

# AALBORG UNIVERSITY

**Examining secular trend and seasonality in count  
data using dynamic generalized linear modelling:  
A new methodological approach to hospital discharge data on  
myocardial infarction**

by

S. Lundbye-Christensen, C. Dethlefsen, A. Gorst-Rasmussen, T. Fischer,  
H.C. Schönheyder, K.J. Rothman and H.T. Sørensen

R-2007-16

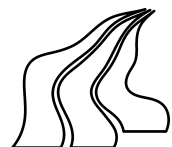
July 2007

DEPARTMENT OF MATHEMATICAL SCIENCES  
AALBORG UNIVERSITY

Fredrik Bajers Vej 7 G ■ DK-9220 Aalborg Øst ■ Denmark

Phone: +45 96 35 80 80 ■ Telefax: +45 98 15 81 29

URL: [www.math.auc.dk/research/reports/reports.htm](http://www.math.auc.dk/research/reports/reports.htm)



# Examining secular trend and seasonality in count data using dynamic generalized linear modelling:

## A new methodological approach to hospital discharge data on myocardial infarction

S Lundbye-Christensen<sup>1</sup>, C Dethlefsen<sup>2</sup>, A Gorst-Rasmussen<sup>1</sup>, T Fischer<sup>3</sup>, HC Schönheyder<sup>4</sup>, KJ Rothman<sup>5,6</sup>, HT Sørensen<sup>5,7</sup>

<sup>1</sup> Department of Mathematical Sciences, Aalborg University, Aalborg, Denmark

<sup>2</sup> Center for Cardiovascular Research, Aalborg Hospital, Aarhus University Hospital, Aalborg, Denmark

<sup>3</sup> Department of Medicine, Viborg Hospital, Denmark

<sup>4</sup> Department of Clinical Microbiology, Aalborg Hospital, Aarhus University Hospital, Aalborg, Denmark

<sup>5</sup> Department of Epidemiology, Boston University School of Public Health, Boston, USA

<sup>6</sup> Department of Medicine, Boston University School of Public Health, Boston, USA

<sup>7</sup> Department of Clinical Epidemiology, Aarhus University Hospital, Aarhus, Denmark

Short title: Trend and seasonality in count data using dynamic models

Correspondence to:  
Søren Lundbye-Christensen  
Department of Mathematical Sciences  
Aalborg University  
Frederik Bajers Vej 7G  
9220 Aalborg Ø  
Denmark  
Telephone: 0045-9635 8860  
Fax: 0045-98158129  
E-mail: [soren@math.aau.dk](mailto:soren@math.aau.dk)

## ACKNOWLEDGEMENTS

We wish to thank Stefan Christensen, Statens Serum Institut, for his input regarding the model fitting procedure.

## ABSTRACT

### **Aims**

Time series of incidence counts often show secular trends and seasonal patterns. We present a model for incidence counts capable of handling a possible gradual change in growth rates and seasonal patterns, serial correlation and overdispersion.

### **Methods**

The model resembles an ordinary time series regression model for Poisson counts. It differs in allowing the regression coefficients to vary gradually over time in a random fashion.

### **Data**

In the period January 1980 to 1999, 17,989 incidents of acute myocardial infarction were recorded in the county of Northern Jutland, Denmark. Records were updated daily.

### **Results**

The model with a seasonal pattern and an approximately linear trend was fitted to the data, and diagnostic plots indicate a good model fit. The analysis with the dynamic model revealed peaks coinciding with influenza epidemics. On average the peak-to-trough ratio is estimated higher using the dynamic model, and gradual changes in seasonal pattern is seen.

### **Conclusion**

Analyses conducted with this model provide detailed insights not available from more traditional analyses. The post-hoc analysis gives ideas to identify possible causal factors and confounders.

Index words: incidence; Kalman filter; overdispersion; peak-to-trough ratio; Poisson; serial correlation; state space model; time series.

## INTRODUCTION

A number of epidemiological studies of cardiovascular disease have identified changes in long-term trends and a distinctive seasonal pattern, with a winter peak. Current statistical methods have been applied to quantify the trend and seasonal patterns<sup>1-7</sup>, but the issue of how to deal with temporally changing seasonal patterns remains unresolved. In this paper, we propose a statistical model that allows for gradual changes in both the seasonal pattern and the shape of the long-term trend.

A cyclic variation in disease incidence implies a cyclic variation in one or more causes. Therefore the analysis of seasonal pattern is important in the search for possible causes. Corresponding changes over time in seasonal pattern in disease and a potential causal factor may indicate a causal relationship. However, interpretation of changes in seasonal pattern is not straight-forward. A strong seasonal pattern can appear to weaken if disease incidence from a non-seasonal cause increases disease occurrence. The role of the seasonal factor may be unchanged, but the change in the extraneous cause could affect it indirectly.

