JOURNAL OF MECHANICS OF CONTINUA AND MATHEMATICAL SCIENCES
www.journalimcms.org

ISSN (Online) : 2454 -7190 Vol.-15, No.-7, July (2020) pp 260-270 ISSN (Print) 0973-8975

NON LINEAR GENERALIZED ADDITIVE MODELS USING LIKELIHOOD ESTIMATIONS WITH LAPLACE AND NEWTON APPROXIMATIONS

Vinai George Biju[1], Prashant CM[2]

[1]Dept. of CSE, CHRIST Faculty of Engineering, Bangalore, India.

[1]Dept. of CSE, Sapthagiri College of Engineering, Bangalore, India.

[2]Dept. of CSE, Acharya Institute of Technology, Bangalore, India.

[1]vinai.george@christuniversity.in, [2]hod-cse@acharya.ac.in

Corresponding Author: Vinai George Biju

https://doi.org/10.26782/jmcms.2020.07.00021

Abstract

The Generalized Additive Model is found to be a convenient framework due of its flexibility in non-linear predictor specification. It is possible to combine several forms of smooth plus Gaussian random effects and use numerically accurate and wide-ranging fitting smoothness estimates. The Newton interpretation of smoothing provides standardized interval approximations. The Model assortment through additional selection penalties and p-value estimates is proposed along with bivariate combination of input variables capturing different non-linear relationship. The proposed extension includes, using non-exponential family distribution, orderly categorical models, negative binomial distributions, and multivariate additive models, log-likelihood based on Laplace and Newton models. The general problem is that there is not one particular architecture do everything with an exponential GAM family.

Keywords : Generalized Additive Model, Diabetic Retinopathy, Laplace, Newton Approximation

I. Introduction

Generalized additive models otherwise expand the standard linear models by making the association with non-linear predictor variables andthe expected value of Y. It implies that an alternate distribution can be connected in addition to the normal distribution for the essential random variants. The Gaussian representations could be used for various statistical applications, but it is to be noted that certain types of problems are not suitable. GAM eases the normal parametric presumption and helps to discover the complexity in association among the dependent and the independent variable that might otherwise be neglected. Most nonparametric approaches such as

thin plate smoothing spline and local regression method do not work well if a great deal of stand-alone non-linear variables is present in the model [XI]. Information sparsity among the input variables in this setting inflates estimation uncertainty. The problem of rapidly increasing variance in the increasing dimension is sometimes called the "dimensionality curse." There are at least two benefits of an additive approximation. First, since every additional term is calculated by means of a simpler univariate when the data is not uniformly approximated and hence the issue of dimensionality is solved out. Second, individual term calculations describe about the dependent variable variation with respective to the independent variables. The generalized additive models were designed in order to extend the additive model to a wide range of distribution families [VI]. Such models presume that the mean of the dependent variable is dependent on a non-linear connected additive predictor through a link function. Generalized additive models allow any member of the exponential distribution family, permitting the probability distribution response to be non-linear. The normal distribution, for example, might not be suitable for modeling discrete data such as counts or restricting data like quantities. Generalized models can therefore be used for a broader range of problems of data science. An additive component, random component and a link function applied to the two components are the features of generalized additive model. In similar situations, the linear and generalized additive models may be used which serve different analytical resolutions. Generalized linear models stress the inference and estimation of model factors, while GAM focus on non-parametrical data exploration. The GAM is seemed to be more fitting for finding the insights of the data set and for visualizing the correlation between the independent variable on the dependent variable. GAM applies local regression methods and the B-Spline with univariate smoothing elements and for bivariate smoothing modules. The generalized cross validation utility has been commonly used as a criterion for smoothing parameters for many non-parametric regression methods [II].

