Downloadable estimates of air pollution for England and Wales and estimation of their health effects

Sujit Sahu

http://www.soton.ac.uk/∼sks/

UNIVERSITY OF Southampton

Collaborators: Sabyasachi Mukhopadhyay, Duncan Lee & Alastair Rushworth

RSS Webinar, February 2018

EPSRC Engineering and Physical Sciences Research Council

Met Office

University of Glasgow

Pollution is still a problem today!


BBC Sign in News Sport Weather iPlayer TV Ra

NEWS UK

Home World UK England N. Ireland Scotland Wales Business Politics Health Education Sci/En

3 April 2014 Last updated at 22:15

Share

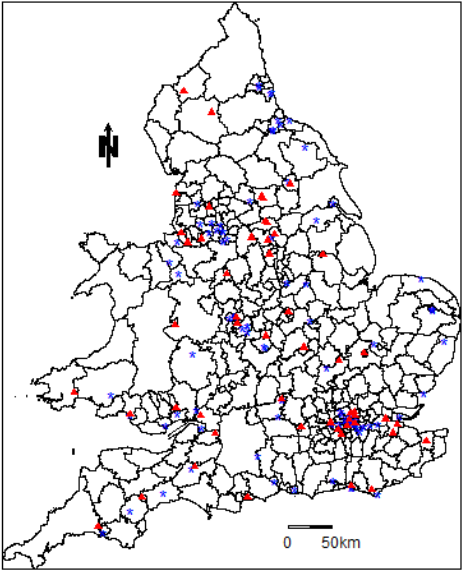
Air pollution: Forecasters hope for cleaner air on Friday

AP

People with lung and heart problems have been advised to avoid strenuous outdoor activity

- London kids on high air pollution: 'Our eyes start stinging' BBC News, 29 January 2017.
- Traffic pollution kills 5,000 a year in UK, says study. BBC News, 17 April 2012.

Automatic Urban and Rural Network (AURN)



- Map of 323 local and unitary authorities in Eng & Wales.
- 144 AURN monitoring locations are blue * and red △.
- Statistical modelling challenge: how do we estimate air pollution at *any* new location?

- So that we may relate health outcome data and pollution.

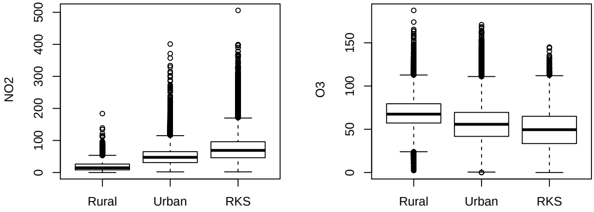
Very sparse air pollution data in the UK

- Monitoring data is very sparse with a lot of missing data.
- Website hosted by DEFRA (Department for Environment Food and Rural Affairs) provides downloadable data.
- Estimates from computer simulation model are biased and not available publicly.

Pollutant	2007	2008	2009	2010	2011	Overall
NO_2	31311	31356	31815	31828	33224	159,534
O_3	22528	19015	18561	18786	19738	98,628
PM_{10}	17783	16939	15240	13968	15297	79,227
$PM_{2.5}$	1754	4121	16725	17667	17910	58,177

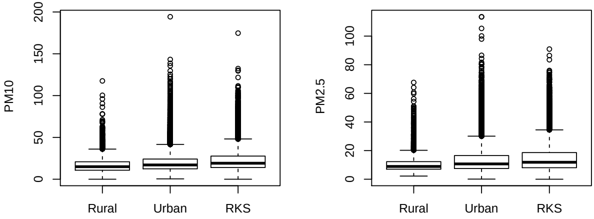
Table: Number of available observations out of the total number of observations in a year, which is 52560 (365×144) for non-leap year and 52704 (366×144) for leap year. A 2008 EU directive triggered $PM_{2.5}$ monitoring in 2009.

How do the data look like?



(a): NO2

(b): O3



(c): PM10

(d): PM2.5

Aims and objectives of our work

1. To model daily levels of four major pollutants namely, NO_2, O_3, PM_{10} and $PM_{2.5}$, for the period 2007–2011.

2. To build up a process based suitable spatio-temporal model that
    1. can handle highly variable and sparse air pollution data.
    2. is more accurate than recently developed methods.
    3. is based on a spatial process which allows us to interpolate at any unobserved location.
    4. allows us to aggregate pollution levels in both space and time.