Existing methods for analysing chronobiological time series of normally distributed and Poisson-distributed observations include Cosinor rhythmometry<sup>8</sup> and the Edwards method<sup>9</sup>. Neither considers the presence of secular trends and other effects, and the seasonal pattern, a simple

sinusoid, remains fixed over time. A generalized linear model for Poisson observations can handle seasonal components via a harmonic representation, and the secular trend via a polynomial or a spline representation<sup>10</sup>. Even more flexible generalized additive models<sup>11</sup> have been used to control for seasonality and trend in studies that link an explanatory variable (air pollution) to morbidity and mortality<sup>12,13</sup>. However, a harmonic representation does not allow for gradual changes over time, while splines and additive models do not recognize the pattern as a seasonal variation.

It is possible to extend the above methods to handle both overdispersion and serial correlation in incidence data. The generalized estimating equation approach<sup>14</sup> offers a method of parameter estimation and hypothesis testing which accounts for overdispersion and correlation. Autoregressive Poisson models have also been suggested<sup>15,16</sup>.

Temporal changes in the seasonal pattern are not easily incorporated into such models, however, and literature on modelling these types of changes is scarce. A simple approach is to apply binning (*i.e.*, division into subintervals) to the time axis and then to analyse subintervals separately<sup>17,18</sup>. Binning of the time axis in combination with spline regression has been used previously to assess coronary mortality<sup>19</sup>. However, the use of binning implies abrupt rather than smooth changes between subintervals, with the risk of residual confounding.

We present a dynamic generalized linear model<sup>20</sup> incorporating a gradual, smooth change in the seasonality pattern over time, as well as overdispersion and serial correlation between observations. To demonstrate its use, we applied the model to count data on acute myocardial infarction taken from a population-based hospital discharge registry. For simplicity, the model does not take into account explanatory variables, but such variables could easily be added.

## METHODS

We consider a sequence of counts of events observed at given times, for example daily occurrence of acute myocardial infarctions in a given region over a given period. Denote by  $y_k$  the count at time point  $t_k$ . The time points need not be evenly spaced. We assume that the count  $y_k$  has a Poisson distribution with the expected number of events (the intensity) being  $\lambda_k$ . Secular trend and seasonality are related to intensity via the formula

$$\log(\lambda_k) = S_k + T_k,$$

where  $S_k$  describes the seasonal component and  $T_k$  describes the secular trend.

We model the seasonal component  $S_k$  at each time point  $t_k$  as a sum of  $p$  sinusoids of the form

$$S_{jk} = a_{jk} \sin(j \cdot 2\pi / T \cdot t_k) + b_{jk} \cos(j \cdot 2\pi / T \cdot t_k), \quad j = 1, \dots, p, \quad (1)$$

where  $T$  is a pre-defined period, for example a year. Over the period from 0 to  $T$ ,  $S_{jk}$  has  $j$  peaks and  $j$  troughs. This yields a flexible model for seasonality patterns in which multiple peaks and troughs are allowed within the same period. Using standard trigonometric formulae, the above equation can be rewritten as  $A_{jk} \cos(j \cdot 2\pi / T \cdot t_k + \phi_{jk})$  where  $A_{jk} = \sqrt{a_{jk}^2 + b_{jk}^2}$  and  $\phi_{jk} = \arctan(a_{jk}/b_{jk})$ . Hence the  $a_{jk}$ 's and  $b_{jk}$ 's govern both the amplitude and phase of the sinusoids describing  $S_k$ .

The secular trend evolves over time as

$$T_k = T_{k-1} + \Delta t_k \alpha_k,$$

where  $\Delta t_k = t_k - t_{k-1}$  is the time elapsed since the last observation and  $\alpha_k$  is the slope. Note that the logarithmic scale corresponds on the original scale to an exponential secular trend.

In the case where  $a_{jk}$ ,  $b_{jk}$ , and  $\alpha_k$  do not depend on  $k$ , the model reduces to a standard generalized linear model<sup>21</sup>. Such a model has been used previously, with binning of the time axis, to study the trend and seasonal pattern of the incidence of acute myocardial infarction<sup>18</sup>. The approach presented in the present paper is more general, in that the parameters  $a_{jk}$ ,  $b_{jk}$ , and  $\alpha_k$  are allowed to change gradually over time in a random manner. Consequently, the intensity  $\lambda_k$  also becomes random. We assume that each of the parameters follows a simple random walk:

$$\begin{aligned} a_{jk} &= a_{j,k-1} + \varepsilon_{jk} \\ b_{jk} &= b_{j,k-1} + v_{jk} \\ \alpha_k &= \alpha_{k-1} + \omega_{jk} \end{aligned}$$

where  $\varepsilon_{jk}$ ,  $v_{jk}$  and  $\omega_k$  are independent normally distributed random variables with mean zero and variances  $\Delta t_k W$  (for  $\varepsilon_{jk}$ ,  $v_{jk}$ ) and  $\Delta t_k V$  (for  $\omega_k$ ). This extension of the generalized linear model to parameters varying over time is a special case of a dynamic generalized linear model<sup>22</sup>. In this model, inference can be performed using the iterated extended Kalman smoother<sup>20</sup>, which is implemented in the package `sspir`<sup>23</sup> for R<sup>24</sup>. The iterated extended Kalman smoother approximates the conditional distribution of  $a_{jk}$ ,  $b_{jk}$ , and  $\alpha_k$ , given all the available information. The conditional distribution can be used subsequently to extract information regarding the seasonal component, the secular trend component, and derived entities. The unknown variance parameters  $V$  and  $W$  must be specified prior to applying the iterated extended Kalman smoother. They can be estimated using standard maximum likelihood methods<sup>25</sup>.

In a generalized linear model in which the parameters  $a_{jk}$ ,  $b_{jk}$  do not depend on time through  $k$ , one must assume a constant seasonal pattern over the entire study period. In contrast, the dynamic model allows for a gradually changing seasonal pattern in which the amplitude and phase of the sinusoids (1) are drifting. This allows investigators to quantify changes over time within seasonal

patterns. The same reasoning applies to the secular trend, where the random walk model for the growth rate  $\alpha_k$  allows the trend component to change over the long term and adapt smoothly to the observations.

There are two additional important consequences of modelling the parameters  $a_{jk}$ ,  $b_{jk}$ , and  $\alpha_k$  as randomly fluctuating. First, we have accounted for overdispersion in the count data by adding an ‘extra layer’ of randomness. Because  $\text{Var}(y_i | \lambda_i) = \lambda_i$ , we have  $\text{Var}(y_i) = \text{E}(y_i) + \text{Var}(\lambda_i)$ , according to the law of total variance. Second, we have allowed for serial correlation in data by using the random walk model for the underlying parameters.

The peak-to-trough-ratio (PTR) of the seasonal component is a measure of the relative risk of disease in the peak period relative to the trough and is often of primary interest when quantifying seasonality in incidence studies. Note that since the PTR is derived from parameters which vary over time, it is itself time-varying and also can be estimated in our dynamic model. It can be interpreted as the peak-to-trough ratio of a window of size  $T$  centred at the current time point  $t_k$ . In the simple case where only one sinusoid is used in (1), it can be calculated directly as  $\text{PTR} = \exp(2A_k)$ , where  $A_k = \sqrt{a_k^2 + b_k^2}$ . In the general case, where  $p$  sinusoids are used, it can be calculated at time  $t_k$  by evaluating the seasonal component  $S_k$  at a grid of time points in a window of width  $T$  centred at  $t_k$  and letting the peak-to-through ratio be the exponential function applied to the range. Prediction intervals for the peak-to-trough ratio can be approximated in either case using all available samples from the conditional distribution.

Concerning model diagnostics, a simple check of model adequacy may be based on diagnostic plots of residuals of the type  $r_k = y_k - \exp(\text{E}(\lambda_k | \text{data}))$ , where  $\text{E}(\lambda_k | \text{data})$  is available from the iterated

extended Kalman smoother<sup>26</sup>. A more detailed discussion of diagnostics for the dynamic generalized linear model is beyond the scope of the present paper.

## MODELLING DATA ON ACUTE MYOCARDIAL INFARCTION

We illustrate the application of the dynamic generalized linear model using data on hospitalizations for acute myocardial infarction in the county of North Jutland, Denmark during the 1980-1999 period. The analysis is based on essentially complete data on discharges from seven local hospitals, available from the County Hospital Discharge Registry. This Registry was established in 1977 and achieved complete coverage of hospital discharges as of 1 January 1980.

All persons with the index episode, defined as the first registered admission to hospital with a diagnosis of acute myocardial infarction (ICD-8: 410-410.99, and ICD-10: I 21-21.9), were included in our model application. During the 1980-1999 period, North Jutland County had an average population of 490,000, corresponding to roughly 10 percent of the Danish population, and 17,989 cases of incident acute myocardial infarction were registered. Observations took the form of daily counts of incident acute myocardial infarctions obtained from Hospital Discharge Registry data from all hospitals in the county. Additional details are provided by Fischer *et al.*<sup>18</sup> We analyzed the total incidence of acute myocardial infarction, without stratification or age standardization.