II. Design of Non-Linear Model Using GAM

Hidden patterns can be identified in the data by flexible predictor functions. Predictor function-regularization helps to avoid overfitting. The use of regularized, non-parametric functions eludes the pitfalls in linear models when dealing with higher order polynomial terms [XII]. This additive modeling process in which the effect of the predictive variables can be apprehended by a smooth function that can be nonlinear depending on the underlying pattern. The nonparametric terminology means that data defines the shape of the predictive functions as opposed to a typically small group of parameters as in case of parametric functions. This can enable the underlying predictive patterns to be more flexibly evaluated without knowing in advance what they are. If a regression model is additive, it doesn't rely on the values of the other variables in the model to interpret the marginal effect of the single variable. However, the ability to control the smoothness of predictor functions is an important feature of GAM. Through changing the degree of smoothness you can prevent wiggly illogical predictor functions [X]. In other words, we may create a predictive relationship essentially smooth in properties, although the dataset can imply it to be noisy. GAM can capture common nonlinear patterns missing from a

conventional linear model [I]. GAM could avoid issues where the nonlinear effects are typically identified by polynomials when parametric regression models are adapted. This leads to complicated model formulations with many terms that are correlated and conflicting outcomes. In addition, choosing the best model involves building a number of transformations, followed by a search algorithm for each predictor to select the best alternative. During model estimation, predictor functions are derived automatically. It is not essential learn what kind of functions that is needed upfront of data analysis. This approach saves time as well as allows finding patterns that a parametric model has skipped [XIII].

Regularization and Smoothing

The GAM system enables to monitor the degree of smoothing for predictor functions in order to prevent overfitting. The predictor variance and bias could be explicitly tackled by regulating the wiggliness in the predictor functions. In addition, the type of penalties used in GAMs relate to the regression using Bayes and to the regularization in L2. Smoother curves have a higher bias, and reduced variance, allowing to clearly stabilizing the bias-variance balance [IX]. Curves having reduced variance are typically more significant in sample testing and validation. Nonetheless, we can skip a major trend if the curve is too smooth.

The key foundations of GAM are smoothing and there exists various classes of smoothers for GAM that is used at a high level: smoothing splines, local regression, P-splines, B-splines and thin plate splines. Loess is part of the smoothers class in the nearest neighborhood. The simplest member of this clan based on loess is namely the running mean smoother which are moving averages and symmetric. Smoothing is attained by sliding a window through the data and calculation of average Y each step on the basis of the closest neighbors and the smoothness degree is defined by the window width [III]. While they are appealing because of their simplicity, mean smoothers have key problems that they perform below par in smoothing especially at data boundaries. This key issue need to be addressed to develop predictive models and therefore more elaborate choices like loess are necessary.

Smoothing splines develop smooth curves entirely in a different way. The smooth function is approximated by reducing the penalized sum of squares instead of using the nearest neighbor approach [IV]. The term enacts smoothness in measurement of the second derivatives ' integrated square. A wiggly curve therefore has significant second derivatives, whereas a straight line has second derivatives of 0. Therefore the squared second derivatives are simply used to calculate the curve's wiggliness. The smoothing parameter controls the balance between model performance and smoothness. The smoothing parameter works differently from the loess span parameter, which governs the window width, although they both serve the same end purpose. A natural cubic slope with knots at each point of data is known as a smoothing spline, has been shown to reduce the penalized sum of squares to a minimum. Nonetheless, smoothing splines pose a significant downside for prediction modeling: when dealing with large models it is not feasible to have knots at each data point [VIII]. Regression splines provide a more convenient alternative to the

smoothing splines. The main advantage of Regression splines is that it could be characterized as a linear arrangement of finite functions that are not governed by on the dependent variable Y.

Modeling Using MGCV

Three main modeling functions exist: the "bam" package in R programming is used for larger datasets and “gamm” for the estimation of widespread mixed additive models via “nlme” package, which provide access to a wide range of random effects and correlated features. The "jagam," package is used for Bayesian stochastic simulation. The‘s’ function is used for isotropic, univariate smoothing of numerous variables and to apply randomized properties. The ‘te’ function specifies the smooths of the tensor product built from any marginal smooths that are individually penalized [IX]. The ‘ti’ function is employed for specifying excluded tensor product interactions, facilitating smooth ANOVA models and their lower order interactions. The initial parameters for these functions are covariates that are smooth. Some additional parameters control the degree of smoothing. The significant fact is that ‘bs’ is a string that shows the category of basis. For example, "ds", "cr" on the spline of Duchon and spline of cubic regression respectively. It could be a vector for the tensor product if various categories of basis for various margins are needed. K is the dimension of basis or the dimension of the basis margin. In the case of tensor it can also be a vector that specifies an area of each margin. In a particular way, m defines the order of penalty and basis. Smoothers being shared has the same smoothing parameter if they are of the same kind of smoother. The “by” parameter of the variable should multiply the smooth or have a separate smooth copy at each point. For the mgcv evaluation functions, only a set of model matrix columns and one or more related penalties are taken into consideration. However, each smooth has a matrix process prediction function, which generates the matrix that maps the smooth coefficients to different covariates [VII].

