Çukurova Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi, 35(2), ss. 271-278, Haziran 2020
Çukurova University Journal of the Faculty of Engineering and Architecture, 35(2), pp. 271-278, June 2020

Air Quality Assessment by Statistical Learning-Based Regularization

Bülent TÜTMEZ[*1]

[1]*İnönü University, Engineering Faculty, Mining Engineering Department, Malatya*

Geliş tarihi: 17.04.2020 *Kabul tarihi: 30.07.2020*

Abstract

PM10 can be stated as a particulate matter smaller than 10 micrometer and it can be suspended in the air. The incremental concentration of PM_{10} affects both human and environment drastically. In this study, an air quality assessment by exhibiting the potential relationships among the secondary indicators and PM_{10} has been focused. For the analyses, statistical learning-based regularization procedures such as Ridge, the Lasso and Elastic-net algorithms have been practiced. In particular, use of Elastic-net algorithm in predicting PM_{10} concentration includes a novelty. As a result of the computational studies, it has been recorded that all the models showed high accuracy capacities. However, the elastic-net model outperforms the other models both accuracy and robustness (stability). Considering the error measurements (MSE and MAPE), the best numerical results have been provided by the Elastic-net model. Use of machine learning-based regularization algorithms in environmental problems can provide accurate model structures as well as generality and transparency.

Keywords: Regularization, Regression, Air quality, PM_{10}

İstatistiksel Öğrenmeye Dayalı Düzenlemeyle Hava Kalitesinin Değerlendirilmesi

Öz

PM_{10}, 10 mikrometreden daha küçük boyutta, havada askıda kalma özelliğine sahip parçacık madde olarak tanımlanabilir. PM_{10}'un çok yüksek konsantrasyonları insan ve çevreyi şiddetli biçimde etkiler. Bu çalışmada, hava kalitesinin değerlendirilmesi amacıyla, ikincil parametreler ile PM_{10} arasındaki ilişkilerin ortaya çıkarılmasına odaklanılmıştır. Analizler için istatistiksel öğrenmeye dayalı düzenleme yöntemleri olan Ridge, Lasso ve Elastic-net yordamlarından yararlanılmıştır. Özellikle Elastic-net yordamının PM_{10} tahmininde kullanımı yenilik taşımaktadır. Hesaplamaların sonucu olarak, bütün modellerin yüksek kestirim kapasitesine sahip oldukları kaydedilmiştir. Bununla birlikte, gerek kestirim başarısı ve gerekse de model gürbüzlüğü (duraylılığı) bakımından Elastic-net modeli diğer yöntemlerle karşılaştırıldığında daha başarılı sonuçlar vermektedir. Model hata ölçümleri (MSE ve MAPE) temel alındığında, en iyi sayısal sonuçlar Elastic-net modeliyle elde edilmiştir. Makine öğrenmesine dayalı düzenleme yordamlarının çevresel problemlerin değerlendirilmesi amacıyla kullanımı başarılı, genelleştirilmiş ve şeffaf model yapılarının oluşturulmasını sağlayabilecektir.

Anahtar Kelimeler: Düzenleme, Bağlanım, Hava kalitesi, PM_{10}

[*]Sorumlu Yazar (Corresponding author): Bülent TÜTMEZ, *bulent.tutmez@inonu.edu.tr*

1. INTRODUCTION

Air quality management has gained critical importance due to urban life and crowded cities. As a result of the industrialization and unrestrained population increase, new paradigms as well as scientific-statistical control mechanisms have accompanied these trends. Among the set points of the process, industry and agricultural works can be highlighted [1].

Air quality is referenced by critical measurements such as particle matters (PM_{10}, $PM_{2.5}$), SO_2, temperature, velocity, humidity, pressure [2]. Like in Turkey, the most countries use these indicators and evaluations are performed both at global (country) and local scales (city, town etc.) periodically. Thus wise, the pollution levels are determined and necessary precautions are taken by public authorities.

