BMC
Medical Research Methodology

RESEARCH ARTICLE Open Access

Using a generalized additive model with autoregressive terms to study the effects of daily temperature on mortality

Lei Yang[1], Guoyou Qin[1], Naiqing Zhao[1*], Chunfang Wang[2] and Guixiang Song[2]

Abstract

Background: Generalized Additive Model (GAM) provides a flexible and effective technique for modelling nonlinear time-series in studies of the health effects of environmental factors. However, GAM assumes that errors are mutually independent, while time series can be correlated in adjacent time points. Here, a GAM with Autoregressive terms (GAMAR) is introduced to fill this gap.

Methods: Parameters in GAMAR are estimated by maximum partial likelihood using modified Newton's method, and the difference between GAM and GAMAR is demonstrated using two simulation studies and a real data example. GAMM is also compared to GAMAR in simulation study 1.

Results: In the simulation studies, the bias of the mean estimates from GAM and GAMAR are similar but GAMAR has better coverage and smaller relative error. While the results from GAMM are similar to GAMAR, the estimation procedure of GAMM is much slower than GAMAR. In the case study, the Pearson residuals from the GAM are correlated, while those from GAMAR are quite close to white noise. In addition, the estimates of the temperature effects are different between GAM and GAMAR.

Conclusions: GAMAR incorporates both explanatory variables and AR terms so it can quantify the nonlinear impact of environmental factors on health outcome as well as the serial correlation between the observations. It can be a useful tool in environmental epidemiological studies.

Background

People have long been interested in potential impacts of temperature on human health [1-3]. Recently, many studies have been done to analyse the way and the extent temperature influences health outcomes [4-6]. Such studies have practical value for the following reasons: knowing the association between environmental factors and health outcomes will help to identify at-risk populations, benefit health department in resource allocation, and provide support for stake holders in prevention [7].

Environmental epidemiology, the investigation of the health risks related to environmental exposures, becomes the main research approach [8]. There are various study designs in environmental epidemiology for estimating health effects of temperature. One of them is model evaluation of time series data to quantify the association between temperature (or other weather/environmental factors etc.) and daily mortality (or hospitalization etc.). These are a type of ecologic study because they analyse daily population-averaged health outcomes and exposure levels [9]. This approach is suitable to study transient acute effects which are due to time-varying exposures [10-14], and the choice of model can have a large influence on the interpretation of environmental effects.

Generalized Linear Model (GLM) [15] and Generalized Additive Model (GAM) [16] are the main models used in environmental epidemiology. When the response variable represents counts (e.g. number of deaths), the model often takes the form of a Poisson regression/additive model with a log link function. The outcome Y_t is assumed to follow a Poisson distribution with mean μ_t which is linked either to a linear combination (in this case, the model is a GLM) or smoothed functions (in

* Correspondence: nqzhao@fudan.edu.cn
[1]Department of Biostatistics, School of Public Health, Fudan University, Shanghai, China
Full list of author information is available at the end of the article

this case, the model is a GAM) of environmental explanatory variables via a log function.

Many studies have identified the nonlinear relations (U, J, or V shaped) of daily mortality on temperature [17-24], which are mainly represented by piecewise linear terms [18-20] or natural cubic splines [21-24]. An assumption is often made that temperature only affects the mortality of the same day or on a single day of lag l [18,20-22], and denoted by the term $temp_{t-l}$ in the model. When l=0, it represents the effect on the same day [18,20-23], when l>0, it represents a lagged effect [18,20,21,23]. An alternative to single lag models is a distributed lag model [25]. It includes multiple lags of temperature over a period of time based on the assumption that the effect of time is distributed over several days into the future. In all the above models, a spline function of time is often used to explain the long time trend, and confounding effects like humidity and air pollution are also controlled by splines.

However, a thorough time series analysis should consider the order of data points and correlation of adjacent points in time [26]. In environmental epidemiological studies, the response variable may also be correlated and it is necessary to embody autocorrelation of the response variable when modelling. But all the above models are standard GLM/GAM that describes how the response variable is stochastically related to explanatory variables without considering how the response can be dependent also on its past values.

In addition, autocorrelation causes trouble in estimation of GLM/GAM, since GLM/GAM essentially requires each observation to be independently distributed. Violation of this assumption can lead to problematic estimates even in very simple cases. For example, if the error terms in a linear regression model are in reality positively autocorrelated, failure to account for this may underestimate the standard errors of the estimated regression coefficients [27].

Another statistical issue often encountered is overdispersion, and many sources of autocorrelation are related to sources of overdispersion [28]. One usual approach to adjust for overdispersion is to specify a dispersion parameter ϕ for estimation [15]. However, if autocorrelation exists, this approach only inflates the variance of the estimate by ϕ but leaves the estimate unchanged, which is an inadequate solution to our issue.

Generalized Estimating Equations (GEE) [29,30] and Generalized Linear Mixed Model (GLMM)/Generalized Additive Mixed Model (GAMM) [31,32], are two extensions from GLM/GAM for grouped or clustered data. GEE and GLMM/GAMM can be implemented in popular software like SAS and R [31-34]. For all these models, a within-group variance-covariance structure can be used to account for the corresponding within-group autocorrelation [31,34]. Time series data, treated as single cluster data, can also be modelled by GAMM. Performance of this degenerative GAMM will be studied in simulation.

In this article, we introduced GAM with Autoregressive terms (GAMAR), which is derived from Generalized Autoregressive Moving Average (GARMA) models [35]. ARMA is a family of models for analysing time series. The notation ARMA(p,q) refers to a model with p autoregressive terms and q moving-average terms [26]. GARMA belongs to the class observation driven models [36] for extending Gaussian ARMA to non-Gaussian settings. GAMAR differs from GARMA in that the linear components in GARMA are generalized to natural splines and MA terms are omitted. The generalization is motivated by modelling nonlinear relationships, while the simplification is justified because AR can be used to approximate MA or ARMA. Additionally, larger AR order in GAMAR will not compromise the estimation of the effects of explanatory variables.

In all, GAMAR has two advantages over GAM: 1) it is a model for generalized time series analysis rather than a probabilistic model like GAM; 2) the AR part of GAMAR can explain the autocorrelation structure of observations. So the Pearson residuals of GAMAR can be closer to white noise than those from GAM, yielding more reliable estimation.

Methods

GARMA

In GARMA, the conditional distribution of each observation y_t, for $t=1,\ldots,n$, given the previous information set $H_t=\{X_1,\ldots,X_t,y_1,\ldots,y_{t-1}\}$ containing past observations y_t and covariate vectors $X_t = (X_{t1},\ldots,X_{tm})$, is assumed to follow the same exponential family distribution [35]. As with the standard GLM, the conditional mean μ_t is related to the variables by a twice-differentiable one-to-one monotonic function g, which is called the link function. However, unlike the standard GLM, the formula here allows autoregressive moving average terms to be included additively [35]:

$$g(\mu_t) = \sum_{i=1}^{m} X_{ti}\beta_i + \sum_{j=1}^{p} c_j\left(g(y_{t-j}) - \sum_{i=1}^{m} X_{t-j,i}\beta_i\right) + \sum_{j=1}^{q} d_j\left(g(y_{t-j}) - g\left(\mu_{t-j}\right)\right),$$

where $\sum_{j=1}^{p} c_j\left(g(y_{t-j}) - \sum_{i=1}^{m} X_{t-j,i}\beta_i\right)$ are autoregressive terms and $\sum_{j=1}^{q} d_j\left(g(y_{t-j}) - g\left(\mu_{t-j}\right)\right)$ are moving average terms.

Count data are often assumed to follow Poisson distribution. And the Poisson GARMA submodel is:

$$\ln(E(y_t)) = \ln(\mu_t) = \sum_{i=1}^{m} X_{ti}\beta_i + \sum_{j=1}^{p} c_j\left(\ln\left(y^*_{t-j}\right) - \sum_{i=1}^{m} X_{t-j,i}\beta_i\right) + \sum_{j=1}^{q} d_j\left(\ln\left(y^*_{t-j}/\mu_{t-j}\right)\right),$$

where $y^*_t = \max(y_t, \tau)$, τ is a positive threshold parameter. Any 0 or negative values of y are replaced by τ, because ln() is not defined for 0 or negative values.

