Science of the Total Environment 628–629 (2018) 217–225

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

Aggregating the response in time series regression models, applied to weather-related cardiovascular mortality

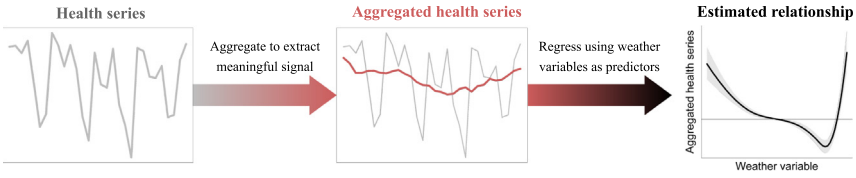
Pierre Masselot [a,*], Fateh Chebana [a], Diane Bélanger [a,b], André St-Hilaire [a], Belkacem Abdous [c], Pierre Gosselin [a,b,d], Taha B.M.J. Ouarda [a]

[a] Institut National de la Recherche Scientifique, Centre Eau-Terre-Environnement, Québec, Canada
[b] Centre Hospitalier Universitaire de Québec, Centre de Recherche, Québec, Canada
[c] Université Laval, Département de Médecine Sociale et Préventive, Québec, Canada
[d] Institut National de Santé Publique du Québec (INSPQ), Québec, Canada

HIGHLIGHTS

- Aggregating the response helps reducing its noise.
- Modelling autocorrelation induced by the aggregation increases the fit of the model.
- The methodology with an aggregated response achieves a better fit than the classical DLNM.
- Using an aggregation with decreasing weights is well adapted to weather-related mortality issues.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 13 September 2017
Received in revised form 4 January 2018
Accepted 2 February 2018
Available online xxxx

Editor: SCOTT SHERIDAN

Keywords:
Time series regression
ARMA
Temporal aggregation
Temporal dependence
Cardiovascular mortality
Temperature

ABSTRACT

In environmental epidemiology studies, health response data (*e.g.* hospitalization or mortality) are often noisy because of hospital organization and other social factors. The noise in the data can hide the true signal related to the exposure. The signal can be unveiled by performing a temporal aggregation on health data and then using it as the response in regression analysis. From aggregated series, a general methodology is introduced to account for the particularities of an aggregated response in a regression setting. This methodology can be used with usually applied regression models in weather-related health studies, such as generalized additive models (GAM) and distributed lag nonlinear models (DLNM). In particular, the residuals are modelled using an autoregressive-moving average (ARMA) model to account for the temporal dependence. The proposed methodology is illustrated by modelling the influence of temperature on cardiovascular mortality in Canada. A comparison with classical DLNMs is provided and several aggregation methods are compared. Results show that there is an increase in the fit quality when the response is aggregated, and that the estimated relationship focuses more on the outcome over several days than the classical DLNM. More precisely, among various investigated aggregation schemes, it was found that an aggregation with an asymmetric Epanechnikov kernel is more suited for studying the temperature-mortality relationship.

1. Introduction

In environmental epidemiology, studies on the health effect of various environmental exposures, often rely on regression models applied to time series data (Gasparrini and Armstrong, 2013). Environmental

* Corresponding author.
E-mail address: pierre.masselot@ete.inrs.ca (P. Masselot).

https://doi.org/10.1016/j.scitotenv.2018.02.014
0048-9697/

exposure variables include atmospheric pollutant levels and temperature, while health issues include various diseases (Barreca and Shimshack, 2012; Blangiardo et al., 2011; Braga et al., 2002; Knowlton et al., 2009; Martins et al., 2006; Nitschke et al., 2011; Szpiro et al., 2014; Yang et al., 2015). In this context, the exposure-response relationship is complex since, among other reasons, the effect of exposure on health issues lasts several days. This is why models have often used exposure windows under the form of moving averages (MA, *e.g.* Armstrong, 2006) or distributed lags (DL, Schwartz, 2000a). In particular, the latter has been extended to deal with nonlinear relationships (distributed lags nonlinear models, DLNM, Gasparrini et al., 2010) which is now widely used in weather-related health population studies (*e.g.* Phung et al., 2016; Vanos et al., 2015; Wu et al., 2013).

The health response, however, is almost always used directly as a daily time series. This could lead to several drawbacks in the regression models of environmental exposure on a health issue. First, the response to an exposure can also be spread across several days (Lipfert, 1993), which means that it would seem more realistic to consider a health time window in response to an associated exposure window. Second, health time series data used in epidemiologic studies are often noisy. The noise can conceal the true signal of the response to an exposure, especially in areas with small populations where the number of cases (mortality or morbidity) is low. Sources of noise include diverse organizational factors such as weekends and holidays (Suissa et al., 2014; Wong et al., 2009), slight changes in the definition of diseases (*e.g.* Antman et al., 2000) as well as behavioral and technological changes. In the end, the noise in the response can reduce the accuracy of the model and the conclusions (*e.g.* Todeschini et al., 2004).

