R package 'MVN'[45]. The test is presented in the following.

Let $X_1, X_2, \ldots, X_n \in \mathbf{R}^d$ be a random sample, where d is the dimension of X_i and n is the number of observations.

The Henze - Zirkler test is based on a nonnegative functional D that measures the distance between two distribution functions and has the property that $D(N_d(0, I_d), Q) = 0$ if and only if $Q = N_d(0, I_d)$, where $N_d(\mu, \sum_d)$ is a d-dimensional normal distribution.

The test statistic $T_\beta(d)$ is defined by:

$$T_\beta(d) = \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n} exp\left(-\frac{\beta^2}{2}\left|Y_j - Y_K\right|^2\right) - 2(1+\beta^2)^{-d/2}\frac{1}{n} \times \sum_{j=1}^{n} exp\left(-\frac{\beta^2}{2(1+\beta^2)}\left|Y_j\right|^2 + (1+\beta^2)^{-d/2}\right), \tag{14}$$

where

$$\beta = \frac{1}{\sqrt{2}}\left(\frac{2d+1}{4}\right)^{1/(d+4)} n^{1/(d+4)}, \tag{15}$$

$$\left|Y_j - Y_k\right|^2 = (X_j - X_k)'\boldsymbol{S}^{-1}(X_j - X_k), \tag{16}$$

$$\left|Y_j\right|^2 = (X_j - \overline{X})'\boldsymbol{S}^{-1}(X_j - \overline{X}) \tag{17}$$

and **S** is the sample covariance matrix of the matrix **X** formed by $X_1, X_2, \ldots, X_n$ as columns.

In the null hypothesis, the test statistic is approximately lognormal distributed. The null hypothesis is rejected if the p-value associated to the test statistics is less than the significance level.

Since the regional series is not multivariate Gaussian, for testing the independence hypothesis of the regional series, the nonparametric dcor t-test[46] of independence has been performed.

The test statistics is:

$$T_n = \sqrt{\nu - 1}\cdot\frac{R_n^*}{\sqrt{1-(R_n^*)^2}}, \tag{18}$$

where n is the sample size, R_n^* is the distance correlation statistics and $\nu = n(n-3)/2$.

The test rejects the null hypothesis at level α if $T_n > c_\alpha$, where c_α is the $(1-\alpha)$ quantile of a Student t distribution with $\nu - 1$ degrees of freedom. For details, the reader may refer to the article of Székely and Rizzo[46].

For testing the hypothesis H_0: *The individual series come from the same population* against the alternative H_a: *The individual series do not come from the same population*, the Kruskall-Wallis test[47] was performed.

The first stage all data is ranked, ignoring the group membership. The rank of the tied values is the average of the ranks they would have received had they not been tied.

The test statistic is defined as:

$$H = (N-1)\frac{\sum_{i=1}^{k} n_i(\overline{r}_{i.} - \overline{r})^2}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(r_{ij} - \overline{r})^2}, \tag{19}$$

where N is the total number of values, k is the number of groups, n_i is the number of observations in the group i, r_{ij} is the rank of observation j from group i, among all observations, $\overline{r}_i$ is the mean of the ranks of the observations in the group i, $\overline{r} = (N+1)/2$.

If the null hypothesis is true, H has approximately a chi-square distribution with $k-1$ degrees of freedom, χ^2_{k-1}. Therefore, the null hypothesis is rejected if $\mathrm{H} > \chi^2_{\alpha;k-1}$, found in the tables of the chi-square distribution at the significance level α (generally set to be 0.05).

If the null hypothesis is rejected, the multiple pairwise comparisons are done using the Dunn's test[48].

Let W_i the i^{th} group's summed ranks and n_i its sample size and $\overline{W_i} = W_i/n_i$. Assign any tied values the average of the ranks they would have received had they not been tied.