GAMAR

GAMAR is derived from GARMA with linear terms replaced by smoothers and MA terms omitted:

$$g(E(y_t)) = g(\mu_t) = \sum_{i=1}^{m} s_i(X_{ti}) + \sum_{j=1}^{p} c_j\left(g(y_{t-j}) - \sum_{i=1}^{m} s_i(X_{t-j,i})\right), \quad (1)$$

where $\sum_{i=1}^{m} s_i(X_{ti})$ are smoothers of covariates, $\sum_{j=1}^{p} c_j\left(g(y_{t-j}) - \sum_{i=1}^{m} s_i(X_{t-j,i})\right)$ are autoregressive terms.

Compared to GAM, (1) allows autoregressive terms to be included additively in the link predictor.

For count data y, we use the Poisson submodel:

$$\ln(E(y_t)) = \ln(\mu_t) = \sum_{i=1}^{m} s_i(X_{ti}) + \sum_{j=1}^{p} c_j\left(\ln\left(y^*_{t-j}\right) - \sum_{i=1}^{m} s_i(X_{t-j,i})\right), \quad (2)$$

here $y^*_t = \max(y_t, \tau)$, τ is a positive threshold parameter.

In our study, we use natural cubic spline (*ns*) as smoother. *ns* is a piecewise-cubic real function on an interval $[a,b]$, where $[a,b]$ is separated by a sequence of k ordered knots: $\alpha = \xi_0 < \xi_1 < \cdots < \xi_{k-1} < \xi_k = b$. *ns* is continuous at interior knots and linear beyond the boundary [16]. Degrees of freedom (*df*) for *ns* equals to the number of subintervals separated by these knots, thus *df* satisfies: $df = k$. *ns* is often represented by linear combination of its B-spline basis [16]. When *df* is specified, the standard practice is to place *df-1* interior knots at evenly spaced intervals in the data to generate the B-spline basis. By using *ns* as smoother in (2), the Poisson GAMAR becomes:

$$\ln(E(y_t)) = \ln(\mu_t) = \sum_{i=1}^{m} ns(X_{it}, df_i) + \sum_{j=1}^{p} c_j\left(\ln\left(y^*_{t-j}\right) - \sum_{i=1}^{m} ns(X_{i-j,t}, df_i)\right).$$

Algorithm

Maximum partial likelihood estimator (MPLE)

For the jointly distributed time series $\{X_t, y_t\}$, $t = 1,\ldots n$, the parameters of GAMAR can be estimated by maximum partial likelihood. The partial likelihood based on y_t for $\{X_t, y_t\}$, $t = 1,\ldots n$ can be expressed as the product of a sequence of conditional likelihoods $f(y_t \mid X^{(t)}, y^{(t-1)}; \theta)$, $t = 1,\ldots n$, where $X^{(t)} = X_1,\ldots X_t$, $y^{(t-1)} = y_1,\ldots y_{t-1}$, that is:

$$PL = \prod_{t=1}^{n} f\left(y_t \middle| X^{(t)}, y^{(t-1)}; \theta\right).$$

This function is referred to as a partial likelihood instead of likelihood because X_t are stochastic [35]. For a Poisson GAMAR, the partial likelihood is:

$$PL = \prod_{t=1}^{n} \frac{\mu_t^{y_t}}{y_t!} e^{-\mu_t},$$

and μ_t can be expanded as $\ln(\mu_t) = \sum_{i=1}^{m} ns(X_{ti}, df_i) + \sum_{j=1}^{p} c_j\left(\ln\left(y^*_{t-j}\right) - \sum_{i=1}^{m} ns(X_{t-j,i}, df_i)\right)$. Since a natural cubic spline is a linear combination of its B-spline basis, it is linear with respect to its parameters, so natural cubic spline functions are like linear terms from the perspective of computation. Therefore,

$$\ln(\mu_t) = \eta_t = \sum_{i=1}^{m} \beta_i X_{ti} + \sum_{j=1}^{p} c_j\left(\ln\left(y^*_{t-j}\right) - \sum_{i=1}^{m} \beta_i X_{t-j,i}\right) \quad (3)$$

where $\theta = (\beta_1\cdots,\beta_m, c_1,\cdots,c_p)^T$ is the model parameter vector.

Modified Newton's method

Maximum partial likelihood estimators are solved by modified Newton's method. This procedure is described in detail in Appendix.

Evaluation of GAMAR by simulation

We conducted two simulation studies, each with 1000 samples, to compare the performance of GAM and GAMAR

for estimation. In simulation study 1, the performance of GAMM is also studied. The R code for simulations and simulation output is included in Additional file 1.

Simulation study 1

The first model:

$$\begin{aligned} &Y_t \sim Poisson(\mu_t) \\ &\ln(\mu_t) = ns(x_t, 5) + a_t \\ &ns(x_t, 5) = \sum_{i=1}^{5} \beta_i s_{i5}(x_t) \\ &a_t = \sum_{i=1}^{3} c_i(\ln(y^*_{t-i}) - ns(x_{t-i}, 5)) \\ &y^*_t = \max(y, \tau), \tau = 0.5. \end{aligned} \tag{4}$$

Here x_t represents a daily averaged temperature series from year 2000–2008 obtained from Shanghai's Meteorological Bureau, and the terms $s_{i5}(x_t)$, i = 1, 2, 3, 4, 5 form the B-spline basis for the natural cubic spline. The parameters are:

$$\beta_0 = 5.02, \beta_1 = -0.35, \beta_2 = -0.36, \beta_3 = -0.38, \beta_4 = -0.33, \beta_5 = -0.15 \text{ and } c_1 = 0.5, c_2 = 0.25, c_3 = 0.12.$$

We used data ranging over time points 1828–3288 (year 2005–2008) to eliminate the impact of the starting value, where the time points stand for the days since Jan 1st 2000.

For two models on Sample 1, the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) [26] of the Pearson residuals [15] are plotted against different lag periods to examine the presence of residual autocorrelation. Estimates of temperature effects from the two models and the true effect were plotted against temperature to show which model provided a better fit.

Besides showing the analysis for a single sample, the averaged results from GAM, GAMM and GAMAR were given, and the statistics calculated were mean estimates, bias, relative error and coverage, short for "coverage rate of confidence interval (CI) on true value". The confidence level was chosen to be 95%, so the coverage of a correct model should be around 95%.

In addition, GAMAR with various AR orders were studied to explore the effects of AR order on estimation. Finally, to verify the consistency of partial maximum likelihood estimation, 1000 samples in the time range of 2558–3288 (2 years), 1828–3288 (4 years), 1097–3288 (6 years) were used for calculation.

Simulation study 2

A second simulation study was performed to study whether the proposed method could approximate a nonlinear curve. We simulated 1000 samples from a model where the covariate is a cosine function of temperature and nonlinear with respect to the parameters.

$$\begin{aligned} &Y_t \sim Poisson(\mu_t) \\ &\ln(\mu_t) = 4.8 + 0.2\cos(\pi(x_t + 3)/28) + a_t \\ &a_t = \sum_{i=1}^{3} c_i\big(\ln(y^*_{t-i}) - (4.8 + 0.2\cos(\pi(x_t + 3)/28))\big) \\ &y^*_t = \max(y, \tau), \tau = 0.5 \end{aligned} \tag{5}$$

Here, the AR terms parameters are identical to (4): c_1 = 0.5, c_2 = 0.25, c_3 = 0.12.

As in simulation study 1, sample statistics were calculated over the time points 1828–3288 (year 2005–2008) for each of the 1000 realizations by GAM and GAMAR. The first sample was analysed just the same way as in simulation study 1, but here the mean and standard deviation of the parameter estimates from the two models for the 1000 samples are presented. To compensate for the absence of true parameters, we compared the two models visually, plotting the mean predicted values from the two models and the true effect. Moreover, the Pearson correlation coefficients of the fitted linear predictors from the two models and true effect were also calculated.

Application to temperature-mortality research

The real example came from three sources: daily mortality of Shanghai in 2001–2004 from Shanghai Municipal Center for Disease Control & Prevention; daily average temperature and humidity in 2001–2004 from Shanghai's Meteorological Bureau; and air pollution data (NO_2, SO_2, PM10) in 2001–2004 from (http://www.envir.gov.cn/airnews/), which was announced by the Shanghai Environmental Monitoring Centre.