In the recent literature, various mathematical-statistical modelling tools have been examined for appraising particle matter concentrations in the air. One of the important indicators, fine particle $PM_{2.5}$ was investigated in different studies. Lai [3] focused on fine particle events and a quality index suggested based on fine particle matter. Nguyen et al. [4] performed a numerical assessment using baseline simulation aerosol effects. Thus, spatio-temporal variations of air quality parameters were appraised. Recently, Yatkin et al. [5] discussed the potential effects of fine particles on small urban domain. For this evaluation, both natural and anthropogenic sources have been utilized.

In parallel to fine particles, relatively coarse particles (PM_{10}) have been handled in various scientific works. One of these works, remote sensing-based estimation was conducted using air station data obtained from Ecuador [6]. In an interesting study, air quality was handled along with chronic stress and potential effects on human health were assessed [7]. Similarly, the influences of particulate matter concentrations (PM_{10}) on tourism have been inspected by a generalized additive model [8]. More recently, a forecasting-based study which uses time series and harmonic regression has been carried out for analyzing the PM_{10} variations in Ankara [9].

In general, the relationships between indicator variables and a target variable are analysed by multivariate regression methods. Although the traditional regression methodologies like linear least squares (LS) generate low bias, these are also sensitive against high variance. The LS fitting is mostly used to obtain a linear structure [10]. On the other hand, more robust tools are required for many problems due to interpretability. In particular, limited number of variables and more generality should be considered by a regression modelling. Besides this, air quality appraising built upon particulate matters includes many complexities due to natural variability and different sourced independent variables and a reliable data analysis should be performed based on reduced variability and high accuracy.

From a statistical point of view, performing high accuracy (regression) or clear pattern recognition (classification) under multi-collinearity conditions has critical importance. To model the complex systems based on reduced variance is also a required conditions for a high level identification. To overcome these problems, Ridge and Lasso (Least absolute shrinkage and selection operator) regularization paths were formulated to structure observations based on optimal coefficients both for regression and classification purposes [11,12]. Recently as a regularized regression method, Elastic-Net has been suggested for eliminating the drawbacks of Ridge and the Lasso techniques [13]. All the methods use shrinkage and regularization, and the relationships are revealed using penalty functions and adaptive parameters.

In this study, PM_{10} concentrations measured by the quality stations in a city are appraised by regularization paths. By this way, the potential complexity and the relationships are assessed by high level regression algorithms. Based on a numerical comparison, the use of relatively new regularization (Elastic-net) in environmental data analysis was objected. Therefore, the results have been given using both performance indicators and magnitude of the coefficients.

2. METHODOLOGY

2.1. Problem statement

Many observations are recorded at air quality stations such as Particulate Matters, SO_2, Temperature, Velocity, Pressure, and Humidity measurements. In general, air condition of a region is described by particulate matter level and distribution [14]. In this process, multivariate and simultaneous interdependent relationships should be focused and analyzed. Due to potential collinearities, high variance and natural uncertainties, more accurate and reliable modelling-classification tools are required.

2.2. Ridge Regression

In a traditional multivariate analysis, a matrix solution is the best way to provide the regression coefficients (Equation 1):

$$\hat{\beta}=\left(X^TX\right)^{-1}X^TY \tag{1}$$

The ridge estimator is constructed using an additive small constant as follows [15] (Equation 2):

$$\widehat{\beta_R}=\left(X^TX+\lambda I_p\right)^{-1}X^TY \tag{2}$$

In Eq (2), λ is a preliminary invariant, employed as a tuning parameter. $\widehat{\beta_R}$ is defined to minimize the penalized sum of squares (Equation 3):

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2$$

$$=RSS+\lambda\sum_{j=1}^{p}\beta_j^2. \tag{3}$$

In eq (3), RSS denotes the residual sum of squares. The term $\lambda\sum_{j=1}^{p}\beta_j^2$ is defined as shrinkage penalty. When $\beta_1,\ldots,\beta_p$ are close to zero, this penalty reduces. In eq. (3), the relative impact of the terms is expressed via the constant λ [16].

2.2. The Lasso

The ridge regression employs all the indicators in the resulting structure. This approach results in limited generality. To eliminate the drawbacks of ridge regression and to increase the model interpretability, the Lasso path was suggested [17].

