5

Generalized Additive Models in Environmental Health: A Literature Review

Jalila Jbilou* and Salaheddine El Adlouni
Université de Moncton, Moncton, Canada

1. Introduction

Time series regression models are especially suitable in epidemiology for evaluating short-term effects of time-varying exposures. Typically, a single population is assessed with reference to its change over the time in the rate of any health outcome and the corresponding changes in the exposure factors during the same period. In time series regression dependent and independent variables are measured over time, and the purpose is to model the existing relationship between these variables through regression methods. Various applications of these models have been reported in literature exploring relationship between mortality and air pollution (Katsouyanni et al. 2009; Wong et al. 2010; Balakrishnan et al. 2011); hospital admissions and air pollution (Peng et al. 2008; Zanobetti et Schwartz 2009; Lall et al. 2011); pollution plumes and breast cancer (Vieira et a. 2005); diet and cancer (Harnack et al. 1997); and mortality and drinking water (Braga et al. 2001). Different time series methods have been used in these studies, i.e. the linear models (Hatzakis et al. 1986) the log-linear models (Mackenbach et al. 1992), the Poisson regression models (Schwartz et al. 2004), and Generalized Additive Models (Dominici 2002; Wood, 2006). The Generalized Additive Models represent a method of fitting a smooth relationship between two or more variables and are useful for complex correlations, that not easily fitted by standard linear or non-linear models.

The present chapter reviews The Genralized Additive Model (GAM), a class of statistical models which have commonly been used in time series regression, specially allowing for serial correlations, which make them potentially useful for environmental epidemiology.

2. Generalized additive models

The classic multiple linear regression model has the form:

$$Y = X\beta + \varepsilon \tag{1}$$

where Y is the response variable, X is the matrix (n×p) of the independent p variables $X_1,\ldots,X_p$, β is the vector of the parameters and ε is the vector of errors normally

* Corresponding Author

distributed with average 0 and variance σ^2 . Consequently, the variable Y is also Normal distribution with $E[Y] = \mu = X\beta$ and the covariance matrix $\sigma^2 I$ (I is the identity matrix). The linear models are central in applied statistics, mainly because of their simple structure and their interpretative ease. However, they present certain limits and are inadequate when the assumption of normality of the response variable is no longer justified. The linear model is extended to the Generalized Linear Model (GLM) to include a large class of the response variable distribution which belongs to the exponential family of distribution. The distribution Y is related to the linear combination of the covariables, $\eta = X\beta$, via the link function g(.), such as $g(\mu) = g(E[Y]) = \eta$.

To introduce more flexibility in the dependence structure between the response variables and covariables, the Generalized Additive Models (GAM), an extension of the GLM, replace the linear dependence functions by more flexible non-linear functions (Hastie and Tibshirani, 1990). The dependences are generally presented by non-parametric smoothing functions. The statistical inference consists on the estimation of the non-linear functions $f(X_j), j = 1,\ldots,p$, for each explicative variable X_j . This allows the identification of the specific form of the effect of each explicative variable on the dependant variable Y.

In practice, the objective is to model the dependence between the response variable, Y, and the explicative variables $X_1,\ldots,X_p$, for three main reasons: the description, the inference and the prediction. The goal is to find an explicit form of the effect $f(X_j)$ of each variable X_j on the variability of Y. The Generalized Additive Model (GAM) can be summarized by the flowing three components:

1. The random component: Y that follows a distribution of the exponential family and the mean and the variance are, respectively, $E[Y] = \mu$ and $\text{var}[Y] = \sigma^2$.
2. The systematic component: the explicative variables $X_1,\ldots,X_p$ that compose the regressor, defined by

$$\eta = \alpha + \sum_{j=1}^{p} f(X_j) \tag{2}$$

3. The link function g(.) is such that $g(\mu) = g(E[Y]) = \eta$, which implies that $E[Y] = \mu = g^{-1}(\eta)$.

The exponential family of distributions contains the Normal, Binominal, Gamma, Poisson, Geometric, Negative Binominal, and Exponential.

