ORİJİNAL ARAŞTIRMA ORIGINAL RESEARCH

Statistical Comparison of the Impacts of Weather Variables and Particulate Matter on Bacterial Density: Methodological Research

Hava Değişkenleri ve Partikül Maddenin Bakteri Yoğunluğu Üzerindeki Etkilerinin İstatistiksel Karşılaştırılması: Metodolojik Araştırma

Betül KAN KILINÇ[a], Büşra ERGİN[b], Meral YILMAZ CANKILIÇ[b]

[a]Department of Statistics, Eskişehir Technical University Faculty of Science, Eskişehir, TURKEY
[b]Department of Biology, Eskişehir Technical University Faculty of Science, Eskişehir, TURKEY

ABSTRACT Objective: In this study, the number of bacteriae in the air at five different regions in Eskişehir (Turkey) according to the settle plate method and some meteorological factors such as temperature and relative humidity were collected and analyzed statistically. The data were used to predict the amount of bacteriae using machine learning algorithms. **Material and Methods:** Air samples were taken from specified sampling points using the settle plate method. The random forest method, multivariate adaptive regression splines, support vector machine and ordinary least squares models were employed to analyze a qualified data set collected through 5 different monitoring locations in Eskişehir city, Turkey. All analyses were conducted using RStudio Software (version 1.2.1335). **Results:** The results revealed that all variables including weather characteristics and particulate matters were found statistically significant on the amount of the bacteria. Among modeling methods, the random forest model overperformed the others in terms of root mean square. **Conclusion:** Results might be useful for researchers studying bacterial density. These findings may provide policy-makers with crucial information for a better climate and health policy development, which helps address the conflict between development and air pollution, may be applicable in Turkey and other countries as well.

Keywords: Bacteria; particulate matter;
random forest method; statistical modelling

ÖZET Amaç: Bu çalışmada, Eskişehir (Türkiye) ilinde 5 farklı bölgede petri kapları kullanılarak; bakteri miktarları, sıcaklık, nem vb. bazı meteorolojik değerler elde edilmiş ve istatistiksel olarak analiz edilmiştir. Elde edilen verilerdeki bakteri miktarı, makine öğrenimi algoritmalarından bazıları kullanılarak tahmin edilmiştir. **Gereç ve Yöntemler:** Petri kabı yöntemi kullanılarak belirlenen örnekleme noktalarından hava örnekleri alınmıştır. Rastgele orman algoritması, çoklu adaptif regresyon eğrileri, destek vektör makinesi ve sıradan en küçük kareler yöntemleri, Eskişehir'deki 5 farklı bölgeden toplanan verileri analiz etmek için kullanılmıştır. Bütün analiz ve hesaplamalarda RStudio kullanılmıştır (versiyon 1.2.1335). **Bulgular:** Sonuçlar incelendiğinde, bütün değişkenlerin; çevre faktörlerin ve kirleticilerin, bakteri yoğunluğu üzerindeki etkisi istatistiksel olarak anlamlı bulunmuştur. Yöntemler arasında, hata kareler ortalamasının karekökü kullanılarak seçim yapıldığında rastgele orman algoritması en iyi sonucu vermiştir. **Sonuç:** Bu sonuçlar bakteri miktarı çalışan araştırmacılar için ileride yararlı olabilir. Bulgular, politika yapıcılara daha iyi iklimlendirme ve sağlık politikası geliştirme için önemli bilgiler sağlayabilir; kalkınma ile hava kirliliği arasındaki çatışmayı çözmeye yardımcı olabilir, benzer koşullar görülen diğer şehir ve ülkelerde de uygulanabilir.

Anahtar kelimeler: Bakteri; partikül madde;
rastgele orman modeli; istatistiksel modelleme

In the literature, there are various researches that examine the airborne particles that may cause allergies, asthma, pathogenic infections in respiratory ailments, respiratory diseases of humans, etc. In the upper air, most bacterial diseases are initiated by humans. Climate change in weather conditions and patterns of particulate matter (PT) in the air may also lead to some changes in human's health. Human's contribution to climate change is evident in the literature for the potential air pollution.[1] As an active role, human beings may control the climate change patterns and bacterial diseases to promote a healthy lifestyle.