Preliminary Modelling

The relatively skewed nature of the response values and its obviously positive amount indicate that some transformation may be appropriate for the use of a Gaussian error model. Taking a Gaussian model without transformation approves the GCV optimization having a successful convergence and the model to have a full rank. The constant assumption of variance presents clear hitches that the variance with the mean increases. The rate at which the variance of data increases with the mean is not easy to assess, but a simple informal approach can be used to obtain some guide, in the absence of a good physical model for the mechanism behind the relationship. In the diversified proportions, the major difference between models is in the 4th root of the response measure in which the model of the transformed data is approximately unbiased. The reaction scale itself is therefore downgrading. The log-gamma model is roughly unbiased just because the overall maximum penalized likelihood isn't necessarily unbiased, but consistent.

Variances of Non-Linear Functions

The prediction matrix parameters β do in fact provide an easy and general method to obtain variance estimates in conjunction with the simulation from the post-distribution of parameters, for any quantity as a consequent from the model, not just the measures which are linear in β. Initially the posterior distribution of the average response variable in the experimental interest region is assessed. As the measure of interest is on a response scale and not the linear predictor scale, the quantity of interest is not linear in β. A prediction matrix is acquired first to address this problem, which records the parameters in order to form the average of the non-linear predictor values. This method of simulation is quite universal: measurements of any predetermined quantity from the fitted model can easily be obtained from the posterior distribution. Therefore, the method is highly efficient in contrast to bootstrapping. One drawback is that the parameters for smoothing are considered to be fixed rather than uncertain in their estimation. This is corrected using the smoothing parameter uncertainty evaluated using maximum likelihood technique [XIV].

Methodology for Likelihood Estimations

It is assumed that the annotations of random variables is indicated as $X_i ... X_n$ and the joint p.d.f. specified as $f(x_i ... x_n; \theta)$ where θ is a vector of unfamiliar constraints of the probability function *f*. The perceived values could be derived into $f(\ )$ and the result is observed as a function of θ. The p.d.f with the statistics worked is identified as "likelihoods" of the factors. Assume that the *n* data x_i which is modelled as annotations of random independent variables X_i by the p.d.f.

$$f(x) = \begin{cases} (\theta+1)x^{\theta} & 0 < x < 1 \\ 0 & otherwise \end{cases} \tag{1}$$

where θ is an indefinite bound. The p.d.f. $f(\ )$ having the perceived data worked is measured as a utility of θ is entitled the likelihood utility of θ.

$$L(\theta) = \prod_{k=1}^{n} (\theta+1) x_k{}^{\theta} \tag{2}$$

The most likely value of the θ is to be evaluated using the hypothesis having $\theta = 0{:}1$ in relation to the unconventional form that the value of θ taking other values. The range of ideals of θ should be consistent with the statistics having a rational likelihood and it should be that the range of ideals has the likelihood at minimum of 10% of the MLE.

Maximum Likelihood with Penalized GAM

$$g(m_i) = P_i\theta + \sum_k NF_{ik} f_k \tag{3}$$

where $y_i \sim Exp(\mu_i, \phi)$. *Let* y_i be the response variable following exponential family of distribution, $\mathbf{P}_i$be a matrix representing parametric model having random coefficients μ, the function f_k be a smooth function of predictors, and NF_{ik} represent the non-linear. The model consist of altering coefficient regression models. The prototype is comprehensive as $g(m_k) = X_k\beta$ where **X** is a matrix holding **P** and the assessed forms of NF_{ik}, and b_k covers μ and the *fk* where $b_k(x)$ represents basis functions.

To elude over fitting, approximation is penalized likelihood by minimise $M_d(\beta) + \sum_j \lambda_j \beta^T S_i \beta$ where is M_d indicates deviance. Asymptotic MSE is improved by cross validation; nevertheless GCV λ converges gradually when compared to maximum likelihood. GCV has superior affinity to several minima and stark under smoothing.