We applied the model described above using four sinusoids ( $p=4$ ), with  $T$  equal to one year (365.25 days). Variances for time-varying parameters were estimated using the maximum likelihood method. The R-package `sspir` was used to perform iterated extended Kalman smoothing.

Uncertainty in the estimated secular trend and seasonal variation was quantified with exact 95%

prediction bands. The peak-to-trough ratio was quantified using 95% prediction bands approximated using samples from the conditional distribution of the parameter based on all available data. We compare results from our dynamic model with those obtained using a generalized linear model, to illustrate the differences between the two approaches.

## RESULTS OF THE DYNAMIC MODELLING

Figure 1 presents a plot of the estimated secular trend component  $T_k$  with 95% prediction bands and the corresponding trend component obtained from the generalized linear model. During the course of the study period, the dynamic model shows deviations from the smooth decline of the generalized linear model. Of particular note are peaks around 1985 and 1989. During the winter season of both years, above-average influenza A activity was recorded in Denmark both by general practitioners and by the Virology Department, Statens Serum Institut (Copenhagen, Denmark). We did not formally test whether these observations were coincidental. Figure 2 shows the estimated seasonal component  $S_k$  in 1985 (solid line) and 1998 (dashed line), compared with the estimate from the generalized linear model (dotted line). The seasonal component for 1985 shows bimodal peaks around May and November and a trough in July. The seasonal component for 1998 looks similar, with peaks and a trough in the same months. However, two additional peaks (March and August) become visible in 1998.

Figure 3 presents a plot of the estimated peak-to-trough ratio (thick line) as a function of time, together with the peak-to-trough ratio obtained from the generalized linear model (thick dashed line). There is an apparent increase in the peak-to-trough ratio during the 1985-1990 period, followed by a decrease in the 1990-1995 period. Throughout the study period, the peak-to-trough

ratio estimated using the dynamic model is larger than that obtained from the generalized linear model.

To assess model fit, we constructed a box plot of the standardized residuals versus binned fitted values (Figure 4). The binning is necessary since residuals from Poisson models with low counts may exhibit a ‘banded’ structure which makes interpretation difficult. The box plot of the standardized residuals versus binned fitted values showed no obvious associations, indicating that the dynamic model adequately describes the observed data (Figure 4).

## DISCUSSION

We have presented a dynamic generalized linear model for Poisson count data (population-based hospital discharge data on acute myocardial infarction), with two evolving components: a seasonal variation and a secular trend.

The strength of our model is its flexibility. It can easily include relevant explanatory variables such as interventions, other seasonal patterns (weekday effects etc.) and covariates which may contribute to variations in seasonality over time.. Each component can be static or dynamic, depending on whether or not its relationship with the outcome varies quantitatively over time. It is possible to assess model fit by means of residual plots. Analyses can be performed with the R-package `sspir`<sup>23</sup> which is freely available from CRAN, [cran.r-project.org](http://cran.r-project.org). Alternatives to `sspir` are the `Ox` package in combination with `SsfPack`<sup>27</sup> or the R-package `StructTS`<sup>28</sup> for cases with simple time structures in normally distributed observations. In general, however, estimation of variance parameters is not

implemented in the latter package. It also can be computationally intensive, especially when the model contains many time-varying terms. A formal test to decide if a covariate enters the model as a static or a dynamic variable, for example along the lines of Nyblom<sup>29</sup>, is not implemented.

The dynamic generalized linear model yields a more detailed picture of a trend component as shown in the myocardial infarction example. The trend of incident cases of acute myocardial infarction analyzed with the static generalized linear model represents an average over the whole study period and results in a smooth line. Using the dynamic model, the trend curve becomes more detailed and shows peaks for the incidence of myocardial infarction around 1985 and 1989.

New hypotheses about associations with possible explanatory factors can be examined by including relevant variables in the model. For example, the model could be used to test the association of peaks in MI incidence with peaks in the incidence of influenza A in the same years.

The seasonal patterns of incident acute myocardial infarction delineated using the linear and the dynamic models look quite similar with peaks around May and November and a trough in July for the years 1985 and 1998. However, the generalized linear model averages data over the entire study period and therefore shows only one, fixed seasonal pattern. In contrast, the dynamic model is capable of identifying small changes in the seasonal pattern appearing between the years 1985 and 1998 (Figure 2).

