# Hierarchical Bayesian spatio-temporal models for air pollution concentrations in the UK

Sujit Sahu

Glasgow: September 2014

http://www.soton.ac.uk/~sks/

EPSRC
Engineering and Physical Sciences Research Council

Met Office

# Summary

- This is a large interdisciplinary research project whose aim is to develop a rigorous statistical framework for estimating the long-term health effects of air pollution.
- The project is run by the University of Southampton, University of Glasgow, the Met Office, and is a collaboration between statisticians, epidemiologists and meteorologists.
- It is funded by the Engineering and Physical Sciences Research Council (EPSRC), and runs for three years starting January 2013.

# 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


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

- The government admits air quality laws will be breached in 15 regions until 2020. BBC News, 6 March 2013.
- Traffic pollution kills 5,000 a year in UK, says study. BBC News, 17 April 2012.

# Talk outline

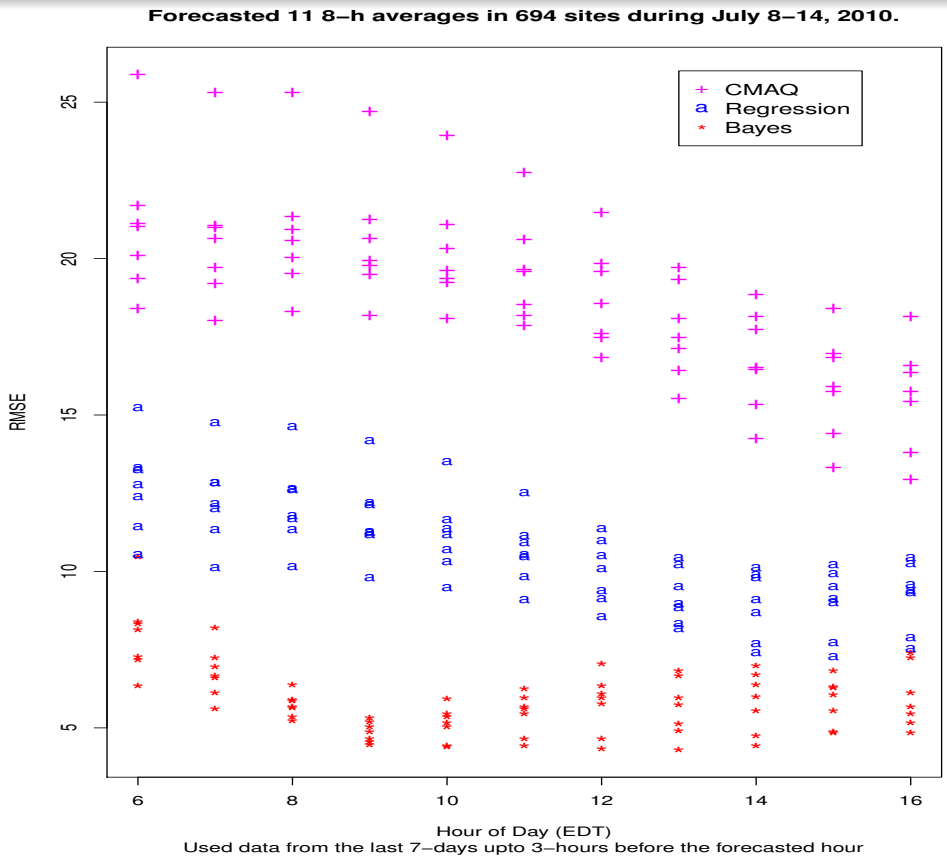
1. Air pollution modelling
    1. Deterministic modelling
    2. Statistical modelling
    3. Our new approach
2. Aggregating air pollution estimates to local authority levels.
3. Conclusions.

# The CMAQ model (deterministic)

- CMAQ = Community Multi-scale Air Quality Model.
- A computer simulation model which produces “averaged” output on 36km, 12km square grid cells.
- Uses variables such as power station emission volumes, meteorological data, land-use, etc. with atmospheric science (appropriate differential equations) to predict pollution levels. Not driven by monitoring station data.
- Outputs are biased but there is no missing data and provides spatial coverage throughout a study region.
- Monitoring data provide more accurate measurements, but are sparse and lots missing!

# Forecasting using Eta-CMAQ

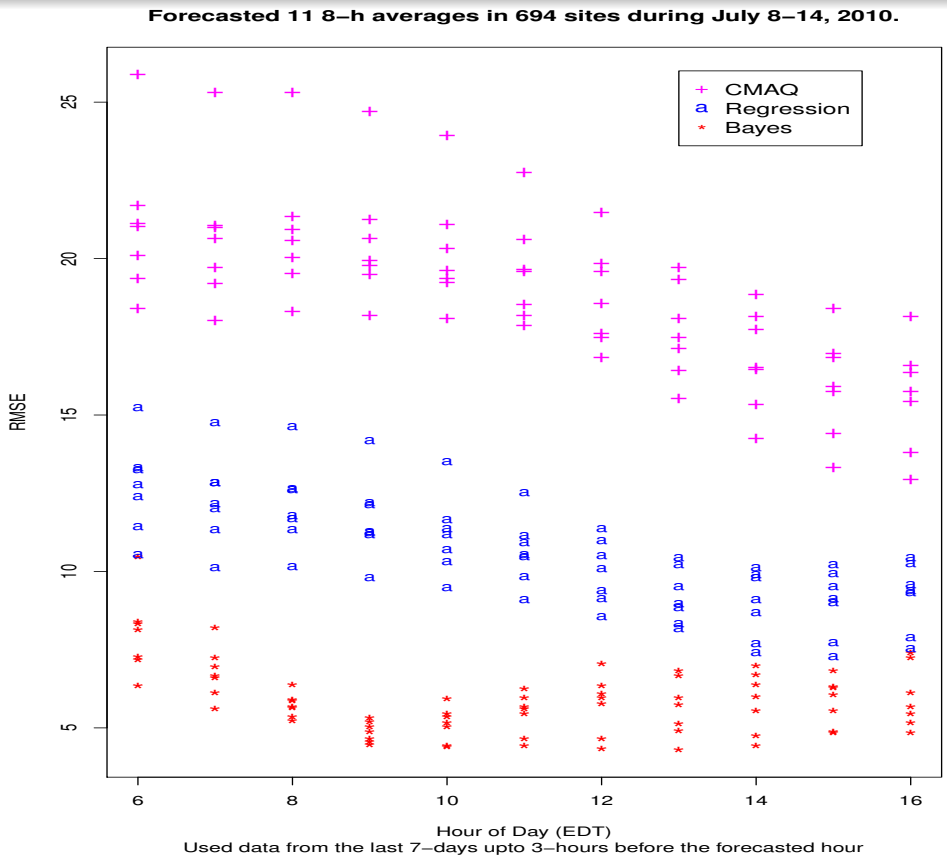
- http://airnow.gov provides forecasts for 8-hour average Ozone concentration level at the current hour.