Correspondence: Betül KAN KILINÇ
Department of Statistics, Eskişehir Technical University Faculty of Science, Eskişehir, TURKEY/TÜRKİYE
E-mail: bkan@eskisehir.edu.tr

PM (PM_{10}, $PM_{2.5}$) in the air is a dynamic component with many different inorganic, biological, chemical and physical properties. Many studies had been carried out in the world, including our country, to understand the possible effects of this PM, which is defined as carcinogenic in humans.[2] In these studies, mostly the non-biological fraction of PM has been addressed. But recently, the focus has been on the biological fraction of PM (bacteria, etc.). Very few studies have investigated the environmental factors influencing the structure of airborne bacterial communities.[3]

Bioaerosols are very small airborne particles (range 0.001 to 100 μm) that biologically originate from plants/animals and may contain living organisms.[4] Therefore, pathogenic and/or non-pathogenic dead or living microorganisms (eg. viruses, bacteria and fungi) can be found in bioaerosols.[5] Bioaerosols are easily migrated from one body to another due to their small size and light weight. Such particles can either be suspended in the air as individual organisms or attached to dust particles or small water droplets.[6] As Jones and Harrison stated, biological materials on the soil constitute approximately 25% of the total PM.[7] International interest in bioaerosols has increased recently, and it is aimed to expand the knowledge pool that includes the identification, amount, distribution and health effects of bioaerosols.[8]

PM has been measured in many studies examining the relationship between air pollution and human health.[9]

Burge and Rogers stated that climate change may alter the timing and abundance of aeroallergens and the growth and distribution of the organisms that produce them.[10]

Bertolini et al. have detailed described of airborne microbial communities in the urban region of Northern Italy and identified variations in bacterial community structure and composition according to the seasons.[3]

Haas et al. carried out his studies on the correlation of the number of particles in the Austrian center with the concentration of bacteria in the air.[11]

Menteşe et al., examined the spatial variation of bioaerosol levels in the city atmosphere of Ankara in their study, and also using culturable methods, *Micrococcus* and *Bacillus* bacteria were identified as the most frequently species in all districts.[12]

Little is known about the microorganisms in the air that represent a relevant part of atmospheric particles and whether these microorganism communities change according to parameters such as meteorological conditions, location, chemical content of the air. Studies on behalf of atmospheric microorganism communities in our country have been limited. However, with the thought that atmospheric microbial communities in urban areas rich in PM have negative effects on human health and the functioning of the ecosystem, interest in research on the atmospheric content of urban areas has been increased.

This research presents a contribution to researches to improve air quality models by estimating microbial pollution in the air with the most appropriate model to be selected according to performance measures. It examines the observed and predicted impacts of changes in major weather variables and transmission of PM_{10} and $PM_{2.5}$. The paper is organized into 5 sections. Section 2 introduces the sampling site and methods. Section 3 focuses on the statistical methods that are applied to the data set: 1) ordinary least squares (OLS) method, support vector regression (SVR), random forest (RF), and multivariate adaptive regression splines. Section 4 discusses the analysis results that may interfere or mediate the impacts of weather variables and PM on the amounts of bacteria. Section 5 summarizes the current research and the last section discusses their limitations and presents some future directions for scientists.