The study of incident cases of acute myocardial infarction is an example of a retrospective analysis. The dynamic generalized linear model approach is equally useful in prospective analyses where the

Kalman filter technique can be used to obtain the predictive distributions necessary for monitoring. The possibility of using this methodology to generate early warnings of epidemic outbreaks of *Mycoplasma pneumoniae* is suggested in an unpublished study by Engebjerg *et al.*<sup>30</sup>

## CONCLUSION

Dynamic generalized linear modelling provides more detailed results than static modelling of secular trend and seasonality in count data. This method can be used to study relevant associations between various sets of data and potentially explanatory factors in both retrospective and prospective study designs.

## FIGURE LEGENDS

Figure 1: The secular trend of the total incidence of myocardial infarction in Northern Jutland and 95% prediction band estimated using a dynamic generalized linear model. Superimposed is the trend estimated using a static generalized model.

Figure 2: The seasonal component estimated for the years 1985 and 1998 using the dynamic generalized linear model. Superimposed is the seasonal component for the entire study period estimated based on the static generalized linear model.

Figure 3: The peak-to-trough ratio with 95% prediction band estimated by using the dynamic generalized linear model. The horizontal line at 1.17 is the peak-to-trough ratio of the seasonal component obtained from the static generalized linear model.

Figure 4: Box plot of residuals versus expected incidence levels obtained from the dynamic generalized linear model. The expected incidence levels are binned in 5 bins from 2 to 4.5.