Autocorrelation of the Pearson residuals from the two models are compared via figure. We also compared their parameter estimates, and their estimated temperature effects via figures. The R code for real example modelling and its outputs is included in Additional file 1.

Results

Simulation results

Simulation study 1

The ACF and PACF plot of the GAM Pearson residuals from the first sample indicate obvious autocorrelation (Figure 1). The Pearson estimate of the dispersion parameter [37] is $\hat{\phi} = 2.675 > 1$, indicating data overdispersion. Since the ACF tails off and PACF cuts off after lag

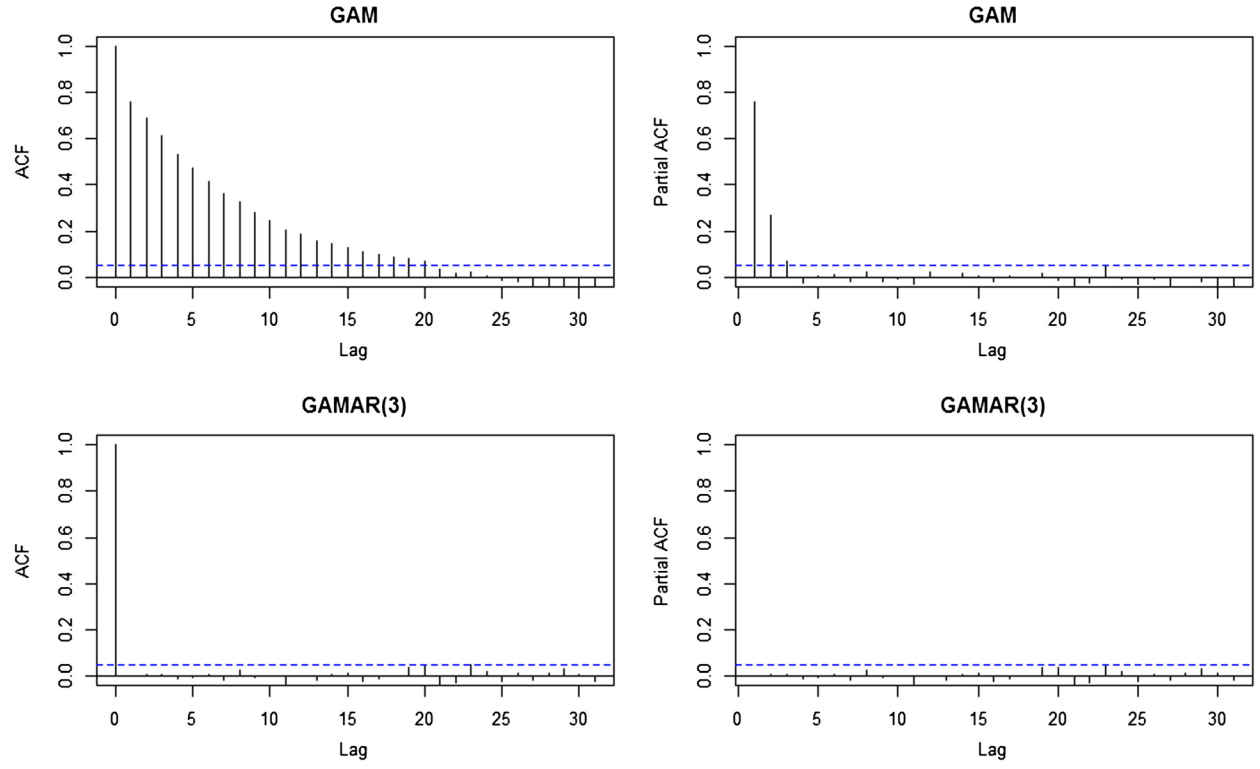

Figure 1 ACF and PACF of GAM and GAMAR(3) for Sample 1 from simulation study 1. These statistics are calculated on Pearson residuals $z_t = \frac{\hat{y}_t - y_t}{\sqrt{\hat{y}_t}}$ from both models.

3, it is natural to use GAMAR(3) instead. The model is given below:

$$\ln(E(y_t)) = ns(x_t, 5) + \sum_{i=1}^{3} c_i(\ln(y^*_{t-i}) - ns(x_{t-i}, 5))$$
$$y^*_t = \max(y, \tau), \tau = 0.5 \quad (6)$$

The ACF and PACF plot of the GAMAR(3) Pearson residuals suggest no autocorrelation, and the dispersion parameter estimate is now $\hat{\phi} = 1.0219 \approx 1$, indicating no overdispersion. Apparently, GAMAR(3) controls autocorrelation and overdispersion simultaneously, and Figure 2 indicates that the predicted spline function $\widehat{ns(x, 5)}$ from GAMAR(3) is much closer to the real model than that from GAM.

Concerning the general performance of GAM and GAMAR, Table 1 shows that the biases of the mean parameter estimates from GAM are almost the same as GAMAR. However, the coverages of the 95% confidence intervals from GAM are far less than 95%, while those from GAMAR are very close to 95%. Also, relative errors of the parameter estimates from GAM are larger than that of GAMAR.

Coverages for GAM corrected for overdispersion [15] are also shown in Table 1 as coverage2. This approach is the same as a standard GAM except that its standard error has been expanded by the dispersion parameter ϕ. As a result, the bias and relative errors remain unchanged, while the

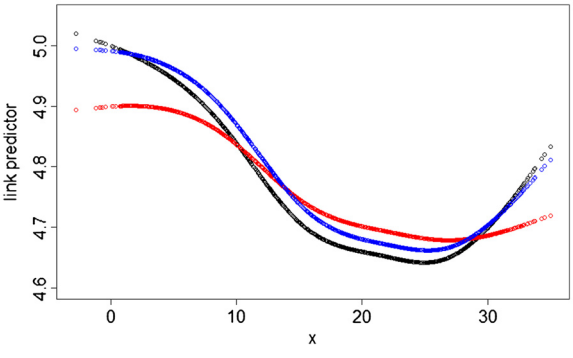

Figure 2 The temperature effects in link scale for Sample 1 from simulation study 1. Black: the true effect, Red: $\widehat{ns(x, 5)}$ from GAM, Blue: $\widehat{ns(x, 5)}$ from GAMAR(3).

Table 1 Results from GAM, GAMM, GAMAR(3) in simulation study 1

		GAM					GAMM (AR(3) correlation structure)				GAMAR(3)			
	TruPar	MeaEst	Bias	RelErr	Coverage	Coverage2	MeaEst	Bias	RelErr	Coverage	MeaEst	Bias	RelErr	Coverage
β_0	5.02	4.9991	−0.0209	0.0127	38	58.8	5.0065	−0.0135	0.0049	92	5.0182	−0.0018	0.0055	94.8
β_1	−0.35	−0.3574	−0.0074	0.1922	33.5	52.4	−0.3576	−0.0076	0.0757	88	−0.3561	−0.0061	0.0686	93.9
β_2	−0.36	−0.3672	−0.0072	0.229	35.3	56.4	−0.3662	−0.0062	0.0843	96	−0.362	−0.002	0.0885	94.6
β_3	−0.38	−0.382	−0.002	0.1531	34.2	55.3	−0.3787	0.0013	0.0608	96	−0.3757	0.0043	0.0733	93.8
β_4	−0.33	−0.3281	0.0019	0.428	38.6	62.6	−0.3418	−0.0118	0.1329	98	−0.3262	0.0038	0.1604	95.5
β_5	−0.15	−0.1495	0.0005	0.4924	35.4	55.0	−0.1581	−0.0081	0.202	92	−0.1472	0.0028	0.2201	93.8
	Mea_co		0.0067	0.2512	35.8	56.8		−0.0077	0.0934	93.7		0.0035	0.1027	94.4
c_1	0.5										0.4982	−0.0018	0.0422	95.3
c_2	0.25										0.2477	−0.0023	0.0947	93.4
c_1	0.12										0.1197	−0.0003	0.1737	94.7
	Mea_ar											0.0015	0.1035	94.5

TruPar=True Parameters= β_i or c_i.
MeaEst=Mean estimates= $\overline{\hat{\beta}}$ or $\overline{\hat{c}_i}$.
Bias= $\overline{\hat{\beta}_i}-\beta_i$ or $\overline{\hat{c}_i}-c_i$.
RelErr=Relative Error= $\overline{\left|\left(\hat{\beta}_i-\beta_i\right)/\beta_i\right|}$ or $\overline{|(\hat{c}_i-c_i)/c_i|}$.
Coverage: the percentage ofestimated 95%CI whichcovers true coefficient in all estimated 95%CI.
Coverage2: coverage from GAM which accounts for overdispersion.
Mea_co: Mean absolute Bias, RelErr, Coverage for parameters of covariates.
Mea_ar: Mean absolute Bias, RelErr, Coverage for parameters of AR terms.

confidence interval is broadened, coverage improved. However, the coverage is still far less than 95%.