- Root Mean Square Errors (RMSE) of real time forecasts.
- The US Environmental Protection Agency has adopted our method.

# Forecasting using Eta-CMAQ

- http://airnow.gov provides forecasts for 8-hour average Ozone concentration level at the current hour.




- Root Mean Square Errors (RMSE) of real time forecasts.
- The US Environmental Protection Agency has adopted our method.

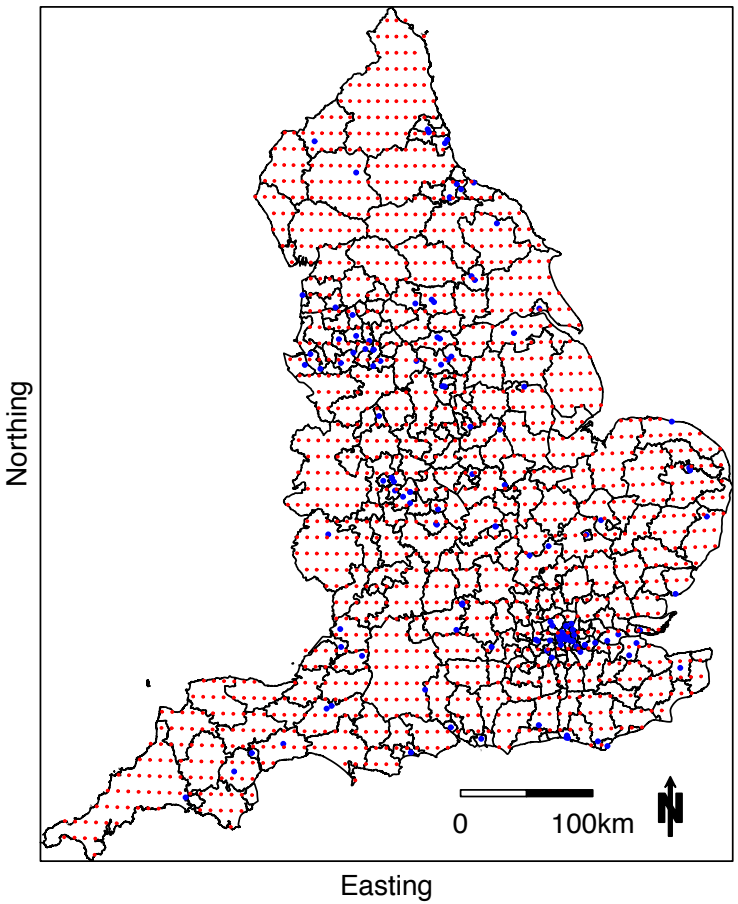
# Recent deterministic modelling of UK data

- CMAQ output are not available for the UK.
- But we have output from a 'similar' model: Air Quality Unified Model (AQUM) (Savage et al., 2013).
- Like CMAQ, AQUM uses atmospheric variables like temperature, humidity, wind speed, wind direction, and also data on emission from various sources.
- The AQUM output (over 12 km grid cells), like those from CMAQ, are not very accurate.
- Also the grid cells are spatially mis-aligned with the local authority boundaries for which we have health outcome data.
- Need to adjust using accurate statistical methods which are capable of estimating uncertainties in the air pollution estimates.

# Air pollution - monitoring data

- There are 166 air pollution monitoring sites in the UK recording hourly measurements (AURN sites).
- We are interested in four most important pollutants: Particulate Matter: $PM_{10}$, $PM_{2.5}$, Ozone($O_3$), and $NO_2$.
- We have downloaded these data for 1826 days in the five years, 2007-2011.
- We work with daily data. Why?
  - We are interested in long term effects not short term (like hourly) extremes or exceedances.
  - UK/EU air pollution regulations are at this time scale.
- Let $Z^{(k)}(\mathbf{s}_j, t)$, $j = 1, \ldots, 166$; $t = 1, \ldots, 1826$ denote these daily data, but on the square root scale to stabilize variance. The super-script $k$ denotes which of the four pollutants we are modelling.

# Air pollution - miss-aligned data and model output




- Map of 323 local authorities in England.
- Red dots define the corners of the 12 km square grid cells where we have AQUM output.
- Blue dots represent the AURN air-quality monitoring sites.

# Mean of different types of sites in UK

- Pollution levels and their averages (shown in the table) vary by site type!

| Site type | Number | $PM_{2.5}$ | $PM_{10}$ | Ozone | $NO_2$ |
|---|---|---|---|---|---|
| Urban Background (UB) | 75 | 12.9 | 19.2 | 59.4 | 46.6 |
| Roadside (RS) | 49 | 14.2 | 21.1 | 50.3 | 74.6 |
| Rural (RL) | 23 | 8.4 | 14.6 | 68.4 | 19.1 |
| Urban Centre (UC) | 24 | 13.8 | 20.1 | 50.3 | 59.4 |
| Urban Industrial (UI) | 10 | 11.3 | 19.9 | 54.9 | 50.3 |
| Suburban (SB) | 19 | 15.1 | 22.8 | 57.4 | 45.6 |
| Kerbside (KS) | 6 | 20.0 | 30.1 | 28.3 | 134.1 |
| Remote (RM) | 5 | NA | 13.8 | 71.7 | 11.2 |
| Airport (AP) | 1 | 13.3 | 19.0 | 53.1 | 63.0 |
| Average pollution | — | 12.97 | 20.27 | 58.50 | 55.77 |

- Note: All measurements are in $\mu g/m^3$ scale.

# Recent Statistical modelling of UK air pollution data

1. Large number of research articles. We mention two most recent work on UK data.
2. Pirani, Gulliver, Fuller, Blangiardo (2014) did spatio-temporal modelling on short-term exposure of $PM_{10}$ in London.
    1. Used data on $PM_{10}$ for 728 days during 2002–2003. Covariates are output of numerical model on a 1km grid, data on emission, temperature etc.
    2. Fitted and compared 5 different regression models.
    3. Did not incorporate spatio-temporal interaction term.
3. Shaddick et al. (2013) used data on annual average of $NO_2$concentration from parts of Europe including UK in 2001.
    1. Spatial model includes various covariates affecting air pollution.
    2. No temporal modelling, hence cannot be used for measuring long term exposure.

# Aims and objectives of our work

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

2. To build up a process based suitable spatio-temporal model that
   1. can handle highly variable 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.

3. To incorporate output of our model (along with their uncertainties) into the model measuring the impact of pollution on human health.