Figure 1

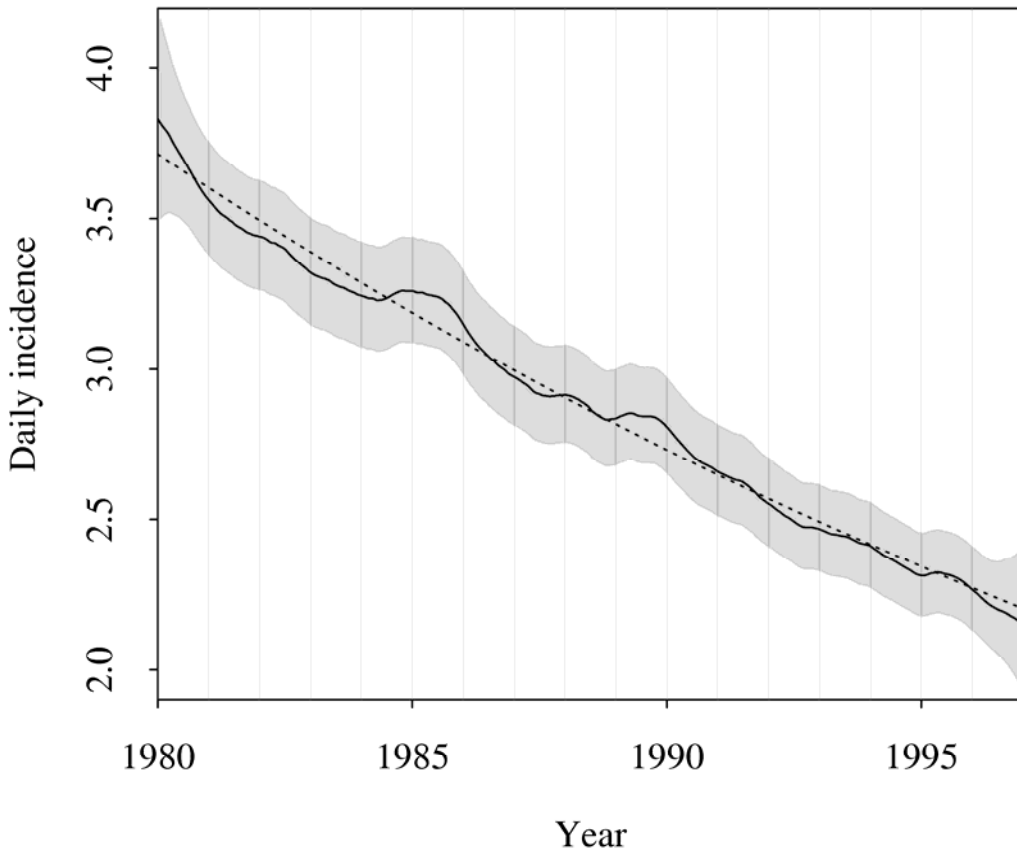


Figure 2

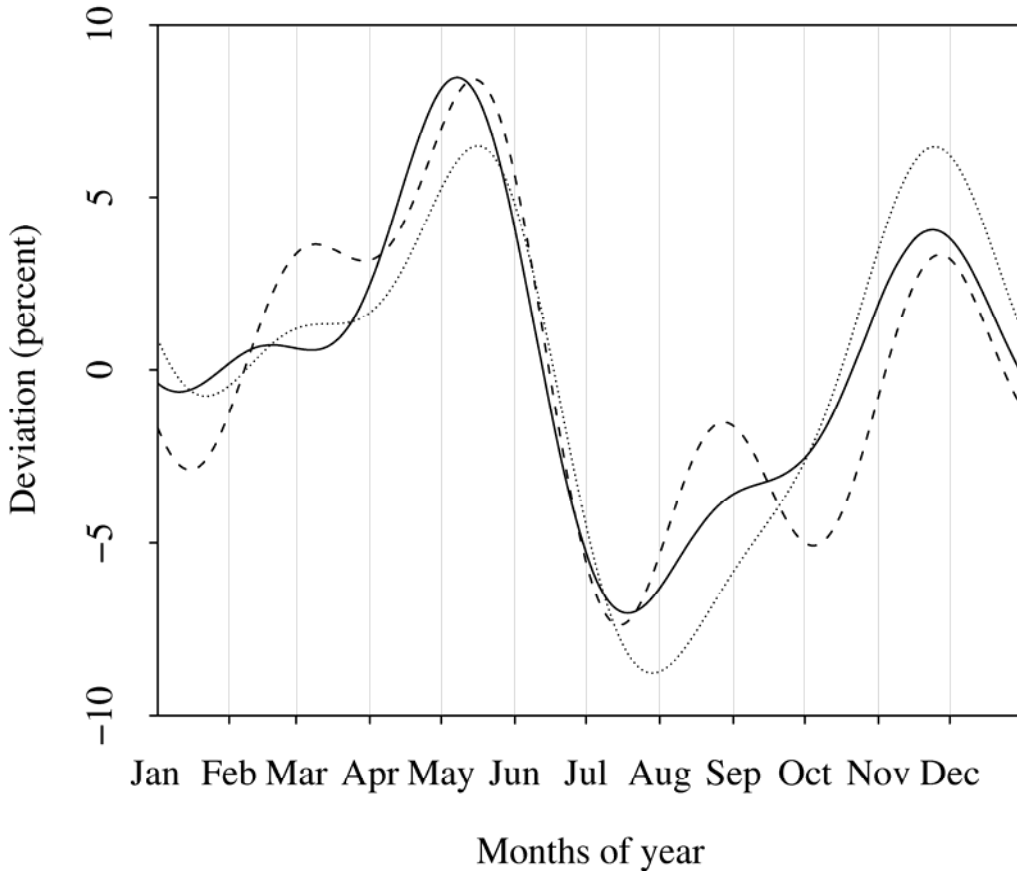


Figure 3

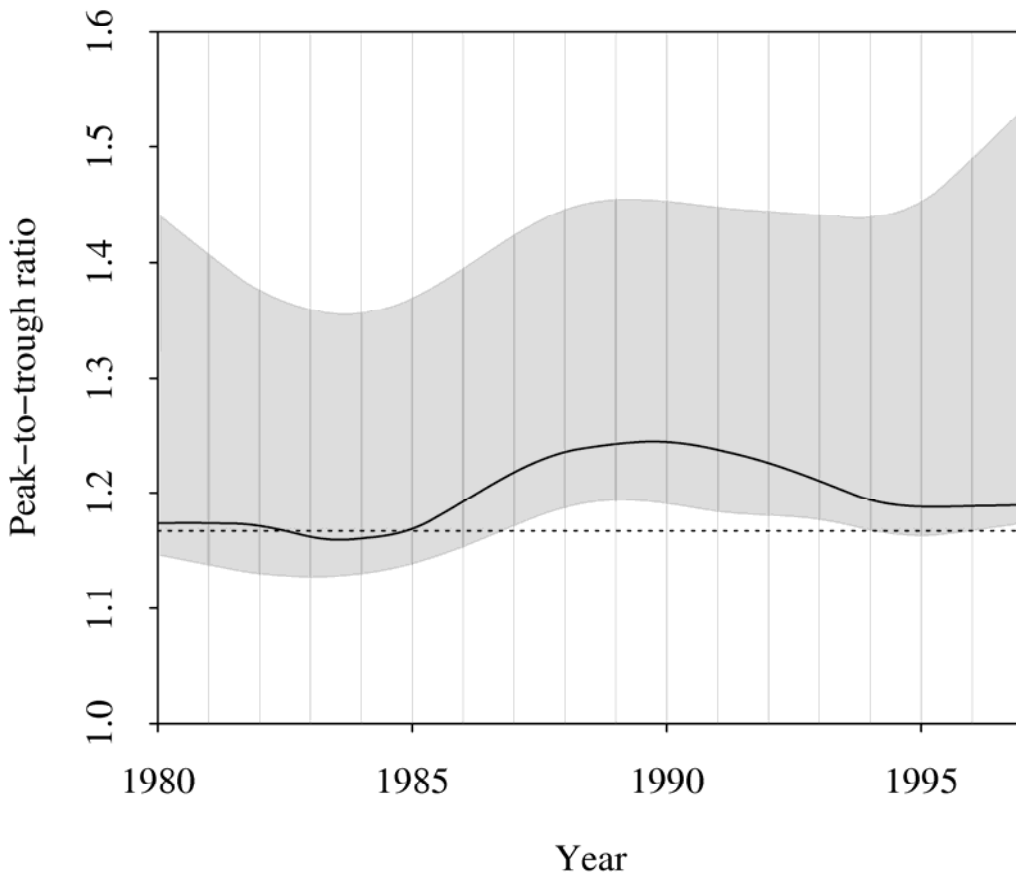
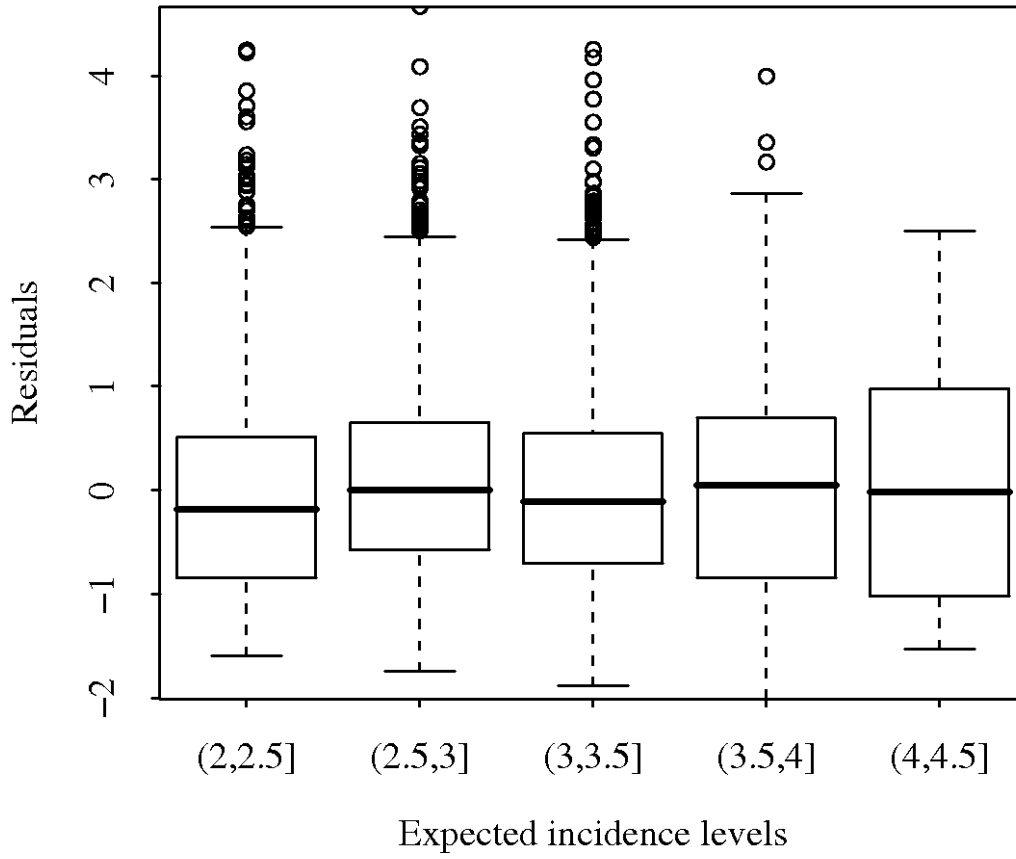


Figure 4



## Reference List

- (1) Arntz HR, Willich SN, Schreiber C, Bruggemann T, Stern R, Schultheiss HP. Diurnal, weekly and seasonal variation of sudden death. Population-based analysis of 24,061 consecutive cases. *Eur Heart J*. 2000;21(4):315-320.
- (2) Boulay F, Berthier F, Sisteron O, Gendreike Y, Gibelin P. Seasonal variation in chronic heart failure hospitalizations and mortality in France. *Circulation*. 1999;100(3):280-286.
- (3) Brennan PJ, Greenberg G, Miall WE, Thompson SG. Seasonal variation in arterial blood pressure. *Br Med J (Clin Res Ed)*. 1982;285(6346):919-923.
- (4) Douglas AS, Dunnigan MG, Allan TM, Rawles JM. Seasonal variation in coronary heart disease in Scotland. *J Epidemiol Community Health*. 1995;49(6):575-582.
- (5) Feigin VL, Anderson CS, Rodgers A, Bennett DA. Subarachnoid haemorrhage occurrence exhibits a temporal pattern - evidence from meta-analysis. *Eur J Neurol*. 2002;9(5):511-516.
- (6) Frost L, Johnsen SP, Pedersen L et al. Seasonal variation in hospital discharge diagnosis of atrial fibrillation: a population-based study. *Epidemiology*. 2002;13(2):211-215.