Laplace Approximate of Likelihood

The function $F(D, \beta)$ indicates the joint distribution data **D** and β representing the parameter. To develop the likelihood, the β term is approximately integrated out. Substitute $F(D, \beta)$ with exponential log *f* of Taylor expansion about β . Estimate is locally sufficient for a Normal distribution to be integrated. The log estimated likelihood is indicated by

$$2I_j = 2I\beta + \log\left(\frac{f}{\phi}\right) - \frac{\beta^T f \beta}{\phi} - \log\left(H_e + \frac{f}{\phi}\right) + M_i \log(2\pi) \tag{4}$$

where $\mathbf{H}_e$is the Hessian matrix of thelog *f* function based on β .

Newton Reweighted Least Squares

Newton reweighted least squares includes minimizing:

$$L = \sum_i d_i (y_i - X_i\beta)^2 + \beta^T L\beta \tag{5}$$

In this case $d_i$could be negative. The prescribed solution $(X_i^T DX_i + L)\beta = X_i^T Dy$ is accustomed to be optimized. The value $\rho = \log(\lambda)$ is approximated using Newton's technique of optimization for the likelihood, l_r. It is to be noted that the ρ vector is parameterized for the log L_+ to be computable. The value for β_λ is solved by Newton reweighted least squares. The derivatives of β_λ is solved using inherent differentiation. Drag the derivatives out of l_r to compute the update for ρ .

Case Study: Diabetic Retinopathy

The data is part of "gss" package in R that provides clinical data as to whether or not diabetic patients have diabetic retinopathy [V]. There are three predictor variables namely: illness year's length (dur), body mass index (bmi), and percent blood-glycosylated hemoglobin (gly). The data reflects an overall hemoglobin concentration of glucose that is below 6 percent for non-diabetics, representing the normal long-term glucose levels.

III. Simulation Results and Analysis

Likelihood provides a good overall results using point estimation approach. The parameter values that optimize the likelihood function would be determined to be most consistent with the data based on the magnitude of the likelihood of evaluating how compatible the parameter values are with the data. The likelihood design is to consider the parameter values to optimize the likelihood and use them as the best parameter estimates. The solution refers to general approaches with the only distinction that it needs to consider more parameters than just one.

Taylor's theorem assures that the Newton equation can measure the optimum in case the original parameter values are close enough to maximum likelihood. Unluckily, this differs between problems. The preliminary approximation if in case is outlying from the likelihood then the Newton iteration may be divergent with calculations actually going beyond the likelihood. However, if log based likelihood is of a difficult form then the Hessian may not be definitely negative in some parameter space regions and therefore the quadratic estimate has no particular limit. In addition, if the initial parameter approximation is unfavorable i.e. both divergence and infinite Hessians can be present in the log-based model.

This could be avoided by considering ϕ not specifying the phase to be taken but merely to describe the path in the space of parameters to look for better parameter values. The calculation of log based characteristics at a couple of locations along the μ_k spectrum typically shows parameter values of a higher probability even if the addition of $\mu_k + \phi$ the chance. In this case, g can be viewed as the direction to find better parameters. This stage comes into role only if the initial value be unsuccessful i.e. the hessian is indefinite and the eigenvalue being positive. If the likelihood is found, this will appreciate the probability by a sufficiently minor step in the direction of g. Because the computational methods of the Newton kind are almost always implemented very easily it also requires numerically measuring the necessary derivatives rather than creating precise expressions for it.

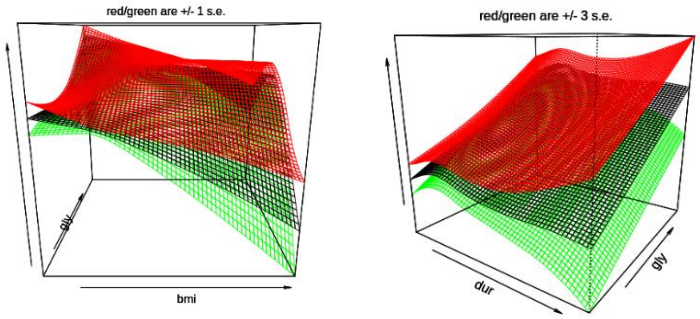


Fig. 1: bmi vs gly and dur vs gly: Non-Linear Predictor Surface with standard errors