MATERIAL AND METHODS

SAMPLING SITE

Located in the city center of Eskişehir in Turkey; air sampling was carried out in different regions with urban density, including 5 monitoring stations such as, Yunus Emre State Hospital (39° 47.6'. 11''K, 30° 29'21.13''D), Train station (39° 46'47.80''K, 30° 30'19.64''D), Atatürk High School (39° 45'58.84''K, 30° 31'33.54''D), Eskişehir Courthouse (39° 46'4.59''K, 30° 32'3. 81''D), and City Hospital (39° 44'58.41''K, 30° 35'41.19''D) given in Figure 1.
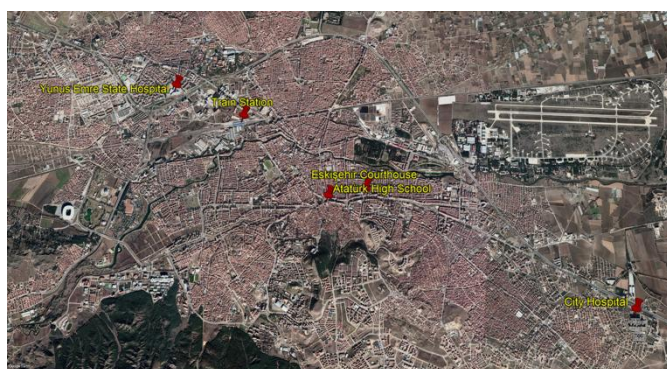

FIGURE 1: Monitoring stations.

Samples were collected in October, November, December, January, February, July, August and September in 2020 (sampling once a month at each station). The samples were not collected from March till August in 2020 due to coronavirus disease-2019 pandemic, a lockdown have been implemented in Turkey.

SAMPLING METHOD AND DETERMINATION OF BACTERIAL LOAD

Air samples were taken from specified sampling points using the settle plate method. According to the method, petri dishes containing solid nutrient media are left open for a certain period of time in the desired environment. Microorganisms carried by the inert particles fall on the surface of the nutrient medium with a certain rate of precipitation. Colony formation is observed in proportion to the level of microbial contamination in the air after incubation of the petri plates.

Mueller Hinton Agar (MHA) was used as the nutrient medium. Cycloheximide was added to the medium in order to prevent colony formation of all eukaryotic organisms except bacteria. Sterile media prepared in 90 mm petri dishes according to the manufacturer's instructions were transported to the sampling points in covered containers. Three petri dishes were sampled for each point. Samples were collected by leaving the lids open for eight hours at a height of 2 meters from the ground. After the specified time, the petri dishes were immediately collected and brought to the laboratory in covered transport containers. Sampled petri plates were incubated at 30-35 °C for 3-5 days for bacteria. Colonies formed after incubation were examined and their numbers were determined. The average of the numbers obtained from 3 petri dishes samples for each point was expressed as colony forming units per petri dish (CFU/petri) for each point in all months.

REGRESSION METHODS

Support vector machines, Bayesian classifiers, decision trees, genetic algorithms, learning with artificial networks, discriminant analysis, RFs, multiple adaptive regression splines (MARS) and k-NN algorithms etc are often used for classification and regression for different field of studies. The following section represents the current methods used in the analysis.

REGRESSION ANALYSIS

Regression analysis is a parametric method that estimates the relationship between one or more independent variables and a dependent variable. The method that is called OLS takes into account that the sum of squared errors is minimized. The objective function for multiple regression is minimized by the OLS method.

$$\sum_i^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y})^2 \qquad (1)$$

Here, e_i is the error term for the regression model that represents the sum of the deviations within the regression line. The dependent variable is called y whereas $\hat{y}$ is the predicted value of y.

RANDOM FOREST MODEL

One of the data mining tools is known as RF algorithm. It is similar to classification trees whereas a tree is constructed based on the splitting rules. Indeed, a large number of trees are constructed in RF algorithm using a randomized number of predictor variables. At each node of the tree, the predictors are selected to find the best split. After reaching the entire trees (default is maximum size of 500), predictions from all trees are combined. Advantage of using RF is that it produces a variable importance plot obtained from predictor variables.[13]

MULTIVARIATE ADAPTIVE REGRESSION SPLINES

MARS is a nonparametric approximation of the relationship between a response and a set of independent variables in reflected pairs of simple linear splines provided by Friedman, in 1991.[14]