For testing the hypothesis that the group A and B come from the same population against the hypothesis that they don't come from the same population, we compute

$$z_i = \frac{\overline{W_A} - \overline{W_B}}{\sigma_i} \tag{20}$$

where:

$$\sigma_i = \sqrt{\left(\frac{N(N+1)}{12} - \frac{\sum_{s=1}^{r}\tau_s^3 - \tau_s}{12(N-1)}\right)\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}, \tag{21}$$

N is the total number of elements in all the k groups,

r is the number of tied ranks across all the groups,

τ_s is the number of elements across all the groups, with the s^{th} tied rank.

If there are no ties, the second term in the first bracket is zero.

Denote by k the number of groups and $z_{1-\alpha/2k}$, the $(1-\alpha/2k)$ point of the standard normal distribution. If $|z_i| < z_{1-\alpha/2k}$, then the null hypothesis can't be rejected.

For checking the existence of outlying means of the individual series, the Mandel's h statistic[49] was used, implemented in 'ILS' package in R[50].

Let k be the number of samples, $\overline{x}_i$, $i = 1, 2, \ldots, k$, their means and $\overline{\overline{x}}$ the overall mean. The Mandel's h test statistics are:

$$h_i = \frac{\overline{x}_i - \overline{\overline{x}}}{[1/(k-1)\sum_{i=1}^{k}(\overline{x}_i - \overline{\overline{x}})^2]^{1/2}} \tag{22}$$

and they have the same distribution for all $i = 1, 2, \ldots, k$.

For example, for $i = k$, the critical values are given by:

$$h_{k;1-\alpha/2} = \frac{(k-1)t_{k-2;1-\alpha/2}}{[k(k-2) + t^2_{k-2;1-\alpha/2}]^{1/2}}, \tag{23}$$

where $t_{k-2;1-\alpha/2}$ is the $(1 - \alpha/2)$ −quantile of the t-distribution with $(k-2)$ degrees of freedom[51].

To test the hypothesis H_0: *The variances of the individual series are identical*, against the alternative one H_a: *At least one of the variances is not identical to the others*, the Brown-Forsythe test[52], implemented in 'lawstat' package in R[53] has been performed.

Let $z_{ij} = |y_{ij} - \overline{y}_j|$,where y_{ij} is the element i in the group j and $\overline{y}_j$ is the median of group j. The Brown – Forsythe test statistic is

$$F = \frac{N-p}{p-1}\frac{\sum_{j=1}^{k} n_j(\overline{z}_{.j} - \overline{z}_{..})^2}{\sum_{j=1}^{k}\sum_{i=1}^{n_j}(z_{ij} - z_{.j})^2}, \tag{24}$$

where N is the total number of observations, k is the number of groups, n_j is the number of observations in group j, $\overline{z}_{.j}$ is the mean of the elements z_{ij} in group j, and $\overline{z}_{..}$ is the overall mean of the z_{ij}.

The test rejects the hypothesis that the variances are equal at the significance level α if $F > F_{\alpha,k-1,n-k}$, where $F_{\alpha,k-1,n-k}$ is the upper critical value of the F distribution with $(k-1)$ and $(n-k)$ degrees of freedom at the significance level of α. Alternatively, the null hypothesis is rejected if the p-value corresponding to the test is less than α.

Availability statement. Data are available in the Supplementary Database1.