# Spatio-temporal auto-regressive models

- General form of spatio-temporal model (Cressie and Wikle, 2011; Banerjee, Carlin and Gelfand, 2004):

$$\begin{aligned} \boldsymbol{Z}_t &= \boldsymbol{O}_t + \boldsymbol{\epsilon}_t, \\ \boldsymbol{O}_t &= \boldsymbol{X}_t\boldsymbol{\beta} + \boldsymbol{\eta}_t, \\ \boldsymbol{\eta}_t &= \rho\boldsymbol{\eta}_{t-1} + \boldsymbol{\omega}_t, \end{aligned}$$

- $\boldsymbol{Z}_t$ is the square-root of observed data from $n$ sites.
- $\boldsymbol{\beta}$ is the regression parameter, $\boldsymbol{X}_t$ design matrix of covariates at time $t$.
- $\boldsymbol{\epsilon}_t$ follows multivariate normal distribution with parameters $(0, \sigma^2_\epsilon \boldsymbol{I}_n)$ independent of $\boldsymbol{\eta}_t$.
- $\boldsymbol{\eta}_t$ is the space-time interaction term, modelled by an auto-regressive Gaussian Process model.

# Proposed Space-time GPP model

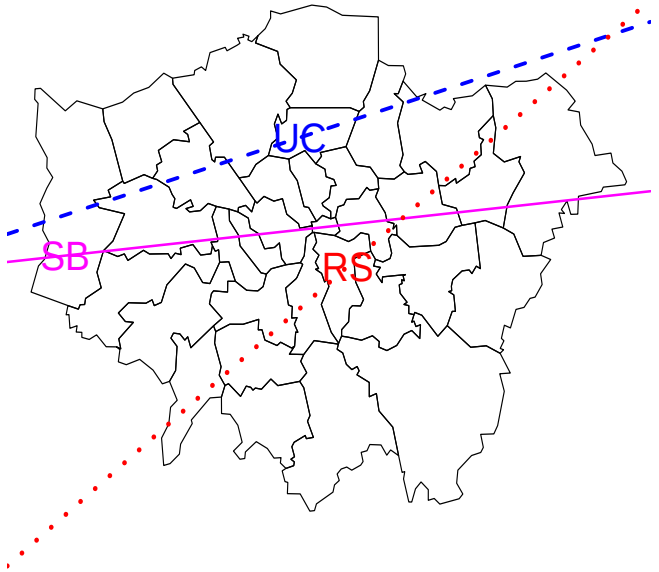
- Sahu and Bakar (2012) extended the auto-regressive models to space-time models based on Gaussian Predictive Processes (GPP).
- But we extend their approach in several ways:
  - adopting anisotropic and non-stationary correlation structure,
  - introducing a further hierarchical model for knot locations based on population density of local authority areas, where
    - knot locations are where the GPP are evaluated.
    - Thus densely populated places should get more knots, since our main interest is to measure impact of pollution on human health.
- Mathematical details are omitted but all models are implemented by extending the `R` package `spTimer` publicly available from CRAN.

# Modelling innovation in the regression part

- We allow site-wise regression lines. If there are $r$ (=9 in our case, UC, SB, .., RS) many type of sites then $\boldsymbol{X}_t\boldsymbol{\beta}$ can be modelled as

$$\boldsymbol{X}_t\boldsymbol{\beta} = \sum_{k=0}^{r} \delta_k(s_i)(\gamma_{0k} + X(s_i, t)\gamma_{1k}),$$

  where $\delta_0(s_i) = 1$ for all $s_i$, $\delta_k(s_i) = 1$, if $s_i$ is of $k$-th type of site, $k = 1, \ldots, r$, $\delta_k(s_i) = 0$, otherwise. $X(s_i, t)$ is AQUM value.
- Different regression lines can be obtained from this general form,
- i.e., one regression line for UC, another for RS, ... so on.



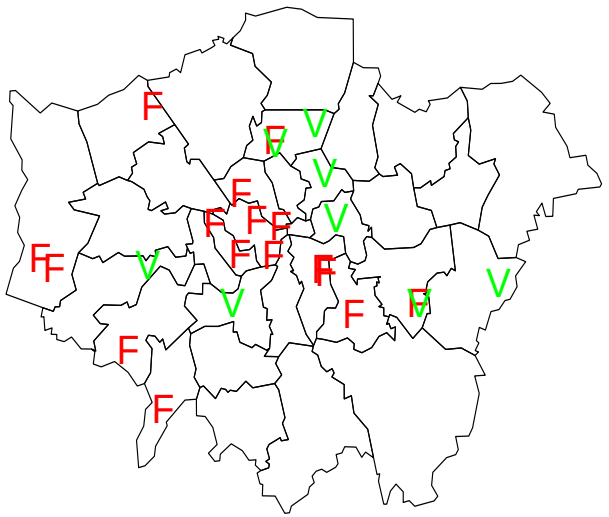

- This versatile all encomapssing model allows, pollutant specific, different regression lines for different site types.

# Catalogue of fitted models

| Model | Linear Part | Knots | Time series | Spatial |
|---|---|---|---|---|
| Model-1 | AQUM | not required | Independent | GP |
| Model-2 | AQUM | not required | AR process | AR |
| Model-3 | AQUM | fixed | AR process | GPP |
| Model-4 | Sitewise Linear | fixed | AR process | GPP |
| Model-5 | AQUM | random | AR process | GPP |
| Model-6 | Sitewise Linear | random | AR process | GPP |
| Model-7 | AQUM | fixed | AR process | GPP (anisotropic) |

# Validation of the models

1. Among the $n$ many monitoring sites we choose at least 10% at random. Denote those as validation sites.
2. Pretend that data at validation sites have not been observed and need to be predicted.
3. Use rest of the data for fitting a model.
4. Predict the pollution values at the validation sites and calculate the RMSE by comparing with the observed data of those sites.
5. The model with the least RMSE is the best.




- F, fitting and V, validation sites in London.

# Results for validating data from whole of UK

Table: Root Mean Square Error (18 validation sites and 148 fitting sites)

| | $PM_{2.5}$ | $PM_{10}$ | Ozone | $NO_2$ |
|---|---|---|---|---|
| SD | 9.55 | 12.0 | 21.82 | 38.06 |
| Kriging | 5.36 | 9.18 | 18.98 | 38.28 |
| AQUM | 8.03 | 14.27 | 19.49 | 36.55 |
| Model-1 | 5.21 | 8.77 | 16.27 | 34.5 |
| Model-2 | 5.26 | 9.10 | 16.9 | 45.3 |
| Model-3 | 4.78 | 7.67 | 12.56 | 24.99 |
| Model-4 | 4.79 | 7.59 | 12.58 | 25.3 |
| Model-5 | 4.79 | 7.59 | 12.58 | 25.3 |
| Model-6 | 4.77 | 7.55 | 12.65 | 26.91 |
| Model-7 | 4.81 | 7.60 | 12.49 | 27.15 |

# Results for validating data from London

Table: Root Mean Square Error (8 validation sites and 21 fitting sites)

| | $PM_{2.5}$ | $PM_{10}$ | Ozone | $NO_2$ |
|---|---|---|---|---|
| SD | 9.82 | 13.40 | 23.97 | 46.84 |
| Kriging | 9.69 | 16.75 | 18.74 | 39.07 |
| AQUM | 8.46 | 13.65 | 18.53 | 33.37 |
| Model-1 | 5.71 | 6.90 | 14.77 | 35.18 |
| Model-2 | 3.95 | 4.90 | 14.11 | 32.37 |
| Model-3 | 3.47 | 3.80 | 12.75 | 24.99 |
| Model-4 | 3.62 | 3.73 | 12.66 | 21.97 |
| Model-5 | 3.47 | 3.79 | 10.05 | 21.22 |
| Model-6 | 3.62 | 3.73 | 10.63 | 21.92 |
| Model-7 | 3.47 | 3.78 | 12.78 | 24.21 |
| *Pirani et al.* | 4.75 | — | — | — |