The non-linear functions $f(.)$ are usually represented by non-parametric dependence functions based on smoothing. The smoothing consists on creating a polynomial function that summarizes the data's tendencies. Some types of smoothing designed to express non-linear relations between the Y variable and the covariates Xj, j = 1… p, of the GAM models are the following: smoothing by scatter plot, parametric regression, mobile average, kernel smoothing and spline smoothing. A spline is a combination of polynomial functions. The knots are the points that mark the transition between the pieces of the polynomials (Eilers and Marx, 1996). The constraints allowing the joining of the polynomial pieces are defined by the number of continuous derivatives from the polynomial to the knots. The most

popular choice of the spline function is the natural cubic spline. It is a polynomial of the 3rd degree whose second derivative is zero at the limits. It offers less flexibility at the limits but this constitutes an advantage since the fit given by the regression spline presents a large variance around the limits (Hastie and Tibshirani, 1990). A smoothing B-spline basis is independent of the response variable Y and depends only on the following information: (i) the extent of the explicative variable; (ii) the number and position of the knots, and (iii) the degree of the B-spline. The properties of the B-splines are:

- It is formed of q + 1 polynomial pieces, each of q degree;
- The polynomial pieces are joined at q knots;
- The derivatives of order one to q at the joining points of the polynomial pieces are continuous;
- The B-spline function is positive on the q + 2 nodes extent and neutral elsewhere;
- It straddles 2q surrounding polynomial pieces (except at the edges);
- For all x, the q + 1 B-spline function are non-null.

One of the main advantages of the generalized additive model (GAM) is that it offers a great flexibility in order to represent the relations between the dependant variable and the explicative variables. Berger et al. (2004) present advantages related to the GAM to describe the relation between the use of the fluoroquinolone antibiotic and the resistance of the Staphylococcus aureus bacteria collected on the adult patients hospitalized for at least 48 hours. The dependant variable Y(t) of the model was the monthly number of cases in which the bacteria collected from the infected patient resisted to the fluoroquinolone and the explicative variables (Xm(t), m=1…p) were the monthly indicators of the antibiotics doses daily administered. The variable Y(t) follows the Poisson distribution P(λ), where the parameter λ corresponds to the average number of the cases per month and is function of the covariates. The link function is the logarithmic function and the regressor has the form $\lambda(t) = a + \sum_{m=1}^{p} f_m(X_m(t))$ in which $f_m(.)$ is a spline function. The results have shown the existence of a significant relation between the use of fluoroquinolone and the resistance of the bacteria.

The GAM models are used in the prognostic analyses of diseases. For example, Gehrmann et al. (2003) explored multiple sclerosis disease in order to identify the variables that have significant effects on the supported progression of the disease, to determine the intensity and the form of these effects and to estimate the survival curves. The use of Generalized Additive Models helped identify that among the available explicative variables; only the level of initial severity and the number of relapses during the twelve months preceding the study had significant effects on the hazard rate. The hazard rate h(t) means the probability of death after the time t, given that the patient has survived up to the time t.

In a study on the failure rate h(t) of patients with breast cancer (Hastie et al., 1992), the GAM model has been considered to identify among the prognostic factors those which presented significant non-linear relations with h(t). These prognostic factors are: the presence or absence of necrosis of the tumor, the size of the tumor, the number of samples examined, the patient's age, the body mass index and the number of days between the surgical intervention and the beginning of the study. Among these variables, the non-significant relation has been identified

with the age and the body mass index. The authors stated that the non-linear modeling had the advantage, firstly, of preventing against the false definition of the model which would lead to incorrect conclusions with regards to the effectiveness of a treatment, and also of provide information on the relation between the prognostic factors and the risk of disease which the standard (linear regression, normal distribution) models do not provide.

The GAM models are also employed in the analyses on the impact of climate and environmental variables on the public's health. In Quebec, a study of the impact of climate variables on mortality was conducted by Doyon et al. (2006). The number of daily deaths was modeled by the Poisson regression with a linking logarithmic function and the explicative climate variables selected were the humidity, the heat threshold and the functions of the average daily temperatures. A similar project carried out on European cities characterized by diverse climatic conditions arrived at the same conclusion of the existence of a significant relation between mortality and the temperature in several cities in Europe (Michelozzi P et al. 2007). The number of deaths and the number of hospital admissions were classified by age groups (15-64 years, 65-74 years, 75 and above years) and by cause (all causes – except death due to external causes –, cardiovascular diseases, cerebrovascular diseases, respiratory diseases, influenza). Considered climate variables are: temperature, dew point, wind speed, wind direction, pressure, total coverage of clouds, solar radiation, precipitations, and visibility. The variables of pollution were SO2, TSP (black smoke), PM10, NO2, and CO. The analysis was done separately for the warm season (April-September) and the cold season (October-March). This provides flexibility for the analysis, allowing the use of different model structures for each season (Terzi and Cengiz, 2009). Recently, Bayentin et al. (2010) used the GAM model to study the association between climate variables and circulatory diseases. The short term effect of climate conditions on the incidence of ischemic heart disease (IHD) over the 1989-2006 period was examined for Quebec's 18 health regions, with control for seasonality and socio-demographic conditions.

