Türkiye Klinikleri Biyoistatistik Dergisi
Turkiye Klinikleri Journal of Biostatistics Turkiye Klinikleri J Biostat. 2021;13(1):57-69

ORİJİNAL ARAŞTIRMA ORIGINAL RESEARCH DOI: 10.5336/biostatic.2020-77137

Comparison of GAM and DLNM Methods for Disease Modeling in Environmental Epidemiology

Çevre Epidemiyolojisinde Hastalıkların Modellenmesi için Kullanılan GAM ve DLNM Metotlarının Karşılaştırılması

Mehmet KARADAĞ[a], Seval KUL[b], Saim YOLOĞLU[c], Mustafa BOĞAN[d], Behçet AL[e]

[a]Department of Biostatistics and Medical Information, Hatay Mustafa Kemal University Faculty of Medicine, Hatay, TURKEY
[b]Department of Biostatistics, Gaziantep University Faculty of Medicine, Gaziantep, TURKEY
[c]Department of Biostatistics and Medical Information, İnönü University Faculty of Medicine, Malatya, TURKEY
[d]Department of Emergency Medicine, Düzce University Faculty of Medicine, Düzce, TURKEY
[e]Department of Emergency Medicine, Gaziantep University Faculty of Medicine, Gaziantep, TURKEY

ABSTRACT Objective: In this study, it was aimed to compare the performance results of the methods modeled by using generalized additive models (GAM) and distributed lag non-linear models (DLNM) methods from real data of three different outcome variables of three separate diseases related to air pollution. **Material and Methods:** The data were retrospectively obtained from three hospitals under the General Secretariat of Gaziantep province public hospitals for a total of 1,916 days between 01 January 2009 and 31 March 2014. Response variables were number of the emergency unit admission, hospitalization and mortality due to asthma, chronic obstructive pulmonary disease (COPD) and pneumonia. The response variables were estimated by GAM and DLNM methods by building four different models and the performances of the models were compared. **Results:** When the estimation performances of GAM and DLNM methods are compared for each of the dependent variables in the prediction of hospitalizations due to asthma, GAM model IV [Akaike Information Criteria (AIC) (4,280.63)] values were found to perform the best. It was observed that DLNM method performed better than GAM in models established for the prediction of almost all other dependent variables. For when compare the odds ratio (OR) plot estimated on particulate matter (PM10); it was seen that GAM method made predictions with lower standard error compared to DLNM methods. **Conclusion:** When the models created with each dependent variable were compared; it was generally observed that superior performance was obtained from the DLNM method. However, the lowest standard error in the OR charts were observed in the models using the GAM method.

Keywords: Generalized additive models;
distributed lag non-linear models;
environmental epidemiology;
epidemiologic modeling

ÖZET Amaç: Bu çalışmada, hava kirliliğiyle ilişkili olduğu düşünülen, 3 farklı hastalığın 3 farklı sonuç değişkeni, gerçek veriler üzerinden genelleştirilmiş eklemeli modeller (GAM) ve dağıtılmış gecikmeli doğrusal olmayan modeller (DLNM) yöntemleri kullanılarak, modellenmesi ve model performanslarının karşılaştırılması amaçlanmıştır. **Gereç ve Yöntemler:** Gaziantep ili Kamu Hastaneleri Genel Sekreterliğine bağlı 3 hastaneden 1 Ocak 2009-31 Mart 2014 tarihleri arasında toplam 1.916 gün boyunca geriye dönük olarak izlenmesiyle elde edilen veriler kullanılarak oluşturuldu. Cevap değişkenleri; astım, kronik obstrüktif akciğer hastalığı (KOAH) ve pnömoni nedeniyle "acil servislerine başvurular", "hastanede yatış" ve "mortalite" sayısı şeklindedir. Tahminlerde GAM ve DLNM yöntemleri kullanılmış, aynı yöntemle kurulan 4 farklı modelden en iyi performansa sahip model, ilgili yöntem için karşılaştırma modeli olarak kullanılmıştır. **Bulgular:** Bağımlı değişkenlerin her biri için GAM ve DLNM metotlarının performansları karşılaştırıldığında, astım tanısı ile hastanede yatan sayısının tahmin edildiği IV. modelde GAM [Akaike Information Criteria (AIC) (4.280,63)] yöntemi en iyi performansı göstermiştir. Geriye kalan bağımlı değişkenlerin neredeyse tümünün tahmini için oluşturulan modellerde ise DLNM yönteminin GAM'den daha iyi performans gösterdiği gözlenmiştir. Partikül maddesi (PM10) üzerinden tahmin edilen göreceli olasılıklar oranı (OR) değerleri için GAM ve DLNM yöntemlerine göre daha düşük standart hataya sahip olduğu görülmüştür. **Sonuç:** İncelenen bağımlı değişkenlerle oluşturulan modeller kıyaslandığında, genelde DLNM yönteminin, GAM yönteminden daha iyi performans gösterdiği gözlenmiştir. Fakat PM10 üzerinden tahmin edilen OR grafiklerinde GAM yönteminin, DLNM yöntemlerinde daha düşük standart hata ve dolayısıyla daha dar güven aralığına sahip tahminler yaptığı görülmüştür.

Anahtar kelimeler: Genelleştirilmiş eklemeli modeller;
dağıtılmış gecikmeli doğrusal olmayan modeller;
çevre epidemiyolojisi;
epidemiyolojik modelleme

Correspondence: Mehmet KARADAĞ
Department of Biostatistics and Medical Information, Hatay Mustafa Kemal University Faculty of Medicine, Hatay, TURKEY/TÜRKİYE
E-mail: mehmet.karadag@mku.edu.tr

Peer review under responsibility of Turkiye Klinikleri Journal of Biostatistics.

Received: 04 Jun 2020 ***Received in revised form:*** 12 Dec 2020 ***Accepted:*** 14 Dec 2020 ***Available online:*** 29 Apr 2021

The population of our cities is increasing day by day. One of the negative consequences of irregular and uncontrolled crowding is increased air pollution.[1] Recently, many scientific studies have been conducted on the negative effects of air pollution on human health and the diseases causing it.[2,3] In studies conducted, a high degree of relationship was observed between air pollution exposure and respiratory diseases.[4] In our country, researching and modeling the effects of environmental factors on public health will provide benefits in many areas from effective use of health institutions' resources to estimating the population at risk of disease and taking precautions.[5]