- (7) Spencer FA, Goldberg RJ, Becker RC, Gore JM. Seasonal distribution of acute myocardial infarction in the second National Registry of Myocardial Infarction. *J Am Coll Cardiol.* 1998;31(6):1226-1233.
- (8) Nelson W, Tong YL, Lee JK, Halberg F. Methods for cosinor-rhythmometry. *Chronobiologia.* 1979;6(4):305-323.
- (9) Edwards JH. The recognition and estimation of cyclic trends. *Ann Hum Genet.* 1961;25:83-87.
- (10) Jensen ES, Lundbye-Christensen S, Samuelsson S, Sorensen HT, Schonheyder HC. A 20-year ecological study of the temporal association between influenza and meningococcal disease. *Eur J Epidemiol.* 2004;19(2):181-187.
- (11) Hastie TJ, Tibshirani RJ. *Generalized Additive Models.* Chapman & Hall; 1990.
- (12) Katsouyanni K, Schwartz J, Spix C et al. Short term effects of air pollution on health: a European approach using epidemiologic time series data: the APHEA protocol. *J Epidemiol Community Health.* 1996;50 Suppl 1:S12-S18.
- (13) Schwartz J, Spix C, Touloumi G et al. Methodological issues in studies of air pollution and daily counts of deaths or hospital admissions. *J Epidemiol Community Health.* 1996;50 Suppl 1:S3-11.

- (14) Zeger SL, Liang K-Y. Longitudinal Data Analysis for Discrete and Continuous Outcomes. *Biometrics*. 1986;42(1):121-130.
- (15) Chan KS, Ledolter J. Monte Carlo EM Estimation for Time Series Models Involving Counts. *Journal of the American Statistical Association*. 1995;90(429):242-252.
- (16) Zeger SL, Qaqish B. Markov Regression Models for Time Series: A Quasi-Likelihood Approach. *Biometrics*. 1988;44(4):1019-1031.
- (17) Crawford VL, McCann M, Stout RW. Changes in seasonal deaths from myocardial infarction. *QJM*. 2003;96(1):45-52.
- (18) Fischer T, Lundbye-Christensen S, Johnsen SP, Schonheyder HC, Sorensen HT. Secular trends and seasonality in first-time hospitalization for acute myocardial infarction--a Danish population-based study. *Int J Cardiol*. 2004;97(3):425-431.
- (19) Seretakis D, Lagiou P, Lipworth L, Signorello LB, Rothman KJ, Trichopoulos D. Changing seasonality of mortality from coronary heart disease. *JAMA*. 1997;278(12):1012-1014.
- (20) Durbin J, Koopman SJ. *Time Series Analysis by State Space Methods*. Oxford University Press; 2001.
- (21) McCullagh P, Nelder JA. *Generalized Linear Models*. Chapman and Hall; 1989.

- (22) West M, Harrison PJ, Migon HS. Dynamic Generalized Linear Models and Bayesian Forecasting. *Journal of the American Statistical Association*. 1985;80(389):73-83.
- (23) Dethlefsen C, Lundbye-Christensen S. Formulating state space models in R with focus on longitudinal regression models. *Journal of Statistical Software*. 2006;16(1):1-15.
- (24) R: A language and environment for statistical computing. 2006.
- (25) Dethlefsen C. *Space time problems and applications*. Ph.D. Thesis: Aalborg University; 2002.
- (26) Jorgensen B, Lundbye-Christensen S, Song XK, Sun L. A longitudinal study of emergency room visits and air pollution for Prince George, British Columbia. *Stat Med*. 1996;15(7-9):823-836.
- (27) Koopman SJ, Shephard N, Doornik JA. Statistical Algorithms for Models in State Space using SsfPack 2.2. *Econometrics*. 1999;2:113-166.
- (28) Ripley BD. Time Series in R 1.5.0. *R News*. 2002;2(2):2-7.
- (29) Nyblom J. Testing for the Constancy of Parameters Over Time. *Journal of the American Statistical Association*. 1989;84(405):223-230.

- (30) Engebjerg, Malene Dahl Skov, Lundbye-Christensen, Søren, Kjær, Birgitte B., and Schönheider, Henrik C. Monitoring Poisson time series using multi-process models. 2006.