3. Parameter estimation

3.1 Local scoring procedure algorithm

The algorithm (presented in Appendix C.1) is summarized as an iterative and weighted process which allows the adjustment of a function f_j, $j = 1 \ldots p$, while keeping the other p-1 dimensions in their actual state. GAM models, in which the iterative algorithm is incorporated in S-Plus, became a popular analytical tool in epidemiology, especially in studies on the effects of environmental variables on public health (Dominici et al,. 2002). However, estimation by this algorithm presents problems of convergence and validity when the weighting matrix W (Appendix C.1) is not diagonal and if the independence hypothesis is not respected. Even if augmenting the number of iterations improves the estimations, the typical estimation errors remain difficult to evaluate and the model's effective dimension is statistically demanding (Wood, 2006). Many authors have suggested more direct approaches to remedy these problems.

3.2 Simultaneous estimation

The most effective way to estimate parameters is the use of a parametric GLM model with a limited number of regression splines or smoothing splines. This reduces the parameter

estimation problem in both cases to that of a GLM model with all its advantages related to the linear dependence functions. Despite the simplicity of the penalized GLM model, in the case of smoothing splines (Hastie and Tibshirani, 1990), the problem of the large system of equations remains. In the case of regression splines, each spline function is the function of the sum of the basis B-spline functions. This situation features the ease of B-spline construction, but the problem of the optimum choice in the position and number of B-spline nodes arises (Hastie and Tibshirani, 1990). Eilers and Marx (1996) have shown that this problem could be avoided by combining the B-splines to a differential penalty. In fact, the penalty is applied directly to the parameters in order to control the roughness of the spline functions. Criterion can be employed for the number of knots and the value of the penalty parameter.

When the P-spline are considered, the GAM has the form $g(\mu) = E(Y) = \alpha + \sum_{j=1}^{p} f_j(X_j)$ with $f_j(X_j) = B_j A_j$ and a response variable distribution belongs to the exponential family. In this section, B_j, $j = 1\ldots p$ is the B-spline matrix (with n_j knots) of $N \times n_j$ dimension, A_j is the n_j-vector of the basic B-spline function coefficients and then represents the part of the the variability of Y explained by X_j. The model can be rewritten as follows:

$$E[Y] = g(\mu) = B\,A \tag{3}$$

where $B = \left[1\ B_1\ B_2 \ldots B_p\right]$ and $A = \left(\alpha, A_1, \ldots, A_p\right)$. We are left with a GLM model and the estimation of the parameters α by maximization of the penalized log-likelihood is done by the penalized GLM Fisher scoring, below, until the desired convergence criterion is obtained.

$$\hat{A}_{t+1} = (B'\hat{W}_t B + P)^{-1} B'\hat{W}_t \hat{z}_t \tag{4}$$

where

$\hat{W} = diag\left\{\dfrac{[h'(\hat{\eta}_i)]^2}{Var(Y_i)}\right\}$, $\hat{z}_i = \hat{\eta}_i + \dfrac{(y_i - \hat{\mu}_i)}{h'(\hat{\eta}_i)}$ and $\boldsymbol{P}$ = blockdiag$(0, \lambda_1 P_1, \ldots \lambda_p P_p)$.

$\boldsymbol{P}$ is the component which summarizes the penalty on the B-spline coefficients of the p covariates and h is the opposite of the linking function g.

The approach assumes that the effect functions f_j of a covariate X_j can be approximated by a polynomial spline written in terms of a linear combination of B-spline basis functions. The crucial problem with such regression splines is the choice of the number and the position of the knots. A small number of knots may lead to a function space which is not flexible enough to capture the variability of the data. A large number of knots may lead to a serious overfiting. Similarly, the position of the knots may potentially have a strong influence on the estimation. A remedy can be based on a roughness penalty approach as proposed by Eilers and Marx (1996).

Smoothing parameters are used to balance the goodness-of-fit and smoothness. A performance measure is used to find the optimum values of the penalties. The number and location of knots are no longer crucial as long as the minimum number of knots is reached. In practice, this approach poses problems to get a solution when the number of the model's smoothing functions is high (Lang and Brezger, 2004). The P-spline approach is easy to conceive and has the advantage of the explicit formula of the estimation matrix and standard errors estimations (Marx and Eilers, 1998). However, the simplicity if reduced if the knots are at unequal distances (Wood, 2006). Thus, despite the advantages of the P-spline approach in the GAM models, the problems of the estimation of the parameters with the *penalized GLM Fisher scoring* algorithm, remains important (Zhao et al., 2006, Wood, 2006, Binder and Tutz, 2006).

