Pollution Effects & Control

Szyszkowicz, et al., J Pollut Eff Cont 2014, 2:1
http://dx.doi.org/10.4172/jpe.1000109

Research Article Open Access

Distributed Lag Models: An Analysis of Milan Mortality Data

Mieczyslaw Szyszkowicz* **and Wesley S Burr**

Population Studies Division, Health Canada, Ottawa, ON, Canada

***Corresponding author:** Mieczysław Szyszkowicz, Population Studies Division, Health Canada, 200 Eglantine Driveway, Ottawa, ON, K1A 0K9, Canada, Tel: (613) 948-4629; Fax: (613) 954-3768; E-mail: mietek.szyszkowicz@hc-sc.gc.ca

Rec date: Feb 07, 2014; **Acc date:** Jun 12, 2014; **Pub date:** Jun 23, 2014

Abstract

The objective of this study is to present a new variation on Distributed Lag Non-Linear Models (DLNMs) for assessing associations between counts of health events and exposure to ambient air pollution. For illustrative purposes, a well-known data set for Milan, Italy was considered. Total Suspended Particulate (TSP) concentrations were used as the air pollution measure, and meteorological data were represented by daily mean temperature and relative humidity. Relative risks (RR) were estimated using Poisson Generalized Linear Models. Two controls for long time-scale variation were considered: a more traditional cubic regression spline smoother, and the more recent case-crossover (CC) control approach. The mortality displacement effect was estimated using DLNMs for a relatively high number of constructed lags. For the considered lags (0–45 days) and the CC approach, three regions were identified: region A (lags 0–7) with RR=1.021 (95 % confidence interval: 1.009, 1.043; region B (lags 8–27) with RR=0.981 (0.965, 0.997); and region C (lags 28–45) with RR=1.018 (1.003, 1.032). The total cumulative risk (regions A + B + C, lags 0–45) gave RR=1.019 (1.001, 1.037). The results were reported for an interquartile range (IQR=86.5) increase in TSP air pollution and are similar in structure to those previously reported, albeit at a significantly reduced level. We attribute the change to the considerable change in long timescale variation left in the residuals, as the clustering effect controls seasonal effects at a much stronger level.

Keywords: Air pollution; Distributed lags; Mortality; Humidity; Temperature; Case-crossover; SAS

Introduction

The aim of this paper is to demonstrate a methodological development for distributed lag models using a case-crossover (clustering) approach to control for long timescale variation. As a demonstrative example, we perform an analysis of the Milan, Italy mortality data set [1] using our new statistical approach. As scientific progress on the methodology and techniques related to air pollution exposure and associated health conditions is continuing, it is interesting to consider new techniques for estimation of associations that are widely known and published in the scientific literature. This study uses distributed lag non-linear models (DLNMs) to effectively represent and quantify associations showing non-linear and delayed effects in time-series data. In particular, we apply DLNMs to quantify mortality displacement as in [1].

This is a methodological paper using real data as a demonstrative example. As health data we are using mortality counts, and relating them to exposure to ambient particulate matter (PM) in the form of Total Suspended Particulate (TSP). Extensive research has demonstrated the associations between exposure to PM and different health conditions related to respiratory and cardiovascular diseases and mortality [2-7]. Thus studies in this domain address an important aspect of environmental epidemiology.

The main purpose of this paper was to compare and contrast two different approaches to the estimation of mortality displacement, both using distributed lag non-linear models. In the first approach, we use case-crossover clustering to account for long timescale variation, and in the second we use a more traditional natural cubic regression spline smoother. These two models are compared using the Milan data set [1], as a well-known previously studied example.

Distributed lag models are a modeling schema for presenting simultaneously both non-linear and delayed dependencies in time-series data [8]. The DLNM methodology in particular is an extension of a statistical regression model for defining the relationship between a set of predictors (such as air pollution) and an outcome (such as health conditions). In this methodology the estimated relationships allow for a temporal structure of dependency. This is useful because in environmental epidemiology a specific occurrence of an exposure event often affects the health outcomes for a lapse of time beyond the event moment.

Materials and Methods

To demonstrate our developments, we used the well-known 1980-1989 Milan, Italy mortality database [1]. The considered data were retrieved from the package SemiPar [9], where they are included for illustrative purposes. In our analysis we considered the time-series data of the form:

(x_t,y_t)=(mortality, airpollution)$_t$, t=1,2,…,3652

i.e., daily mortality and daily mean air pollution for 3,652 consecutive days. We assumed that the data used are close to their original representation, i.e., to the data described in the published work on mortality displacement, an assumption backed by our Table 1 and the similar Table 1 in [1] . The mortality data are daily counts derived from death certificates and are restricted to residents of Milan who died from natural causes (International Classification of Disease revision 9, cases 1-799). As was the case for the authors of the publication containing the original data, we also used Total Suspended Particulates (TSP) as our ambient air pollution measure. The imputed

(19% originally missing) version of the TSP data was provided by the SemiPar package, and thus used. Meteorological data were expressed as daily mean temperature and relative humidity. In the considered data set from SemiPar, there were no missing values for any of the considered data variables.

Variable	Average	Range	SD	Median	IQR
Mean temperature	14	-6.1 – 31.5	8.1	13.7	14
Mean relative humidity	62	0 – 99.7	17	62.3	24
TSP	136.5	3.5 – 529.5	76	117.1	87
Total mortality	32	Oct-66	7.7	31	11

Table 1: Summary statistics of Milan mortality data (SD –standard deviation, IQR–interquartile range (75th percentile–25th percentile)).

We analyzed the Milan mortality data using a Poisson Generalized Linear Model (GLM) [10] framework, with computation and estimation done in the R programming language [11] using the DLNM package [8,12]. The script used in R is presented in the Appendix 1. As our data are time-series, we applied two approaches to adjust for time-based variation (across 3652 days). In both approaches the model used was matched to that of [1] as closely as possible. Thus, each model contains a DLNM function of TSP, as well as spline functions of temperature and relative humidity (with 4 *df* each) and indicator variables for high temperatures, as in [1]. The remaining terms deal with the day-of-week effect and long time-scale variation, and it is here that the approaches diverge.

	Cluster: 2 weeks		Cubic Spline, 33 df	
Region/Lag	RR	95% CI	RR	95% CI
A: 0 - 15	1.012	0.998, 1.026	1.04	1.021, 1.056
B: 16 - 20	0.993	0.988, 0.999	1.003	0.997, 1.009
C: 21 - 45	1.014	0.998, 1.030	1.029	1.006, 1.054
A: 0-7	1.021	1.009, 1.033	1.035	1.021, 1.050
B: 8-27	0.981	0.965, 0.997	1.016	0.996, 1.036
C: 28-45	1.018	1.003, 1.032	1.021	1.002, 1.041
A+B+C: 0-45	1.019	1.001, 1.037	1.074	1.035, 1.113

Table 2: Results: decomposition of the TSP effect.

For our first approach, we applied a case-crossover clustering approach, and grouped the data by 14-day periods, with each group considered as an individual cluster with two measurements for the same day of the week ([13,14] for details on why this is a sensible choice). For verification purposes, we repeated our analysis and defined clusters by 21-day periods instead. In the latter situation we had 3 data points for each day of week. We did this twice by grouping days into clusters in descending (from 1 to 3,652) and ascending (from 3,652 to 1) temporal order. In our second approach, we used a natural cubic regression spline smoother to control for time with 33 degrees of freedom (*df*), i.e., the effective *df* used in [1]. This approach is widely used in environmental epidemiology, although traditionally more *df* are used. This lower *df* choice was made in an attempt to match the model of [1] as closely as possible, although we used a fixed-*df* spline smoother rather than the penalized smoother used there. In the second approach we also included a day-of-week term, which was not necessary in the clustering approach.

We considered a high number of possible lags (0-45 days) for TSP, the same number chosen by Zanobetti et al. in their initial analysis [1]. The full specifications of the DLNM + GLM model used are presented in the R script in the Appendix 1.

Results

The results are organized and presented in three figures and two tables. Table 1 presents summary statistics on mortality and environmental characteristics (TSP, temperature, and relative humidity). The table appears to be nearly identical to a similar summary provided by Zanobetti et al. [1], (Table 1). The results for an increase in one Inter-Quartile Range (IQR) of TSP (86.9 for this dataset) are presented in Table 2 as relative risks (RRs) with their corresponding 95% Confidence Intervals (95% CI). The results are shown for three regions (lag ranges) labeled as A, B, and C. The regions were determined on the basis of the values of individual relative risks as in [1]. The table shows both the results using the regions A, B and C as determined in [1], i.e., A being lags 0-15, B being lags 16-20, and C being lags 21-45, and also for regions as determined by our first, new, approach, using the case-crossover clustering control for time, with the determination made in the same way: positive RRs (A, lags 0-7), negative RRs (B, lags 8-27), and positive RRs (C, lags 28-45).

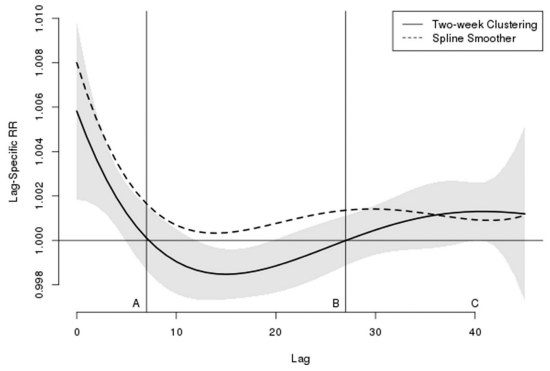

Figure 1: Effects of exposure to TSP on relative risk (RR) of mortality along the considered lags, for regions A, B, and C with borders where RR=1.0. The solid line is our "Approach 1", accounting for long timescale variation using the case-crossover approach, while the dashed line is our "Approach 2", using natural cubic regression splines for time-based smoother.

The results shown in Table 2 were obtained using the code from Appendix A, estimating a GLM for each model described above, with our two approaches for controlling long timescale variation The results for the smooth function of time model are not identical to those previously estimated in [1], but they do share common structure: positive and significant for regions A, C and the total cumulative effect, and not significant for region B. This holds for both the region classification scheme for [1] and the classification obtained from our clustering time-control approach. The total cumulative RR=1.074 (95% CI: 1.035, 1.113) is higher than that obtained in [1], but we have replaced the penalized spline smoothers with fixed-df smoothers, so

some variation between the two is expected. For our new approach, GLMs are estimated with two-week clusters acting in place of the smooth function of time. Each cluster contains two of each of the days of week (see R code in Appendix 1, and [13,14] for details on choosing two-weeks versus three or more). Under regions defined as in [1], regions A, RR=1.012 (95% CI: 0.998, 1.026), and C, RR=1.014 (95% CI: 0.998, 1.030) cease to be positively significant, while region B becomes negative and significant, RR 0.993 (95% CI: 0.988, 0.999). When restricted by its own positive/negative relative risk status, chosen as in [1], the results return closer to the state previously observed, with regions A, RR=1.021 (95% CI: 1.009, 1.033), and C, RR=1.018 (95% CI: 1.003, 1.032) being positively significant. However, region B remains negatively significant, RR=0.981 (95% CI: 0.965, 0.998). The total cumulative risk is positively significant, albeit much smaller than observed using a smooth function of time, with RR=1.019 (95% CI: 1.001, 1.037).

Figure 1 shows the estimated RRs for mortality (with 95% CIs along the 45 lags) for an increase in concentration of TSP of one IQR, showing both the two-week clustering approach (solid line) and the spline function of time approach (dashed line). The behavior follows that of [1], (Figure 2), with the modifications discussed above. The differences between the two are discussed further below. Figure 2 shows the cumulative effects of the same unit exposure. In Figure 1 we marked the regions A, B, and C, where the RRs have the same sign (positive or negative). In addition, we used vertical lines to emphasize their boundaries.

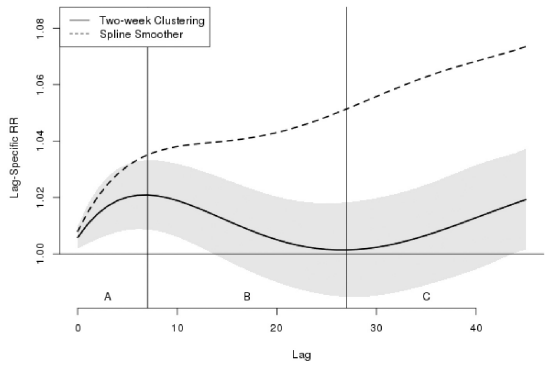

Figure 2: Total cumulative relative risk for mortality of exposure to TSP, along the considered lags for regions A, B and C as in Figure 1. Again, the solid line is the case-crossover result, and the dashed line is the spline smoother result.

Discussion and Conclusions

In this work we analyzed mortality displacement with the goal of estimating the effect of air pollution, adjusting for this mechanism using the DLNM technique. We used data from the public domain (provided by the second author of [1]) and although the set used appears to be not quite identical to that used by Zanobetti et al. [1], it is close enough for the results to be comparable. Our results are quite different for the case where we used case-crossover (clustering) models with clusters composed of 14 and 21 days. In these models "time" is managed by cluster structures with two and three points for each day of the week, respectively. The overall effect (A+B+C; RR=1.019, 95% CI: 01.001, 1.037) is positive and statistically significant, but significantly lower than [1], which gave RR=1.067 (1.038, 1.096). We obtained negative and statistically significant results for region B under both divisions; that of [1] and the same algorithm applied to our approach. The regions are quite different between the two, with our approach having a significantly larger number of lags with negative risk. This can partially be attributed to the overall reduction in estimated risk (shifting the entire curve downward) and partially to a shape change which we suspect to be data-driven.

In our second (baseline) approach, we used a more traditional GLM with a cubic regression spline controlling for long timescale variation (using 33 df to match [1]). We obtained somewhat similar results to those published previously (Table 2) [1], albeit with increased levels. As the models used were not identical (we used fixed-df splines versus the penalized option), and as the authors of [1] did not use the slightly more flexible DLNM framework (as it had not been developed), the variation between the results is understandable. The relative risks for the second approach remain comparable to that of [1] despite the difference in implementation, allowing us to compare our two implemented models more directly (and thus, by proxy, to the previously published results). As our second approach is nearly that of [1], by dropping the smooth function of time and replacing the mortality counts by their clusters instead (our first approach), we obtain different results which can be attributed to the change in time control.

In this study, positive and statistically significant short-term effects on daily mortality in Milan, Italy were observed in relation to exposures to ambient air pollution. The DLNM technique allows us to effectively quantify the effects of delayed exposure over time. The goal of this work was to highlight a new technique (case-crossover time control within a DLNM framework) and to use it in the domain of air pollution and related health effects. In addition, we provided a SAS procedure (PHREG) which implements the case-crossover technique (Appendix 2). As was recently observed, some realizations of this methodology may generate bias for different options in the procedure for ties [15]. In the presented code each case and its controls are considered as separate strata. As was investigated in the work [15] such approach generates proper results for all options.

As the new case-crossover time control approach provides reduced results to those using a more traditional time smoother, an open question for future research is what effect causes the reduction. As the total cumulative effect results in [1] (and those provided by our second approach with a 33 df time smoother) are more positive than those of our first (CC) approach, our current hypothesis is that whatever portion of the time variation is controlled for by the CC approach (essentially, removed from consideration) for the Milan data is more strongly related than the residuals. The approximate 33 df used for the time smoother in [1] fails to control for seasons in any significant way [16], so we suspect that a seasonal association between TSP and mortality may be responsible for the difference. Note that this decrease in risk is not generalizable, as it will depend on the particular city of interest and its geography and climate–some cities may well have an increase in total cumulative effect when seasonality is controlled more completely.

References

1. Zanobetti A, Wand MP, Schwartz J, Ryan LM (2000) Generalized additive distributed lag models: quantifying mortality displacement. Biostatistics 1: 279-292.

2. Pope CA , Thun MJ, Namboodiri MM, Dockery DW, Evans JS, et al. (1995) Particulate air pollution as a predictor of mortality in a prospective study of U.S. adults. Am J Respir Crit Care Med 151: 669-674.
3. Katsouyanni K, Touloumi G, Spix C, Schwartz J, Balducci F, et al. (1997) Short-term effects of ambient sulphur dioxide and particulate matter on mortality in 12 European cities: results from time series data from the APHEA project. Air Pollution and Health: a European Approach. BMJ 4314: 1658-1663.
4. Katsouyanni K, Touloumi G, Samoli E, Gryparis A, Le Tertre A, et al. (2001) Confounding and Effect Modification in the Short-Term Effects of Ambient Particles on Total Mortality: Results from 29 European Cities within the APHEA2 Project. Epidemiology 12: 521-531.
5. Dominici F, Peng RD, Bell ML, Pham L, McDermott A, Zeger S, et al. (2006) Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases. JAMA 295: 1127–1134.
6. Sicard P, Lesne O, Alexandre N, Mangin A, Collomp R (2011) Air Quality Trends and Potential Health Effects - Development of an Aggregate Risk Index. Atmospheric Environment 45: 1145-1153.
7. WHO (2004) Meta-analysis of time-series studies and panel studies of Particulate Matter (PM) and Ozone (O3). WHO task group. WHO/EURO 04/5042688.
8. Gasparrini A (2011) Distributed Lag Linear and Non-Linear Models in R: The Package dlnm. J Stat Soft 43: 1-20.
9. Wand M (2012) SemiPar: Semiparametric regression. R package version 1: 0-4.
10. Nelder J, Wedderburn R (1972) Generalized linear models. Journal of the Royal Statistical Society Series A 135: 370-384.
11. Refman R (1999) A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
12. Gasparrini A, Armstrong B (2012) Distributed lag non-linear models in R: the package dlnm. London School of Hygiene and Tropical Medicine, UK.
13. Szyszkowicz M (2006) Use of generalized linear mixed models to examine the association between air pollution and health outcomes. Int J Occup Med Environ Health 19: 224-227.
14. Szyszkowicz M, Tremblay N (2011) Case-crossover design: air pollution and health outcomes. Int J Occup Med Environ Health; 24: 249-255.
15. Wang SV, Coull BA, Schwartz J, Mittleman MA, Wellenius GA (2011) Potential for bias in case-crossover studies with shared exposures analyzed using SAS. Am J Epidemiol 174: 118-124.
16. Burr, Samuel W (2012) Air Pollution and Health: Time Series Tools and Analysis.Queen's University, Kingston, Ontario.