The results of GAMM with AR(3) correlation structure and GAMAR(3) are very similar, but the former model is much more time consuming. This disparity will be described in discussion. Hence results from only 50 samples fitted by GAMM are summarized in Table 1.

To examine how the AR order in GAMAR influences the results, we also fitted the GAMAR(1), GAMAR(2), GAMAR(4), and GAMAR(5). As Table 1 in Additional file 2 shows, the coverage grows and relative error drops as the AR order increases from 1 to 3. For models with AR order larger than 3, there is little difference in coverage, relative error and bias among them.

The results for different time ranges are listed in Table 2 in Additional file 2, from which we can see that the larger the number of time-points, the lower the bias and relative errors, and the higher the coverage. This illustrates the asymptotic unbiased (consistency) property of MPLE.

Simulation study 2

In this simulation study, The ACF and PACF plot of the GAM Pearson residuals from the first sample also reveal obvious autocorrelation (Figure 3). Just as in simulation study 1, we can see that ACF tails off and PACF cuts off after lag 3 for GAM, suggesting AR(3) would be suitable. In contrast, the ACF and PACF of GAMAR(3) on the same data are both very close to 0. And Figure 4 indicates that the predicted spline function $\widehat{ns}(x,5)$ from GAMAR(3) is much closer to the real model than that from GAM.

In Figure 5, we can see the mean estimated temperature effects from GAMAR(3) is closer to the true effect than that from GAM. Meanwhile, the Pearson correlation coefficients between estimated temperature effect and true effect are also different (GAM: 0.9341; GAMAR(3): 0.9980), which means GAMAR(3) provides a better fit. Also, the standard deviations of estimates from GAM are larger than those from GAMAR (Table 2).

Table 2 Results from GAM and GAMAR(3) in simulation study 2

		GAM		GAMAR(3)			
	TruPar	MeaEst	Sd	MeaEst	Sd2	DifEst	DifSd
		5.006	0.0842	5.0246	0.0371	−0.0186	0.0471
		−0.2812	0.0864	−0.2806	0.0307	−0.0006	0.0557
		−0.3894	0.1074	−0.3859	0.0412	−0.0035	0.0662
		−0.4379	0.1011	−0.4321	0.0399	−0.0058	0.0612
		−0.3975	0.0825	−0.3888	0.0407	−0.0087	0.0418
		−0.4536	0.1886	−0.4442	0.0754	−0.0094	0.1132
		−0.2757	0.1001	−0.2646	0.0448	−0.0111	0.0553
c_1	0.5			0.4978	0.0260		
c_2	0.25			0.2482	0.0287		
c_3	0.12			0.1196	0.0261		

TruPar=True Parameters= c_i.
MeaEst=Mean estimates= $\overline{\hat{\beta}}$ or $\overline{\hat{c}_i}$.
Sd= Standard deviation of 1000 $\hat{\beta}$ and $\hat{c}_i$.
DifEst=MeaEst of GAM- MeaEst of GAMAR(3).
DifSe=Sd of GAM-Sd of GAMAR(3).

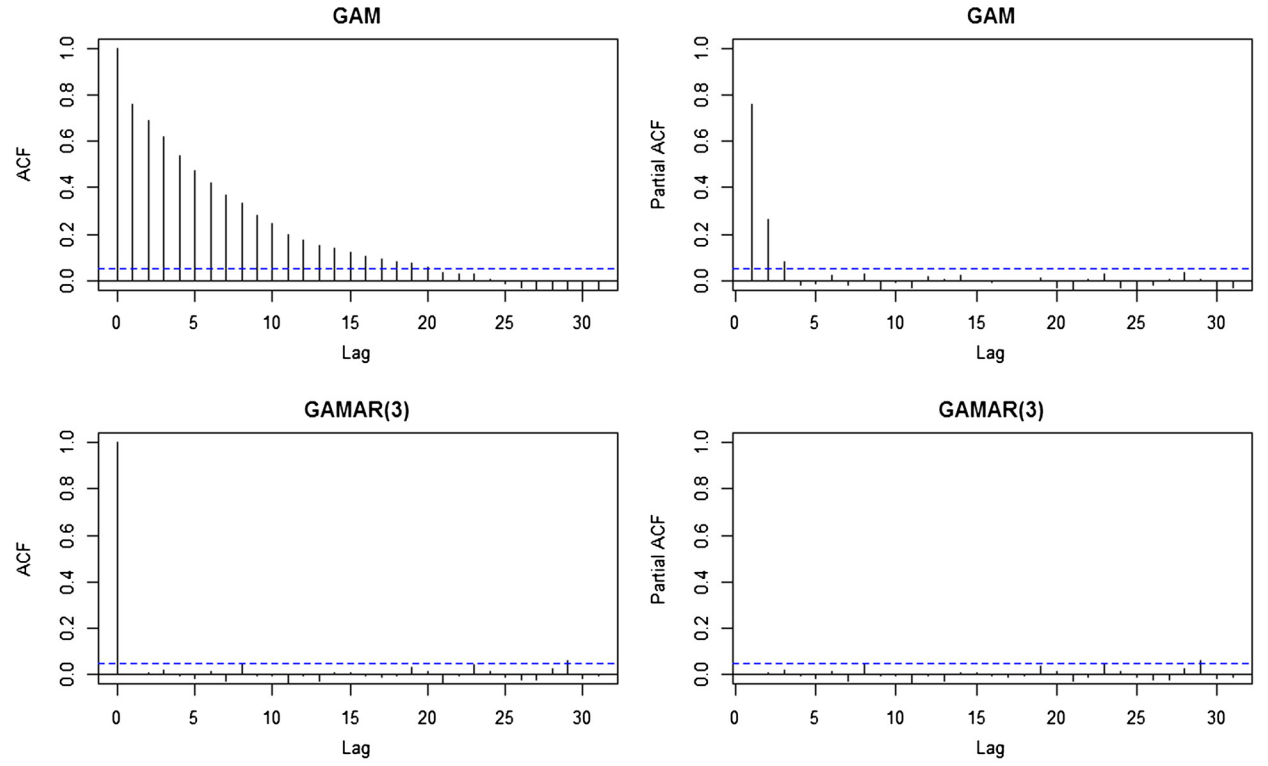

Figure 3 ACF and PACF of GAM and GAMAR(3) for Sample 1 from simulation study 2. These statistics are calculated on Pearson residuals $z_t = \frac{\hat{y}_t - y_t}{\sqrt{\hat{y}_t}}$ from both models.

Application to temperature-mortality research

In the real case analysis, GAM is first used. The long time trend of original observations is unobvious as shown in the left side of Figure 6. So a rather small df: 2, is used to control this secular trend. The daily mortality after adjustment is in the right side of Figure 6.

And the complete GAM is given below:

$$\begin{aligned} E(y_t) &= \mu_t = \exp(\eta_t) \\ \eta_t &= \beta_0 + ns(t) + ns(temp_{t-lag1}) \\ &\quad + ns(temp_{t-lag2}) + ns(pres_t) + ns(humi_t) \\ &\quad + ns(no2_t) + ns(so2_t) + ns(pm10_t) \\ &\quad + w_t(week_t) \end{aligned} \tag{7}$$

Lagged days of two temperature terms and df of all remaining natural spline functions are first determined in a sequence to minimize AIC. Then this set of parameters is used as starting value to find the final parameters which minimize AIC locally. The final parameters are: $lag1 = 4$, $lag2 = 10$, $\text{dftemp}_{t-lag1} = 5$, $\text{dftemp}_{t-lag2} = 4$, dfpres = 2, dfhumi = 2, dfno2 = 3, dfso2 = 2, dfpm10 = 3

The two lagged temperature terms separately represent short term temperature effect and long term temperature effect. The term $w_t(week_t) = \sum_{i=1}^{6} \beta_i I_i(week_t)$ stands for week effect, where $week_t$ represents corresponding day in a week for date t, and I_i $(week_t)$, i = 1,2,3,4,5,6 are indicator functions for the number of a day in a week. For each function, if $week_t = i$, I_i $(week_t) = 1$; if $week_t \neq i$, I_i $(week_t) = 0$.

