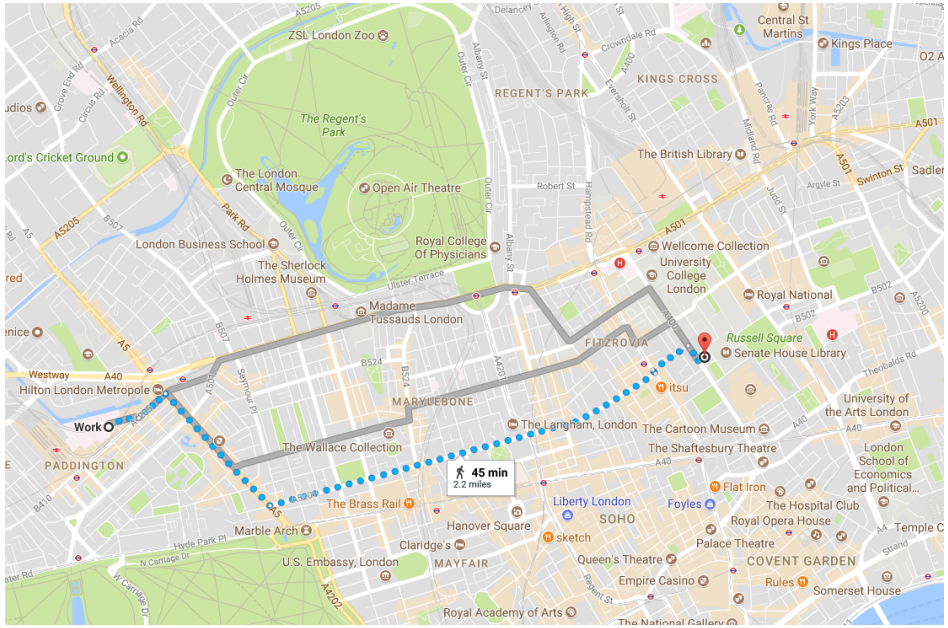
Statistical modelling and challenges in environmental epidemiology

Marta Blangiardo

Imperial College London
MRC-PHE Centre for Environment and Heath
m.blangiardo@imperial.ac.uk

LSHTM, 17^{th} October 2017

Who I am Senior Lecturer in Biostatistics
Imperial College, Department of Epidemiology and Biostatistics
MRC-PHE Centre for Environment and Health



What I do Focus of my research is on the development of statistical methods to answer applied epidemiological questions.

Environmental Epidemiology

- Environmental epidemiology focuses on linking environmental hazards (exposures) to health outcomes.
  - Two main ingredients
    - Environmental exposure
      $\Rightarrow$ a continuous field over the study area.
    - Health outcomes
      $\Rightarrow$ cohort / administrative data.

Environmental Epidemiology

- Environmental epidemiology focuses on linking environmental hazards (exposures) to health outcomes.
  - Two main ingredients
    - Environmental exposure
      $\Rightarrow$ a continuous field over the study area.
    - Health outcomes
      $\Rightarrow$ cohort / administrative data.
  - Study designs typically used:
    - Cohort studies - Focus on long-term effects (individual data);
    - Small area studies - Focus on long-term effects (aggregated data);
    - Time-series studies - Focus on short-term effects (aggregated data);

Environmental Epidemiology

- Environmental epidemiology focuses on linking environmental hazards (exposures) to health outcomes.
  - Two main ingredients
    - Environmental exposure
      $\Rightarrow$ a continuous field over the study area.
    - Health outcomes
      $\Rightarrow$ cohort / administrative data.
  - Study designs typically used:
    - Cohort studies - Focus on long-term effects (individual data);
    - Small area studies - Focus on long-term effects (aggregated data);
    - Time-series studies - Focus on short-term effects (aggregated data);

Spatial and temporal dependencies are keys.

Why accounting for spatial and temporal dependencies is important

Spatial (temporal) patterns suggest that observations close to each other have more similar values than those far from each other.

Are we explicitly interested in the spatial pattern of disease risk?

Do we want to evaluate temporal trends for each area?

Why accounting for spatial and temporal dependencies is important

Spatial (temporal) patterns suggest that observations close to each other have more similar values than those far from each other.

Hypothesis generating

Are we explicitly interested in the spatial pattern of disease risk?

Do we want to evaluate temporal trends for each area?