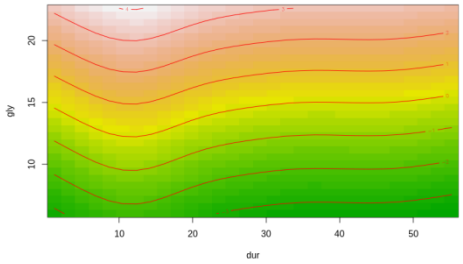


Fig. 2: A dur vs gly Non-Linear contour

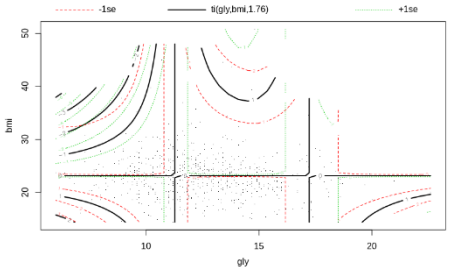
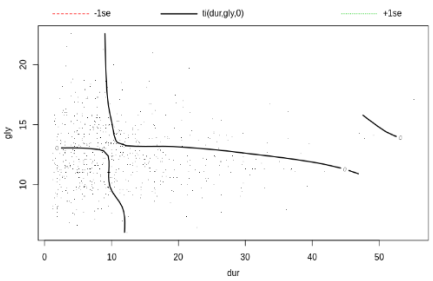


Fig. 3: B Smoothing Function for bmi vs gly



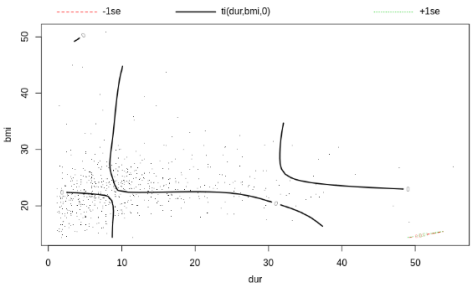


Fig. 4: Pairwise Smoothing Function for (dur vs gly) and (dur vs bmi)

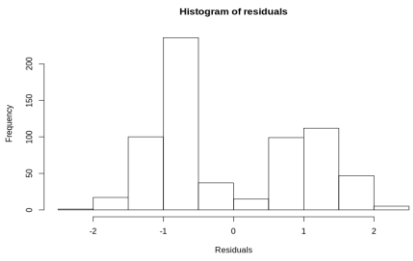


Fig. 4.A: Histogram of Residuals

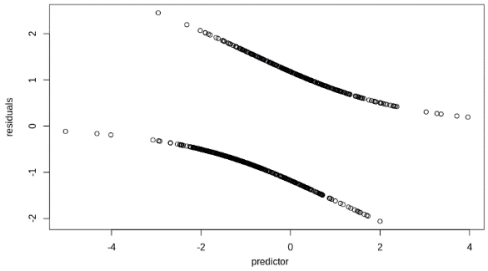


Fig. 4.B: Predictor vs Residual

Table 1: Statistical validation for Smoothing Variables

	k'	edf	k-index	p-value
s(dur)	9.00E+00	3.45E+00	1	0.52
s(gly)	9.00E+00	9.89E-01	0.96	0.13
s(bmi)	9.00E+00	2.36E+00	0.98	0.34
ti(dur,gly)	8.10E+01	2.10E-04	0.98	0.28
ti(dur,bmi)	8.10E+01	2.17E-04	1.01	0.62
ti(gly,bmi)	8.10E+01	1.76E+00	0.95	0.07

The GAM is found to be a convenient framework because of the flexibility in non-linear predictor specification. It is possible to combine several forms of smooth plus Gaussian random effects and use numerically accurate and wide-ranging approaches for fitting smoothness estimates. The Bayesian interpretation of smoothing provides standardized interval approximations. The Model assortment through additional selection penalties or p-value estimates is proposed over theAIC criteria which is otherwise prone to errors. Figure 1 indicates that “bmi vs gly” models the non-linear predictor surface with least standard error. The Figure 2.A indicates “dur vs gly” capturing marginal non-linearity when compared to other combination of variables. The Figure 2.B captures multiple nonlinear relationships indicating more confidence in “gly and bmi” to be used as a tensor product when compared to other variable combinations which can be seen in the Figure 3. GAM is executed with full convergence after 15 iterations. Gradient range was found to be [-3.065529e-05, 1.057611e-05] with score 385.7915 and scale 1. The Hessian is found to be positive definite and the eigenvalue was found to have a range of [2.734779e-07, 0.9274749]. Table I indicates the tensor product interaction with ‘gly’ and ‘bmi’ is statistically significant.The Histogram and QQ graph in Figure 4A and 4B is very evident to be a line indicating that the distribution is marginally normal. Variance is also suggested to be around constant as the mean rises. It is seen to have a positive non-linear correlation for response against formfitting values with a great deal of dispersion. Clearly, the likelihood model is preferably used to the model established on the norm of the transformed data if the response scale is of prime interest.