Figure 7 shows that PACF all exceed 95% CI bounds for the autocorrelations (blue dashed line) [38] lags less than 5, and are contained within the bounds for lags larger than 4. So GAMAR(4) is then chosen to fit the data

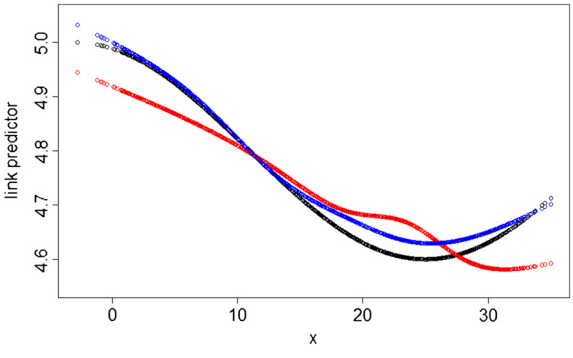

Figure 4 The temperature effects in link scale for Sample 1 from simulation study 2. Black: the true effect, Red: $\widehat{ns(x,6)}$ from GAM, Blue: $\widehat{ns(x,6)}$ from GAMAR(3).

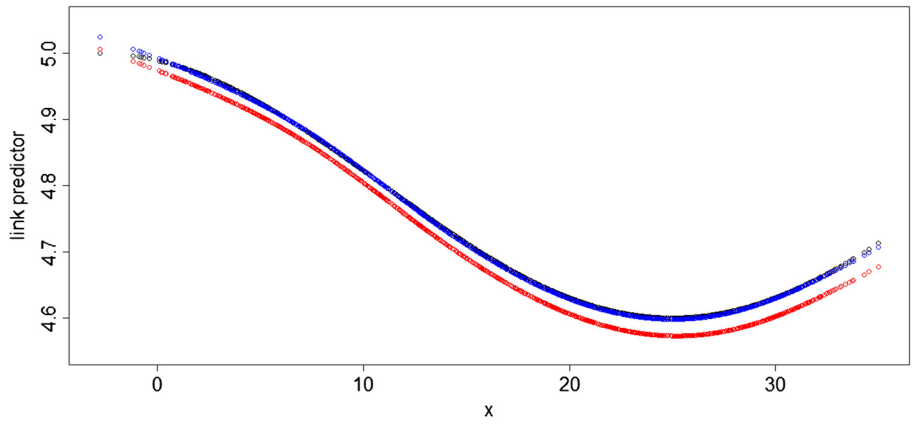

Figure 5 The averaged temperature effects in link scale from Simulation study 2. Black: the true effect, Red: $\widehat{ns}(x,6)$ from GAM, Blue: $\widehat{ns}(x,6)$ from GAMAR(3).

as described below:

$$
\begin{aligned}
E(y_t) &= \mu_t = \exp(\eta_t) \\
\eta_t &= f(x_t) + \sum_{i=1}^{4} c_i\left(\ln\left(y^*_{t-i}\right) - \eta_{t-i}\right) \\
f(x_t) &= \beta_t + ns(t) + ns(temp_{t-4}) + ns(temp_{t-10}) \\
&\quad + ns(pres_t) + ns(humi_t) + ns(no2_t) \\
&\quad + ns(so2_t) + ns(pm10_t) + w_t(week_t)
\end{aligned}
\tag{8}
$$

All ACF and PACF of Pearson residuals from (8) are now below 0.1, thus Pearson residuals from GAMAR(4) are close to white noise. The estimated coefficients of two lagged temperature terms and AR terms are given in Table 3. In Figure 8 and Figure 9, we can see that the estimated effects of $temp_{t\text{-}4}$, $temp_{t\text{-}10}$ from two models are different.

Discussion

In environmental epidemiological studies, the meteorological/environmental influences on human health indicators are often explained in a modelling framework. And the predominately used models are GLM/GAM. In our research, the drawbacks of GLM/GAM are studied, and their melioration: GAMAR, is given. Simulation studies reveal that GAMAR is more suitable when observations are autocorrelated.

While the true AR order is known in simulation studies, AR order needs to be determined in real case. Three enlightening perspectives into this issue are given below:

1. In time series analysis, ACF and PACF are good indicators of the order of process. CIs of an uncorrelated series are often treated as criterion for selecting orders [38]. For example, when PACF exceed the limit of CI at certain lag, autoregressive term at that lag needs to be modelled. This approach is used to preliminarily determine the AR order in the real case study.
2. Significance of AR terms can also be considered in modelling. If p-value of AR(n_0+1) from GAMAR (n_0+1) is larger than 0.05, while p-value of AR(n_0) from GAMAR(n_0) is less than 0.05, then GAMAR (n_0+1) is unnecessary and GAMAR(n_0) is a favorable model. To illustrate it, we also use GAMAR(5) to the real data, and find p-value of AR(5) to be 0.1460>0.05.
3. The goal of introducing AR terms is to control autocorrelation. So if Pearson residuals of GAMAR (n_0) show little sign of autocorrelation, or the PACF is within the CI for all lags, then the AR terms are adequate. In real data case, this requirement is justified.

In application, the first and second advices are practical methods to determine AR order preliminarily, and the third can be used to finally justify the choice. There have been extensive discussions about choosing AR orders [39].

We also want the covariate parameter estimation to be robust with respect to different AR orders. Simulation study provides some information for this issue: while the real model is GAMAR(3) in simulation study 1. GAMAR(1), GAMAR(2), GAMAR(4), GAMAR(5) are also used for estimation. From Table 1 in Additional file we can see that GAMAR(1), GAMAR(2) fit data much better than GAM; while they differ little from GAMAR (3). Since the true AR(3) parameter is rather small (0.12), probably when the effect of neglected AR term is small, covariate parameter estimation won't differ much. When the AR order is larger than the real, the estimation results are almost the same as GAMAR(3). In all,