Why accounting for spatial and temporal dependencies is important

Spatial (temporal) patterns suggest that observations close to each other have more similar values than those far from each other.

Hypothesis generating

Are we explicitly interested in the spatial pattern of disease risk?

Do we want to evaluate temporal trends for each area?

Is the spatial clustering and/or temporal trend a nuisance quantity that we wish to take into account but are not explicitly interested in?

$\rightarrow$ Spatial regression / time series regression.

Why accounting for spatial and temporal dependencies is important

Spatial (temporal) patterns suggest that observations close to each other have more similar values than those far from each other.

Hypothesis generating

Are we explicitly interested in the spatial pattern of disease risk?

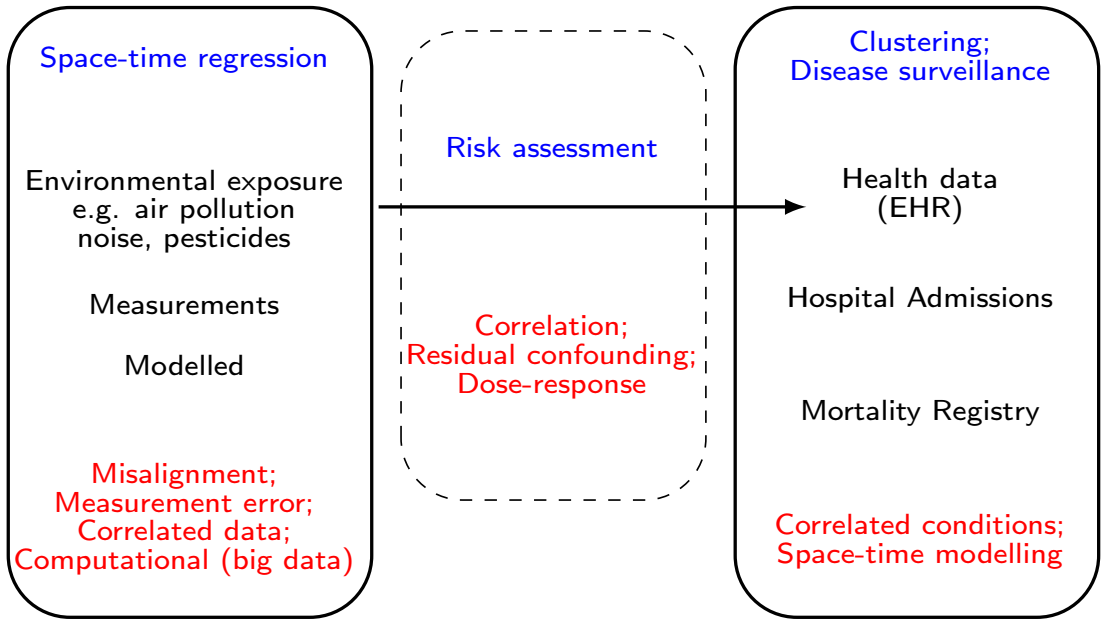
Do we want to evaluate temporal trends for each area?

Risk assessment

Is the spatial clustering and/or temporal trend a nuisance quantity that we wish to take into account but are not explicitly interested in?

$\rightarrow$ Spatial regression / time series regression.

Data/modelling/challenges in environmental epidemiology



Data integration, Uncertainty propagation & feedback; Missing data

In this talk

- Two main areas:
  ⇒ disease surveillance (space-time modelling);
  ⇒ risk assessment (residual confounding and correlated exposures).

Small area modelling framework

When dealing with aggregated data at the small area level

- The easiest model consists in specifying the observed counts of events for area i $(i = 1, \ldots, N)$ as

$$O_i \sim \text{Poisson}(\lambda_i E_i)$$

  - $\text{SMR}_i = \frac{O_i}{E_i}$ is the MLE for λ_i
  - SMR_i very imprecise for rare diseases and/or areas with small populations, λ_i estimated variance is $\frac{O_i}{E_i^2}$
    $\Rightarrow$ Highlights extreme risk estimates based on small numbers.