3.3 Bayesian method

The Bayesian approach is essentially based on the concept that the parameters to be estimated are not constants but are considered as random variables. Bayesian statistical inference is based on the posterior distributions of the parameters, which combine the prior information and observed one from the sample. In the case of the GAM models, we wish to estimate the parameter α and the functions $f_1,\ldots,f_p$. One of the advantages of the Bayesian approach compared to the penalized GLM Fisher scoring algorithm is the fact that the uncertainty related to the variance of the components is taken into account trough the posterior distribution of the parameters (Fahrmeir and Lang, 2001, Zhao et al., 2006). In practice, the analytical form of the posteriori distribution is rarely available and then it s difficult to extract their characteristic for risk assessment purposes. The Markov chain Monte Carlo procedure (MCMC), allows to obtain all these characteristics by simulating samples from the posterior distribution and thus to deduce parameter estimators, the quantiles and associated risk as well as estimator uncertainty. More details on the MCMC approach and their convergence diagnostics are studied in El Adlouni et al. (2006).

In the case of the P-spline functions, the parameters $\boldsymbol{a}$ of the GAM model, in equation (2), are a random variables. The penalties based on the finite differences of the B-spine coefficients are replaced by their stochastic equivalent which correspond to a random walks of order one or two, defined by

$$a_{j\rho} = a_{j,\rho-1} + u_{j\rho}\text{, or } a_{j\rho} = 2a_{j,\rho-1} - a_{j,\rho-2} + u_{j\rho} \tag{5}$$

with $u_{j\rho} \sim N(0,\tau_j^2)$ and the initial values a_{j1}, a_{j2} are constants. The level of smoothing is thus controlled by the variance parameter τ_2, which must also be estimated. Lang and Brezger (2004) suggest a prior distribution of the parameters a_j of the form:

$$a_j \left| \tau_j^2 \right. \propto \frac{1}{(\tau_j^2)^{rk(K_j)/2}} \exp\left(-\frac{1}{2} a_j' K_j a_j \right) \tag{6}$$

where K is the penalty and depends on the smoothing function f_j and on the nature of the X_j variable. The prior distribution of the parameter τ_2 is an Inverse Gamma distribution

$IG(c_j, d_j)$, where c_j, d_j are the hyper-parameters and are usually given by prior knowledge on the variables. It is however necessary to perform a sensitivity analysis on the prior choice.

The posteriori distribution of the model has the following form:

$$p(\alpha, a_1, \tau_1^2, ..., a_p, \tau_p^2 | y) \propto L(y, \alpha, a_1, \tau_1^2, ..., a_p, \tau_p^2)$$

$$\propto \prod_{j=1}^{p} \frac{1}{(\tau_j^2)^{rk(K_j)/2}} \exp\left(-\frac{1}{2\tau_j^2} a_j' K_j a_j\right) \prod_{j=1}^{p} (\tau_j^2)^{-a_j - 1} \exp\left(-\frac{b_j}{\tau_j^2}\right) \quad (7)$$

All the inference is based on the posterior distribution. The MCMC algorithm can be performed to estimate the empirical posterior distribution and the predictive distribution of the quantile to deduce the risk values.

The assumptions of the Bayesian estimation model are completed by the following conditional independence assumptions:

a. For all explicative variables and f_j parameters, the observations Y_i are conditionally independent.
b. The prior distributions of the parameters are conditionally independent.
c. The priori distribution of the fixed effects and variances τ_j^2, $j = 1, \ldots, p$ are mutually independent.

4. Performance measure

In order to select the smoothing penalty and the number of knots that leads to the most adequate fit some performance measures are used. The most used performance measures are the Akaïke information criterion (*AIC*) and the generalized cross-validation (*GCV*). They are based on the deviance statistic (or the statistical likelihood ratio) that, for a counting GAM model (the case of the Poisson distribution), is obtained by the following formula:

$$D(y; \hat{\mu}) = 2\sum_{i=1}^{n} [y_i \ln(y_i / \hat{\mu}_i) - (y_i - \hat{\mu}_i)] \quad (8)$$

The Akaïke information criterion developed by Akaike (1973) measures the quality of the model fit to observed data series. It is the function of the deviance function $D(y; \mu)$ and is obtained by the following formula:

$$AIC = \frac{1}{n}[D(y; \hat{\mu}) + tr(R)\phi] \quad (9)$$

where $tr(R)$ the sum of the diagonal elements of the matrix R of the weighted additive-fit operator of the last iterations in the estimation process, and ϕ the scale parameter.

The generalized cross-validation for the smoothing penalty is obtained by the following formula:

$$GCV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{y_i - \hat{f}_\lambda(x_i)}{1 - tr(S_\lambda)/n}\right\}^2 \tag{10}$$

Where S is the smoother. For the GCV of the model, the corresponding criterion is based on :

$$GCV = \frac{\frac{1}{n}D(y;\hat{\mu})}{[1 - tr(R)/n]} \tag{11}$$

R is the weighted additive-fit operator of the last iteration in the estimation of the model.