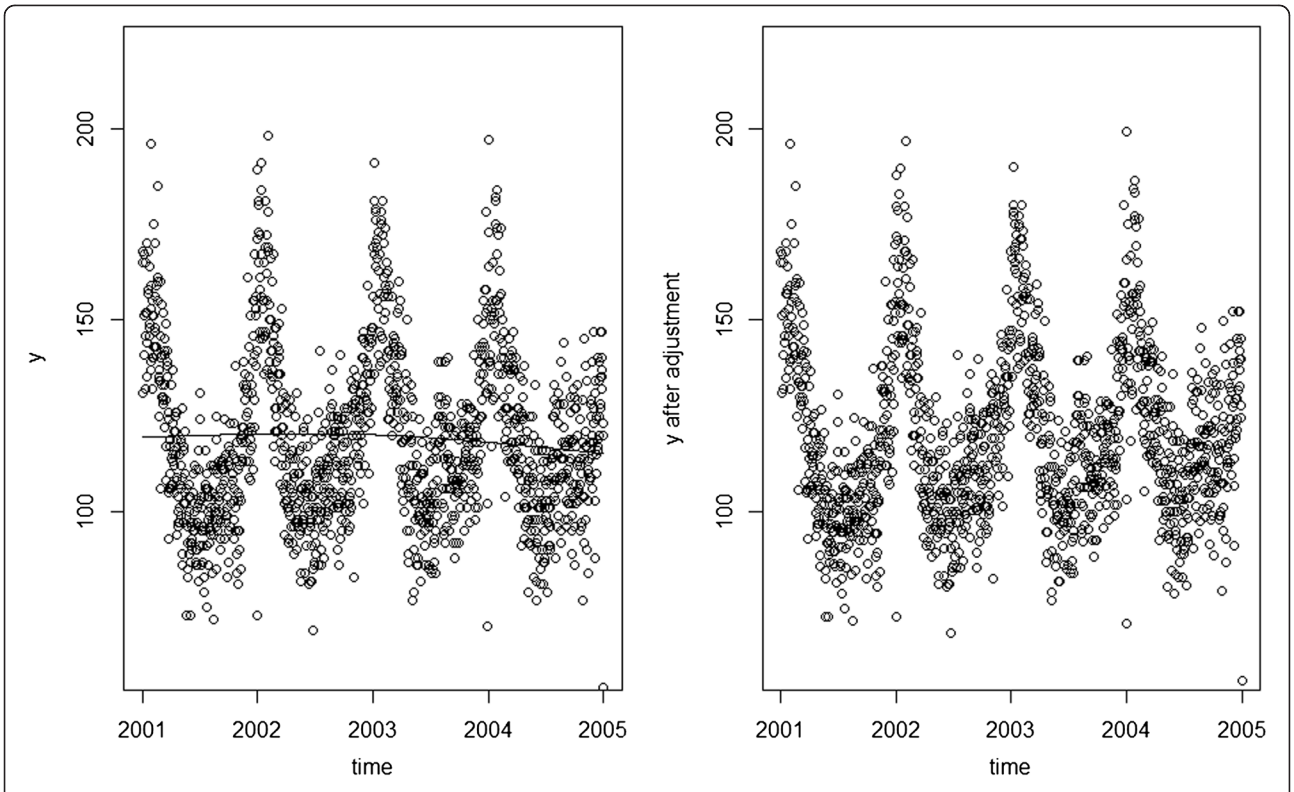

Figure 6 Mortality before and after adjustment for secular trend. Left: the original mortality, Right: the mortality adjusted for secular trend.

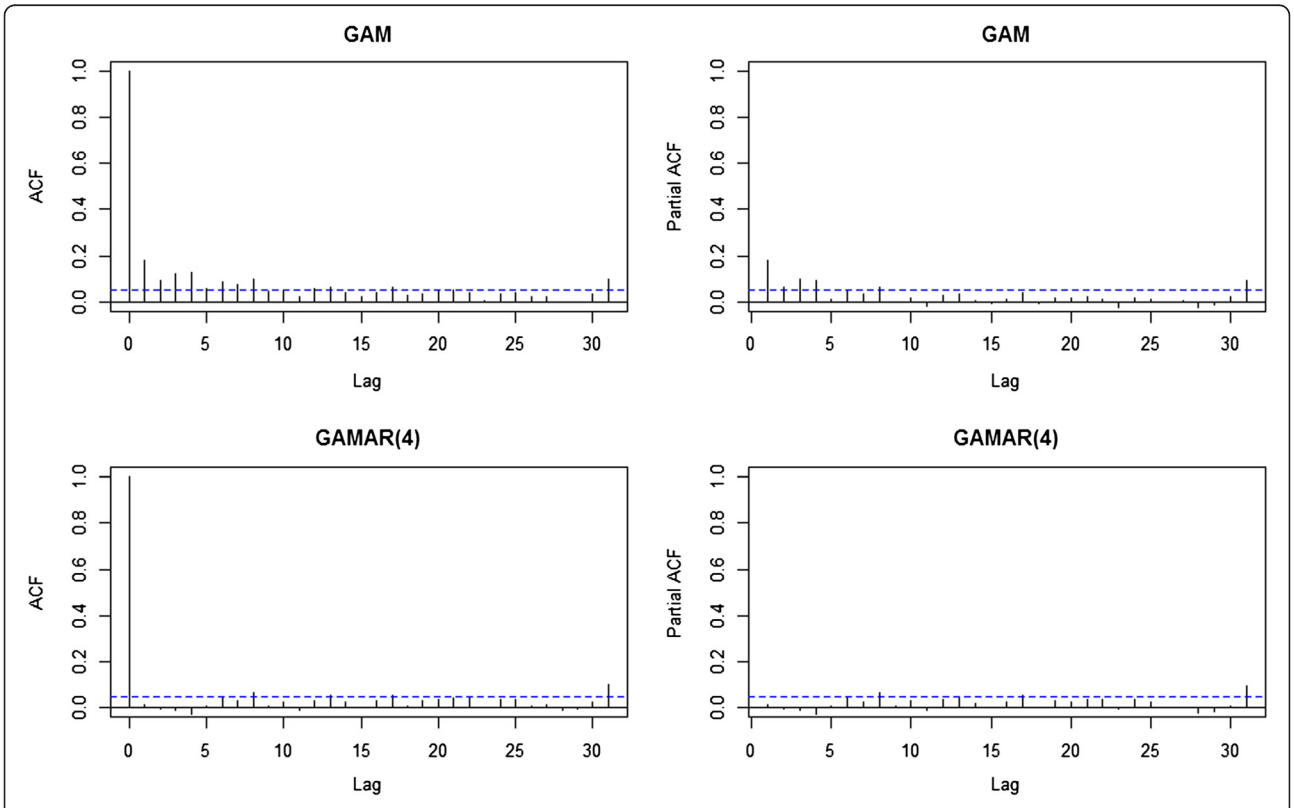

Figure 7 ACF and PACF of GAM and GAMAR(4) for real case. These statistics are calculated on Pearson residuals $z_t = \frac{\hat{y}_t - y_t}{\sqrt{\hat{y}_t}}$ from both models.

Table 3 temperature effects and AR estimates from GAM and GAMAR(4)

	GAM				GAMAR(4)			
	Estimate	Se	Z	Pr(>|z|)	Estimate	Se	Z	Pr(>|z|)
ns(temp1,5)1	0.2443	0.0212	−11.5349	0.0000	−0.1947	0.0254	−7.6621	0.0000
ns(temp1,5)2	0.2668	0.0281	−9.4929	0.0000	−0.2290	0.0322	−7.1108	0.0000
ns(temp1,5)3	0.3278	0.0258	−12.6878	0.0000	−0.2989	0.0290	−10.3113	0.0000
ns(temp1,5)4	0.3422	0.0495	−6.9160	0.0000	−0.2898	0.0568	−5.1052	0.0000
ns(temp1,5)5	0.2254	0.0283	−7.9578	0.0000	−0.2569	0.0320	−8.0179	0.0000
ns(temp2,4)1	0.2355	0.0197	−11.9372	0.0000	−0.1833	0.0248	−7.3986	0.0000
ns(temp2,4)2	0.1599	0.0225	−7.1233	0.0000	−0.1430	0.0263	−5.4312	0.0000
ns(temp2,4)3	0.2472	0.0443	−5.5748	0.0000	−0.1978	0.0526	−3.7603	0.0002
ns(temp2,4)4	0.0752	0.0251	−3.0007	0.0027	−0.0559	0.0297	−1.8838	0.0596
AR1					0.1426	0.0211	6.7521	0.0000
AR2					0.0773	0.0223	3.4664	0.0005
AR3					0.1179	0.0221	5.3466	0.0000
AR4					0.1259	0.0223	5.6398	0.0000

Estimate: estimate for a parameter.
Se: Standard Error for a parameter.
Z=Estimate/Se, which approximately follows N(0,1).
Pr(>|z|): the probability of obtaining Z at least as extreme as the one that was actually observed, assuming that the true value is 0. This time P value is derived from N(0,1).

covariate parameter estimation is robust with respect to disturbance of AR order.

GAMM can also give good estimates as shown in Table 1. However, while the speed of GAMM in fitting short time series is acceptable, say 2.05 seconds for sample with 100 observations and 10.17 seconds for sample with 200 observations (both executed in server), the time grows nonlinearly with respect to the length of time series and thus GAMM would become computationally formidable when time series data is long. For example, one sample of 1461 observations in simulation study 1 would take about 1 hour and 7 minutes for GAMM, while only 0.02 seconds for GAMAR (both executed in server): the former is 2×10^5 times the latter. Thus GAMM is much more computing intensive than GAMAR when sample size is large. Therefore, we only calculated 50 samples for GAMM in simulation study 1.