Small area modelling framework

When dealing with aggregated data at the small area level

- The easiest model consists in specifying the observed counts of events for area i $(i = 1, \ldots, N)$ as

$$O_i \sim \text{Poisson}(\lambda_i E_i)$$

  - $\text{SMR}_i = \frac{O_i}{E_i}$ is the MLE for λ_i
  - SMR_i very imprecise for rare diseases and/or areas with small populations, λ_i estimated variance is $\frac{O_i}{E_i^2}$
    $\Rightarrow$ Highlights extreme risk estimates based on small numbers.

- SMR in each area is estimated independently
  $\rightarrow$ makes no use of risk estimates in other areas of the map, even though these are likely to be similar.

- Ignores possible spatial correlation between disease risk in nearby areas

Small area modelling framework

When dealing with aggregated data at the small area level

- The easiest model consists in specifying the observed counts of events for area i $(i = 1, \ldots, N)$ as

$$O_i \sim \text{Poisson}(\lambda_i E_i)$$

  - $\text{SMR}_i = \frac{O_i}{E_i}$ is the MLE for λ_i
  - SMR_i very imprecise for rare diseases and/or areas with small populations, λ_i estimated variance is $\frac{O_i}{E_i^2}$
    $\Rightarrow$ Highlights extreme risk estimates based on small numbers.

- SMR in each area is estimated independently
  $\rightarrow$ makes no use of risk estimates in other areas of the map, even though these are likely to be similar.

- Ignores possible spatial correlation between disease risk in nearby areas

Bayesian 'smoothing' estimators in a hierarchical formulation.

Hierarchical modelling for small area data

Poisson-logNormal model

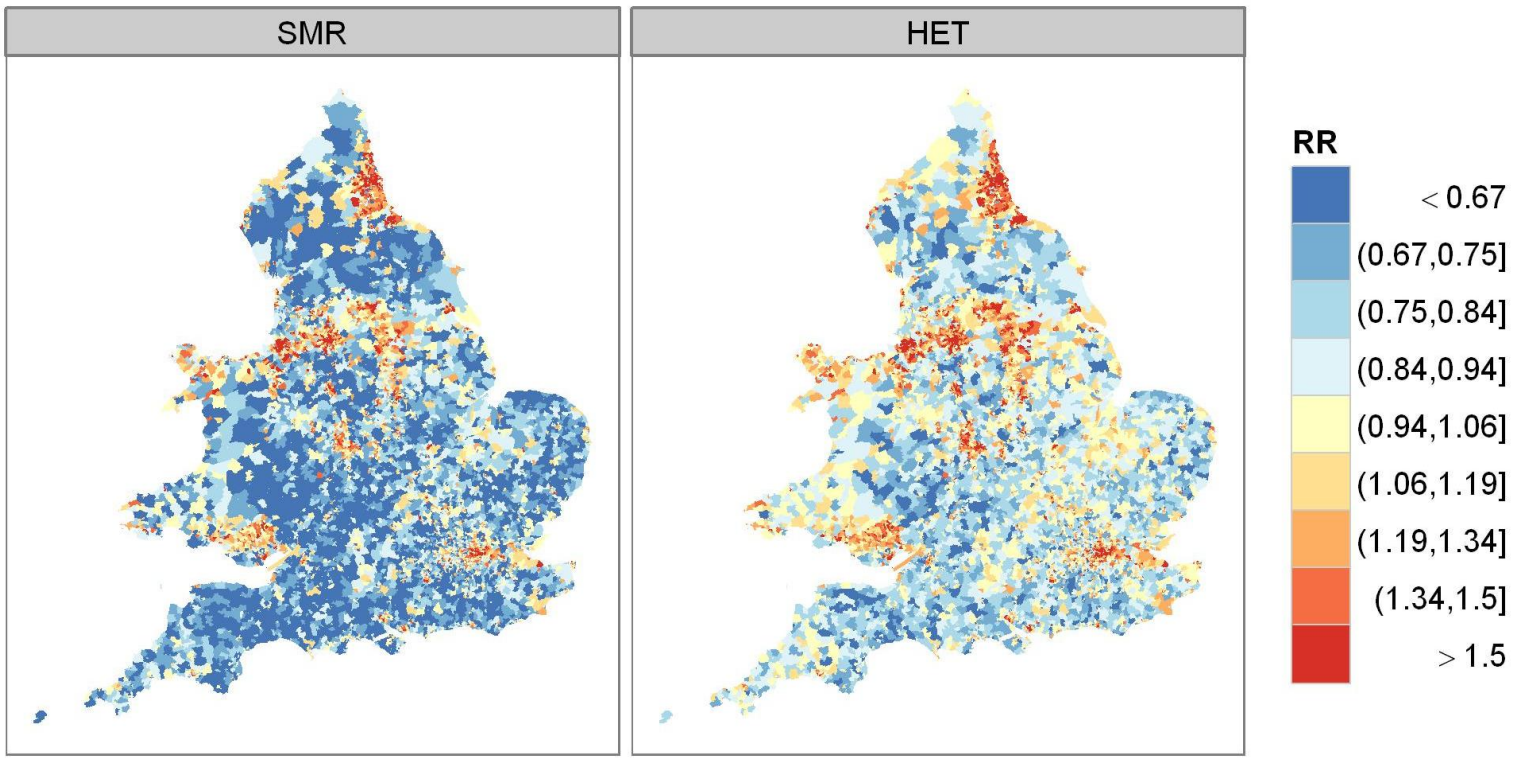
$$\begin{aligned} O_i &\sim \text{Poisson}(\lambda_i E_i) \\ \log \lambda_i &= \alpha + v_i \\ v_i &\sim \text{Normal}(0, \sigma_v^2) \end{aligned}$$