References

1. Kondratatyev, K. I., Ivlev, L. S., Krapivin, V. F. & Varostos, C. A. *Atmospheric aerosol properties. Formation, processes and impact* (Springer, 2006).
2. Ginoux, P., Prospero, J. M., Gill, T. E., Hsu, N. C. & Zhao, M. Global-scale attribution of anthropogenic and natural dust sources and their emission rates based on MODIS Deep Blue aerosol products. *Rev. Geophys.* **50**, RG3005, https://doi.org/10.1029/2012RG000388 (2012).
3. Haywood, J. M. *et al*. Radiative properties and direct radiative effect of Saharan dust measured by the C-130 aircraft during SHADE: 1. Solar spectrum. *J. Geophys. Res.* **108**(D18), 8577, https://doi.org/10.1029/2002JD002687 (2003).
4. Smoydzin, L., Fnais, M. & Lelieveld, J. Ozone pollution over the Arabian Gulf – Role of meteorological conditions. *Atmos. Chem. Phys. Discuss.* **12**, 6331–6361, https://doi.org/10.5194/acpd-12-6331-2012 (2012).
5. Mahowald, N. M. *et al*. The size distribution of desert dust aerosols and its impact on the Earth system. *Aeolian Res.* **15**, 53–71, https://doi.org/10.1016/j.aeolia.2013.09.002 (2014).
6. Brunerkreef, B. & Holgate., S. T. Air pollution and health. *Lancet* **360**, 1233–1242, https://doi.org/10.1016/S0140-6736(02)11274-8 (2002).
7. Giannadaki, D., Pozzer, A. & Lelieveld, J. Modeled global effects of airborne desert dust on air quality and premature mortality. *Atmos. Chem. Phys.* **14**, 957–968, https://doi.org/10.5194/acp-14-957-2014 (2014).
8. Grini, A., Zender, C. S. & Colarco, P. Saltation sandblasting behavior during mineral dust aerosol production. *Geophys. Res. Lett.* **29**(18), https://doi.org/10.1029/2002GL015248 (2002).
9. Shao, Y., Raupach, M. R. & Findlater, P. A. The effect of saltation bombardment on the entrainment of dust by wind. *J. Geophys. Res.* **98**, 12719–12726, https://doi.org/10.1029/93JD00396 (1993).
10. Astitha, M., Lelieveld, J., Abdel Kader, M., Pozzer, A. & de Meij, A. Parameterization of dust emissions in the global atmospheric chemistry-climate model EMAC: impact of nudging and soil properties. *Atmos. Chem. Phys.* **12**, 11057–11083, https://doi.org/10.5194/acp-12-11057-2012 (2012).
11. Spyrou, C. *et al*. Modeling the radiative effects of desert dust on weather and regional climate. *Atmos. Chem. Phys.* **13**, 5489–5504 (2013).
12. Chen, S. *et al*. T. Comparison of dust emission, transport, and deposition between the Taklimakan Desert and Gobi Desert from 2007 to 2011, 2017, Science China Earth Sciences, 60, 1-1, doi:1.01007/s11430-016-9051-0 (2017).
13. Prospero, J. M. Saharan dust transport over the North Atlantic Ocean and Mediterranean: an overview in *The Impact of Desert Dust Across the Mediterranean* (eds Guerzoni, S., Chester, R.) 133–151 (Kluwer Academic Publishers, 1996).
14. Prospero, J. & Lamb, P. African droughts and dust transport to the Caribbean: climate change implications. *Science* **302**, 1024–1027, https://doi.org/10.1126/science.1089915 (2003).
15. Levin, Z., Teller, A., Ganor, E. & Yin, Y. On the interactions of mineral dust, sea-salt particles, and clouds: A measurement and modeling study from the Mediterranean Israeli Dust Experiment campaign. *J. Geophys. Res.* **110**, D20202, https://doi.org/10.1029/2002GL016055 (2005).
16. Teller, A., Xue, L. & Levin, Z. The effects of mineral dust particles, aerosol regeneration and ice nucleation parameterizations on clouds and precipitation. *Atmos. Chem. Phys.* **12**, 9303–9320, https://doi.org/10.5194/acp-12-9303-2012 (2012).
17. Karydis, V. A., Kumar, P., Barahona, D., Sokolik, I. N. & Nenes, A. On the effect of dust particles on global cloud condensation nuclei and cloud droplet number. *J. Geophys. Res.* **116**, D23204, https://doi.org/10.1029/2011JD016283 (2011).
18. Alfaro, S. C. & Gomes, L. Modeling mineral aerosol production by wind erosion: Emission intensities and aerosol size distributions in source areas. *J. Geophys. Res.* **106**, 18075–18084, https://doi.org/10.1029/2000JD900339 (2001).