MARS models do not assume any underlying relationship between the dependent variable and the predictors. This relationship is constructed by an adaptive fitting procedure where piecewise polynomials of degree q (splines) are driven from regressing the dependent variable onto predictors. The MARS algorithm adaptively selects the basis function set by 2 iterative approaches; forward and backward selection. It uses the residual squared error in iterations to compare the partition points. The criterion used to set the final model is a modified generalized cross validation (GCV) of the one first proposed by Craven and Wahba in 1979.[15]

In MARS, the multivariate regression model is formed as given in Eq. (2),

$$y = \beta_o + \sum_{m=1}^M \beta_m h_m(x_m) + \epsilon \qquad (2)$$

where the $h_m\ (x_m)$ are terms based on one or 2 predictors. These terms are created using a basis of Hockey Stick functions. For a single numeric predictor (x) are given by one of the two piecewise linear function defined below is called Hockey Stick functions,

$$(x - t)_+ = \begin{cases} (x_i - t_i) \; if \; x_i > t \\ 0 \qquad\quad if \; x_i \leq t \end{cases}$$

$$(t - x)_+ = \begin{cases} (t - x_i) \; if \; x_i < t \\ 0 \qquad\quad if \; x_i \geq t \end{cases} \tag{3}$$

terms are chosen from set of basis functions in pairs. These terms will consist of terms created from both forms $(X_j - t)_+$ and $(t - X_j)_+$ that are shown in [Figure 2](#).
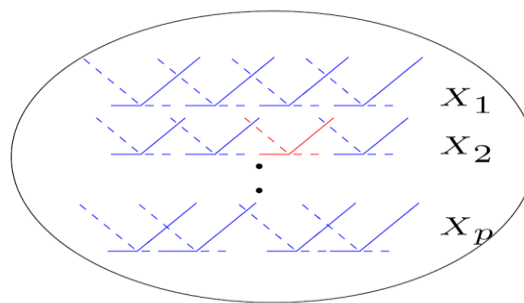

FIGURE 2: Basis functions.

Generation of the next term will either be another pair of basis functions based on predictor $(X_j, j = 1, ..., p)$, interactions terms based on another predictor (X_j) with the term $(X_j - t)_+$, or interaction terms based on (X_j) formed with $(t - X_j)_+$. This process carries on until either the maximum number of terms (M) is reached. The GCV criterion is also used to decide if a new term should be added or not. The GCV criterion given in Eq (4) penalizes models with too many terms.

$$GCV(\lambda) = \frac{\sum_{i=1}^{n}\left(y_i - \hat{f}_\lambda(x_i)\right)^2}{\left(1 - \frac{C(\lambda)}{n}\right)^2} \tag{4}$$

Here $C(\lambda)$ equals a model complexity measure which is the number of effective parameters in the model.[13]

SUPPORT VECTOR REGRESSION

SVR gives you the flexibility to define the error range. Therefore, the model is fitted to the data as long as the errors are within a range. The objective function of SVR is to minimize the model coefficients. The algorithm works for all data points as best as possible. However, some of the observations may fall outside the margins, as such it is needed to choose a tolerant value for errors.[16]

The principle of SVR is based on defining a hyperplane which can distinguish the 2 classes from each other in an optimal way. While it is possible to distinguish linearly separable data from a plane of the size to which they belong, it is possible to separate the nonlinearly separable data by a hyperplane by moving it to a higher dimension than its size.

When binary classification is considered for a data set that can be linearly separated, there are infinite numbers of hyperplanes that can distinguish it. While SVR creates the decision surface, it

tries to maximize the distance between the 2 classes. There is only one hyperplane with a maximum boundary between these planes. The hyper-planar optimal hyperplane, which subtracts the maximum to the maximum, is called the optimal hyperplane. The points that limit the border width are called "support vectors". The support vector algorithm tries to minimize the training error while classifying with the separator hyperplane with the largest boundary width. This method aims to have a linear discriminant function with the largest marginal difference between the classes. After the decision function is determined, the class to which the new instance belongs is determined according to the value it receives in the function. If the dataset cannot be separated linearly, the data are transported to another higher dimension where they can be linearly separated, and the classification is done in that space.