In order to assess a more realistic relationship between an exposure and a health issue as well as reduce the noise impact in the health response, it is proposed to consider an aggregation window over time in the health response also, in addition to the exposure. More precisely, moving aggregation is considered here, *i.e.* the time step of data points in the obtained series remains the same, in opposition to aggregation where the time step of data points is reduced (*e.g.* from daily values to monthly values).

Aggregating the response series is expected to have two advantages: (1) better representing the spread of the health response to an exposure and (2) reducing the noise in the health series. Indeed, aggregated series are less sensitive to random perturbation in the data. An aggregated response should make regression models more robust to variations induced by noise, leading to more reliable relationship estimates. This idea is consistent with the results of Cristobal et al. (1987) in a non-time series context, which showed that pre-smoothing a response variable to remove noise leads to consistent estimates with low variance in linear regression. In a similar study, Sarmento et al. (2011) concluded that regression models are more robust to noise when both the response and the exposure are aggregated.

There have been few preceding cases of aggregated responses (Roberts, 2015; Sarmento et al., 2011; Schwartz, 2000b), but the regression models applied did not account for the specificities of an aggregated response. These specificities include the presence of extra autocorrelation in the residuals and a modification of their distribution. Therefore, the objective of the present paper is to introduce a general methodology dealing with an aggregated response. The methodology allows the use of a DLNM with an aggregated response and deals with the autocorrelation created by the aggregation. The exposure-response surface of a DLNM with aggregated response is then compared to the surface of a classical DLNM in order to assess the impact on the estimated relationship. In past studies, only the moving average (Roberts, 2015; Sarmento et al., 2011) and Loess (Schwartz, 2000b) have been considered to aggregate the response. In the present paper, other aggregations are considered, in particular Nadaraya-Watson kernel smoothing (Nadaraya, 1964; Watson, 1964) with different kernels including the Epanechnikov kernel (Epanechnikov, 1969) and an asymmetric kernel proposed in Michels (1992).

The paper is organized as follows. Section 2 introduces the proposed methodology for an aggregated response. Section 3 illustrates the methodology and its benefits by applying it on a weather-related cardiovascular mortality case. The methodology is first compared to models with a non-aggregated response and then, different aggregation strategies are compared. The results are discussed in Section 4 and the conclusions are presented in Section 5.

2. Methods

This section introduces the statistical methodology consisting in 1) performing a temporal aggregation on the response time series y_t; and 2) modelling the aggregated response $\tilde{y}_t$ according to an exposure x_t through a regression model.

2.1. Aggregation of the response

The temporal aggregations considered in the proposed methodology are all local, *i.e.* the aggregation $\tilde{y}_t$ of a time series y_t depends only on a subset of observations close to the current observation. In particular, linear local aggregations can be expressed as:

$$\tilde{y}_t = \sum_{i \in I} w_i y_{t+i} \tag{1}$$

where the w_i are the weights attributed to each observation and I is the aggregation window. The most common aggregation is the H-day centered moving average (MA) where $w_i = 1/H$, $I = \{-\frac{H-1}{2}; \ldots; \frac{(H-1)}{2}\}$ and H (odd number) is the size of the window I. The main issue with MA is that the attributed weights are constant, giving equal importance to all the observations covered by the window I, even the farthest ones. This can result in small distortions in $\tilde{y}_t$ (Schwartz et al., 1996). The most common alternative aggregation is the Nadaraya-Watson kernel smoothing (Nadaraya, 1964; Watson, 1964) which generalizes the MA with the w_i following a deterministic function $K(.)$ called "kernel". This allows for the weights to be non-constant. Although a number of kernels exist, the most popular is the Epanechnikov kernel (Epanechnikov, 1969) designed to minimize the squared error of the fit for a fixed window I. In addition, the asymmetric kernel of Michels (1992) is considered (thereafter called the Michels kernel). Unlike the MA and the Epanechnikov kernels, the Michels kernel is asymmetric (*i.e.* $w_i \neq w_{-i}$), giving more weights to future values ($i > 0$) in order to better represent physiological adaptation.

Classical aggregation includes past values ($i < 0$) to compute $\tilde{y}_t$. However, in the context of the response of a regression model, it seems more logical to explain only the future health response at time t. Therefore, all three aggregations discussed above (MA, Epanechnikov and Michels kernels) are considered using only future values (*i.e.* $I = \{0; \ldots; H-1\}$). The MA for future values attributes constant weights, the Epanechnikov kernel attributes decreasing weights from $i = 0$ to $i = H-1$ and the Michels kernel attributes the maximum weight few days after the current one. Their shape is shown in Fig. 1.

2.2. Regression model with aggregated response

The second step of the methodology is to use the aggregated response $\tilde{y}_t$ as the response of the general regression model:

$$\tilde{y}_t = \beta x_t + \epsilon_t \tag{2}$$

where β is the regression coefficient and ϵ_t is the residual of the regression. The aggregated response $\tilde{y}_t$ raises two questionings for the model in Eq. (2): the probability distribution of residuals and the temporal dependence created by the aggregation.

When using a linear aggregation such as in Eq. (1), which is the case here, the distribution of ϵ_t can generally be considered Gaussian. This is