There are many methods capable of modeling the trend of the answer variable with explanatory variables. Generalized linear model (GLM) is one of this models.[6] Additive modeling methods have become a suitable alternative to linear regression when there is no clear linear relationship between the response variable and estimators in the data we analyzed. The generalized additive model (GAM), which was introduced by Hastie and Tibshirani in 1990, is a model that makes inferences towards researchers in many environmental epidemiology.[7,8] GAM models, which are frequently used in the modeling of disease numbers, allowing the variable in the trend to be added to the model with the smoothing function, are a more advanced form of GLM models.[9] One important reason researchers choose GAM method is that they can attenuate solid linear assumptions between estimators and response variables.[2] GAM is a generalized linear model with a linear predictor involving a sum of smooth functions of covariates.[3] In fact, this method uses the smoothing curves to model the relationship between the dependent variable and the response.[1] A GAM such as a GLM uses a link function to establish the relationship between the mean of the response variable and a smoothed function of the explanatory variable(s). The most important feature of the GAM method is its ability to solve the linear and non-monotonous relationship between the independent variable set and the response variable.[4] This capability of GAM is based on the fact that it directly identifies the nature of the relationship between the explanatory variable set and the response variable instead of considering some parametric relationships in the data.[5]

In recent years, one of the statistical methods developed to measure the effects of air pollution on health is the distributed lag nonlinear models (DLNM) approach.[10,11] Recent studies have shown that a period lasting a few days the effects of extreme temperatures or high levels of pollution on human health are more than the effects of exposure on the same day.[12,13] Distributed lag non-linear models (DLNMs) are a modeling tool used to detect potentially non-linear and delayed dependencies.[14] In 2010, thanks to studies Gasparini has develop DLN and DLNM methodology in epidemiology studies. The term lag is expressed as on the scale of time between exposure and outcome. DLNM term, which is borrowed from on the time series analysis literature, correspond the time lull between the exposure event and the outcome when evaluating the delay of the effect within this time framework, the expose-outcome relationship can be defined as: a specific exposure events yields effects on various future outcomes.[10] DLNM is a method of modeling to define structures dependency nonlinear delay, also called the exposure-response relationship to the delay structure.[14] The DLNM method was developed in a framework that modeling both the exposure-response relationship and the past-time effect at the simultaneous.[11] The methodology of the DLNM is based on a two-dimensional space, which is called cross-basis, can define both the estimation space, and the relationship shape of the resulting lag effect dimension at the same time.[11] From this point of view, DLNM has become a new model that provides more flexible variations as well as creating a unified framework for the model range that have previously has been used in this setting.[10]

MATERIAL AND METHODS

This study was approved by İnönü University Scientific Research Ethics Committee (Health Sciences Non-entrepreneurial Clinical Research Ethics Committee) with 2018/2-8 protocol number. Our observations included “admission to the hospitals” and “hospitalization” and “mortality” statuses of patients aged 16 and above, who have applied to the Emergency Service of Dr. Ersin Arslan Training and Research Hospital, 25 December Public Hospital and Gaziantep University Medical Faculty Hospital, all part of the Public Hospitals of Gaziantep Province. Data were formed by using retrospective data for a total of 1916 days between 01 January 2009 and 31 March 2014. In studies on environmental epidemiology related to air pollution, models are generally designed on estimation of the number of patients.[15] In addition, in the models used, the response variable is presented as the daily admissions to the service, hospitalization or mortality.[16] Our dependent variables were "admission to emergency unit", "hospitalization" and "mortality" due to asthma, chronic obstructive pulmonary disease (COPD) and pneumonia. Factors related to air pollution that is thought to adversely affect human health are included in the model.[17] In this study, the average daily temperature (°C), PM10 (particulate matter) (µg/m^3), presence of desert storm, and the occurrence of the event on weekdays or weekends were added to the models as predictors.

GAM

In general, the model is as follows;

$$g(\mu_i) = X_i^* \theta + s_1(x_{1i}) + s_2(x_{2i}) + s_3(x_{3i}, x_{4i}) + \cdots \tag{1}$$

Where $\mu_i \equiv \mathbb{E}(Y_i)$, $Y_i \sim$ Exponential family (μ_i, ϕ). Y_i is dependent variable. X_i^* is a row of model matrix , θ is parameter vector, and s_i are the smoothing functions of dependent variables.[1,6] The s_i function in the model (1) is either a non-parametric smoothing function or has a regression spline structure.[2] In this type of models to cope with to over-fitting problem, are estimated by penalized maximum likelihood estimation.[7]

$$l(\eta) - \frac{1}{2} \textstyle\sum_j \lambda_j \int \left[s_j^{''}(x)\right]^2 dx. \tag{2}$$

Where$l$ is log likelihood of the linear estimator. The total term part is the wigglines measurement of the components of the GAM function.[3]λ_j are smoothing parameters, the trade-off between goodness of fit of the model and model smoothness is controlled the with λ_j. Practically Penalized Iteratively Reweighted Least Squares (P-IRLS) method is used instead of the Least Square Maximum Likelihood method when the model is fitted.[8] For instance, given the likelihood upon, at the t^{th} P-IRLS iteration the following penalized sum of squares would be minimized in the matter of η to discover the $(t+1)^{th}$ forecast of the linear predictor, $\eta^{[t+1]}$

$$\left\| W^{[t]} \left(z^{[t]} - \eta \right) \right\|^2 + \textstyle\sum_j \lambda_j \int \left[s_j^{''}(x)\right]^2 dx. \tag{3}$$

Where $W^{[t]}$ are iterative weights and $z^{[t]}$ are pseudo data, and are given by $W_{ii}^{[t]} = 1 / \sqrt{g^{'}(\mu_i^{[t]})^2 V_i^{[t]}}$ and $z_i^{[t]} = \eta_i^{[t]} + g^{'}(\mu_i^{[t]})(y_i - \mu_i^{[t]})$, where $V_i^{[t]}$ is proportional to variance of Y_i with respect to present estimate $\mu_i^{[t]}$.[7] $z^{[t]}$ is the vector of pseudodata. If there are categories variable in a model, the pseudo data should be used to insert model.[9]

The apparent difficulty of using the penalized likelihood approach is to the estimation of the smoothing parameter (λ). λ is a non-negative parameter selected by data analysis. $\lambda \rightarrow \infty$ give rise to a straight-line estimate for s, while λ=0 results in an un-penalized piecewise linear regression estimate.[3]

SMOOTHING FUNCTIONS

In models that explorer the effects of air pollution on health, trend estimation was made by adding the time variable to the model with the smoothing function.[4,8] GAM are an extension of GLMs. GAMs are sum of smoothing functions structures to the covariates.[9] There are many ways to add predictors to models with the help of smoothing function. Natural cubic spline is an example of this structures.[3] Natural Cubic Spline is a piecewise-cubic function in the range of [a, b]. This [a, b] range was divided regularly with k piece $\alpha = \xi_0 < \xi_1 < \cdots < \xi_{k-1} < \xi_k = b$. In our study, the time argument was added to the model by using the Natural Cubic Spline (ns) smoothing function.[9] The degree of freedom for ns is equal to the number of sub-intervals divided by the k-node, so the degree of freedom in the smoothing functions is determined by k.