While we use natural cubic spline as smoother in the model, penalized spline and smoothing spline can also be implemented. With natural splines, one constructs a spline basis with knots at fixed locations throughout the range of the data. Smoothing splines and penalized splines have circumvented the problem of choosing the knot locations by constructing a very large spline basis and then penalizing the spline coefficients to reduce the effective number of df [40]. Despite the flexibility the penalized way provided, [40] identified both fully parametric and nonparametric methods can perform well in similar studies. Thus natural spline smoother can still address many practical problems. In addition, our model

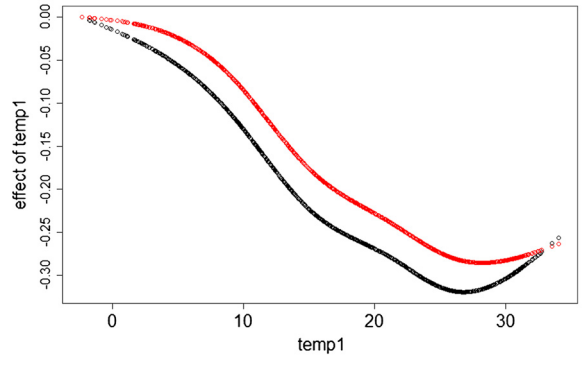

Figure 8 Effects of $temp_{t-4}$. Black: $\widehat{ns(temp_{t-4},6)}$ from GAM, Red: $\widehat{ns(temp_{t-4},6)}$ from GAMAR(4).

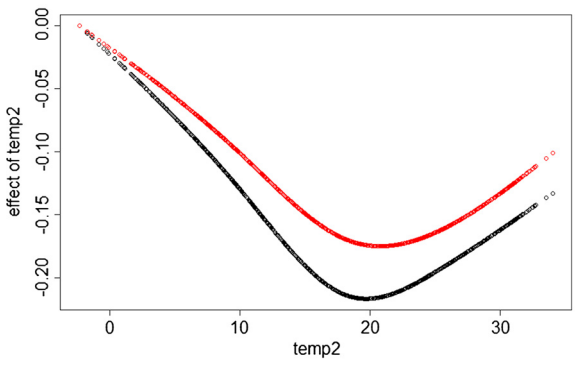

Figure 9 Effects of $temp_{t-10}$. Black: $\widehat{ns(temp_{t-10},6)}$ from GAM, Red: $\widehat{ns(temp_{t-10},6)}$ from GAMAR(4).

can be straightforwardly extended to the nonparametric way by including the penalty.

For natural spline with B-spline basis, selecting the df is of essential importance for application. A general approach is to use a data-driven method and to select the number of df which optimizes a particular criterion [40], like AIC. We used AIC to determine df of all splines except that of time in the real case. Another strategy is to use a df based on background knowledge or previous studies. For example, natural spline of time is chosen to represent the long term trend among different years, and itself shouldn't contain any yearly fluctuation. Whether df meets the requirement is judged by comparing the trend of observations before and after adjustment, as well as the shape of the spline function visually in Figure 6. Section 2.1 of [40] gives a full treatment of this issue.

Besides including lagged temperature effects additively in the model, [41,42] have developed a family of distributed lag non-linear models (DLNM), which can simultaneously represent non-linear exposure-response dependencies and delayed effects. Still, AR terms can be aptly added to DLNM just as what we've done to GAM.

Finally, the model can also be applied in other areas. Researchers can use GAMAR for their specific research purposes. The most immediate extensions are other environmental epidemiological studies, specifically studies quantifying relationships between air pollution and mortality. Air pollution functions in a different way from temperature to impact human health, and there are also some differences between their modelling: 1) the air pollution-mortality relation is often simplified as linear, for such simplicity facilitates parameter estimation and interpretation; 2) distributed lag model is often used, since the effects of air pollution can last for many days; 3) cumulative effect: the total impact of pollution of a certain day over a period of following days, represented by the sum of parameters at all lags, is of great interest [9]. Such differences pose new questions for further study, like assessing the impact of AR terms on estimated cumulative effect rather than a single parameter.

Conclusions

This article proposes GAMAR for fitting time series data with explanatory variables and autoregressive terms. Two simulation studies with functions that approximate the response from the real example showed that GAMAR performed better than GAM. In the real example, there was residual autocorrelation with GAM, but little sign of autocorrelation with GAMAR. Also, different estimates of the temperature effects were obtained with GAMAR and GAM.

Appendix

Modified Newton's method

Given partial likelihood, maximum partial likelihood estimators are solved by a modified Newton's method. For Newton's method, the iteration goes:

$$\theta_{m+1} = \theta_m - \left(\frac{\partial^2 \ln(PL)}{\partial\theta_i \partial\theta_j}\right)^{-1} \frac{\partial \ln(PL)}{\partial\theta}|_{\theta=\theta_m},$$

until convergence.

For a modified Newton's method, the iteration goes:

$$\theta_{m+1} = \theta_m + \Gamma_*^{-1}(\theta_m) \frac{\partial \ln(PL)}{\partial\theta}|_{\theta=\theta_m}, \quad (9)$$

until it convergence.

Here $\Gamma(\theta) = -\frac{\partial^2 \ln(PL)}{\partial\theta\partial\theta^T}$, which is the information matrix, and $\Gamma_*^{-1}(\theta_m)$ is a modified version of $\Gamma^{-1}(\theta_m)$.

In (9):

$$\frac{\partial \ln(PL)}{\partial\theta_i} = \frac{\partial \sum_{t=1}^{n} (y_t \ln(\mu_t) - \mu_t - \ln(y_t!))}{\partial\theta_i} = \sum_{t=1}^{n} (y_t - \mu_t) \frac{\partial\eta_t}{\partial\theta_i},$$

$$\Gamma(\theta) = -\frac{\partial \sum_{t=1}^{n} (y_t - \mu_t) \frac{\partial\eta_t}{\partial\theta_i}}{\partial\theta_j} = \begin{pmatrix} A & B \\ B^T & C \end{pmatrix}.$$

Since $\ln(\mu_t) = \eta_t = \sum_{i=1}^{m} \beta_i X_{ti} + \sum_{j=1}^{p} c_j \left(\ln\left(y^*_{t-j}\right) - \sum_{i=1}^{m} \beta_i X_{t-j,i} \right)$, then:

$$\frac{\partial\eta_t}{\partial\beta_i} = X_{ti} - \sum_{r=1}^{p} c_r X_{t-r,i}, \frac{\partial\eta_t}{\partial c_i} = \ln\left(y^*_{t-i}\right) - \sum_{k=1}^{m} \beta_k X_{t-i,k},$$

$$\frac{\partial^2\eta_t}{\partial\beta_i\partial\beta_j} = 0, \frac{\partial^2\eta_t}{\partial c_i \partial c_j} = 0, \frac{\partial^2\eta_t}{\partial\beta_i \partial c_j} = -X_{t-j,i}.$$

So

$$A = \left(\sum_{t=1}^{n} \mu_t (X_{ti} - \sum_{r=1}^{p} c_r X_{t-r,i})(X_{tj} - \sum_{r=1}^{p} c_r X_{t-r,j}) \right)_{mm},$$

$$B = \left(\sum_{t=1}^{n} \left(\mu_t (X_{ti} - \sum_{r=1}^{p} c_r X_{t-r,i}) \left(\ln\left(y^*_{t-j}\right) - \sum_{k=1}^{m} \beta_k X_{t-j,k} \right) + (y_t - \mu_t) X_{t-j,i} \right) \right)_{mp,}$$

$$C = \left(\sum_{t=1}^{n} \left(\ln\left(y^*_{t-i}\right) - \sum_{k=1}^{m} \beta_k X_{t-i,k} \right) \left(\ln\left(y^*_{t-j}\right) - \sum_{k=1}^{m} \beta_k X_{t-j,k} \right) \right)_{pp}.$$

And the modified inverse matrix $\Gamma_*^{-1}(\theta_m)$ is defined as: if $\Gamma(\theta)$ is reversible, then $\Gamma_*^{-1}(\theta) = \Gamma^{-1}(\theta)$;

If $\Gamma(\theta)$ is irreversible. And its eigenvalue are λ_1, λ_2,. . .,λ_{mp}, then we can find orthogonal matrix P, which satisfies:

$$P^T\Gamma(\theta)P = diag(\lambda_1, \ldots, \lambda_{mp}).$$

Let $\lambda_i^* = \max(\lambda_i, \delta)$, $\delta = 0.01$, $i = 1, 2, \cdots, mp$, then:

$$\Gamma_*^{-1}(\theta) = Pdiag\left(\lambda_1^{*-1}, \ldots, \lambda_{mp}^{*\ -1}\right)P^T.$$

Such procedure ensures $\Gamma_*^{-1}(\theta_m)$ to be positive definite.

Additional files

Additional file 1: R code for this study. This file contains core R codes for this study, including the function of GAMAR, data generation in simulation studies, data fitting in simulation studies, real case analysis (including the procedure to choose the parameters) and generation of tables and figures. This file also contains a brief description of every R program.