3. To integrate output from a computer simulation model AQUM (Air Quality Unified Model) on a 12-kilometer grid.

Spatio-temporal auto-regressive models

- General form of spatio-temporal model (books by Cressie and Wikle, 2011 and Banerjee, Carlin and Gelfand, 2015):

$$\begin{aligned} Z(\mathbf{s}, t) &= \mu(\mathbf{s}, t) + \epsilon(\mathbf{s}, t), \\ \mu(\mathbf{s}, t) &= \mathbf{x}(\mathbf{s}, t)'\boldsymbol{\beta} + \eta(\mathbf{s}, t), \\ \eta(\mathbf{s}, t) &= \rho\,\eta(\mathbf{s}, t-1) + \omega(\mathbf{s}, t), \end{aligned}$$

- $Z(\mathbf{s}, t)$ is the square-root of observed data at site $\mathbf{s}$ and time t.
- $\boldsymbol{\beta}$ is the regression parameter, $\mathbf{x}(\mathbf{s}, t)$ is the covariate vector.
- $\epsilon(\mathbf{s}, t)$ is the white noise $N(0, \sigma_\epsilon^2)$, e.g. accounting for measurement error.
- $\eta(\mathbf{s}, t)$ is the space-time interaction term, modelled by an auto-regressive Gaussian Process model.

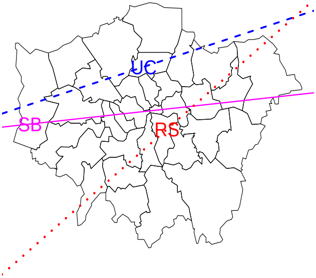
Modelling the regression part, $\mathbf{x}(\mathbf{s}, t)'\boldsymbol{\beta}$

- We allow site-wise regression lines. Have 3 site types: Rural, Urban and Road Side. With $x(\mathbf{s}_i, t)$ as the AQUM value, we assume:

$$\mathbf{x}(\mathbf{s}_i, t)'\boldsymbol{\beta} = \sum_{k=0}^{2} \delta_k(s_i)\left(\beta_{0k} + \beta_{1k} X(s_i, t)\right),$$

where $\delta_0(\mathbf{s}_i) = 1$ for all $\mathbf{s}_i$, and for $k = 1, 2$, $\delta_k(\mathbf{s}_i) = 1$, if $\mathbf{s}_i$ is of k-th type of site, $\delta_k(\mathbf{s}_i) = 0$, otherwise.

- Different regression lines can be obtained from this general form,
- i.e., one regression line for Rural, another for Urban, and another for Road Side.



- This versatile all encompassing model allows, pollutant specific, different regression lines for different site types.

Modelling the space-time interaction term, $\omega(\mathbf{s}, t)$

- We use an extended space-time model based on Gaussian Predictive Processes (GPP).
- We have added further flexibility into the model by improving the knot-selection process in the GPP method.
- The extension allowed us to have more knots in the densely populated areas leading to better estimation in those neighbourhoods.
- Details are omitted but all models are implemented by extending the `R` package `spTimer` publicly available from CRAN.

Results for NO_2 and O_3 model validation

NO_2: Fitting N = 92,440, validation N=67,094, SD=37.19					
Model	**RMSPE**	**MAPE**	**Bias**	**Coverage (%)**	R^2
Simple Kriging	32.87	22.88	2.56	69.59	0.53
Linear model	30.46	19.63	–5.09	94.43	0.60
Best model	17.65	12.99	0.41	97.42	0.89
O_3: Fitting N = 58,900, validation N=39,728, SD=22.23					
Simple Kriging	13.30	9.86	–2.95	78.25	0.80
Linear model	16.0	12.42	8.47	93.86	0.69
Best model	10.17	7.59	0.07	91.72	0.89

Table: Assessment of predictive performance for a range of models for NO_2 and O_3. R^2 denotes the sample correlation coefficient between the predictions and actual observations.

Results for PM_{10} and $PM_{2.5}$ model validation

PM_{10}: Fitting N = 46,894, validation N=32,333, SD=11.98					
Model	**RMSPE**	**MAPE**	**Bias**	**Coverage (%)**	R^2
Simple Kriging	7.34	4.75	–0.75	64.96	0.77
Linear model	9.98	6.74	–1.74	93.70	0.61
Best model	5.48	3.56	–0.65	90.03	0.81
$PM_{2.5}$: Fitting SS = 35,791, validation SS=22,386, SD=9.52					
Model	**RMSPE**	**MAPE**	**Bias**	**Coverage (%)**	R^2
Simple Kriging	4.63	2.96	–0.72	67.84	0.81
Linear model	8.03	5.30	–1.87	92.73	0.60
Best model	4.30	2.66	–0.97	82.38	0.85

Table: Assessment of predictive performance for a range of models for PM_{10} and $PM_{2.5}$. R^2 denotes the sample correlation coefficient between the predictions and actual observations.