DLNM

The time series outcomes are Yt and DLNM general model represent with equation (4):

$$g(\mu_t) = \alpha + \sum_{j=1}^{J} s_j(x_{tj};\beta_j) + \sum_{j=1}^{K} \gamma_k u_{tk} \tag{4}$$

Where Y_t is the dependent variable with $t = 1, \ldots, n$, $\mu \equiv E(Y)$, g is a monotonic link function and comes from a distribution belonging to the exponential family.The function s_j is smooth function of covariate x_{tj} and the parameter vector is β_j. The variables u_{tk} contain other predictors with linear effects specified by the related coefficients y_k. The s_j functions are similar to the smoothing functions used in GAM models.[10]

The response variable Y_t is daily death count in hospitals, assumed to arise from a overdispersed Poisson distribution with $E(Y) = \mu, V(Y) = \emptyset\mu$, and a canonical long-link in.[15] Here $\mu_t \equiv E(Y)$, where g is a uniform link function and Y is assumed to have an exponential distribution.[6,18] f_j indicates smoothed relationships between the linear estimators defined by the smoothing function and the parameter vector β_j and the variable x_j.[19] u_k variables, γ_k coefficients include other linear estimators whose relationships are defined.[20] f_j functions in the DLNM structure are similar to nonparametric methods such as GAM.[15,20] In time series analyses of environmental factors the outcomes Y_t are commonly daily counts, assumed to originate from a so-called overdispersed Poisson distribution with E(Y)=μ, V(Y)=∅μ and a canonical log-link in.[20]

EFFECTIVE DEGREES OF FREEDOM

The correct assignment of the degree of freedom is an important criterion for the smoothing parameter to better fit a model to the data.[11] Also, the effective degree of freedom (EDF) is an indication of the amount of smoothing. Accordingly, a smoothing operator with f_λ and τ is the degree of freedom of the function; $f_\lambda = \sum_j \lambda_j f_j$. One of the estimation approaches of EDF is calculated by the equality (5) and algorithm (6)

$$\tau = iz\{(X^TWX + f_\lambda)^{-1}X^TWX\} \tag{5}$$

An alternative approach to estimate the optimal EDF; $A = X(X^TX + f_\lambda)^{-1}X^T$ and $F = (X^TX +$ $f\lambda)-1X7X$ are effect matrices. Let's consider the mean of error squares for a model:

$$\mathbb{E}(\|y - Ay\|^2) = \sigma^2\{n - 2iz(A) + iz(AA)\}b^Tb \tag{6}$$

Here $b = \mu - A\mu$ is the bias of smoothing, and $\hat{b} = \hat{\mu} - A\hat{\mu}$ the estimate of b. It is also possible to reach variance prediction. $\sigma^2 = \frac{\|y - Ay\|^2 - \hat{b}^T\hat{b}}{n - 2iz(A) + iz(AA)}$ In the variance equation here, $\tau = 2iz(A) - iz(AA) = 2iz(F) + iz(FF)$,which is in the denominator, is the EDF of a model. In this approach, matrix A is considered to have an effect on the line matrix equation $H = X(X^TX)^{-1}X^T$. Thus, such a value of λ is assigned that the trace of the obtained matrix A equals the desired degree of freedom.[21] There are many different methods to make the optimal EDF decision. In addition, it was found that EDF assignment was made based on the number of seasons, trend structure and other similar structures of the variable to be estimated with the help of the smoothing function.[22]

In our study, different GAM models were tested in order to suggest the GAM model that can make the closest estimate to the true response variable, and the predictive performances of these models were compared. Below are three different model attempts for the dependent variable. These four model structures are applied in all dependent variables. Model structures are as follows:

$$\begin{aligned} ModelI\!: \log\{\mathbb{E}[Dependent\ variable_i]\} &= (Day\ of\ the\ week_i) + (Storm_i) + (Pm10_i) + (Pm10_{1i}) + (Pm10_{2i}) + (Pm10_{3i}) \\ &+ (Pm10_{4i}) + (Temperature_i) + Time_i \end{aligned}$$

$$ModelII: \log\{\mathbb{E}[Dependent\ variable_i]\}$$
$$= (Day\ of\ the\ week_i) + (Storm_i) + (Pm10_i) + (Pm10_{1i}) + (Pm10_{2i}) + (Pm10_{3i}) + (Pm10_{4i}) + (Temperature_i) + f(Time_i, df = 2, \text{"ns"})$$

$$ModelIII: \log\{\mathbb{E}[Dependent\ variable_i]\}$$
$$= (Day\ of\ the\ week_i) + (Storm_i) + (Pm10_i) + (Pm10_{1i}) + (Pm10_{2i}) + (Pm10_{3i}) + (Pm10_{4i}) + (Temperature_i) + f(Time_i, df = 6, \text{"ns"})$$

$$ModelIV: \log\{\mathbb{E}[Dependent\ variable_i]\}$$
$$= (Day\ of\ the\ week_i) + (Storm_i) + (Pm10_i) + (Pm10_{1i}) + (Pm10_{2i}) + (Pm10_{3i}) + (Pm10_{4i}) + (Temperature_i) + f(Time_i, df = 12, \text{"ns"})$$

In the first model, all estimators were added to the model without using the smoothing function, and the value of the PM10 variable on its day and the four-day delays (lags) were taken into account. In practice, a total of 21 seasons were observed in our time series. In the models, the time variable is added to the model with the help of the smoothing function and the degree of freedom is considered to be 21. However, in the scatter plot of the 21-node estimator and dependent variable prediction, it was observed that there was an over fitting problem due to the large degree of freedom value. Choosing the required smoother level for the predictor is the critical step in implementing GAMs and DLNM. This is ideally done by defining the level of smoothing by applying the principle of efficient degrees of freedom. Between the total number of observations and the total number of degrees of freedom used when fitting the model, a reasonable balance must be maintained.[12] With reference to previous studies, the degree of freedom of the time variable for Model II was determined as two.[23] In the III. model, six degrees of freedom was used to form it, since Wang et al. argued that it will increase the model prediction success on smoothing functions.[24] In the model IV, the degree of freedom was corrected with a value of 12, based on previous studies.[25] Also, the performance of three different smoothing functions, which are frequently used in environmental epidemiology studies in models, has been tested (Cyclic cubic regression splines, Natural Cubic splines (NS) and a Thin Plate Regression Splines and NS was used in the model because its best performance was observed in the NS function structure. These four corrected models were applied for all dependent variables in our study. It was also tested for GAM and DLNM methods and compared as to which models and methods performed better, based on Akaike Information Criteria (AIC), Adjusted Akaike Information Criterion (AICc), Bayesian Information Criterion (BIC) and Adjusted R^2 model performance criteria. It is a function in the mgcv package in the Anova gam R package program. This function used to test the statistical significance of one or more fit gam models is anova.gam. While this method is tested for a single model, it tests the significance of each parameter and smoothing term with the wald statistical method, and when it is examined for more than one model, it tests whether the performance of the fit models shows a statistically significant difference.[11] When new data are assigned and deleted instead of the observations containing 106 missing data, the prediction performances of all models are compared. Three structures are considered here. First, the model was fitted while there were missing observations, secondly, the model was fit by making assignments (imputation methods: mean, median and on a algorithm) instead of the missing, and finally the model was fit by deleting the missing observations. At the end of these model performance comparisons (with AIC, BIC and R^2). Similar performance results (AIC, BIC and R^2) were obtained from that three scenario. Finally, the initial data, which was 1916 days in total, was finally analyzed as 1,810 days. In the R package program, "mgcv" and "dlnm" packages were used for DLNM to analyze Generalized Additive Models. The R codes were provided in the additional file.