REGRESSION METRIC

It is very often to use mean square error (MSE) loss to evaluate the predictive model and root MSE (RMSE) is used to report its performance. The RMSE can be calculated as given in Eq. (5),

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2} \qquad (5)$$

DATA ANALYSIS

The previous studies have found that long-term climate warming tends to several infectious diseases.[17-19] These lead us to consider PM_{10} and $PM_{2.5}$ were as air pollutants and the possible weather conditions such as daily temperature, humidity and pressure as climate effects that may have impacts on human diseases. Those were provided as the independent variables in the analysis as weather variables that may also tend to increase the amounts of bacteria which was a count data. Air sampling is used to monitor the populations of airborne micoorganisms. As it helps scientists to inform the public about air quality, the bacteria data collected from 5 different monitoring stations were used as dependent variable in the analysis and totally 40 measurements used to explain how much of these potential effects are statistically significant. In order to carry out statistical analysis, it was thought to increase the number of samples and analyzes were made based on the daily amounts of bacteria (Mha), humidity (%), temperature (°C) and pressure (mBar) values obtained from 5 selected stations (Eskişehir, Turkey).

TABLE 1: The variables.

Variables	Minimum	Q_1	Mean	SD	Q_3	Maximum
Temperature	-1.00	6.00	16.75	8.74	28.00	33.00
Pressure	840.20	850.40	856.90	9.26	865.90	873.00
Humidity	26.86	28.16	32.99	8.74	37.09	48.62
PM_{10}	1.00	10.75	20.50	11.69	30.25	40.00
$PM_{2.5}$	1.03	2.69	18.64	15.71	29.84	50.55
Bacteria	7	90	154.3	94.17	229.8	358.0

SD: Standard deviation.

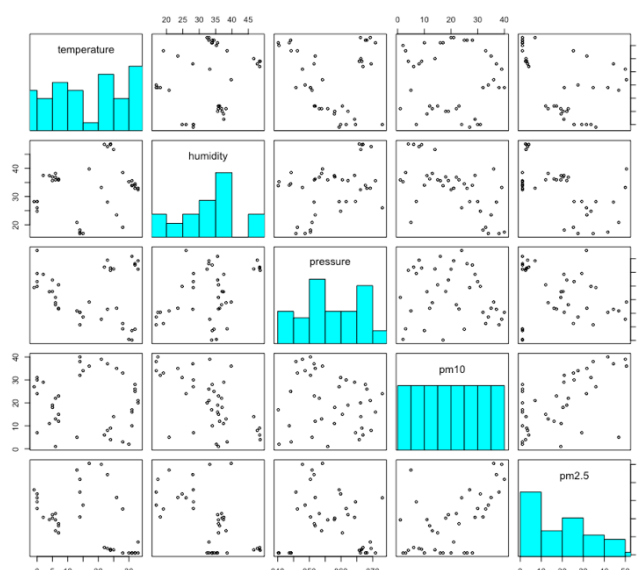The summary of the variables used in the analysis is given in Table 1.

FIGURE 3: Distribution of the variables.

The pair distribution of the variables are represented in Figure 3.

RESULTS

The proposed models statistically examined the effects of climatic conditions such as temperature, humidity and pressure and PM on the amounts of bacteria. The performance of the methods to estimate the amounts of bacteria, were obtained in terms of RMSE. Due to the difference in scales, all variables are centered and scaled.

OLS, SVR, MARS and RF models were implemented using RStudio. The optimum parameters were selected for RF and SVR by performing a search for all possible combinations of these parameters. RMSE were used in order to evaluate the performance of four methods. All of the independent variables found to be statistically significant in each model. The RMSE values of the four methods were obtained as given in Table 2.

TABLE 2: Performance of methods.

Method	Ordinary least squares	Support vector regression	Multiple adaptive regression splines	Random forest
Root mean square error	81.297	59.111	69.447	37.861

As it can be seen from the Table 2, the performance of the RF algorithm is more successful among others.