The Lasso model shrinks down the model coefficients. It establishes models that simultaneously employ regularization to conduct feature selection [18]. The estimated Lasso coefficients $\widehat{\beta_L}$ can be stated as (Equation 4):

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\left|\beta_j\right|$$

$$=RSS+\lambda\sum_{j=1}^{p}\left|\beta_j\right| \tag{4}$$

The Lasso path utilizes an l_1 (norm) penalty in place of an l_2. In parallel to ridge analysis, the selection of a reliable λ has also determinative importance.

2.2. Elastic-net

Even though the Lasso is an effective variable selection method, it may contain several drawbacks [13]. If the number of variables is bigger than the number of measurements, the Lasso can select at most *N* variables. In the same condition, when the variables are correlated, ridge regression outperforms the Lasso. As an integration of the Lasso and ridge, elastic net performs best, because it obtains a strong combination of sparsity and regularization [19]. The following structure shows the objective function of the model (Equation 5):

$$L(\lambda_1,\lambda_2,\beta)=\left|y-X\beta\right|^2+\lambda_2|\beta|^2+\lambda_1|\beta|_1 \tag{5}$$

In Eq. (5), λ_1,λ_2 are fixed and non-negative. For $\alpha\epsilon[0,1)$ the elastic-net penalty can be provided (Equation 6 and 7):

$$\alpha=\lambda_2/(\lambda_1+\lambda_2), \tag{6}$$

$$(1-\alpha)|\beta|_1+\alpha|\beta|^2 \quad (7)$$

In the regularization system, the l_1 part of the penalty forms a sparse model. However, the quadratic part of the penalty performs the l_1 part more stable. Elastic-net regularization consists of two stages [13]:

- For each fixed λ_2, the ridge regression coefficients are determined,

- The Lasso-type shrinkage along the Lasso coefficient determination path is performed.

3. IMPLEMENTATION

3.1. Data Set and Structure Identification

The air quality data set includes the observations recorded in Eskisehir city by national authority within an action plan [20]. The data set covers the temporal measurements within the period February 2007 - December 2013. Due to the practical problems encountered in measurement processes [20] and outlier values, some of data were not able to consider. The data set comprising of 75 average values covers particulate matter (PM_{10}), SO_2 and meteorological parameters such as temperature, pressure, humidity and velocity [21]. The parameters considered are the general air quality assessment parameters referred in literature [2].

To reveal the pattern of the relationships in the data set, a series of bivarite median plots have been constructed as in Figure 1. The fences separate the observations. The bags consist of 50% of all observations. By this structure, the outliers can be exhibited. The relationships between each indicator variable and PM_{10} indicate that temperature, velocity and humidity effects are similar and high variability. There are no outliers. On the other hand, both pressure and SO_2 produce lower variance and some outliers. It should be noticed that the outlier provided by SO_2 bivarite plot may be resourced from recording or extra ordinary situation.

3.2. Results and Discussion

The substantial parts of the computations have been performed via the packages in R [22] such as glmnet [23] and Caret [24]. Data scaling in the regularization algorithms have been performed by [15] (Equation 8):

$$\tilde{x}_{ij}=\frac{x_{ij}}{\sqrt{\frac{\sum_{i=1}^{n}\left(x_{ij}-\bar{x}_j\right)^2}{n}}} \quad (8)$$