5. Confounding variables, concurvity, and interaction

5.1 Confounding variables

Confounding is potentially present in all observational studies. A confounding factor in the field of environmental health refers to a situation in which an association between an exposure (i.e. air pollution) and a health outcome (i.e. morbidity or mortality) is distorted because it is mixed with the effect of a third variable – the confounding variable (i.e. humidity). The confounding variable is related to both the exposure and the outcome. The distortion introduced by a confounder can lead to an overestimation (positive confounding, affecting the outcomes in the same direction as the exposure under study) or underestimation (negative confounding, affecting the outcomes in the opposite direction of the exposure under study) of the association between exposure and outcome. Confounding variables can be controlled for by using of one or more of a variety of techniques that eliminate the differential influence of the confounder. For example, if one group is mostly females and the other group is mostly males, then the gender may have a differentially effect on the outcome. As a result, we will not know whether the outcome is due to the treatment or due to the effect of gender. If the comparison groups are the same on all extraneous variables at the start of the experiment, then differential influence is unlikely to occur. The control techniques are essentially attempts to make the groups similar or equivalent. Confounding variables are to be differentiated from intermediating or latent variables that are part of the causal pathway between the exposure and the outcome (Budtz-Jorgensen et al., 2007).

Peng et al. (2006) identified two types of confounding variables: those that are measured and are already included in the model, and those that are not. They propose as an adjustment to this problem the inclusion of a non-linear function of actual and future data in the model. In the study of the relation between air pollution and mortality, the non-measured confounders are the factors that influence the mortality in the same way as the air pollution variables (Peng et al., 2006). These factors produce seasonal effects and long-term tendencies on the mortality which deforms the relation between the mortality and the air pollution (i.e. Influenza epidemics and pulmonary infections). In these situations, the inclusion of the variable "time" helps to reduce the bias caused by these factors.

Other processes for managing the confounding effects are the methods of sampling: specification and matching. Specification is the scheme that specifies the value of the potential confounding and excludes other values (i.e. non smokers only in the study). This method of sampling allows focusing solely on the subjects of the study in question but does not enable the generalization of results. The matching consists on grouping the subjects with similar values of the confounding variable. It has the advantage of eliminating the influences of the confounding with important effects and of improving the precision (strength) by balancing the number of cases and controlling each layer.

5.2 Concurvity

The non-linear dependence that remains between the covariates is referred to as the concurvity in the GAM models by analogy to the co-linearity in GLM models. Researchers (Ramsay et al., 2003) insist that a certain degree of concurvity exists in every epidemiological time series, especially when the time is included in the model as a confounding variable. The main problem caused by concurvity in a GAM model is the presence of a bias in the model, more specifically the overestimation of the parameters and the underestimation of the standard errors. The use of asymptotically unbiased estimator of standard errors introduced by Hastie and Tibshirani (1990) and demonstrated by Dominici et al. (2003) does not solve the bias problem. The consequence of this is the inflation of type I errors in the signification tests, resulting in the conclusion of the presence of significant effect (Ramsay et al., 2003).

Several approaches have been proposed to control the problem of concurvity in time series. One method of estimation of the variation, based on the bootstrap parametric, has also produced biased results based on simulations by Ramsay et al. (2003). These recommend instead the use of parametric models such as the GLM model with natural splines (Dominici et al., 2002). He (2004) suggests the use of a non parametric model GAM to explore data in a primary level of analysis and when the appropriate variables are retained, to pursue the analysis with a parametric model GLM with natural splines, all while keeping the same degree of smoothing.

Figueras et al. (2005) developed the conditional bootstrap method in order to control the effect of the concurvity. In this type of bootstrap, B bootstrap replicates are generated. In each of these, the values of the independent variables are the same as those of the observed data, with only the values of the response variable being varied from replicate to replicate. The value assumed by the outcome in each observation is conditional (hence the technique's name) upon the values of the set of independent variables in said observation. The conditional Bootstrap approach has been tested on simulated data and leads to good results.

5.3 Interactions in the GAM model

The interaction within a statistical model denotes the effect of two or more variables, which is not simply additive. In other words, the effect is due to the combination of two or more variables in the model. A consequence of the interaction between two variables is that the

effect of a variable depends on the value observed for the other one. A form of interaction often found in bibliography is the modification of the effect. The modification of the effect happens when the statistical measure of the association between the explicative variable X_1 and the response variable Y depends on the level of another variable X_2, known as the effect modifier. The extent of the relationship depending on the value of the effect modifier contributes to the improvement of the model fit. In the field of environmental health, this allows us to identify the most vulnerable groups to a particular condition (Wood, 2006; Bates and Maechler, 2009).