RESULT

From a total of 564,019 patients observed, 213,559 (37.87%) patients were diagnosed with asthma, 150,999 (26.77%) patients were diagnosed with COPD, and 199,461 (35.36%) patients with pneumonia. The descriptive statistics for PM10 and average temperature and values for 1,810 days, which included dust storms on 81 (4.50%) days, and weekdays on 1,290 (71.27%) days, are summarized in Table 1.

TABLE 1: Descriptive statistics for PM10 values and temperature.

Variables (n=1,810)	Mean±SD	*Minimum*	*M [Q_1 Q_3]*	*Maximum*	*IQR*
PM_{10} (µg/m^3)	88.6±58.7	11.0	72.0 [48.0 115.0]	631.0	67.0
Temperature (°C)	15.7±9.1	-2.6	14.5 [7.6 24.5]	34.9	16.9

PM10: Particulate matter; SD: Standart deviation; M: Median; IQR: İnterquartile range (Q_3-Q_1).

When Table 1 is examined, it was observed that PM10 (86.6±58.7) and the highest 631.0 µg/m^3 value was present. The average temperature (15.7±9.1) and the highest was 34.9 °C.

Total numbers of events for each outcome are given in Table 2. Total number for emergency room visit was very high for each outcomes and number of the mortality due to the asthma was quite low.

TABLE 2. Total numbers of events for each outcomes.

	Asthma	COPD	Pneumonia
Admission to emergency unit	211,352	145,004	188,401
Hospitalization	2,178	4,374	10,625
Mortality	29	1,621	435

COPD: Chronic obstructive pulmonary disease.

In order to suggest the GAM and DLNM models that can make the closest estimation to the real answer variable, four different models were determined, and the estimation performances of the related models were compared. Dependent variables in the models are the daily values of the number of patients admitted to the emergency and related services of the hospitals due to the cases of asthma, COPD and pneumonia, the number of patients hospitalized and the number of patients who have died. Four models were analyzed with GAM and DLNM and model prediction performance values such as AIC, corrected AIC, BIC and corrected R^2 were examined.

The performance values of the GAM and DLNM methods established to estimate the daily number of patients diagnosed with asthma as a result of admission to the hospital and the difference in statistics compared to the Model I are presented in Table 2. Model IV, which has a degree of freedom 12 with NS smoothing function in both GAM and DLNM methods, gave the best results. DLNM models with AIC (29570.14), Corrected AIC (29,540.25), BIC (29,611.43) and corrected R^2 (0.7912) values were observed to perform better.

When the patients hospitalized with the diagnosis of asthma are examined, the best performance in GAM and DLNM methods was observed in Model IV. It was observed that GAM models reached the best values in terms of AIC (4,280.63), corrected AIC (4,280.73) and corrected R^2 (0.4647). It was observed that GAM models had significance with Model I on the other hand, it was observed that statistical significance could not be achieved in the difference of DLNM models with Model I (p>0.05) (Table 3).

When the patients who died due to the diagnosis of asthma were examined, the best performance in GAM and DLNM methods was observed in Model III. DLNM models were observed to reach the best values in terms of AIC (290.45), corrected AIC (290.51), BIC (317.99) and corrected R^2 (0.0272). In GAM and DLNM methods, it was observed that significance in differences could not be achieved as compared with Model I (p>0.05) (Table 3).

TABLE 3. Performance results of models for the relationship between asthma disease and air pollution.

Variables	Models		AIC	AICc	BIC	R^2	p value
Admission to emergency unit	*I*	*GAM*	35,792.53	32,792.63	35,842.04	0.7135	-------
		DLNM	35,800.13	35,795.45	35,851.64	0.7025	-------
	II	*GAM*	35,623.01	35,623.11	35,672.52	0.7176	**0.001**
		DLNM	35,625.05	35,623.04	35,680.13	0.7170	**0.001**
	III	*GAM*	32,095.70	32,095.80	32,142.21	0.7602	**0.001**
		DLNM	35,096.70	32,095.73	32,163.18	0.7601	**0.001**
	IV	*GAM*	29,574.22	29,574.32	29,623.73	0.7911	**0.001**
		DLNM	29,570.14	29,540.25	29,611.43	0.7912	**0.001**
Hospitalization	*I*	*GAM*	4,422.92	4,423.02	4,472.43	0.4378	-------
		DLNM	4,420.32	4,422.95	4,450.43	0.4370	-------
	II	*GAM*	4,399.32	4,399.42	4,449.37	0.4332	**0.001**
		DLNM	4,399.00	4,399.35	4,426.83	0.4330	0.094
	III	*GAM*	4,322.10	4,322.20	4,371.61	0.4554	**0.001**
		DLNM	4,334.28	4,334.31	4,361.79	0.4514	0.113
	IV	*GAM*	4,280.63	4,280.73	4,330.14	0.4647	**0.001**
		DLNM	4,285.11	4,285.14	4,312.62	0.4633	0.095
Mortality	*I*	*GAM*	292.62	292.72	342.13	0.0178	-------
		DLNM	290.48	292.65	320.13	0.0170	-------
	II	*GAM*	291.37	291.47	340.88	0.0206	0.092
		DLNM	291.40	291.43	318.91	0.2040	0.094
	III	*GAM*	290.49	290.56	339.97	0.0271	0.115
		DLNM	290.45	290.51	317.99	0.0272	0.113
	IV	*GAM*	291.54	291.64	341.05	0.0212	0.095
		DLNM	291.54	261.57	319.05	0.0210	0.095

R^2: Corrected coefficient of determination; GAM: Generalized additive models; DLNM: Distributed lag non-linear models; AIC: Akaike information criterion; AICc: Corrected Akaike information criterion; BIC: Bayesian information criterion.

When the performance values for the emergency room visit due to COPD were analyzed, it was observed that in the GAM and DLNM methods, Model IV was superior to the performance values of the other previous three models. There were statistically significant differences in each DLNM model as compared to Model I ($p<0.001$). When DLNM and GAM model IV performances were compared, it was observed that DLNM method gave superior results in terms of AIC (21,432.48), corrected AIC (21,432.58), and BIC (21,481.99) (Table 4).

When the inpatients (hospitalized patients) who have been diagnosed with COPD are examined, the best performance in GAM and DLNM methods was observed in the Model IV. DLNM models were observed to achieve higher performances compared to GAM in terms of AIC (6,170.66), corrected AIC (6,170.76) and BIC (6,220.17). It was observed that the differences of DLNM models with Model I were statistically significant (Table 4).

When the patients who died with a COPD diagnosis are examined, the best performance in GAM and DLNM methods was observed in the Model IV. It has been observed that DLNM models reach the best values in terms of AIC (4,285.33), corrected AIC (4,285.43) and BIC (4,334.84). Statistically significant differences were observed in GAM and DLNM methods as compared with Model I ($p<0.05$) (Table 4).

TABLE 4. Performance results of models for the relationship between COPD disease and air pollution.

Variables	Models		AIC	AICc	BIC	R^2	p value
Admission to emergency unit	*I*	*GAM*	24,015.12	24,016.15	23,988.48	0.6045	------
		DLNM	23,737.69	23,737.79	23,787.20	0.6030	------
	II	*GAM*	23,690.82	23,690.92	23,740.33	0.6124	**0.001**
		DLNM	23,542.09	23,542.19	23,591.60	0.6673	**0.001**
	III	*GAM*	22,556.72	22,556.82	22,606.23	0.6520	**0.001**
		DLNM	22,444.06	22,444.16	22,493.57	0.6963	**0.001**
	IV	*GAM*	21,542.05	21,542.15	21,591.56	0.6854	**0.001**
		DLNM	21,432.48	21,432.58	21,481.99	0.6826	**0.001**
Hospitalization	*I*	*GAM*	6,699.61	6,499.71	6,549.12	0.4333	------
		DLNM	6,468.90	6,469.00	6,518.41	0.4348	------
	II	*GAM*	6,495.28	6,495.38	6,544.79	0.4282	0.196
		DLNM	6,461.58	6,461.68	6,511.09	0.4282	**0.002**
	III	*GAM*	6,209.43	6,209.53	6,258.94	0.4869	**0.001**
		DLNM	6,197.05	6,197.15	6,246.56	0.4853	**0.001**
	IV	*GAM*	6,177.99	6,178.09	6,227.50	0.4919	**0.001**
		DLNM	6,170.66	6,170.76	6,220.17	0.4909	**0.001**
Mortality	*I*	*GAM*	4,460.61	4,460.71	4,510.12	0.0261	------
		DLNM	4,462.28	4,462.38	4,511.79	0.0170	------
	II	*GAM*	4,334.99	4,335.09	4,384.50	0.0868	**0.001**
		DLNM	4,337.18	4,337.28	4,386.69	0.0780	**0.001**
	III	*GAM*	4,298.02	4,298.12	4,347.53	0.1010	**0.001**
		DLNM	4,294.34	4,294.44	4,343.85	0.0966	**0.010**
	IV	*GAM*	4,290.68	4,290.78	4,340.19	0.1055	**0.001**
		DLNM	4,285.33	4,285.43	4,334.84	0.1018	**0.002**

COPD: Chronic obstructive pulmonary disease; R^2: Corrected coefficient of determination; GAM: Generalized additive models; DLNM: Distributed lag non-linear models; AIC: Akaike information criterion; AICc: Corrected Akaike information criterion; BIC: Bayesian information criterion.

When the performance values for the emergency room visit due to pneumonia were examined, statistically significant differences were found in the GAM and DLNM methods compared to Model I and that the model IV was superior to the performance values of the other previous three models ($p<0.001$). Comparing their performance over GAM and DLNM over Model IV, GAM method has been observed to give superior results in terms of AIC (51,557.48), corrected AIC (51,557.58), BIC (51,606.99) and R^2 (0.4571).

When the patients hospitalized with the diagnosis of pneumonia are examined, the best performance in GAM and DLNM was observed to be the Model IV. DLNM models were observed to achieve higher performances compared to GAM in terms of AIC (10,740.25), corrected AIC (10,740.35) and BIC (10,789.76). When patients who died were examined, it was observed that the performance of the DLNM III model structure was superior as compared to the other models (Table 5).

TABLE 5. Performance results of models for the relationship between pneumonia disease and air pollution.

Variables	Models		AIC	AICc	BIC	R^2	p
Admission to emergency unit	*I*	*GAM*	66,642.09	66,642.19	66,691.60	0.2069	-------
		DLNM	69,690.15	69,690.25	69,739.66	0.2056	-------
	II	*GAM*	68,052.08	68,052.18	68,101.59	0.2327	**<0.001**
		DLNM	68,170.45	68,170.55	68,219.96	0.2308	**<0.001**
	III	*GAM*	59,048.34	59,048.44	59,097.85	0.3277	**<0.001**
		DLNM	59,383.44	59,383.54	59,432.95	0.3272	**<0.001**
	IV	*GAM*	51,557.48	51,557.58	51,606.99	0.4571	**<0.001**
		DLNM	51,889.13	51,889.29	51,938.64	0.4515	**<0.001**
Hospitalization	*I*	*GAM*	14,062.34	14,062.44	14,111.85	0.0445	-------
		DLNM	13,987.60	13,987.70	14,037.11	0.0467	-------
	II	*GAM*	14,062.11	14,062.21	14,111.62	0.0444	0.132
		DLNM	13,987.61	13,987.71	14,037.12	0.0467	0.196
	III	*GAM*	11,310.45	11,310.55	11,359.96	0.2850	**<0.001**
		DLNM	11,321.20	11,321.00	11,370.71	0.2810	**<0.001**
	IV	*GAM*	10,755.71	10,755.81	10,805.22	0.3822	**<0.001**
		DLNM	10,740.25	10,740.35	10,789.76	0.3811	**<0.001**
Mortality	*I*	*GAM*	2,116.22	2,116.32	2,165.73	0.0065	-------
		DLNM	2,104.43	2,104.46	2,131.94	0.0100	-------
	II	*GAM*	2,068.85	2,068.95	2,118.36	0.0354	**<0.001**
		DLNM	2,061.81	2,061.84	2,089.32	0.0354	**<0.001**
	III	*GAM*	2,063.22	2,063.32	2,113.96	0.0399	**<0.001**
		DLNM	2,051.24	2,051.27	2,078.75	0.0442	**0.010**
	IV	*GAM*	2,064.62	2,064.72	2,114.21	0.0391	**<0.001**
		DLNM	2,053.17	2,053.20	2,080.68	0.0415	**0.002**

R^2: Corrected coefficient of determination; GAM: Generalized additive models; DLNM: Distributed lag non-linear models; AIC: Akaike information criterion; AICc: Corrected Akaike information criterion; BIC: Bayesian information criterion.

In models with the best predictive success in three diseases modeled with GAM and DLNM methods, the predicted OR and confidence intervals of PM10 main effect and delays over an increase of 1 mg/m^3 increase are examined in figures. DLNM and GAM for lag0-4 also gave meaningful results in emergency room visit due to asthma (p<0.05), while GAM value with OR=1.0011 (95% CI=1.0001-1.0021) for Lag0 PM10 predictor was found to be significant; while it was observed that the prediction with DLNM model with OR = 1.0001 (95% CI=0.9999-1.0020) was not significant (Figure 1). Similar results were observed in models between COPD reference lag0, lag4 and ex (death) lag3, admission toe emergency unit due to Pneumonialag 0-1 and lag3 and hospitalized lag3 PM10 value and dependent variable (Figure 2). When the relationship between the COPD inpatient variable and the PM10 Lag0-4 asthma inpatient was modeled with GAM, the OR values could not be found significant, while the prediction made with the DLNM model was significant (Figure 1, Figure 2, Figure 3).

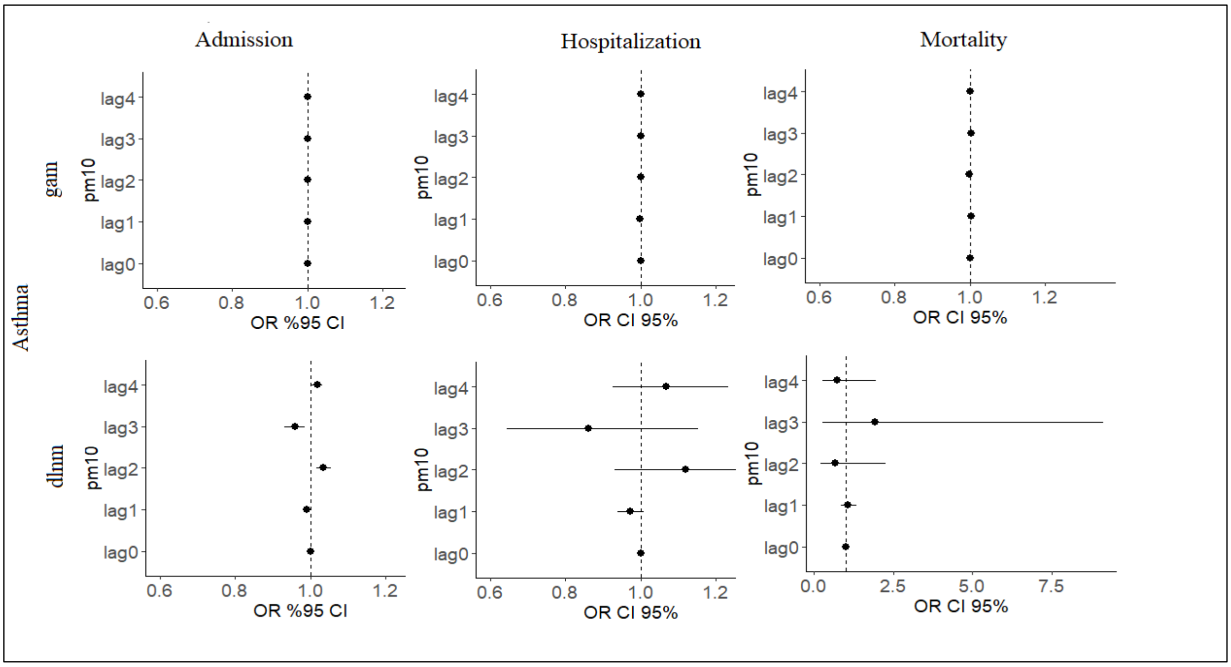

FIGURE 1: Odds ratio (OR) estimated values of 1 mg/m³ increase of particulate matter (PM10) variable with generalized additive models and distributed lag non-linear models for asthma disease. CI: Confidence interval.

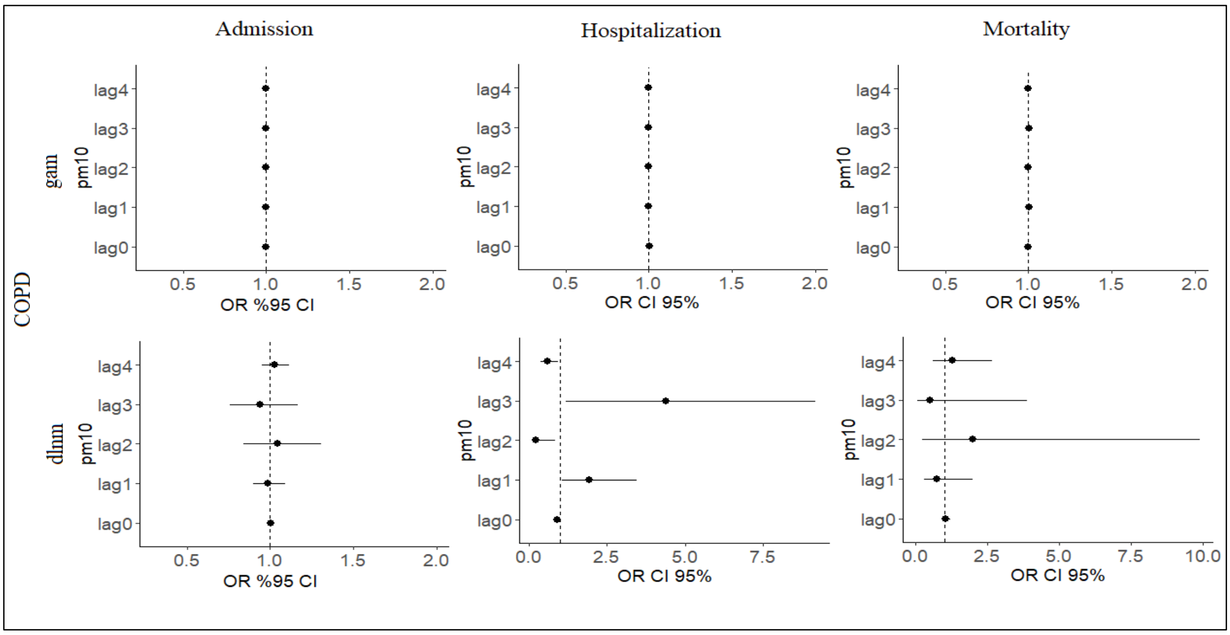

FIGURE 2: Odds ratio (OR) estimated values of 1 mg/m³ increase of particulate matter (PM10) variable with generalized additive models and distributed lag non-linear models for COPD disease. CI: Confidence interval.

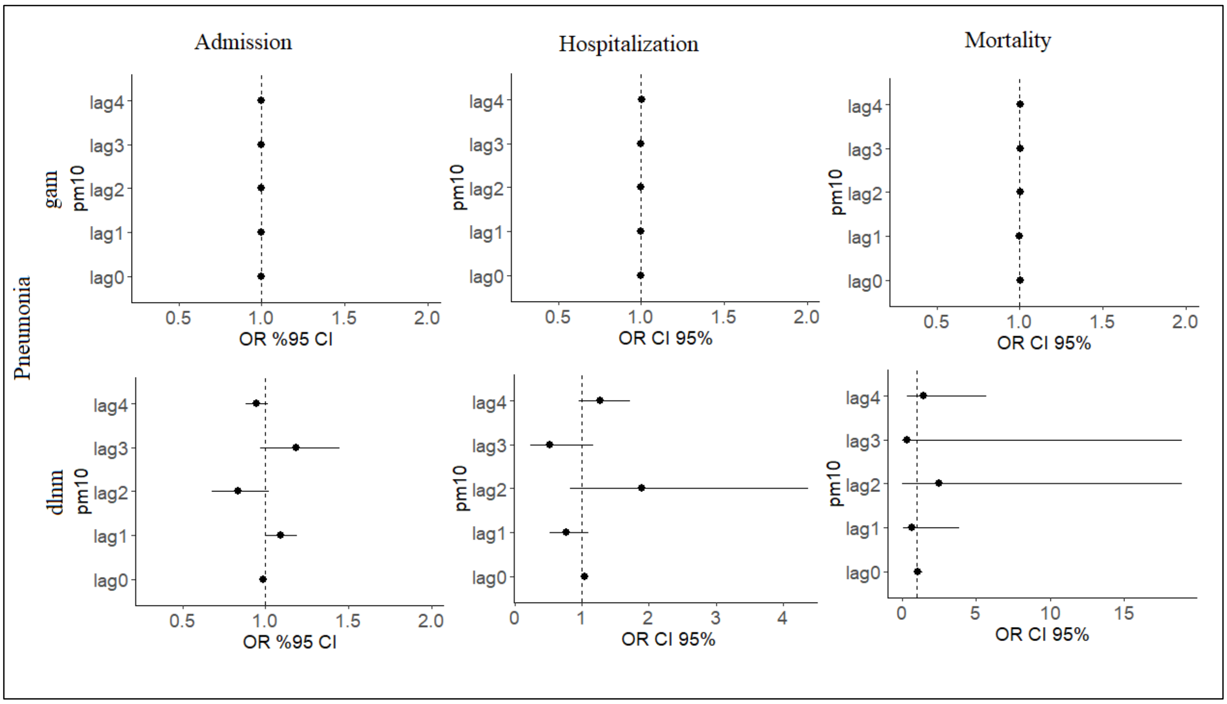

FIGURE 3: Odds ratio (OR) estimated values of 1 mg/m³ increase of particulate matter (PM10) variable with generalized additive models and distributed lag non-linear models for pneumonia disease. CI: Confidence interval.

DISCUSSION

Gaziantep is located in the southeast of Turkey; and it is exposed to the hot and dry weather during the summer. In addition, this region is seriously affected by the dust carried by desert storms from Arabia and Africa. The temperature accompanying the dust of desert storms that occur at frequent intervals, poses an important threat to human health.[16] For this reason, it was thought that Gaziantep would be a suitable example to examine the effects of environmental factors on human health.

When the estimation performances of GAM and DLNM methods are compared for each of the dependent variables; for the evaluation of seasonal and environmental effects on the numbers of people who have come to the hospital due to asthma; it is observed that DLNM Model IV with (AIC (29,570.14), BIC (29,611.43) and corrected R^2 (0.7912), has shown a higher performance as compared to GAM. In the prediction of patients hospitalized with the diagnosis of asthma, GAM IV model (AIC (4,280.63), corrected AIC (4,280.73) and corrected R^2 (0.4647)) values were found to perform better. DLNM Model III (AIC (290.45), corrected AIC (290.51) and corrected R^2 (0.0210)) in the prediction of asthma patients who have died. When the OR plot estimated on PM10 main effects and delays results were evaluated in Figure 2; In estimating the response variable of emergency visit due to asthma, it was seen that GAM method made predictions with lower standard error and thus a narrower confidence interval in DLNM methods.

In addition, it was observed that DLNM and GAM prediction directions were different in Lag3 and Lag4 estimates for emergency room visit due to the asthma. Another study similar to the results of our study, was conducted by Gasparrini in 2011. In the study, they concluded that GAM models with smoothing functions based on the penalize spline structure in DLNM related application would be a good alternative to DLNM.[19] Masselot et al. have seen that the prediction of the classic DLNM method that they have used, in which they have examined the relationship between weather values and cardiovascular mortality, has increased its success in prediction as compared to other models that they used.[26]

Gasparrini established a general conceptual and statistical framework for modeling exposure–lag–response associations, built upon the DLNMs in 2014. The same time the methodology is illustrated with an example application to cohort data and validated through a simulation study. The results of this simulation analysis are consistent with previous studies on one-dimensional models for correlations of exposure-lag-response assuming a linear relationship between exposure-response. The findings of this cohort data analysis show that a model with a B-spline cross-basis function is the best fitting option for both AIC and BIC. This model uses 9 df in total for expressing the bidimensional association.[27]

During the evaluation of environmental factors and seasonal effects on the number of patients admitted to the hospital for COPD, DLNM IV model performed better than other GAM models. When the OR estimates are analyzed, it is observed that DLNM estimates have a higher number of significant estimates, but the width of confidence intervals is high.

GAM IV model was superior in the prediction of patients presenting with a diagnosis of pneumonia as compared to other models. It was observed that DLNM model showed superior results in patients who were hospitalized due to pneumonia and died. While no significance was observed in any of the OR values estimated with DLNM, Lag1 as a reference variable predicted by GAM and Lag3 as a horizontal variable showed significance.

When the COPD inpatient variable and lag0-4 PM10 independent variable were modeled with DLNM, all of the OR results obtained were significant, whereas GAM models contained a value of OR confidence intervals.

In terms of model performance criteria, DLNM IV Model (AIC (29,570.14), BIC (29,611.43) and corrected R^2 (0.7912)) was observed to perform better than the DLNM III model GAM method, predicting the patients who died from asthma. In the prediction of patients hospitalized with the diagnosis of asthma, GAM (AIC (4,322.10), corrected AIC (4,322.20) and corrected R^2 (0.4554)) values were found to perform better. When the OR plot which was estimated on PM10 main effects and delays was evaluated in Figure 1, Figure 2, Figure 3; in estimating the response variable of emergency room visit due to asthma, it was observed that GAM method made predictions with a lower standard error and thus a narrower confidence interval in DLNM methods.