Summary of RMSEs for daily data for London only

Model	RMSPE	MAPE	Bias	R^2	Cover (%)
PM_{10}: Fitting N = 11,828, validation N=1,393					
Best model	3.81	2.85	0.87	0.85	89.37
Pirani et al 2014	4.75	–	–	0.63	–

Table: Model validation and model choice measures for PM_{10} using 24 fitting and 5 validation sites within London only.

Pirani, et al (2014): *J. of Exposure Science and Environmental Epidemiology*, 319-327.

Aggregating to local authority area, $\mathcal{A}_k$

- We define average pollution:

$$v(\mathcal{A}_k, t) = \frac{1}{|\mathcal{A}_k|} \int_{\mathbf{s} \in \mathcal{A}_k} \mu^2(\mathbf{s}, t) d\mathbf{s}, \tag{1}$$

where $\mu(\mathbf{s}, t)$ is the true unobserved concentration at location $\mathbf{s}$ and at time t.

- We estimate it by block average as follows:

$$\hat{v}(\mathcal{A}_k, t) = \frac{1}{N_k} \sum_{j=1}^{N_k} \mu^2(\mathbf{s}^*_{kj}, t), \tag{2}$$

where $\mu(\mathbf{s}^*_{kj}, t)$ is a prediction of the pollution concentration at location $\mathbf{s}^*_{kj}$, all within the areal unit $\mathcal{A}_k$, from the air pollution model at time t.

Alignment continued...

- Here N_k is the number of grid 1 km^2 corners within the LA $\mathcal{A}_k$.
- To obtain $\mu^2(\mathbf{s}^*_{kj}, t)$ we use AQUM values at 1 km^2 (i.e. very high) resolution.
- These finer resolution AQUM values will strengthen the accuracy of the prediction maps.
- Note μ^2 because of the square-root transformation used to model pollution concentration.
- Surely, $\hat{v}(\mathcal{A}_k, t)$ will have uncertainty from the estimated $\mu(\mathbf{s}^*_{kj}, t)$.
- How can we propagate that uncertainty to the health outcome model?

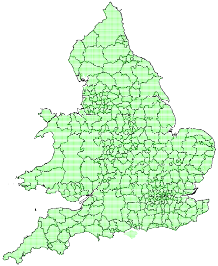
MCMC to the rescue:

- Imagine that the we have L MCMC samples $\mu^{(\ell)}(\mathbf{s}^*_{kj}, t)$, for $\ell = 1, \ldots, L$.
- Then, we form

$$v^{(\ell)}(\mathcal{A}_k, t) \;=\; \frac{1}{N}\sum_{j=1}^{N} \mu^{(\ell)2}(\mathbf{s}^*_{kj}, t).$$

- The health outcome model is also implemented by MCMC.
- Our proposal then is to use the $v^{(\ell)}(\mathcal{A}_k, t)$ in the ℓth iteration of the health outcome model.
- This allows us to propagate uncertainty from the air pollution model to the health outcome model.

Aggregating to local and unitary authority (LUA) areas



- Map of 346 LUAs in England and Wales.
- A 1-kilometer square grid (151,248 green dots) is superimposed.
- Average air pollution in an LUA is the block average of the pollutions in the green dots falling within that LUA.
- Our best Bayesian model is used to interpolate (model based Kriging) the air pollution at the green dots.
- Thus we produce air pollution estimate at any LUA at any time point (daily, monthly, annual)!