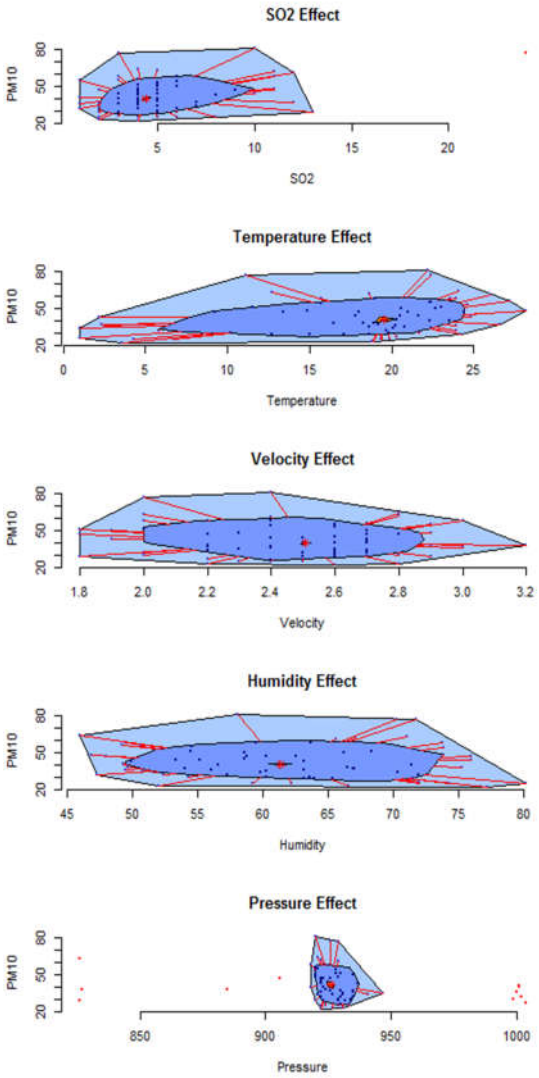

Figure 1. Bivariate plots for relationships

To determine the initial constants like λ, a grid was structured. using 6 x 75 matrix, with 6 rows (5 indicator variables plus intercept) and 75 columns (number of observations).