CONCLUSION

In this study, two models investigating the relationship between air pollution and respiratory diseases are compared. As a result of comparison, it is aimed to compare models' performances and propose suitable models to researchers. Firstly, independent from the methods, the performance of the models were better with degrees of freedom 6 and 12 in terms of AIC, BIC and R^2. Second, it was observed that the performance of DLNM was slightly better than GAM for all outcomes. Furthermore, confidence intervals of OR estimates made with DLNM methods on PM10 estimation over the dependent variable were wider than GAM, in other words, GAM models made predictions with lower standard errors. Finally, when number of the event corresponding to each day is small (in our case for mortality due to the respiratory disease), GAM performs better estimations than DLNM. So for rare events GAM model is advised.

Source of Finance

During this study, no financial or spiritual support was received neither from any pharmaceutical company that has a direct connection with the research subject, nor from a company that provides or produces medical instruments and materials which may negatively affect the evaluation process of this study.

Conflict of Interest

No conflicts of interest between the authors and / or family members of the scientific and medical committee members or members of the potential conflicts of interest, counseling, expertise, working conditions, share holding and similar situations in any firm.

Authorship Contributions

***Idea/Concept:** Mehmet Karadağ; **Design:** Mehmet Karadağ, Seval Kul; **Control/Supervision:** Saim Yoloğlu, Seval Kul; **Data Collection and/or Processing:** Behçet Al, Mustafa Boğan; **Analysis and/or Interpretation:** Seval Kul, Mehmet Karadağ; **Literature Review:** Mehmet Karadağ, Mustafa Boğan; **Writing the Article:** Mehmet Karadağ; **Critical Review:** Seval Kul.*

REFERENCES

1. Sundell J. On the history of indoor air quality and health. Indoor Air. 2004;14 Suppl 7:51-8. [Crossref] [PubMed]
2. Hales S, Blakely T, Woodward A. Air pollution and mortality in New Zealand: cohort study. J Epidemiol Community Health. 2012;66(5):468-73. [Crossref] [PubMed] [PMC]
3. Artun GK, Polat N, Yay OD, Üzmez ÖÖ, Arı A, Tuygun GT, et al. An integrative approach for determination of air pollution and its health effects in a coal fired power plant area by passive sampling. Atmos Environ. 2017;150:331-45. [Crossref]
4. Guarnieri M, Balmes JR. Outdoor air pollution and asthma. Lancet. 2014;3;383(9928):1581-92. PMID: ; PMCID: [Crossref] [PubMed] [PMC]
5. Patz JA, Engelberg D, Last J. The effects of changing weather on public health. Annu Rev Public Health. 2000;21:271-307. [Crossref] [PubMed]
6. McCullagh P. Generalized linear models. Eur J Oper Res. 1984;16(3):285-92. [Crossref]
7. Zuur A, Ieno EN, Smith GM. Additive and generalised additive modelling. Analyzing Ecological Data. 2nd ed. New York: Springer; 2007. p.120-33.
8. Wood S, Augustin NH. GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. Ecol Model. 2002;157(2-3):157-77. [Crossref]
9. Hastie TJ, Tibshirani RJ. Smoothing. Generalized Additive Models. 1nd ed. New York: Routledge; 1990. p.38-65.
10. Austin PC. A comparison of regression trees, logistic regression, generalized additive models, and multivariate adaptive regression splines for predicting AMI mortality. Stat Med. 200710;26(15):2937-57. [Crossref] [PubMed]
11. Wood S. Generalized Additive Models: An Introduction with R. 2nd ed. Florida: CRC Press; 2017. p.11-138. [Crossref]
12. Guisan A, Edwards TC, Hastie T. Generalized linear and generalized additive models in studies of species distributions: setting the scene. Ecol Model. 2002;157(2):89-100. [Crossref]
13. Yee TW, Mitchell ND. Generalized additive models in plant ecology. J Veg Sci. 1991;2(5):587-602. [Crossref]
14. Wood SN. Stable and efficient multiple smoothing parameter estimation for generalized additive models. J Am Stat Assoc. 2004;99(467):673-86. [Link]
15. Hastie T, Tibshirani R. Generalized additive models: some applications. J Am Stat Assoc. 1987;82(398):371-86. [Link]
16. Bayram H, Bogan M, Kul S, Oktay MM, Akpinar-Elci M, Al B, et al. Effects of desert dust storms and meteorological variables on emergency room visits and hospitalization due to COPD in South East Turkey. Am J Respir Crit Care Med. 2015;191:A6167. [Link]
17. Xia Y, Tong H. Cumulative effects of air pollution on public health. StatMed. 2006;25(20):3548-59. [Crossref] [PubMed]
18. Dobson AJ. An Introduction to Generalized Linear Models, 2nd ed: Florida: CRC Press; 2010. p.11-90.
19. Gasparrini A. Distributed lag linear and non-linear models in R: the package dlnm. J Stat Softw. 2011;43(8):1-20. [Crossref] [PubMed] [PMC]
20. Gasparrini A, Armstrong B, Kenward MG. Distributed lag non-linear models. Stat Med. 2010;20;29(21):2224-34. [Crossref] [PubMed] [PMC]
21. Rodriguez G. Smoothing and non-parametric regression. Springer; 2001:1-12. [Link]
22. Ma W, Sun X, Song Y, Tao F, Feng W, He Y, et al. Applied mixed generalized additive model to assess the effect of temperature on the incidence of bacillary dysentery and its forecast. PLoS One. 2013;29;8(4):e62122. [Crossref] [PubMed] [PMC]
23. Yang L, Qin G, Zhao N, Wang C, Song G. Using a generalized additive model with autoregressive terms to study the effects of daily temperature on mortality. BMC Med Res Methodol. 2012;12(1):165. [Crossref] [PubMed] [PMC]
24. Wang Q, Gao C, Wang H, Lang L, Yue T, Lin H, et al. Ischemic stroke hospital admission associated with ambient temperature in Jinan, China. PLoS One. 2013;19;8(11):e80381. [Crossref] [PubMed] [PMC]
25. Peng RD, Dominici F, Louis TA. Model choice in time series studies of air pollution and mortality. J R Stat Soc A Stat. 2006;169(2):179-203. [Crossref]
26. Masselot P, Chebana F, Bélanger D, St-Hilaire A, Abdous B, Gosselin P, et al. Aggregating the response in time series regression models, applied to weather-related cardiovascular mortality. Sci Total Environ. 2018;628:217-25. [Crossref] [PubMed]
27. Gasparrini A. Modeling exposure-lag-response associations with distributed lag non-linear models. Stat Med. 2014;28;33(5):881-99. Erratum in: Stat Med. 2014;28;33(5):900. [Crossref] [PubMed] [PMC]