# CRPS and coverage

Table: CRPS and nominal coverage for 95% prediction intervals using Model-5 for 5 years daily data from London

| Pollutant | crps | coverage | SD |
|---|---|---|---|
| $NO_2$ | 14.0 | 95.3 | 47.4 |
| Ozone | 7.3 | 80.0 | 24.0 |
| $PM_{10}$ | 2.0 | 98.4 | 13.4 |
| $PM_{2.5}$ | 1.7 | 87.9 | 9.8 |

Table: CRPS and nominal coverage for 95% prediction intervals using Model-5 for 5 years daily data from whole UK

| Pollutant | crps | coverage | SD |
|---|---|---|---|
| $NO_2$ | 13.9 | 95.2 | 38.1 |
| Ozone | 6.9 | 84.6 | 21.8 |
| $PM_{10}$ | 3.6 | 85.3 | 12.0 |
| $PM_{2.5}$ | 2.1 | 95.2 | 9.5 |

# Illustration: Estimate of parameters of Model-6 for $PM_{10}$ in London

| Parameter | Mean | Median | SD | 2.5% | 97.5% |
|---|---|---|---|---|---|
| AQUM Intercept | 3.64 | 3.66 | 0.1 | 3.4 | 3.8 |
| AQUM slope | 0.06 | 0.06 | 0.01 | 0.04 | 0.08 |
| UB Intercept | 0.32 | 0.32 | 0.02 | 0.28 | 0.36 |
| UB Slope | –0.006 | –0.006 | 0.002 | –0.01 | –0.0027 |
| RS Intercept | 0.37 | 0.37 | 0.02 | 0.34 | 0.41 |
| RS Slope | –0.005 | –0.005 | 0.002 | –0.008 | –0.0015 |
| UC Intercept | 0.37 | 0.37 | 0.02 | 0.34 | 0.41 |
| UC Slope | 0.34 | 0.34 | 0.02 | 0.31 | 0.38 |
| SB Intercept | 0.36 | 0.36 | 0.02 | 0.33 | 0.40 |
| SB Slope | –0.005 | –0.005 | 0.001 | –0.008 | –0.002 |
| KS Intercept | 1.33 | 1.33 | 0.02 | 1.29 | 1.39 |
| KS Slope | –0.016 | –0.01 | 0.002 | –0.02 | –0.01 |
| $\rho$ | 0.09 | 0.08 | 0.05 | 0.03 | 0.21 |
| $\sigma^2_\epsilon$ | 0.11 | 0.11 | 0.004 | 0.11 | 0.12 |
| $\sigma^2_\eta$ | 0.52 | 0.23 | 0.85 | 0.13 | 3.03 |
| $\phi$ | 0.0035 | 0.0008 | 0.006 | 0.0001 | 0.02 |

# Further validation results.

- We also compare the best statistical Model-6 with the AQUM outputs using:
  - one site at a time leave-out cross-validation RMSE.
  - But we only validate the sites with at least 30% observations to have stable RMSE.

Table: 115 RMSEs for $NO_2$ in the UK

| Models | Minimum | Mean | SD | Maximum |
|---|---|---|---|---|
| AQUM | 8.07 | 33.60 | 21.24 | 134.59 |
| Model-6 | 10.28 | 22.69 | 12.55 | 87.43 |

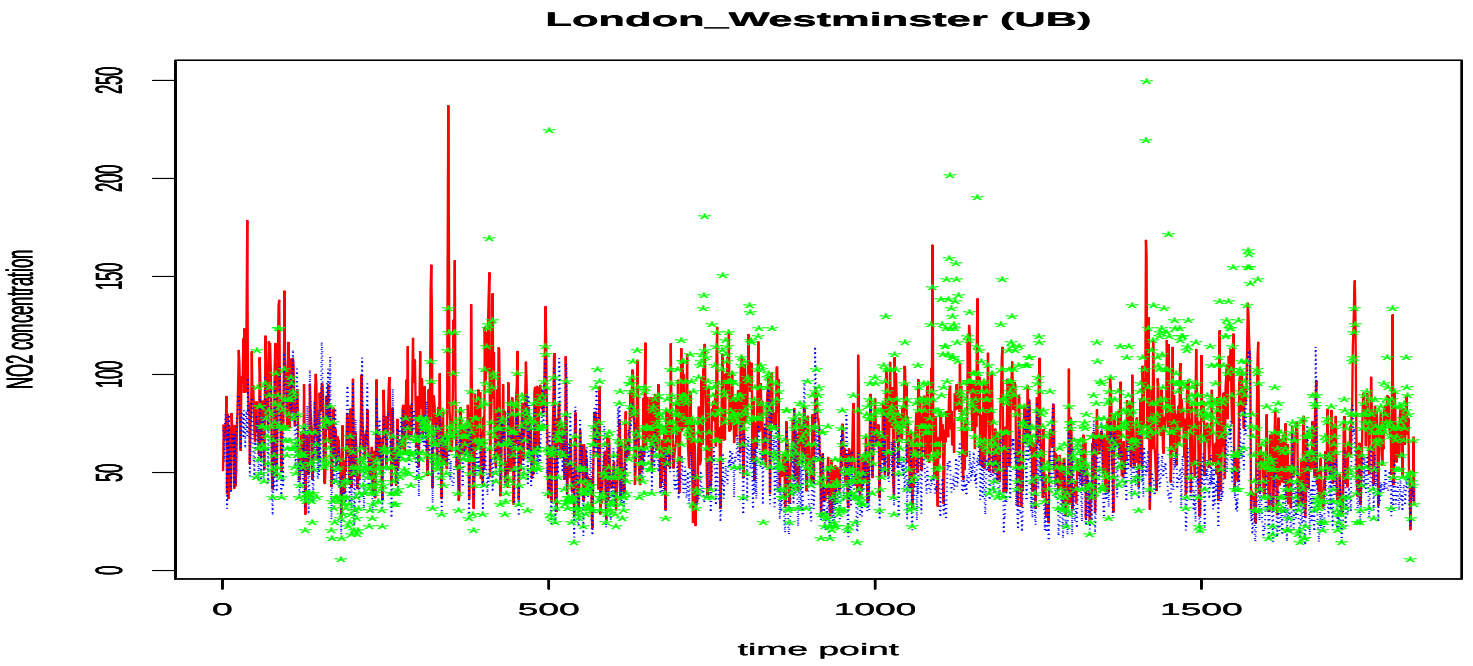
# Summary of Cross-validation RMSEs for England data