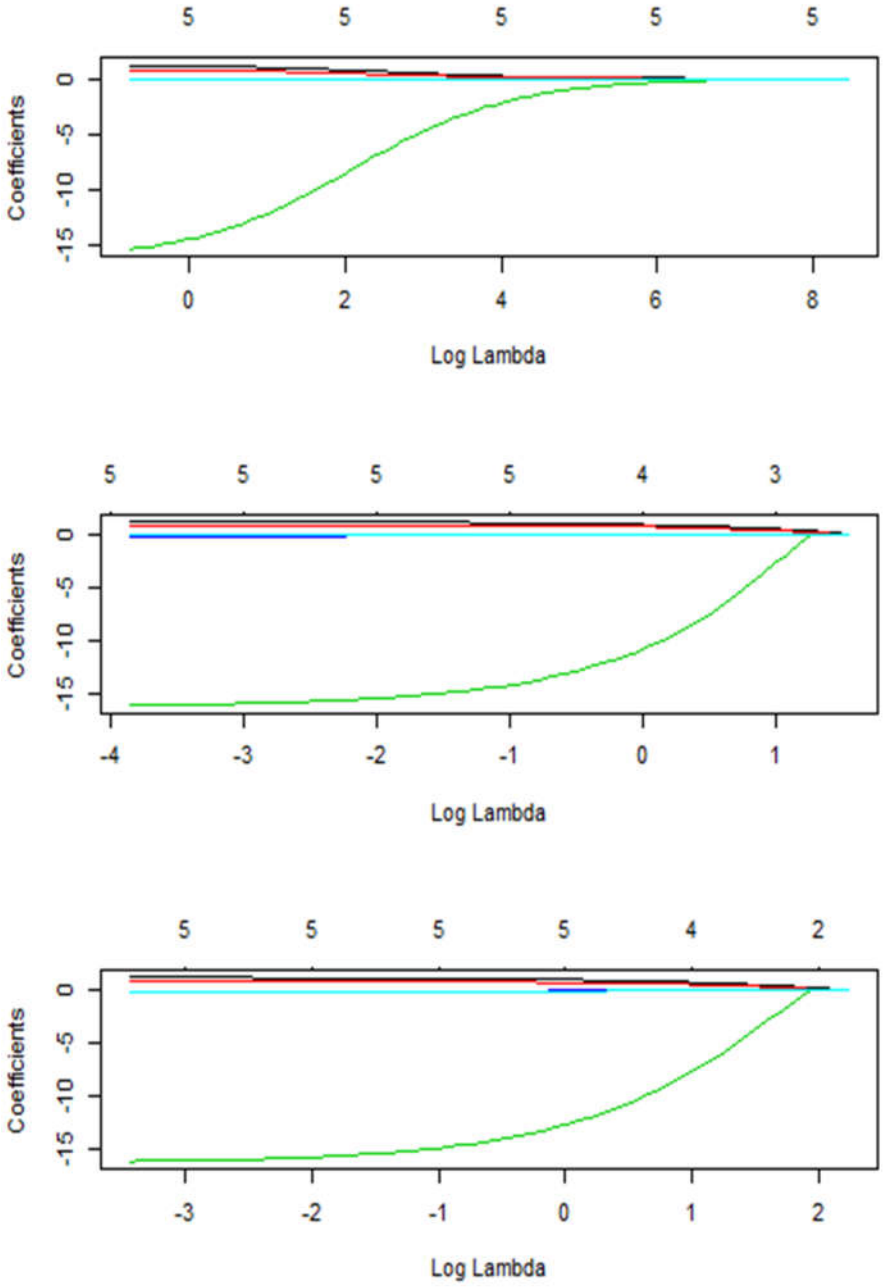

Figure 2. Coefficients provided by training models

To provide the model parameters and performance measures, first data set was splitted into two groups: 75% (Training) and 25% (Testing). Figure 2 indicates the optimized model coefficients

obtained by the λ grid values. The numbers on the plots refer the number of indicator variables using for determining the coefficients. These reference model parameters have been obtained using training observations.

In the second step, the critical parameter λ was optimized by ten-fold cross validation. Figure 3 indicates the cross validation-based parameter optimization structure.