V. Conclusion

A common framework is build that is as practically useful framework using the multiple GAM extensions based on Laplace and Newton models. The multi-penalty smoothing approach is proposed that makes numerically accurate estimation of the smoothing function offering an extensive series of smoothness. The inferential

machinery is expanded for smooth configurations to include a stronger AIC model framework. This approach enables computationally accurate inferencing mechanisms using the GAM extensions. The approach also provides the benefit of allowing access for exponential family GAMs to the inferential machinery as proposed. The downside of this technique is that it requires derivatives of the likelihood of the parameters in the higher order to be evaluated. The simple penalty approach offers a wide range of seamless smoothing elements like random Gaussian model. The AIC measure estimates for the penalized regression is yet to be worked on in the near future.

VI. Acknowledgement

The authors would like to acknowledge Sapthagiri College of Engineering, Bangalore, VTU Research center for the support towards conduction of the research work.

References

I. Baquero OS, Santana LM, Chiaravalloti-Neto F. Dengue forecasting in São Paulo city with generalized additive models, artificial neural networks and seasonal autoregressive integrated moving average models. PloS one. 2018; 13 (4).

II. da Silva Marques D, Costa PG, Souza GM, Cardozo JG, Barcarolli IF, Bianchini A. Selection of biochemical and physiological parameters in the croaker Micropogoniasfurnieri as biomarkers of chemical contamination in estuaries using a generalized additive model (GAM). Science of The Total Environment. 2019 Jan 10; 647: 1456-67.

III. Diankha O, Thiaw M. Studying the ten years variability of Octopus vulgaris in Senegalese waters using generalized additive model (GAM). International Journal of Fisheries and Aquatic Studies. 2016; 2016: 61-7.

IV. Falah F, Ghorbani Nejad S, Rahmati O, Daneshfar M, Zeinivand H. Applicability of generalized additive model in groundwater potential modelling and comparison its performance by bivariate statistical methods. Geocarto international. 2017 Oct 3; 32 (10): 1069-89.

V. Gu C. Smoothing spline ANOVA models: R package gss. Journal of Statistical Software. 2014 Jun 30; 58 (5):1-25.

VI. Hastie T, Tibshirani R. Generalized additive models for medical research. Statistical methods in medical research. 1995 Sep; 4(3):187-96.

VII. Jiang Y, Gao WW, Zhao JL, Chen Q, Liang D, Xu C, Huang LS, Ruan LM. Analysis of influencing factors on soil Zn content using generalized additive model. Scientific reports. 2018 Oct 22; 8(1):1-8.

VIII. Li S, Zhai L, Zou B, Sang H, Fang X. A generalized additive model combining principal component analysis for PM2. 5 concentration estimation. ISPRS International Journal of Geo-Information. 2017 Aug; 6 (8): 248.

IX. Matsushima S. Statistical learnability of generalized additive models based on total variation regularization. arXiv preprint arXiv:1802.03001. 2018 Feb 8.

X. Pedersen EJ, Miller DL, Simpson GL, Ross N. Hierarchical generalized additive models: an introduction with mgcv. Peer J Preprints; 2018 Nov.

XI. Ravindra K, Rattan P, Mor S, Aggarwal AN. Generalized additive models: Building evidence of air pollution, climate change and human health. Environment international. 2019 Nov 1; 132: 104987.

XII. Tanskanen J, Taipale S, Anttila T. Revealing hidden curvilinear relations between work engagement and its predictors: Demonstrating the added value of generalized additive model (GAM). Journal of Happiness Studies. 2016 Feb 1; 17 (1): 367-87.

XIII. Wood SN. Generalized additive models: an introduction with R. Chapman and Hall/CRC; 2017 May 18.

XIV. Yoon H. Effects of particulate matter (PM10) on tourism sales revenue: A generalized additive modeling approach. Tourism Management. 2019 Oct 1; 74: 358-69.