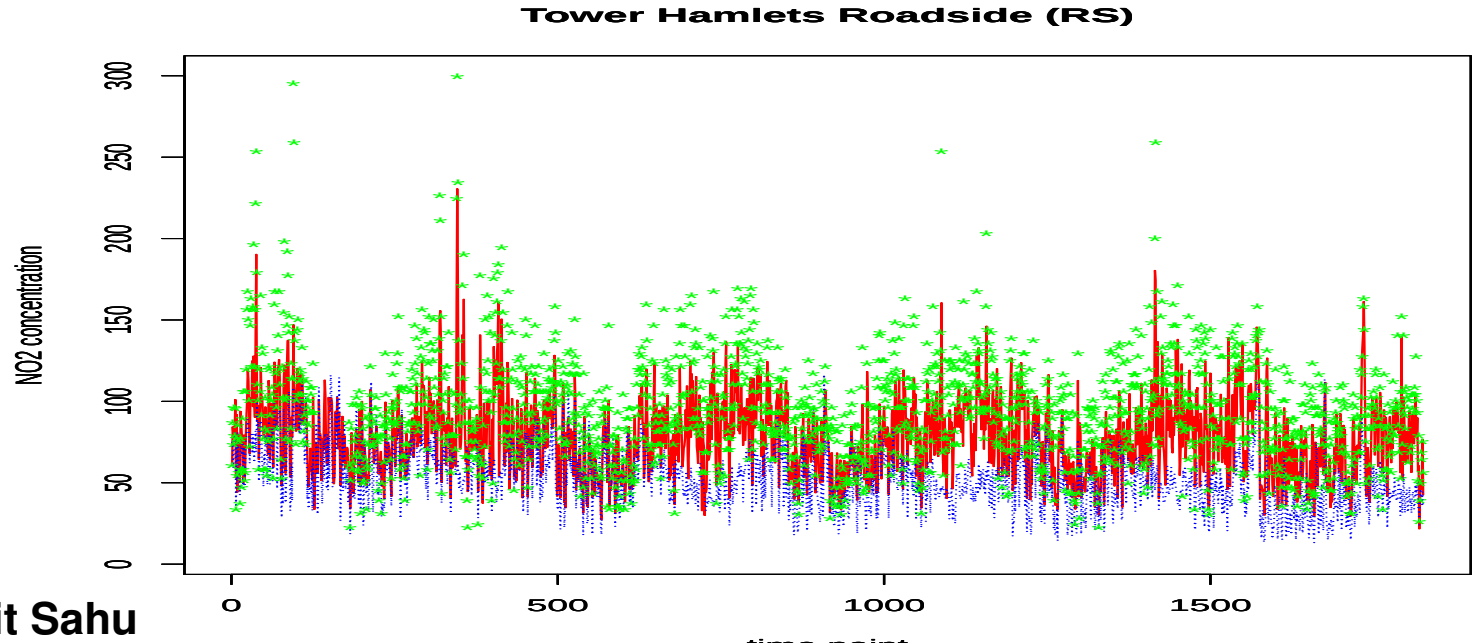
| Models | Minimum | Mean | SD | Maximum |
|---|---|---|---|---|
| 52 RMSEs for $NO_2$ | | | | |
| AQUM | 10.48 | 29.93 | 14.47 | 134.59 |
| Model-6 | 10.28 | 19.16 | 9.16 | 56.99 |
| 31 RMSEs for Ozone | | | | |
| AQUM | 15.44 | 19.30 | 3.27 | 39.02 |
| Model-6 | 7.32 | 10.82 | 2.99 | 18.30 |
| 34 RMSEs for $PM_{10}$ | | | | |
| AQUM | 9.96 | 13.65 | 2.11 | 28.05 |
| Model-6 | 4.51 | 4.19 | 1.10 | 7.73 |
| 30 RMSEs for $PM_{2.5}$ | | | | |
| AQUM | 5.76 | 8.31 | 1.15 | 10.88 |
| Model-6 | 3.15 | 4.32 | 0.73 | 5.93 |

# Summary of Cross-validation RMSEs for London data

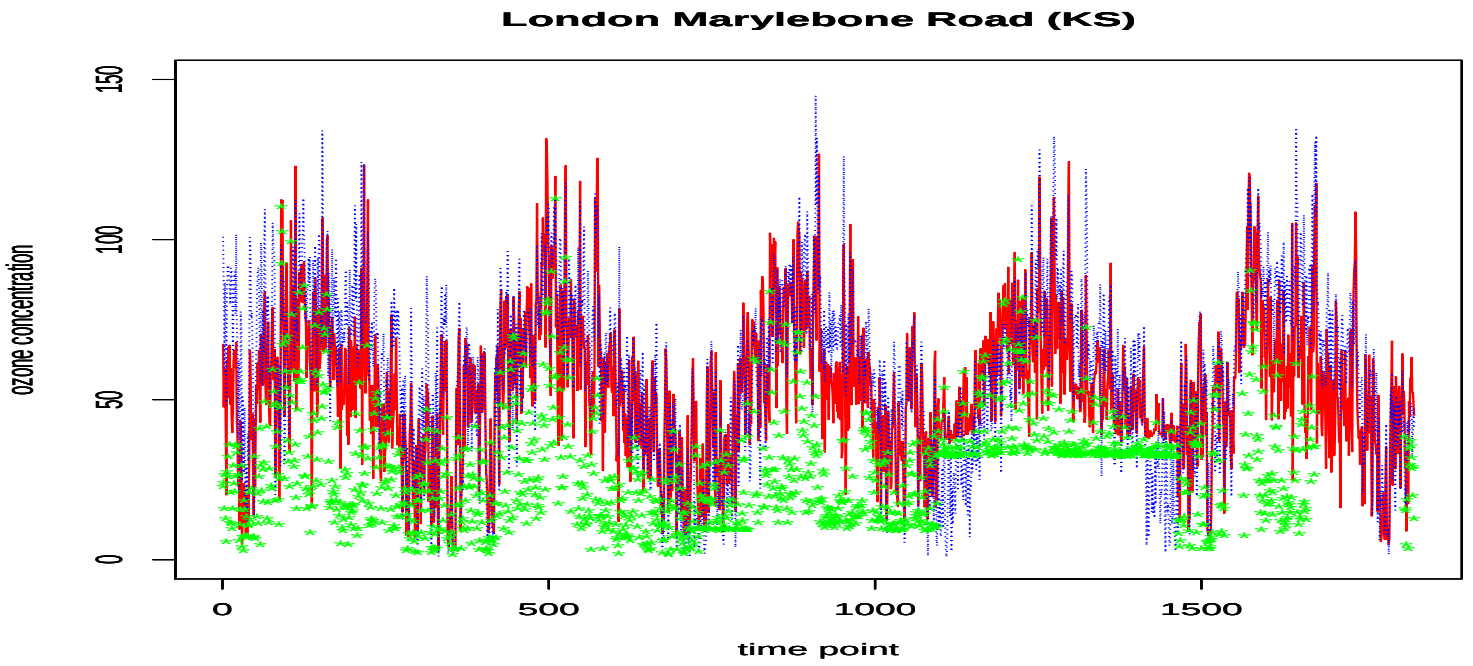
| 17 RMSEs for $NO_2$ | | | | |
|---|---|---|---|---|
| Models | Minimum | Mean | SD | Maximum |
| AQUM | 20.22 | 43.66 | 32.18 | 134.59 |
| Model-6 | 15.60 | 33.44 | 25.32 | 97.53 |
| 12 RMSEs for Ozone | | | | |
| AQUM | 15.71 | 19.02 | 7.91 | 39.02 |
| Model-6 | 6.36 | 11.02 | 7.69 | 31.60 |
| 8 RMSEs for $PM_{10}$ | | | | |
| AQUM | 11.36 | 15.95 | 5.82 | 28.05 |
| Model-6 | 3.70 | 6.98 | 4.97 | 16.87 |
| 7 RMSEs for $PM_{2.5}$ | | | | |
| AQUM | 7.56 | 5.80 | 0.85 | 9.94 |
| Model-6 | 3.07 | 3.59 | 0.36 | 4.97 |