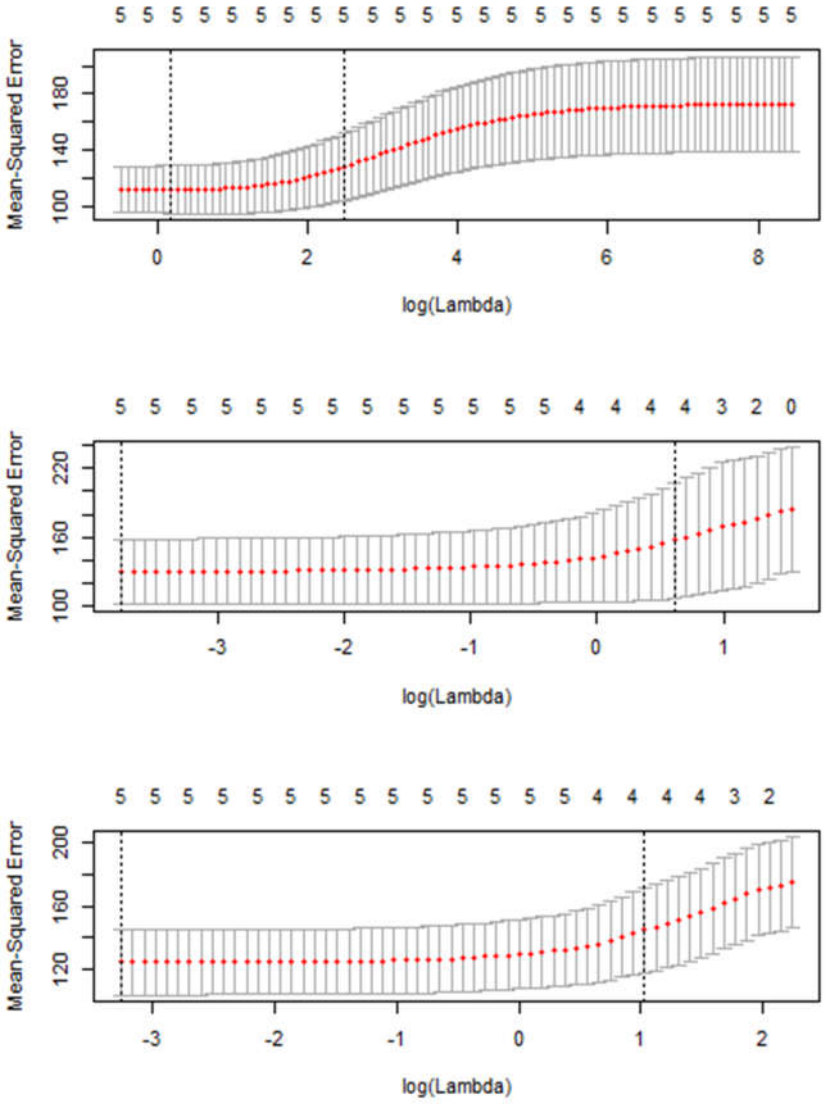

Figure 3. Cross Validation-based MSEs

In consequence of the simulations, the optimum tuning value has been provided by the smallest cross-validation error. In order to obtain the determinative parameter λ against MSEs, ten-fold cross validation has been conducted.

The model optimizations (final coefficients) are given in Table 1. The case studies showed that all the models have notable estimation capacities. The magnitudes of the coefficients include some potential for an explanation. As seen in Table 1,

the Elastic-net model explains the relationships via relatively smaller magnitudes comparing with the other models. This point has importance to analyse the degree of the variability.

Providing an effective bias-variance trade-off permits to minimize the model's total error [25]. From an error-based analysis, the best accuracy has been provided by the elastic-net model. Table 2 summarizes the testing model performances based on Mean Squared Error (MSE) and Mean Absolute Percentage Error (MAPE) measures. The lower MSE and MAPE refer the better accuracy. It should be noticed that even though the Ridge model employs 5 indicator variables, not only the Lasso but also the Elastic-net models use 4 independent variables in identification. This means that both the models have more generality comparing with the ridge model structure.

Table 1. Optimized model coefficients

Model	Equation
Ridge	PM_{10} = 95.96+0.61SO_2+0.23Tem-3.18Vel-0.14Hum-0.05Pre
Lasso	PM_{10} = 88.33+1.18SO_2+0.28Tem-0.20Hum-0.05Pre
Elastic-Net	PM_{10} = 41.68+0.78SO_2+0.23Tem-0.02Hum-0.01Pre

Table 2. MSE and MAPE performances on testing data

Model	MSE	MAPE
Ridge	146.32	0.253
Lasso	145.44	0.246
Elastic-Net	142.01	0.232

One of the main motivations of this study was specify the effects of air quality model parameters on particulate matter. Table 1 indicates that SO_2 and temperature parameters have additive effects. However, the rest of the parameters such as velocity, humidity and pressure have reducing effects.

Although the elastic-net model seems the best model from Table 2, the estimation capacities of the regularization paths are also very close. It should be noticed that both the Lasso and the elastic-net use limited parameters and these have more general structures than the Ridge model. Besides these, as a hybrid model structure, elastic-net confirmed more technical superiorities:

– Eliminating limitation on the number of selected variables;
– Stabilizing the l_1 regularization path.

4. CONCLUSIONS

Particulate matter (PM) addresses particles suspended in the air. The incremental concentration of PM_{10} and its detrimental effects on human and environment have gained attention in the world.
This study focused on revealing the potential relationships among the secondary air quality indicators and PM_{10} concentrations. For this purpose, high level regression procedures such as Ridge, the Lasso and Elastic-net regularization algorithms have been utilized. The case studies showed that all the models have huge capacities to specify the relationships. In particular, the elastic-net path can be suggested for the system including high number of variables. Due to generality and transparency, this hybrid model can be suggested for analyzing spatial-environmental processes.

5. REFERENCES

1. Mallik, C., 2019. Anthropogenic Sources of Air Pollution, in Air Pollution: Sources, ed. Impacts and Controls, Saxena, P., Naik, V., CABI. New Delhi.
2. Radzka, E., Rymuza, K., 2019. The Effect of Meteorological Conditions on PM10 and PM2.5 Pollution of the Air. Rocznık Ochrona Srodowiska 21(1), 611-628.
3. Lai, L.W., 2016. Public Health Risks of Prolonged Fine Particle Events Associated with Stagnation and Air Quality Index Based on Fine Particle Matter with Diameter <2.5 mμ in the Kaoping Region of Taiwan. Int. J. of Biometeorology, 60(12), 1907-1917.
4. Nguyen, G.T.H., Shimadera, H., Uranishi, K., Matsuo, T., Kondo, A., Thepanondh, S., 2019. Numerical Assessment of PM2.5 and 0-3 Air Quality in Continental Southeast Asia:

Baseline Simulation and Aerosol Direct Effects Investigation. Atmospheric Environment, 219, 117064.

5. Yatkin, S., Gerboles, M., Belis, C.A., Karagulian, F., Lagler, F., Barbiere, M., Borowlak, A., 2020. Representativeness of an Air Quality Monitoring Station for PM2.5 and Source Apportionment Over a Small Urban Domain. Atmospheric Pollution Research, 11(2), 225-233.

6. Alvarez-Mendoza, C.I., Teodoro, A.C., Torres, N., Vivanco, V., 2019. Assessment of Remote Sensing Data to Model PM10 Estimation in Cities with a Low Number of Air Quality Stations: A Case of Study in Quito. Ecuador, Environments, 6(7), 85.

7. Petrowski, K., Bastianon, C.D., Buhrer, S., Brahler, E., 2019. Air Quality and Chronic Stress a Representative Study of Air Pollution ($PM_{2.5}$, PM_{10}) in Germany. J. Occupational and Environmental Medicine, 61(2), 144-147.

8. Yoon, H., 2019. Effects of Particulate Matter (PM10) on Tourism Sales Revenue: a Generalized Additive Modelling Approach. Tourism Management, 74, 358-369.

9. Akdi, Y., Okkaoglu, Y., Golveren, E., Yucel, M.E., 2020. Estimation and Forecasting of PM10 Air Pollution in Ankara Via Time Series and Harmonic Regressions. Int. J. Environmental Science and Technology, https://doi.org/10.1007/s13762-020-02705-0.

10. Draper, N.R., Smith, H., 1998. Applied Regression Analysis, Wiley, USA.

11. Saleh, A.K.M.E., Arashi, M., Kibria, B.M.G., 2019. Theory of Ridge Regression with Applications, John Wiley & Sons, USA.

12. Tutmez, B., 2018. Bauxite Quality Classification by Shrinkage Methods, Journal of Geochemical Exploration, 191, 22-27.

13. Zou, H., Hastie, T., 2005. Regularization and Variable Selection Via the Elastic Net. Journal of the Royal Statistical Society, Series B:301-320.

14. Megaritis, A.G., Fountoukis, C., Charalampidis, P.E., Pilinis, C., Pandis, S.N., 2013. Response of Fine Particulate Matter Concentrations to Changes Ofemissions and Temperature in Europe. Atmos. Chem. Phys., 13, 3423–3443.

15. James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An Introduction to Statistical Learning, Springer, New York.

16. Dorugade, A.V., 2014. New Ridge Parameters for Ridge Regression. Journal the Association of Arab Universities for Basic and Applied Sciences, 15(1), 94-99.

17. Hastie, T., Tibshirani, R., Wainwright, M., 2015. Statistical Learning with Sparsity, CRC Press, Boca Raton.

18. Kuhn, M., Johnson, K., 2013. Applied Predictive Modelling, Springer, New York.

19. Khan, M.H.R., Anamika, B., Tamanna, H., 2019. Stability Selection for Lasso, Ridge and Elastic net Implemented with AFT Models, Statistical Applications in Genetics and Molecular Biology, 18(5), 10.1515/sagmb-2017-0001.

20. ÇŞB., 2014. Eskişehir İli Temiz Hava Eylem Planı, THEP (2014-2019), Eskişehir. (in Turkish).

21. Tutmez, B., 2019. Multivariate Statistical Control of Air Quality. 2. International Mersin Symposium, Mersin, 370-381.

22. R Development Core Team, 2008. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (ISBN 3-900051-07-0).

23. Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization Paths for Generalized Linear Models Via Coordinate Descent. J. Statistical Softwares, 33, 1–22.

24. Kuhn, M., 2008. Building Predictive Models in R Using the Caret Package. Journal of Statistical Software 28(5), 1-26.

25. Alpaydın, E., 2010. Introduction to Machine Learning, the MIT Press, Cambridge.