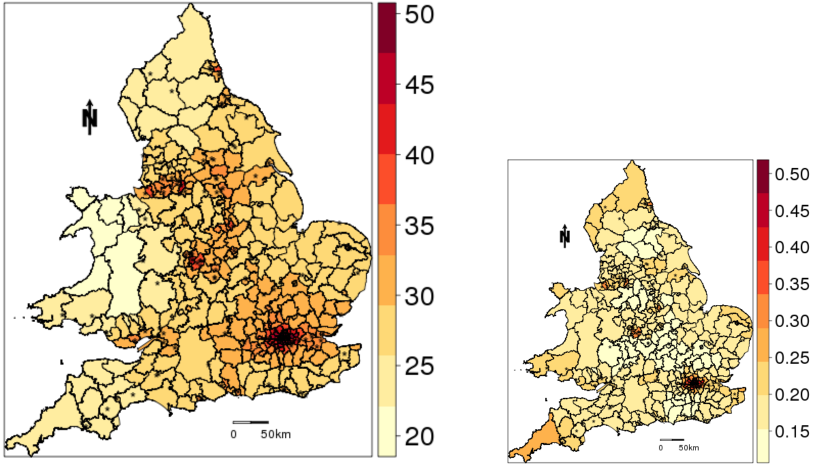


Figure: Local authority-wise annual prediction plot for NO_2 and their standard deviations (right panel) for 2011. Annual limit value of 40 is exceeded in most cities.

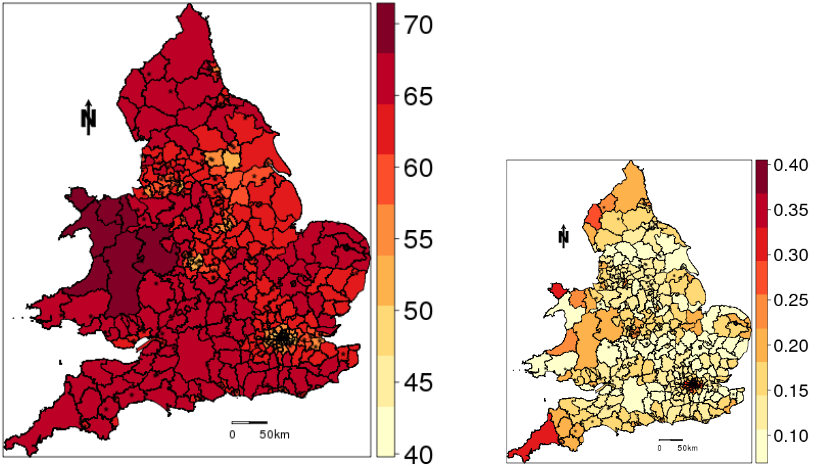


Figure: Local authority-wise annual prediction plot for O_3 and their standard deviations (right panel) for 2011. Rural areas have higher O_3 levels than urban areas.

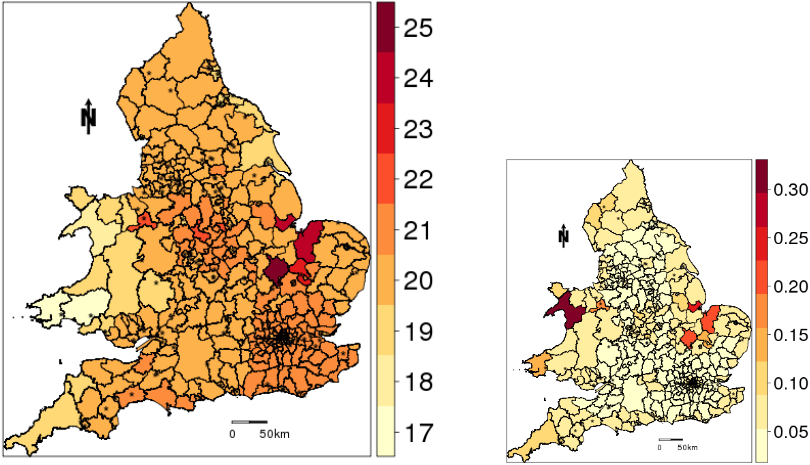


Figure: Local authority-wise annual prediction plot for PM_{10} and their standard deviations for 2011. Cities and suburbs have higher levels.

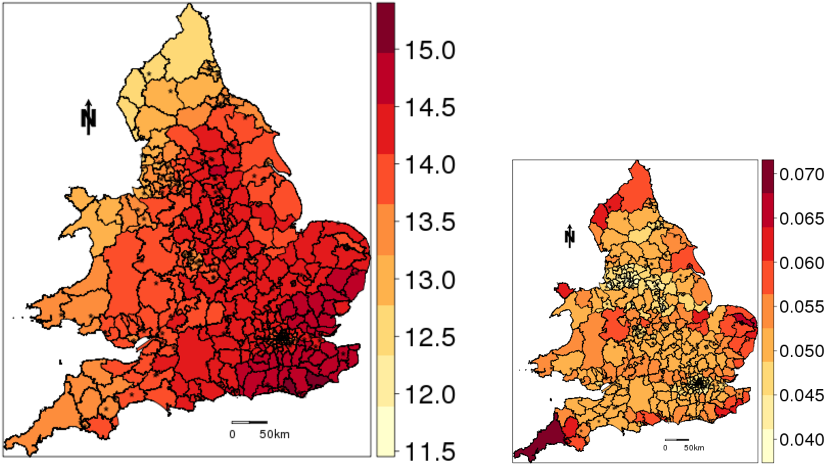


Figure: Local authority-wise annual prediction plot for $PM_{2.5}$ and their standard deviations for 2011. Cities and suburbs, especially in the South-East, have higher levels.