DISCUSSION

There are some studies which investigated the environmental factors influencing the structure of airborne bacterial communities. In these, the bacterial community has been classified using *polymerase chain reaction* method.[3] However in this study, not only the environmental factors but also the PM influencing the bacterial density has been investigated. In another study, Schwab et al. studied the gram-negative bacteria

montly and investigated the weather effects and conclued that where the warmer the temperature, there the more bacteria was.[20] In another study, nitrogen oxide, sulfur dioxide, volatile organic compounds, dioxins, and polycyclic aromatic hydrocarbons, heavy metals, PM were all considered air pollutants and climate change resulting from environmental pollution is shown as some of the effects on many dieseases.[21] The other study showed that airborne bacteria, fungi, and actinobacteria associated PM_{10} and $PM_{2.5}$, as well as their relationships with gaseous pollutants, O-3, SO_2 and NO_2, and meteorological factors could be used to determine the quality of PM. According to the results of the study, temperature was found significantly correlated with both PM_{10} and $PM_{2.5}$.[22] Seasonal variations to investigate fungal ait quality was studied by Wu et al. In their study, a multivarite regression model was used to predict the fungal dispersion associated with humidity, wind speed and temperature. In addition to that, the model proposed that temperature was negatively affected fungal dispersion whereas humidity and wind speed supported it.[23] These findings may provide policy-makers with crucial information for a better climate and health policy development, which helps address the conflict between development and air pollution, may be applicable in Turkey and other countries as well.

CONCLUSION

This study, evaluates and compares the performance of **OLS**, SVR, RF algorithm and multivariate adaptive regression splines to predict the amounts of bacteria at different monitoring stations. Data at intersections was collected and used to develop the models. Due to the worldwide pandemic, the measurements were collected from November 2019 to February 2020 and July-August 2020. The variables used were found statistically significant.

Hopefully, this study can serve as a support for more cross-disciplinary collaborations leading to mapping out the impacts of climate variable changes and PMs and the change in the amounts of bacteria.

Authorship Contributions

Idea/Concept: Betül Kan Kılınç, Büşra Ergin, Meral Yılmaz Cankılıç; Design: Betül Kan Kılınç, Meral Yılmaz Cankılıç; Control/Supervision: Betül Kan Kılınç, Meral Yılmaz Cankılıç; Data Collection and/or Processing: Büşra Ergin, Betül Kan Kılınç; Analysis and/or Interpretation: Betül Kan Kılınç; Literature Review: Büşra Ergin, Betül Kan Kılınç; Writing the Article: Betül Kan Kılınç, Meral Yılmaz Cankılıç, Büşra Ergin; Critical Review: Betül Kan Kılınç, Meral Yılmaz Cankılıç; Financial Resource: Bu çalışma, Eskişehir Teknik Üniversitesi (ESTÜ, Bilimsel Araştırmalar ve Projeler Komisyonu (BAP) tarafından fonlananan bir projenin çıktısıdır, kişiye ait bir fonlama bulunmamaktadır; Materials: Meral Yılmaz Cankılıç.

REFERENCES

1. Wu X, Lu Y, Zhou S, Chen L, Xu B. Impact of climate change on human infectious diseases: empirical evidence and human adaptation. Environment International. 2016;86:14-23. [Crossref] [PubMed]

2. Kan Kılınç B. The effects o industry increase and urbanization on air pollutants in Turkey: a nonlinear air quality model. Applied Ecology and Environmental Research. 2019;17(4):9889-903. [Crossref]

3. Bertolini V, Gandolfi I, Ambrosini R, Bestetti G, Innocente E, Rampazzo G, et al. Temporal variability and effect of environmental variables on airborne bacterial communities in an urban area of Northern Italy. Appl Microbiol Biotechnol. 2013;97(14):6561-70. [Crossref] [PubMed]

4. Georgakopoulos DG, Despres V, Frohlich-Nowoisky J, Psenner R, Ariya PA, Posfai M, et al. Microbiology and atmospheric processes: biological, physical and chemical characterization of aerosol particles. Biogeosciences. 2009;6(4):721-37. [Crossref]