6. Conclusion

Environmental health research is becoming a cornerstone for supporting evidence-based (informed) decision making in healthcare services and management. Providing evidence through robust and relevant epidemiologic studies in environmental health research may be improved through an adequate utilization of statistics methods. In this chapter, we reviewed the Generalized Additive Models and the most used estimating methods and presented their advantages and limits. Knowing this, researchers should take into account these aspects when it is time to define exposures and outcomes, to map spatial variations, to design epidemiologic studies' conceptual frameworks and to select suitable estimating models. These critical aspects are of central importance for developing clinical and public health decision making to reduce the burden of environment impacts on individual and population health. Moreover, using accurate and relevant methods, i.e. GAM, in environmental epidemiology studies is a cornerstone for developing effective actions that may help save cost and improve decision making performance.

Improvements will be seen also in clinical practices through a better understanding and the integration in medical decisional algorithms of the effects of long term exposition to specific environmental factors. These effects are translated into risks of occurrence and prognosis of sensitive diseases (i.e. breast cancer, lung cancers). Spreading GAM method utilization in environmental epidemiology through a clinical perspective is highly recommended to develop effective decisional tools that may greatly improve personalized medicine. Moreover, GAM method may help to better manage follow-up of patients exposed to long term medications and reduce side-effects and complications. This review highlights the utility of Generalized Additive Model (GAM) for risk assessment (such as breast cancer) related to environmental factors and explored the use of the GAM for risk assessment in the presence of multiple non-linear effects. The selection and the estimation of the parameters and non-linear functions (B-Splines and P-splines) are essential for an adequate estimate of the risk. Next research should explore how GAM models may help the development of relevant risk assessment tools that may be integrated in personalized medical decision making algorithms. The GAM will allow the integration of environmental factors and others health determinants in clinical algorithms that may help improve the personalization of healthcare delivery. These algorithms will be implemented in public health programs (i.e. personalization of breast cancer screening based on women individual risk) and clinical algorithms (i.e. for patients with a diagnosis of breast cancer the personalization of follow-up will be based on the surveillance of relevant factors such as the biomarkers, the clinical signs and the exposition to environmental factors).

This chapter presented the potential of the Generalized Additive Model (GAM) for environmental studies. Generalized additive models (GAMs) are a generalization of generalized linear models (GLMs) and constitute a powerful technique to capture nonlinear relationships between explanatory variables and a response variable. Selection of the best parameter estimation methods, control for confounding variables and concurvity aims to reduce bias and improve the use of the GAM model. Moreover, when using the GAM model in environmental health, and for an adequate interpretation of the outputs, socio-economic and demographic parameters should be considered.

7. Acknowledgement

The authors would like to thank the CIHR Team on Familial breast cancer at Université Laval (QC) leaded by Dr Jacques Simard; and also the Consortium national de formation en santé-Université de Moncton (NB) for the financial support they provided to prepare and publish this chapter.

8. References

Akaike H. (1973). Information theory as an extension of the maximum likelihood principle. Second International Symposium on Information Theory (B. N. Petrov, et F. Csaki), pp. 267-281, Akademiai Kiado, Budapest.

Balakrishnan K, Ganguli B, Ghosh S, Sankar S, Thanasekaraan V, Rayudu VN, Caussy H; HEI Health Review Committee. Short-term effects of air pollution on mortality: results from a time-series analysis in Chennai, India. *Res Rep Health Eff Inst.* 2011 Mar;(157):7-44.

Bates D. and M. Maechler (2009). lme4: Linear mixed-effects models using S4 classes. URL http://CRAN.R-project.org/package=lme4. R package version 0.999375-31.

Bayentin, L., S. El Adlouni, T.B.M.J. Ouarda, P. Gosselin, B. Doyon and F. Chebana (2010). Spatial variability of climate effects on ischemic heart disease hospitalization rates for the period 1989-2006 in Quebec, Canada. International Journal of Health Geographics, 9:5. doi:10.1186/1476-072X-9-5.

Ballester F., Rodríguez P., Iñíguez C., Saez M., Daponte A., Galán I., Taracido M., Arribas F., Bellido J., Cirarda F.B., Cañada A., Guillén J.J., Guillén-Grima F., López E., Pérez-Hoyos S., Lertxundi A. and Toro S. (2006). Air Pollution and cardiovascular admissions association in Spain: results within the EMECAS Project. Epidemiol. Community Health 60: 328-336.