Estimating health effects (Lee et al 2016, Biostatistics)

- Let Y_{kt} denote the number of hospitalisation in the kth local authority $\mathcal{A}_k$ in the tth month.
- $k = 1, \ldots, 323$ local authorities in England
- $t = 1, \ldots, 60$ months in five years, 2007-2011.

$$\begin{aligned} Y_{kt} &\sim \text{Poisson}\left(E_{kt} R_{kt}\right) \\ \log(R_{kt}) &= \alpha + \beta_1 \hat{v}_{kt} + \beta_2 \texttt{jsa}_{kt} + \beta_3 \texttt{house}_{kt} + \psi_{kt} \end{aligned}$$

- E_{kt} is directly standaridised hospitalisation (age and sex) counts nationally.
- R_{kt}: Relative risk,
- $\hat{v}_{kt}$; pollution estimate.
- $\texttt{jsa}_{kt}$: Average job seekers allowance.
- $\texttt{house}_{kt}$: Average house price.
- ψ_{kt}: space-time random effect.

Results from the health outcome model

	RR	Lower 2.5%	Upper 97.5%	Pollutant SD
NO_2	1.028	1.021	1.033	16.07
PM_{10}	1.026	1.011	1.039	4.90
$PM_{2.5}$	1.006	0.993	1.020	4.11
O_3	0.997	0.994	0.999	7.30

Table: Estimated health effects from each pollutant for a range of models. All results are presented as relative risks for a one standard deviation increase in pollution.

- An estimated 2.8% increased risk of hospitalisation due to one sd increase in exposure to NO_2.
- Implies 17,000 extra hospital admissions per year, as there are around 613,000 admissions per year in England.
- This implies a potential spending of **£4.76 million** assuming a week's stay on average.

Conclusions

1. We have developed pollutant specific models which worked well for *a*ll four important pollutants, PM_{10}, $PM_{2.5}$, O_3, NO_2.
2. Our models fill up the sparsity of the observed air quality data by integrating output from the AQUM which are available over a fine grid.
3. Our models also improve similar other modelling attempts, e.g. Pirani et al (2014). We are not aware of any similar study offering high quality air pollution estimates along with their individual error error bars.
4. We are able to estimate pollution levels, along with their uncertainties, at any desired level of adiministrative geography.
5. We are able to measure long term exposure since we have modelled daily data for a 5 year period for whole of UK, for all four pollutants.

Pollution data available

1. Exposure estimates, and their uncertainties, from our best model:
    1. for all four pollutants
    2. at both daily and annual time scales
    3. for the five years 2007-2011
    4. at the 151,248 1-kilometer grid points
    5. and also for all the local authorities in England and Wales
2. are available online. Total size is about 64GB.
3. From my website http://www.soton.ac.uk/~sks/.
4. Thus we provide the most accurate empirically verified air pollution estimates at 1-kilometer grid in E & W.

Future data usage

Possibilities are endless!

- Government and regulatory bodies can use the data to evaluate post-hoc compliance to air pollution standards in even un-monitored areas all over England and Wales.
- Compliance can be evaluated at any socio-economic-politico geographic scale: i.e. post-code, local authority area, LSOA, electoral wards etc.
- Researchers from both academic and government agencies such as the Public Health England can link air pollution to a range of health out-come data.
- For example, colleagues in UCL are associating air pollution levels with the millenium cohort data on children's mental health.

Future work

For example

- Improve the models by further methodological development, e.g. multivariate models for the four pollutants.
- Obtain similar exposure estimates for 2012-2017 possibly using new and improved models
- Develop on-line tools (apps) to deliver data sets on the fly!
  - for user defined geographies and coarser time domains, e.g. monthly, quarterly etc.
- Evaluate health impact using rigorous epidemiological studies.

- Please email me (*S.K.Sahu@soton.ac.uk) if you have queries about the data sets.*