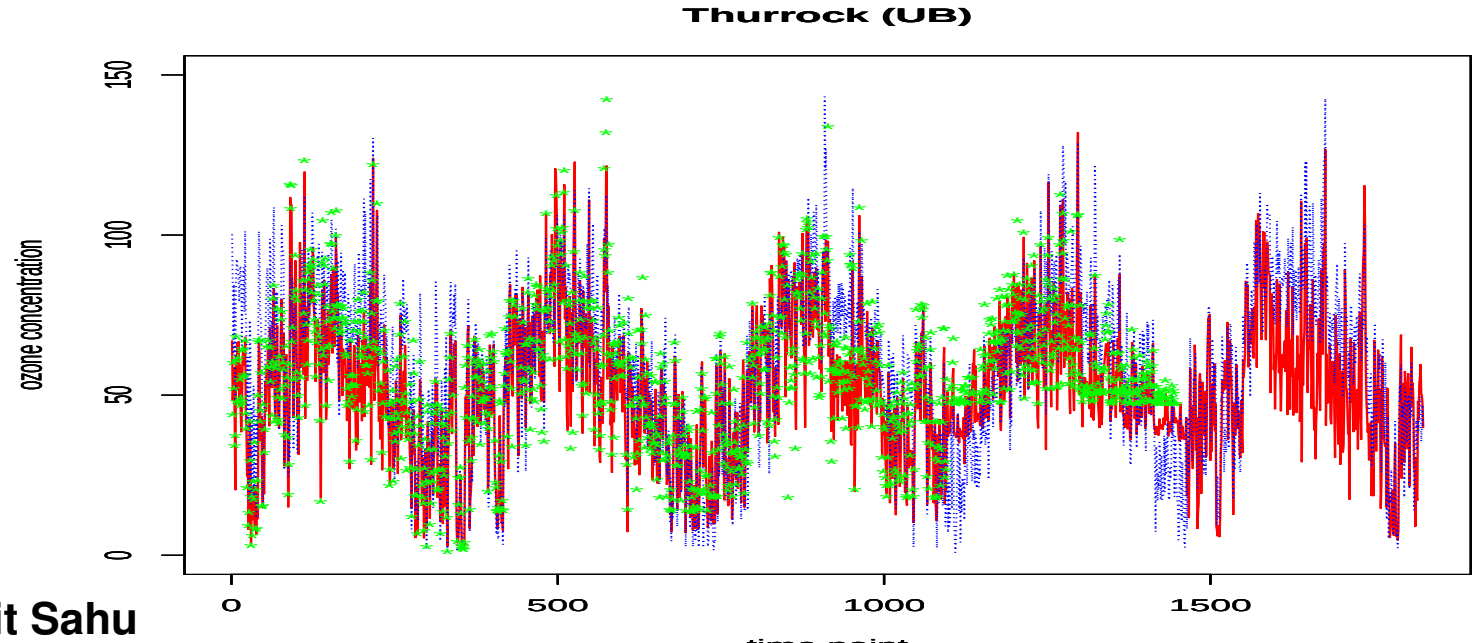
# Comparison of site-wise validation: $NO_2$