Berger P., L. Pascal, C. Sartor , J. Delorme, P. Monge, C. P. Ragon, M. Charbit, R. Sambuc and M. Drancourt (2004). Generalized Additive Model demonstrates fluoroquinolone use/resistance relationships for Staphylococcus aureus. European Journal of Epidemiology, 19, pp. 453-460.

Binder H. and G. Tutz (2006). Fitting Generalized Additive Models: A comparison of Methods. Universität Freiburg i. Br., Nr. 93.

Braga, A.L., Zanobetti, A. and Schwartz J (2001), The time course of weather related deaths, Epidemiology, 12, 662-667.

Budtz-Jorgensen E., N. Keiding, P. Grandjean and P. Weiher (2007). Confounder selection in environmental epidemiology: Assessment of health effects of prenatal mercury exposure. Ann. Epidemiol. 17(1), pp 27-35.

Dominici F., A. McDermott and T.J. Hastie (2003). Issues in Semi-parametric Regression with Applications in Time Series Models for Air pollution and Mortality. Rapport de recherche. http://biosun01.biostat.jhsph.edu/~fdominic/trevorpaper.pdf

Dominici F., McDermott A., Zeger S. L. and J.M. Samet (2002). On the use of generalized additive models in time-series studies of air pollution and health. *American Journal of Epidemiology*. 156(3), pp 193-203.

Doyon B., D. Bélanger and P. Gosselin (2006). Effets du climat sur la mortalité au Québec méridional de 1981 à 1999 et simulations pour des scénarios climatiques futurs. *Rapport de recherche*, INSPQ.

Eilers P. H. C. and B. D. Marx (1996). Flexible Smoothing with B-splines and Penalties. Statistical Science, Vol. 11, No. 2, pp. 89-121.

El Adlouni, S., Favre, A.C. and Bobée, B. (2006). Comparison of methodologies to assess the convergence of Markov Chain Monte Carlo methods. Computational Statistics and Data Analysis 50(10): 2685-2701.

Fahrmeir L. and S. Lang (2001). Bayesian inference for generalized additive mixed models based on Markov random field priors. *Appl. Statist.*, 50, Part 2, pp. 201-220.

Figueiras A., J. Roca-Pardiñas and C. Cadarso-Suárez (2005). A bootstrap method to avoid the effect of concurvity in generalized additive models in time series of air pollution. Journal of Epidemiology and Community Health. 59, pp. 881-884.

Gehrmanu U., B. Hellriegel, A. Neiss and L. Fahrmeir (2003). Analysis of the time to sustained progression in Multiple Sclerosis using generalized linear and additive models. Discussion paper 354-SFB 386- LMU Munich.

Hastie T., L. Sleeper and R. Tibshirani (1992). Flexible covariate effects in the proportional hazard model. Breast Cancer Research and *Treatment*, 22, pp. 241-250.

Hastie T. and R. Tibshirani (1990). Generalized Additive Models. *Chapman & Hall*, New-York.

Hastie T. and R. Tibshirani (1993). Varying-coefficient models. *Journal of the Royal Statistical Society B*, 55, pp. 757-796.

Harnack, L., G. Block, et al. (1997). "Association of Cancer Prevention-Related Nutrition Knowledge, Beliefs, and Attitudes to Cancer Prevention Dietary Behavior." *Journal of the American Dietetic Association* 97(9): 957-965.

Hatzakis, A., Katsouyanni, K., Kalandidi, A., Day, N. and Trichopoulos, D. (1986) Short-term effects of air pollution on mortality in Athens. Int. J. Epidemiol., 15, 73-81.

He S.(2004). Generalized additive models for data with concurvity: statistical issues and a novel fitting approach. Rapport de recherche. *http://etd.library.pitt.edu/ETD/available/etd-12022004-103805/unrestricted/ShuiHe.pdf*

Katsouyanni K, Samet JM, Anderson HR, Atkinson R, Le Tertre A, Medina S, Samoli E, Touloumi G, Burnett RT, Krewski D, Ramsay T, Dominici F, Peng RD, Schwartz J, Zanobetti A; HEI Health Review Committee. Air pollution and health: a European and North American approach (APHENA). *Res Rep Health Eff Inst*. 2009 Oct;(142):5-90.

Lall R., Ito K. and G.D. Thurston (2011). Distributed Lag Analyses of Daily Hospital Admissions and Source-Apportioned Fine Particle Air Pollution. *Environ Health Perspect.* April; 119(4): 455–460.

Lang, S. and A. Brezger (2004): Bayesian P-Splines. Journal of Computational and Graphical Statistics, 13, 183-212.

Liang K.Y. and S. L. Zeger (1986). Longitudinal data analysis using general linear models. *Biometrika*, 73, 13-22.