19. Namikas, S. L. Field measurement and numerical modelling of aeolian mass flux distributions on a sandy beach. *Sedimentology* **50**, 303–326, https://doi.org/10.1046/j.1365-3091.2003.00556 (2003).
20. Huang, J., Wang, T., Wang, W., Li, Z. & Yan, H. Climate effects of dust aerosols over East Asian arid and semiarid regions. *J. Geophys. Res. Atmos.* **119**(19), 11398–11416, https://doi.org/10.1002/2014JD021796 (2014).
21. Bergametti, G. & Gillette, D. A. Aeolian sediment fluxes measured over various plant/ soil complexes in the Chihuahuan desert. *J. Geophys. Res.* **115**, 1–17, https://doi.org/10.1029/2009JF001543 (2010).
22. Ganor, E., Osetinsky, I., Stupp, A. & Alpert, P. Increasing trend of African dust, over 49 years, in the eastern Mediterranean. *J. Geophys. Res.* **115**, D07201, https://doi.org/10.1029/2009JD012500 (2010).
23. Mahowald, N. M. *et al*. Observed 20th century desert dust variability: impact on climate and biogeochemistry. *Atmos. Chem. Phys.* **10**, 10875–10893, https://doi.org/10.5194/acp-10-10875-2010 (2010).
24. Yoon, J., von Hoyningen-Huene, W., Kokhanovsky, A. A., Vountas, M. & Burrows, J. P. Trend analysis of aerosol optical thickness and Angström exponent derived from the global AERONET spectral observations. *Atmos. Meas. Tech.* **5**, 1271–1299, https://doi.org/10.5194/amt-5-1271-2012 (2012).
25. Hsu, N. C. *et al*. Global and regional trends of aerosol optical depth over land and ocean using SeaWiFS measurements from 1997 to 2010. *Atmos. Chem. Phys.* **12**, 8037–8053, https://doi.org/10.5194/acp-12-8037-2012 (2012).
26. De Meij, A., Pozzer, A. & Lelieveld, J. Trend analysis in aerosol optical depths and pollutant emission estimates between 2000 and 2009. *Atmos. Environ.* **51**, 75–85, https://doi.org/10.1016/j.atmosenv.2012.01.059 (2012).
27. Dubovik, O. *et al*. Variability of Absorption and Optical Properties of Key Aerosol Types Observed in Worldwide Locations. *J. Atmos. Sci*. **59**, 590–608, doi:10.1175/1520-0469(2002)059<0590:VOAAOP>2.0.CO;2 (2002).
28. Ridley, D. A., Heald, C. L., Kok, J. F. & Zhao, C. An observationally constrained estimate of global dust aerosol optical depth. *Atmos. Chem. Phys.* **16**, 15097–15117, https://doi.org/10.5194/acp-16-15097-2016 (2016).
29. Smirnov, A. *et al*. Atmospheric Aerosol Optical Properties in the Persian Gulf Region, *J. Atmos. Sci*. **59**, 620–634, doi:10.1175/1520-0469(2002)059<0620:AAOP IT>2.0.CO;2 (2002).
30. Kim, J. *et al*. Consistency of the aerosol type classification from satellite remote sensing during the Atmospheric Brown Cloud–East Asia Regional Experiment campaign. *J. Geophys. Res.* **112**, D22S33, https://doi.org/10.1029/2006JD008201 (2007).
31. Miller, R. L. *et al*. Mineral dust aerosols in the NASA Goddard Institute for Space Sciences ModelE atmospheric general circulation model. *J. Geophys. Res.* **111**, D06208, https://doi.org/10.1029/2005JD005796 (2006).
32. Dusek, U. *et al*. Size matters more than chemistry for cloud-nucleating ability of aerosol particles. *Science* **312**(5778), 1375–1378, https://doi.org/10.1126/science.1125261 (2006).
33. Reid, E. *et al*. Characterization of African dust transported to Puerto Rico by individual particle and size segregated bulk analysis. *J. Geophys. Res -. Atmos.* **108**(D19), 8591, https://doi.org/10.1029/1200dJD002935 (2003).
34. Barbulescu, A. A new method for estimation the regional precipitation. *Water Resour. Manag.* **30**(1), 33–42, https://doi.org/10.1007/s11269-015-1152-2 (2016).