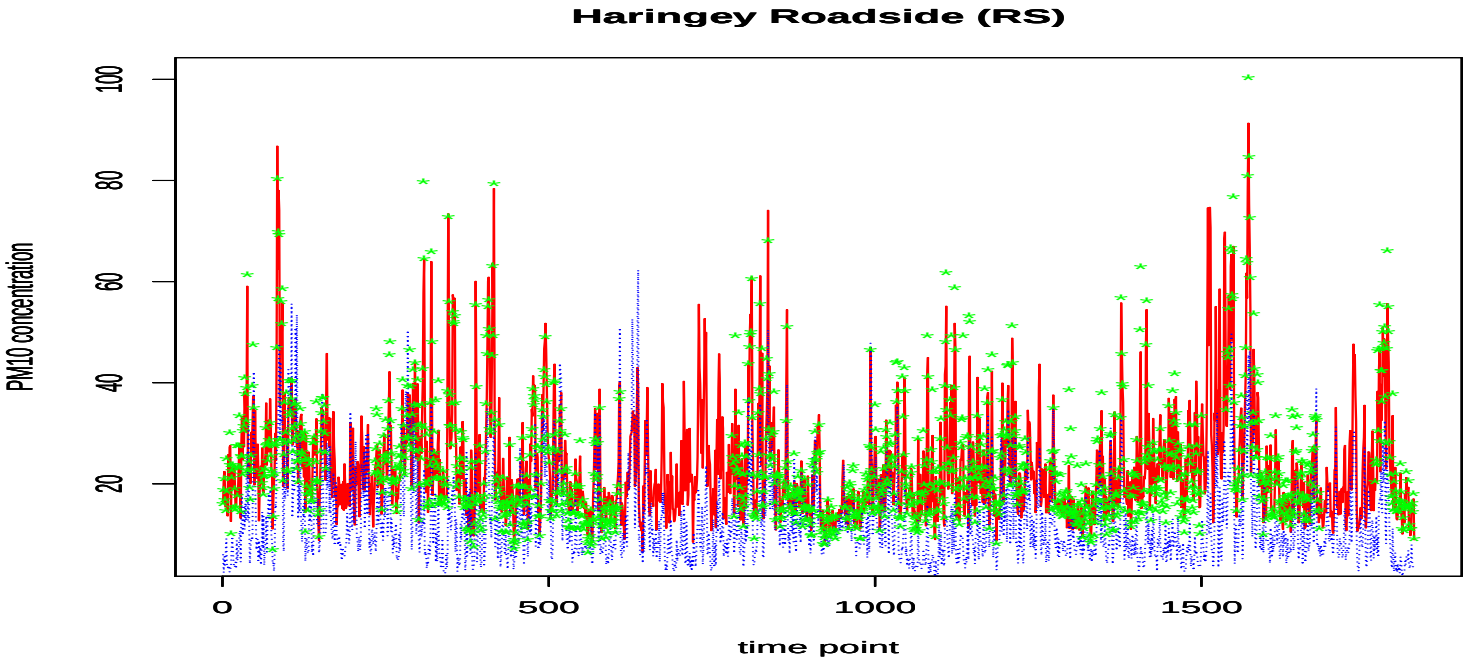



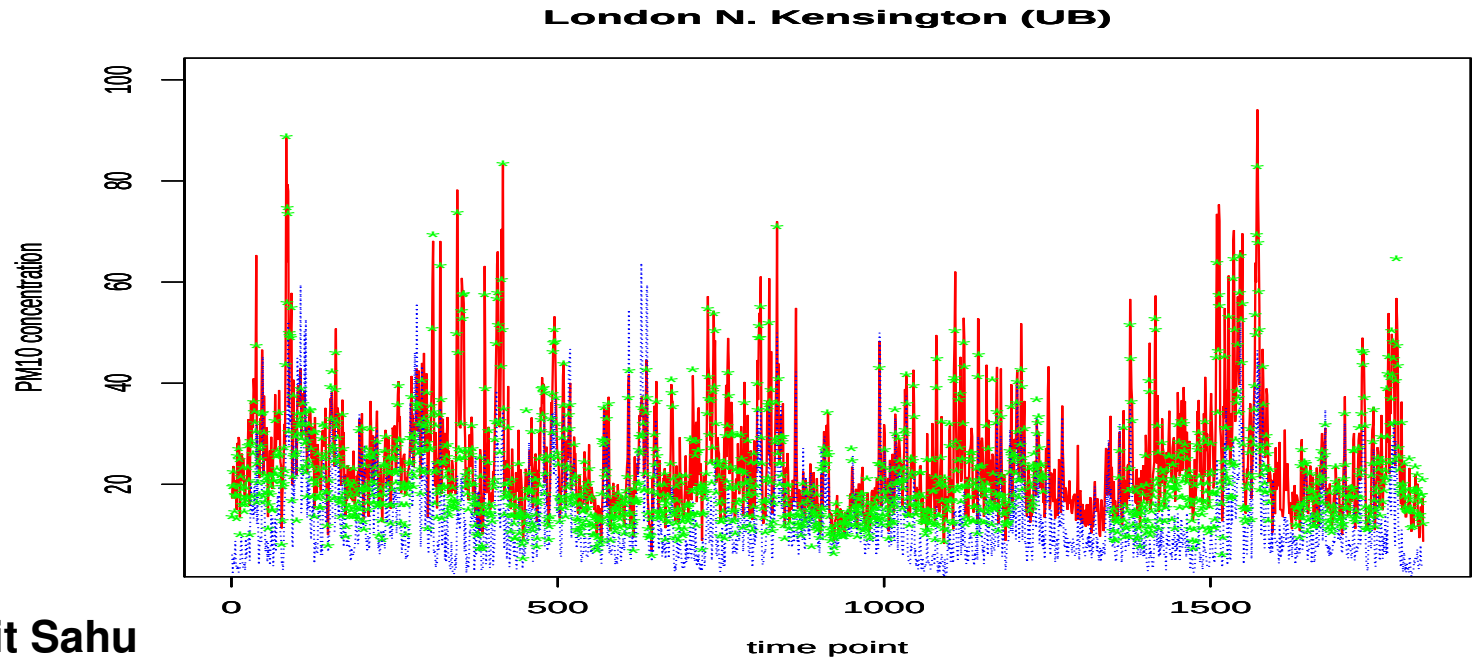
# Comparison of site-wise validation: Ozone