Mackenbach, J., A. Kunst, and C. Looman (1992). Seasonal variation in mortality in The Netherlands. Journal of Epidemiology and Community Health 46, 261–265.

Marx B.D. and P. H. C. Eilers (1998). Direct generalized additive modeling with penalized likelihood. *Computational Statistics & Data Analysis*, 28, pp.193-209.

McCullagh P. and J. A. Nedler (1989). Genralised linear models. *Chapman&Hall.*

Michelozzi P et al. (2007) Assessment and prevention of acute health effects of weather conditions in Europe, the PHEWE project: background, objectives, design. Environ Health; 6:12.

Peng R.D., F. Dominici and A. L. Thomas (2006). Model choice in time series studies of air pollution and mortality. *J. R. Statist. Soc. A*, 169, *Part 2*, pp. 179-203.

Peng R.D., Chang H.H., Bell M.L., McDermott A., Zeger S.L., Samet J.M. and F. Dominici (2008) Coarse Particulate Matter Air Pollution and Hospital Admissions for Cardiovascular and Respiratory Diseases Among Medicare Patients *JAMA*;299(18):2172-2179.

Ramsay T.O., R.T. Burnett and D. Krewski (2003). Exploring bias in generalized additive models for spatial air pollution data. *Environmental Health Perspectives*, 111(10), 1283-1288.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics,*11(4), pp 735-757.

Schwartz J., J. M. Samet and J. A. Patz (2004). Hospital admissions for heart disease: the effects of temperature and humidity. *Epidemiology*, 15(6), pp. 755-761.

Terzi Y. and M. A. Cengiz (2009). Using of generalized additive model for model selection in multiple poisson regression for air pollution data. Scientific Research and Essay Vol.4 (9), pp. 867-871.

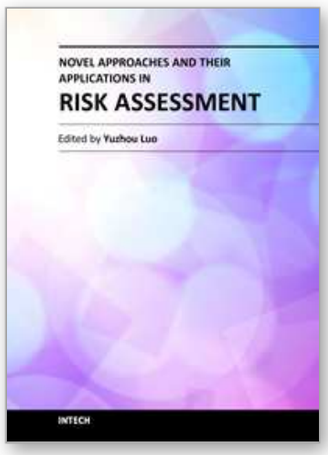
Vieira V., Webster Thomas., Weinberg J., Aschengrau A. and Ozonoff D. (2005) Spatial analysis of lung, colorectal, and breast cancer on Cape Cod: An application of generalized additive models to case-control data. *Environmental Health: A Global Access Science Source.* 4:11

Wood, S. N. (2006). Generalized Additive Models, An introduction with R. Chapman & Hall/ CRC.

Wong CM, Thach TQ, Chau PY, Chan EK, Chung RY, Ou CQ, Yang L, Peiris JS, Thomas GN, Lam TH, Wong TW, Hedley AJ; HEI Health Review Committee. Part 4. Interaction between air pollution and respiratory viruses: time-series study of daily mortality and hospital admissions in Hong Kong. *Resp Rep Health Eff Inst.* 2010 Nov;(154):283-362.

Zanobetti, A. and J. Schwartz. (2009). A Novel Approach to Estimate Distributed Lag Model Between Hospital Admissions and Ozone: A Multi-City Time Series Analysis. *Epidemiology*: November - Volume 20 - Issue 6 - p S62

Zhao Y., J. Staudenmayer, B.A. Coull and M. P. Wand (2006). General design bayesian generalized linear mixed models. *Statistical Science*. 21(1), pp 35-51.



Novel Approaches and Their Applications in Risk Assessment

Edited by Dr. Yuzhou Luo

ISBN 978-953-51-0519-0
Hard cover, 344 pages
Publisher InTech
Published online 20, April, 2012
Published in print edition April, 2012

Risk assessment is a critical component in the evaluation and protection of natural or anthropogenic systems. Conventionally, risk assessment is involved with some essential steps such as the identification of problem, risk evaluation, and assessment review. Other novel approaches are also discussed in the book chapters. This book is compiled to communicate the latest information on risk assessment approaches and their effectiveness. Presented materials cover subjects from environmental quality to human health protection.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Jalila Jbilou and Salaheddine El Adlouni (2012). Generalized Additive Models in Environmental Health: A Literature Review, Novel Approaches and Their Applications in Risk Assessment, Dr. Yuzhou Luo (Ed.), ISBN: 978-953-51-0519-0, InTech, Available from: http://www.intechopen.com/books/novel-approaches-and-their-applications-in-risk-assessment/generalized-additive-models-in-environmental-health-a-review-of-litterature

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen