Supplementary material for: Generalized additive models for gigadata

Simon N. Wood[0], Zheyuan Li[1], Gavin Shaddick[1] and Nicole H. Augustin[1]
[0] School of Mathematics, University of Bristol, Bristol, U.K.
[1]Mathematical Sciences, University of Bath, Bath, U.K.
simon.wood@bristol.ac.uk

May 24, 2016

1 Animations

The animations `duchon.mp4` and `tps.mp4`, accompanying this document, show model predicted daily log black smoke levels every 20 days, for the duration of the data. `duchon.mp4` is for the model recommended in the paper, while `tps.mp4` is the variant using thin plate splines for spatial effects. The latter shows serious artefacts in the last decade. The grey dots shown in the animations are locations of stations active in the year being plotted, while the black dots show inactive station locations. Predictions are not made for most of Scotland, because of the limited coverage over the full timespan of the data, for an area with quite distinct physical geography relative to the rest of the UK.

2 Further dataset details

The UK has a long history of monitoring air quality. The first nationwide air quality monitoring network was the 'National Survey' which started operation after the 1956 Clean Air Act and measured black smoke (BS) and sulphur dioxide Loader (2002). Together with its successors, it measured BS from 1961 to 2006. The network eventually ceased operation as the UK introduced new measuring methods in compliance with EU Directives 80/8/779/EEC Fowler et al. (2006), but a new black carbon network Fuller and Connolly (2012) continued monitoring black smoke after this point. Black smoke remains of interest as it is shown to be association with adverse health effects (for example, Hoek et al. (2002); Elliott et al. (2007)) and data arising from this network has been the subject of a a number of studies including Elliott et al. (2007); Fanshawe et al. (2008); Dadvand et al. (2011); Gulliver et al. (2011); Shaddick and Zidek (2014). However, none of these consider the modelling of daily measurements, the importance of which is perhaps best summarised by Perry et al. (2009): "Daily data sets are of particular interest for analyse of extremes, as well as the investigation of daily relationships with other factors." In addition to driving the acute effects of air pollution on health,

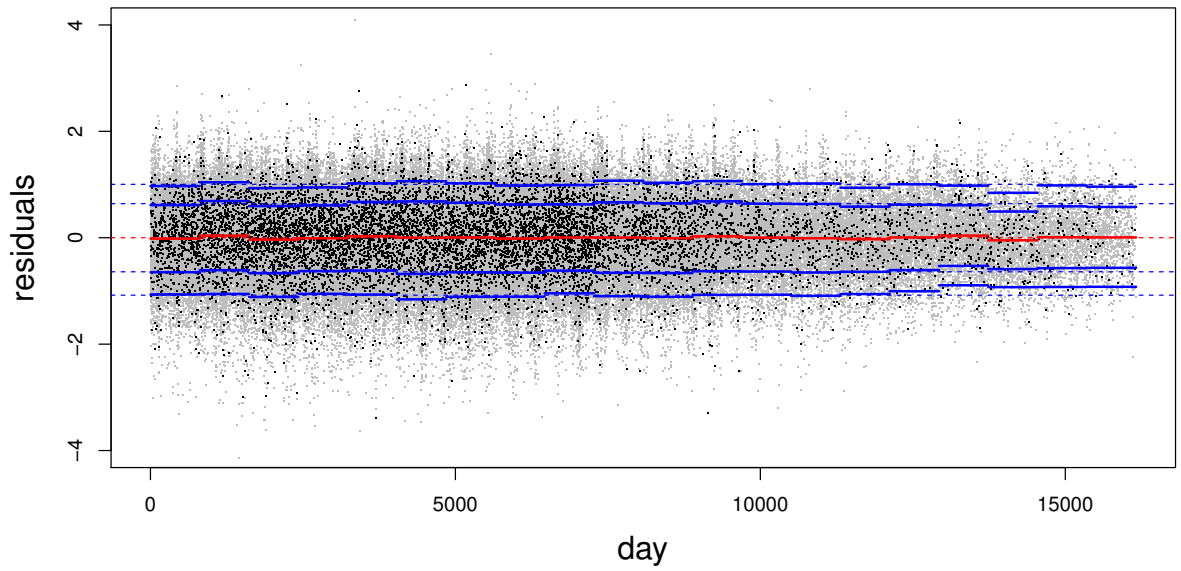

Figure 1: Residuals against time. The grey dots are a random sample of 200000 residuals and the black dots a random sample of 20000 to give a clearer picture of distribution. The red dashed line is the zero line, while the piecewise constant red lines show means of residuals within temporal blocks. The inner blue dashed lines are at $\pm$ one standard deviation, computed for the whole dataset, while the piecewise constant blue lines show $\pm$ one standard deviation computed within 20 time blocks. The outer blue lines are the equivalent for 5%and 95% quantiles.

short-term high exposures may aggravate chronic disease, an effect that will be masked when using aggregate measures of exposures, e.g. annual averages.

Historical daily measurements of BS data can be downloaded from the DEFRA data archive (`http://uk-air.defra.gov.uk/data`). This contains ca. 9.5 million daily measurements from a total of 2874 sampling locations within Great Britain, measured between the 3^{rd} October 1961 and the 31^{st} December 2005. Each site is classified as either industrial, residential, commercial, rural, open field or mixed areas, with possible subclasses according to local land use(Loader (2002), Section 4.4). Site locations are recorded with a precision of $100\mathrm{m}$. Elevation for all sites was obtained from the Ordnance Survey's Terrain 50 modelof Terrain-50 (2015), which provides altitude on a $50\mathrm{m} \times 50\mathrm{m}$ grid over Great Britain. Daily minimum, mean and maximum temperatures together with monthly total rainfall werer obtained from the Met Office's UKCP09 gridded dataset (Perry and Hollis, 2006; Perry et al., 2009).

2.1 Missing values and data transformation

Not all data will be used for modelling. UKCP09 data are not available for 12 sites scattered along the coast near Newcastle-upon-Tyne, probably because they are located too close to the sea. Due to their special locations, imputation of missing UKCP09 data is not reliable, hence these sites are excluded from analysis. Among the 2862 sites to be examined, 482 have missing site types. Excluding those sites might lead to substantial information loss, hence it is decided to label them as a "missing" type to proceed.

Raw black smoke observations bs are non-negative integers, highly skewed. The mean of black smoke is 47.2, while the median is 21. We worked with $\texttt{logbs} = \log(\texttt{bs} + \epsilon)$ to stabilise model residual variability. The choice of ϵ is essentially arbitrary, as long as it is not too big compared with bs. It was chosen as 1 for our analysis, so that logbs has a minimum of 0. After such transform, the mean and median is 3.13 and 3.09, respectively.

3 Residuals and effects

Figure 1 shows residuals against time, illustrating that there seems to be little systematic trend in the residual pattern over time, although there is some annual variability visible. Figure 2 shows the weather effects, random effects and main effects of time variables, for model (1). The overall long term decline is clearly visible, along with the annual and weekly cycles. Both (cube root transformed) rainfall and elevation (height) have strong negative effects on black smoke levels. Low min and max temperatures have a strong positive effect on pollution levels, with the maximum apparently more important than the minimum. The lagged temperature effects are less clearly interpretable.

4 Preferential sampling

As described in the main paper, we investigated the potential for preferential sampling to bias our results (noting that our set up, of a slowly evolving fixed network where stations may be dropped on the basis of long term average response levels at the station location, does not fall within the framework of Diggle et al., 2010, and the mechanism for potentially introducing bias is therefore different and more limited). As the basis for simulating data susceptible to preferential sampling related problems, the model

$$\begin{aligned}\log(\texttt{bs}_i) = f_2(\texttt{doy}_i) + f_3(\texttt{dow}_i) + f_4(\texttt{y}_i, \texttt{doy}_i) + f_6(\texttt{doy}_i, \texttt{dow}_i) \\ + f_7(\texttt{n}_i, \texttt{e}_i) + f_9(\texttt{n}_i, \texttt{e}_i, \texttt{doy}_i) + f_{10}(\texttt{n}_i, \texttt{e}_i, \texttt{dow}_i) \\ + f_{11}(\texttt{h}_i) + \alpha_{k(i)} + b_{\text{id}(i)} + e_i\end{aligned}$$

was fit to the data from 1967, when the network's spatial coverage was at its most complete, while levels were high and the spatial pattern complex. A long term linear temporal trend was then added to this model matching the trend in the real data. From this fitted model we could simulate datasets broadly resembling the real dataset, but with a strong spatial pattern persisting over years. Over time the preferential sampling evident in the real network evolution preferentially removes stations from the regions that have low pollution in the simulation. Hence our model fit to simulation data following the same evolution as the real network, will be susceptible to the same adverse effects of preferential sampling that might effect its fit to the real data. By maintaining the strength of the spatial pattern over time, we heighten the potential for bias to be introduced in later years, when the reduced network will force the modelling results to be somewhat reliant on the model spatial smoothness assumptions (by contrast if the true fields attenuated and became constant over space, later in the data, then the smoothness assumptions would be too compatible with the 'true model' to cause bias).

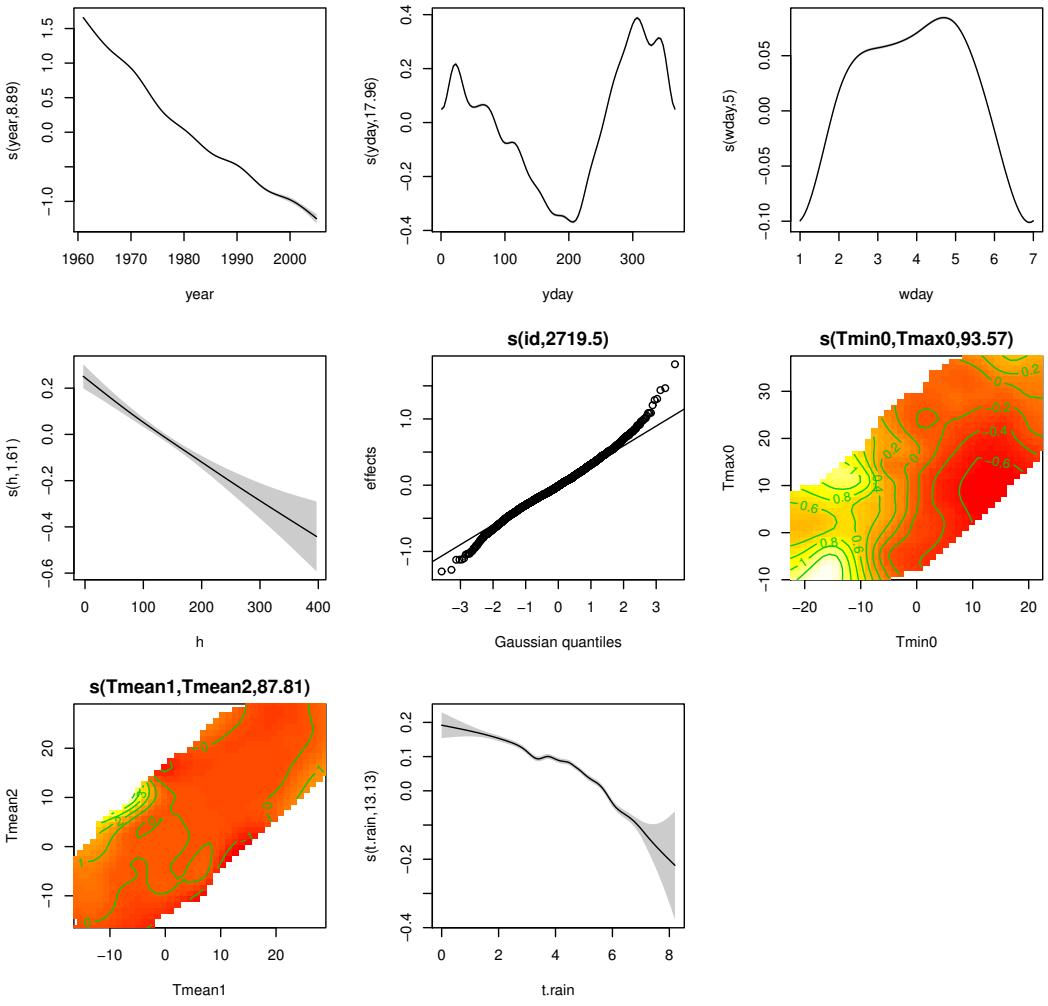

Figure 2: Smooth main effects of time, height, weather variables, and a QQ-plot of station random effects.

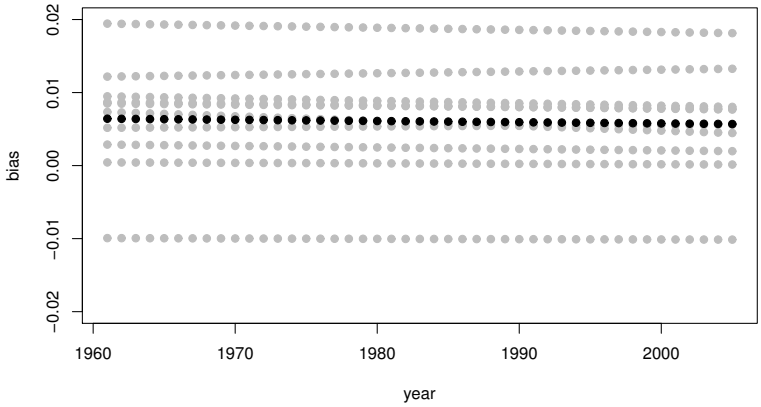

Figure 3: Average bias by year for each of 10 replicates of the simulation described in the text (grey), indicating the potential for preferential sampling bias in our results. The black dots show the mean bias over all 10 replicated, by year. To help interpret the size of the bias, recall that the average log black smoke started at around 5 in 1961 (peak around 7), and that a log error of 0.006 corresponds to a percentage error of 0.6% on the original scale.

Notice that we can not assess the potential for preferential sampling to bias our results by simulating directly from our full model fit to the whole real dataset: if we did that then the prior smoothness assumptions would be exactly compatible with the 'true' model from which we simulated the data, and there would be no mechanism for preferential sampling to introduce bias. Notice also that, in principle, preferential sampling could affect the network design's coverage of covariates other than space and time, but this does not appear to happen to any substantial extent for the black smoke data - coverage of other covariates remains fairly complete even late in the data set.

Data were simulated from the model using the real network design, and the full model from the main paper was then fitted to the simulated data. From this full fitted model we then predicted at each of the locations at which monitoring stations were present in 1967, for every day of every year (from 1961 to 2005). i.e. the observed evolution of the network was used in the fit data, but the predict data was a static network with good spatial coverage and no drop out. We then compared the predictions to the simulation truth. Figure 3 illustrates the average bias (predicted minus truth) observed each year in 10 replicates of this simulation. There is no evidence for a systematic trend in bias, and the bias itself appears to be very small. It appears that despite the clear evidence for some preferential sampling in the data, the potential for this to introduce substantial bias into the model results is small.

References

Dadvand, P., S. Rushton, P. J. Diggle, L. Goffe, J. Rankin, and T. Pless-Mulloli (2011). Using spatio-temporal modeling to predict long-term exposure to black smoke at fine spatial and temporal scale. *Atmospheric environment 45*(3), 659–664.

Diggle, P. J., R. Menezes, and T.-l. Su (2010). Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 59*(2), 191–232.

Elliott, P., G. Shaddick, J. C. Wakefield, C. de Hoogh, and D. J. Briggs (2007). Long-term associations of outdoor air pollution with mortality in great britain. *Thorax 62*(12), 1088–1094.

Fanshawe, T., P. Diggle, S. Rushton, R. Sanderson, P. Lurz, S. Glinianaia, M. Pearce, L. Parker, M. Charlton, and T. Pless-Mulloli (2008). Modelling spatio-temporal variation in exposure to particulate matter: a two-stage approach. *Environmetrics 19*(6), 549–566.

Fowler, D., N. Cape, et al. (2006). A review of the UK urban network for measurement of Black Smoke, SO_2 and NO_2: summary report. Technical report, Department of Environment, Food and Rural Affairs, Nobel House, 17 Smith Square, London, SW1P 3JR.

Fuller, G. and E. Connolly (2012). *Reorganisation of the UK Black Carbon network*. Nobel House, 17 Smith Square, London, SW1P 3JR: Kings College London and Department of Environment, Food and Rural Affairs.

Gulliver, J., C. Morris, K. Lee, D. Vienneau, D. Briggs, and A. Hansell (2011). Land use regression modeling to estimate historic (1962- 1991) concentrations of black smoke and sulfur dioxide for great britain. *Environmental science & technology 45*(8), 3526–3532.

Hoek, G., B. Brunekreef, S. Goldbohm, P. Fischer, and P. A. van den Brandt (2002). Association between mortality and indicators of traffic-related air pollution in the netherlands: a cohort study. *The lancet 360*(9341), 1203–1209.

Loader, A. (2002). *Instruction manual: UK Smoke and Sulphur Dioxide Network*. Culham Science Centre: Netcen, AEA Technology.

of Terrain-50, C. (2015). *OS Terrain 50: user guide and technical specification*. Adanac Drive, Southampton, SO16 0AS: Ordnance Survey.

Perry, M. and D. Hollis (2006, 6). The generation of monthly gridded datasets for a range of climatic variables over the United Kingdom.

Perry, M., D. Hollis, and M. Elms (2009, 6). The generation of daily gridded datasets of temperature and rainfall for the uk.

Shaddick, G. and J. V. Zidek (2014). A case study in preferential sampling: Long term monitoring of air pollution in the uk. *Spatial Statistics 9*, 51–65.