5. Mandal J, Brandl H. Bioaerosols in indoor environment-a review with special reference to residential and occupational locations. Open Environ Biol Monit J. 2011;4:83-96. [Crossref]

6. Van Leuken JPG, Swart AN, Havelaar AH, Van Pul A, Van der Hoek W, Heederik D. Atmospheric dispersion modelling of bioaerosols that are pathogenic to humans and livestock-A review to inform risk assessment studies. Microb Risk Anal. 2016;1:19-39. [Crossref] [PubMed] [PMC]

7. Jones AM, Harrison RM. The effects of meteorological factors on atmospheric bioaerosol concentrations--a review. Sci Total Environ. 2004;326(1-3):151-80. [Crossref] [PubMed]

8. Kim KH, Kabir E, Jahan SA. Airborne bioaerosols and their impact on human health. J Environ Sci (China). 2018;67:23-35. [Crossref] [PubMed] [PMC]

9. Brunekreef B, Holgate ST. Air pollution and health. Lancet. 2002;360(9341):1233-42. [Crossref] [PubMed]

10. Burge HA, Rogers CA. Outdoor allergens. Environ Health Perspect. 2000;108 Suppl 4(Suppl 4):653-9. [Crossref] [PubMed] [PMC]

11. Haas D, Galler H, Luxner J, Zarfel G, Buzina W, Friedl H, et al. The concentrations of culturable microorganisms in relation to particulate matter in urban air. Atmos Environ. 2013;65:215-22. [Crossref]

12. Menteşe S, Rad YA, Arısoy M, Güllü G. Ankara şehir atmosferinde biyoaerosol seviyelerinin mekansal değişimi [Spatial variation of bioaeorol levels in the urban atmosphere of Ankara]. Ekoloji. 2009;19:21-8. [Crossref]

13. Hastie T, Tibshirani R, Friedman JH. The Elements of Statistical Learning. 2nd ed. New York: Springer; 2008. [Crossref]

14. Friedman JH. Multivariate adaptive regression splines. Annals of Statistics. 1991;19(1):1-67. [Crossref]

15. Craven P, Wahba G. Smoothing noisy data with spline functions. Numer Math. 1978;31:377-403. [Crossref]

16. James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning: with Applications in R. New York: Springer, Ch.10 Boosting and Additive Trees; 2018. p.337-72.

17. Epstein PR. Climate and health. Science. 1999;285(5426):347-8. [Crossref] [PubMed]

18. Rodó X, Pascual M, Doblas-Reyes FJ, Gershunov A, Stone DA, Giorgi F, et al. Climate change and infectious diseases: can we meet the needs for better prediction? Clim Chang. 2013;118:625-40. [Crossref]

19. Ostfeld RS, Brunner JL. Climate change and Ixodes tick-borne diseases of humans. Philos Trans R Soc Lond B Biol Sci. 2015;370(1665):20140051. [Crossref] [PubMed] [PMC]

20. Schwab F, Gastmeier P, Meyer E. The warmer the weather, the more gram-negative bacteria - impact of temperature on clinical isolates in intensive care units. PLoS One. 20145;9(3):e91105. [Crossref] [PubMed] [PMC]

21. Manisalidis I, Stavropoulou E, Stavropoulos A, Bezirtzoglou E. Environmental and health impacts of air pollution: a review. Front Public Health. 2020;8:14. [Crossref] [PubMed] [PMC]

22. Alghamdi MA, Shamy M, Redal MA, Khoder M, Awad AH, Elserougy S. Microorganisms associated particulate matter: a preliminary study. Sci Total Environ. 2014;479-80:109-16. [Crossref] [PubMed]

23. Wu DM, Zhang Y, Qin WW, Tian Y, Li AG, Hou LA, et al. Assessment of seasonal variations in concentration, particle-size distribution, and taxonomic composition of airborne fungi in a courtyard space. Atmospheric Pollution Research. 2021;12(2):113-21. [Crossref]