35. Rokach, L. & Maimon, O. Clustering methods, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.149.9326&rep=rep1&type=pdf.
36. Hansen, P. & Nagai, E. Analysis of Global k-means, an Incremental Heuristic for Minimum Sum of Squares Clustering. *J. Classif.* **22**, 287–310, https://doi.org/10.1007/s00357-005-0018-3 (2005).
37. Everitt, B. S., Landau, S., Leese, M. & Stahl, D. *Cluster analysis*, 5th edition. (Chichester, 2011).
38. Xu, R. & Wunsch, D. Survey of Clustering Algorithms. *IEEE T. Neural Networ.* **16**(3), 647–678, https://doi.org/10.1109/TNN.2005.845141 (2005).
39. MODIS, https://neo.sci.gsfc.nasa.gov/view.php?datasetId=MODAL2_M_AER_RA.
40. Anderson, T. W. & Darling, D. A. *A Test of Goodness of Fit. J. Am. Stat. Assoc.* **49**(268), 765–769 (1954).
41. Jarque, C. M. & Bera, A. K. Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Econ. Lett.* **6**(3), 255–259, https://doi.org/10.1016/0165-1765(80)90024-5 (*1980*).
42. Shapiro, S. S., Wilk, M. B. & Chen, V. A. *Comparative Study of Various Tests for Normality. J. Am. Stat. Assoc.* **63**, 1343–1372, https://doi.org/10.1080/01621459.1968.10480932 (1968).
43. Wuertz, D., Setz, T. & Calabi, Y. *Rmetrics - Markets and Basic Statistics*, https://cran.r-project.org/web/packages/fBasics/fBasics.pdf (2015).
44. Henze, N. & Zirkler, B. A class of invariant consistent tests for multivariate normality. *Comm. Statist. Theory Methods* **19**, 3595–3617 (1990).
45. Korkmaz, S., Goksuluk, D. & Zararsiz, G. MVN: An R Package for assessing multi-variate Normality, https://cran.r-project.org/web/packages/MVN/vignettes/MVN.pdf (2018).
46. Szekely, G. & Rizzo, M. The distance correlation t-test of independence in high dimension. *J. Multivariate Anal.* **217**, 193–213, https://doi.org/10.1016/j.jmva.2013.02.012 (2013).
47. Hollander, M. & Wolfe, D. A. *Nonparametric Statistical Methods*. (John Wiley & Sons, 1973).
48. Dunn, O. J. Multiple comparisons using rank sums. *Technometrics* **6**(3), 241–252, https://doi.org/10.1080/00401706.1964.10490181 (1964).
49. Mandel, J. The validation of measurement through interlaboratory studies. *Chemom. Intell. Lab.* **11**, 109–119, https://doi.org/10.1016/0169-7439(91)80058-X (1991).
50. ILS: interlaboratory Study, https://cran.r-project.org/web/packages/ILS/index.html.
51. Wilrich, P.-T. Critical values of Mandel's h and k statistics, the Grubbs and the Cochrane test statistics. *ASTA-Adv. Stat. Anal.* **94**(1), 1–10, https://doi.org/10.1007/s10182-011-0185-y (2011).
52. Brown, M. B. & Forsythe, A. B. Robust tests for equality of variances. *J. Am. Stat. Assoc.* **69**(346), 364–367, https://doi.org/10.1080/01621459.1974.10482955 (1974).
53. Garwirth, J. L. *et al*. Tools for Biostatistics, Public Policy, and Law, https://cran.r-project.org/web/packages/lawstat/lawstat.pdf (2017).

Acknowledgements

This project was financially funded by The Research Office of Zayed University, United Arab Emirates (Project No. R 17081).

Author Contributions

A.B. introduced the methods and wrote the article performed the statistical analysis and mathematical modeling, Y.N. retrieved the data and wrote the introduction F.H. contributed to the State of the art and finding the references, Y.N. and F.H. reviewed the final form of the article before the submission.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-018-27727-0.

Competing Interests: The authors declare no competing interests.