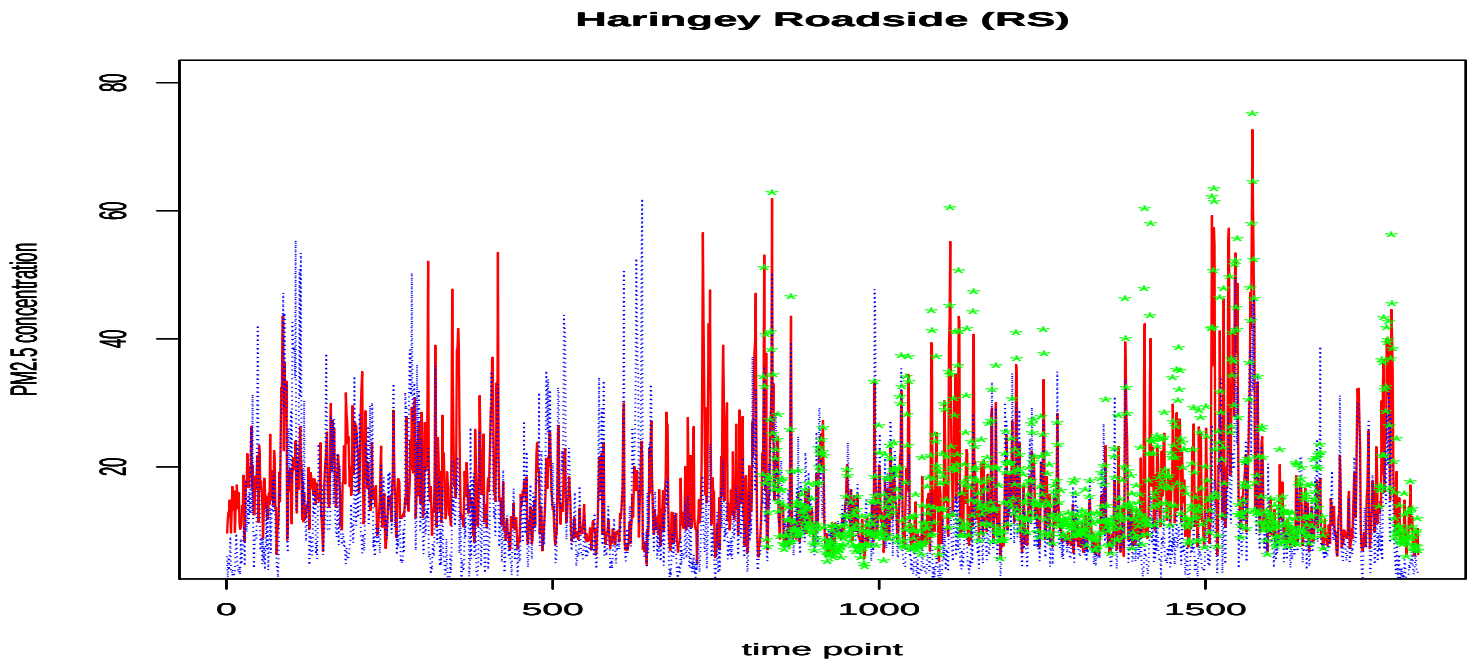



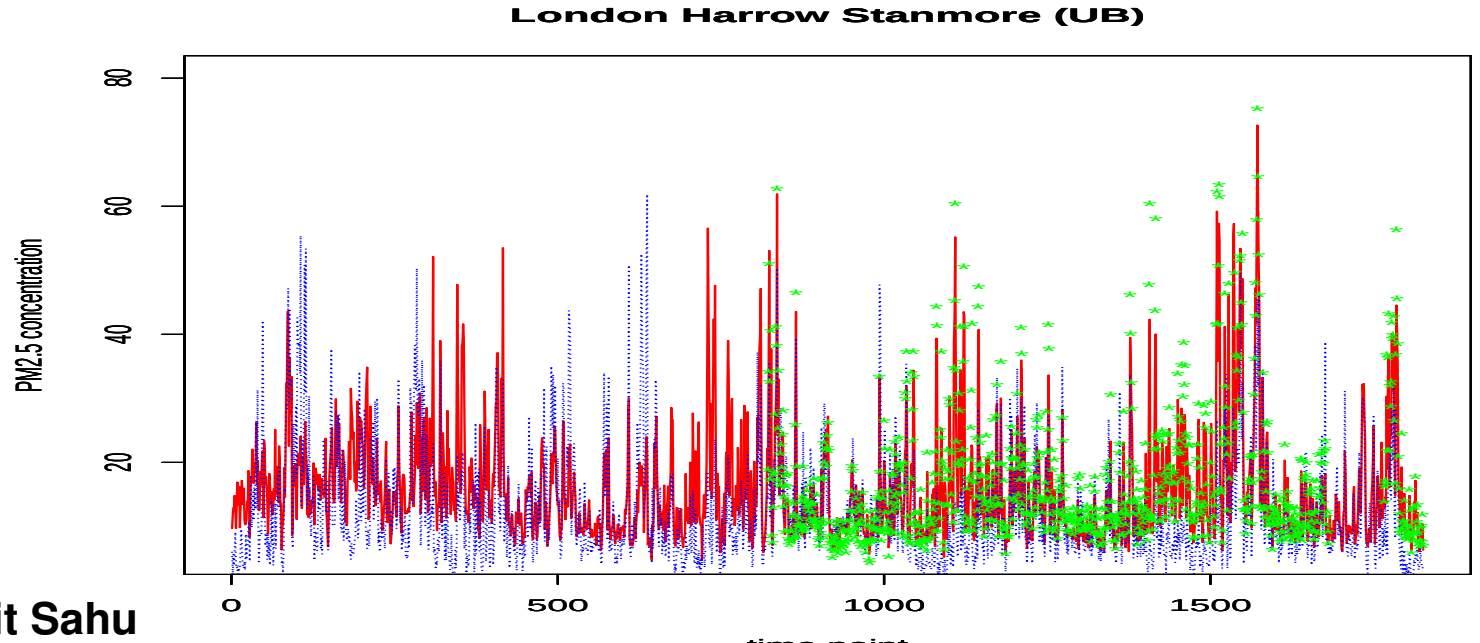
# Comparison of site-wise validation: $PM_{10}$



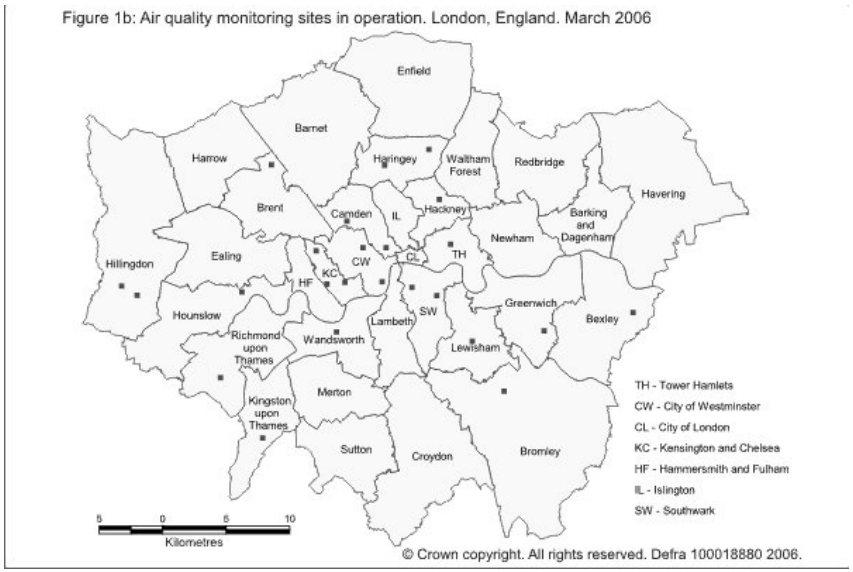



# Comparison of site-wise validation: $PM_{2.5}$



# Different spatial scales for air pollution and health outcome data




- The health model is based on the irregular spatial units $\mathcal{A}_k, k = 1, \ldots, n$.
- The air pollution model is for the data from the $J$ monitoring stations.

- When modelling spatio-temporal data, the time unit in the health model (annual, monthly) can be different from the time unit (hourly, daily) of the air pollution data.

- Therefore, we propose methods that will allow us to aggregate high resolution spatio-temporal data.

# Spatial alignment:

We define average pollution:

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

where $\mu(\mathbf{s})$ is the true unobserved concentration at location $\mathbf{s}$ and at time $t$.
We estimate it by block average as follows:

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

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 $t$.

- Note $\mu^2$ because of the square root transformation used to model pollution concentration.

# Alignment continued...

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

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

- Now, recall that the health outcome model is also implemented by MCMC.
- Our proposal then is to use the $v^{\ell}(\mathcal{A}_k, t)$ in the $\ell$th iteration of the health outcome model.
- This allows us to propagate uncertainty from the air pollution model to the health outcome model.
- Duncan's talk continues from here...

# Conclusions

1. We have proposed a number of non-stationary, anisotropic models which worked well for *a*ll four important pollutants, $PM_{10}$, $PM_{2.5}$, Ozone, $NO_2$.
2. Our approach does not need pollutant specific considerations.
3. AQUM outputs are better than others but clearly improved by using a *single* model, as shown by cross-validation studies both for UK and London data.
4. These models also improve similar other modelling attempts (e.g. Pirani et al.).
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, unlike other studies.

# Discussion

1. *Statistical models have the added advantage* of producing the correct prediction uncertainty in air pollution estimates, which are required by the health outcome model.
2. MCMC sampling also enables us to estimate the uncertainties in the spatial (point to local authority level) and temporal (daily to monthly or annual) aggregates.
3. Statistical modelling is preferable since the models provide a complete description of the data, not only the summaries and averages, but also the variability of the data.
4. All the assumptions in the modelling are also explicit which enables their scrutiny and can suggest to re-modelling for further improvement.
5. By building a statistical model we are able to account for (rather than 'adjust for') the effects of all the variables (e.g. site types).