Additional file 2: Additional Tables. This file contains 2 tables related to simulation study 1.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

LY conducted the analysis and drafted the manuscript. GQ participated in discussions and provided advice. NZ inspired and tutored LY, and reviewed the paper. CW and GS collected the mortality data for real case study. All authors read and approved the final manuscript.

Acknowledgements

This research is supported by a project 30972551 from National Natural Science Foundation of China (NSFC).

Author details

[1]Department of Biostatistics, School of Public Health, Fudan University, Shanghai, China. [2]Vital Statistics Division, Shanghai Municipal Center for Disease Control & Prevention, Shanghai, China.

Received: 30 January 2012 Accepted: 9 October 2012
Published: 30 October 2012

References

1. Ellis FP, Nelson F: **Mortality in the elderly in a heat wave in New York City, August 1975.** *Environ Res* 1978, **15**:504–512.
2. McKee CM: **Deaths in winter: can Britain learn from Europe?** *Eur J Epidemiol* 1989, **5**(2):178–182.
3. Ballester F, Corella D, Perez-Hoyos S, Saez M, Hervas A: **Mortality as a function of temperature. A study in Valencia, Spain, 1991–1993.** *Int J Epidemiol* 1997, **26**(3):551–561.
4. Ha J, Shin Y, Kim H: **Distributed lag effects in the relationship between temperature and mortality in three major cities in South Korea.** *Sci Total Environ* 2011, **409**(18):3274–3280.
5. Yu W, Guo Y, Ye X, Wang X, Huang C, Pan X, Tong S: **The effect of various temperature indicators on different mortality categories in a subtropical city of Brisbane, Australia.** *Sci Total Environ* 2011, **409**(18):3431–3437.
6. Basu R, Malig B: **High ambient temperature and mortality in California: Exploring the roles of age, disease, and mortality displacement.** *Environ Res* 2011, **111**(8):1286–1292.
7. Patz JA, Engelberg D, Last J: **The effects of changing weather on public health.** *Annu Rev Publ Health* 2000, **21**:271–307.
8. Baker DANM: *Environmental Epidemiology: Study Methods and Application*. Oxford: Oxford University Press; 2008.
9. Peng RD, Dominici F: *Statistical methods for environmental epidemiology in R*. New York: Springer; 2008.
10. Ostro BD, Roth LA, Green RS, Basu R: **Estimating the mortality effect of the July 2006 California heat wave.** *Environ Res* 2009, **109**(5):614–619.
11. Hertel S, Le Tertre A, Jockel KH, Hoffmann B: **Quantification of the heat wave effect on cause-specific mortality in Essen, Germany.** *Eur J Epidemiol* 2009, **24**(8):407–414.
12. Goldberg MS, Gasparrini A, Armstrong B, Valois MF: **The short-term influence of temperature on daily mortality in the temperate climate of Montreal, Canada.** *Environ Res* 2011, **111**(6):853–860.
13. Hajat S, Kovats RS, Atkinson RW, Haines A: **Impact of hot temperatures on death in London: a time series approach.** *J Epidemiol Community Health* 2002, **56**(5):367–372.
14. Ballester J, Robine JM, Herrmann FR, Rodo X: **Long-term projections and acclimatization scenarios of temperature-related mortality in Europe.** *Nat Commun* 2011, **2**:358.
15. McCullagh P, Nelder JA: *Generalized Linear Models*. London: Chapman and Hall/CRC; 1989.
16. Hastie TJ, Tibshirani RJ: *Generalized Additive Models*. London: Chapman and Hall/CRC; 1990.
17. Braga AL, Zanobetti A, Schwartz J: **The effect of weather on respiratory and cardiovascular deaths in 12 U.S. cities.** *Environ Health Perspect* 2002, **110**(9):859–863.
18. Carson C, Hajat S, Armstrong B, Wilkinson P: **Declining vulnerability to temperature-related mortality in London over the 20th century.** *Am J Epidemiol* 2006, **164**(1):77–84.
19. Muggeo VM: **Modeling temperature effects on mortality: multiple segmented relationships with common break points.** *Biostatistics* 2008, **9**(4):613–620.
20. Kim H, Ha JS, Park J: **High temperature, heat index, and mortality in 6 major cities in South Korea.** *Arch Environ Occup Health* 2006, **61**(6):265–270.
21. Curriero FC, Heiner KS, Samet JM, Zeger SL, Strug L, Patz JA: **Temperature and mortality in 11 cities of the eastern United States.** *Am J Epidemiol* 2002, **155**(1):80–87.
22. Ren C, Williams GM, Morawska L, Mengersen K, Tong S: **Ozone modifies associations between temperature and cardiovascular mortality: analysis of the NMMAPS data.** *Occup Environ Med* 2008, **65**(4):255–260.
23. Hoffmann B, Hertel S, Boes T, Weiland D, Jockel KH: **Increased cause-specific mortality associated with 2003 heat wave in Essen, Germany.** *J Toxicol Environ Health A* 2008, **71**(11–12):759–765.
24. Anderson BG, Bell ML: **Weather-related mortality: how heat, cold, and heat waves affect mortality in the United States.** *Epidemiology* 2009, **20**(2):205–213.
25. Schwartz J: **The distributed lag between air pollution and daily deaths.** *Epidemiology* 2000, **11**(3):320–326.
26. Hamilton JD: *time series analysis*. Princeton, NJ: Princeton University Press; 1994.
27. Kutner M, Nachtsheim C, Neter J, Li W: *Applied Linear Statistical Models Fifth Edition*. New York: McGraw-Hill/Irwin; 2005.
28. Barron DN: **The analysis of count data: overdispersion and autocorrelation.** *Sociol Methodol* 1992, **22**:179–220.
29. Zeger SL, Liang KY: **Longitudinal data analysis for discrete and continuous outcomes.** *Biometrics* 1986, **42**(1):121–130.
30. Liang KY, Zeger SL: **Longitudinal data analysis using generalized linear models.** *Biometrika* 1986, **73**(1):13–22.
31. Pinheiro JC, Bates DM: *Mixed-Effects Models in S and S-plus*. New York: Springer; 2000.
32. Wood SN: *Generalized Additive Model: an introduction with R*. New York: Chapman and Hall/CRC; 2006.
33. Lin X: *SAS Macro GAMM1 to fit generalized additive mixed models using smoothing splines.*: Harvard University; http://scholarjsagotskydev.iq.harvard.edu/amaity/software/gamm1.
34. *The GLIMMIX Procedure.*: SAS Institute inc; 2006. support.sas.com/rnd/app/papers/glimmix.pdf.
35. Benjamin MA, Rigby RA, Stasinopoulos DM: **Generalized autoregressive moving average models.** *J Am Stat Assoc* 2003, **98**(461):214–223.
36. Cox DR, Gudmundsson G, Lindgren G, Bondesson L, Harsaae E, Laake P, Juselius K, Lauritzen SL: **Statistical analysis of time series: some recent developments [with discussion and reply].** *Scand J Stat* 1981, **8**(2):93–115.

37. Ruoyan M: *Estimation of Dispersion Parameters in GLMs with and without Random Effects.*: Stockholm University; 2004. http://www2.math.su.se/matstat/reports/serieb/2004/rep5/report.pdf.
38. Box G, Jenkins GM, Reinsel G: *Time Series Analysis: Forecasting & Control*. 3rd edition.: Prentice Hall; 1994.
39. Zhang S, Qi L: *shi jian xu lie fen xi jian ming jiao cheng*. Beijing: Beijing jiaotong university press; 2003.
40. Peng RD, Dominici F, Louis TA: **Model choice in time series studies of air pollution and mortality.** *J R Stat Soc A Stat Soc* 2006, **169**(2):179–203.
41. Gasparrini A, Armstrong B, Kenward MG: **Distributed lag non-linear models.** *Stat Med* 2010, **29**(21):2224–2234.
42. Gasparrini A: **Distributed lag linear and non-linear models in R: The package dlnm.** *J Stat Softw* 2011, **43**(8):1–20.

doi:10.1186/1471-2288-12-165
Cite this article as: Yang *et al.*: Using a generalized additive model with autoregressive terms to study the effects of daily temperature on mortality. *BMC Medical Research Methodology* 2012 **12**:165.