where

- O_i, E_i: observed and expected nb of cases in area i
- $\lambda_i = \exp(\alpha + v_i)$: RR in area i compared with expected risk based on age and sex of population
- Parameters v_i: **area-specific random effects**
- residual RR $= \exp(v_i)$

Lung cancer incidence in males, 1985-2009, England and Wales (I)

RR estimates using 2 methods



SMRs and smoothed RRs

Local spatial dependency

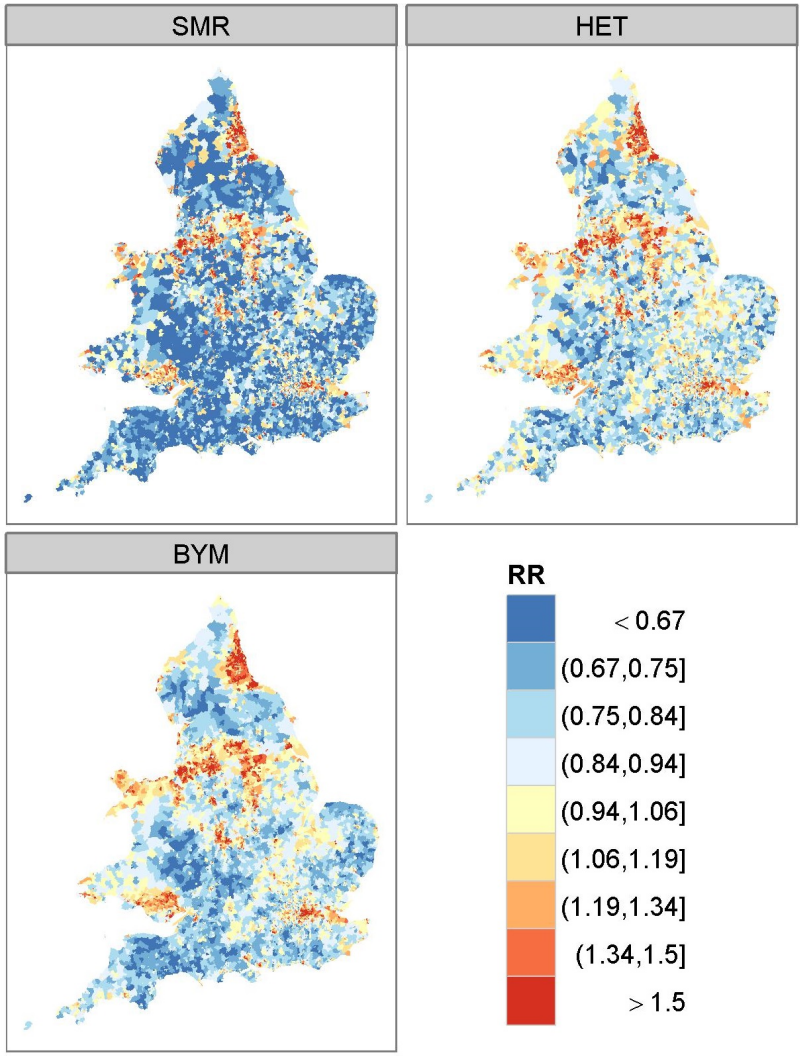
To account for local dependency it is possible to add a spatial structure in the model:

Convolution model

$$
\begin{aligned}
O_i &\sim \text{Poisson}(\lambda_i E_i) \\
\log \lambda_i &= \alpha + v_i + u_i \\
v_i &\sim \text{Normal}(0, \sigma_v^2) \\
u_i \mid u_{-i} &\sim \text{Normal}\left(\frac{\sum_{k=1}^{n} w_{ik} u_k}{\sum_{k=1}^{n} w_{ik}}, \frac{\sigma_u^2}{\sum_{k=1}^{n} w_{ik}} \right)
\end{aligned}
$$

- u_i follows a conditional autoregressive specification (CAR); it assumes that only neighbouring areas contribute to the distribution of area i ($w_{ik} = 1$).
- The combination of v_i and u_i guarantees local and global smoothing (BYM model).

Residual RR of lung cancer incidence in males, 1985-2009, England and Wales II



SMR: non smoothed RR

HET: non spatially smoothed residual RR $\exp(v)$

BYM: spatially and non spatially smoothed residual RR $\exp(v + u)$

From spatial to spatio-temporal modelling

- The hierarchical structure can be extended to incorporate time into a space-time model.
- The stability (or not) of the spatial pattern can aid interpretation.
- The specific space-time components of the model can potentially pinpoint unusual / emerging hazards.

Spatio-temporal hierarchical model

$$\begin{aligned} O_{it} &\sim \text{Poisson}(\lambda_{it} E_{it}) \\ \log \lambda_{it} &= \alpha + v_i + u_i + \phi_t + \psi_{it} \end{aligned}$$

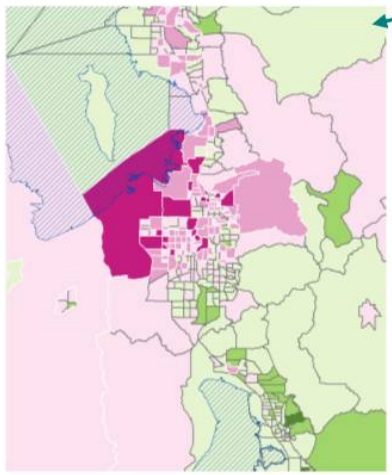
Temporal trend and space-time interaction:

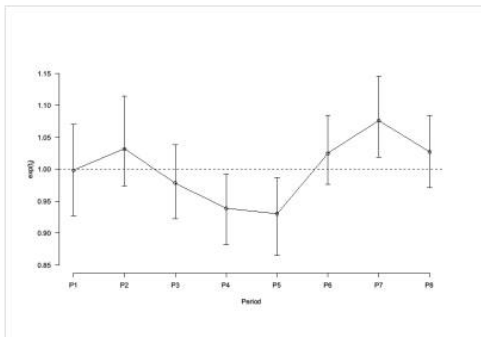
- Temporal trends: $\phi_t \sim \text{Normal}(\phi_{t-1}, \sigma^2_\phi)$
- Space-time interaction: $\psi_{it} \sim \text{Normal}(0, \sigma^2_\psi)$

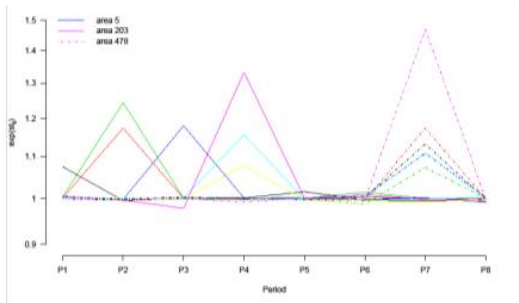
Schematic representation

Noise model: Poisson/Binomial

Latent structure: Space + Time + Interactions







joint Bayesian estimation

Inference

Disease Surveillance

Disease surveillance: why

Surveillance is a key practice within the NHS to ensure that the right information is available at the right time to inform public health decisions and interventions.

It is the ongoing systematic data collection, analysis, interpretation and dissemination of information in order for action to be taken.

Appropriate for a wide range of problems:

- to investigate the epidemiology of a public health issue;
- to assess an impact of an intervention/policy;
- to provide early warning detection.

Disease surveillance: examples

Several areas where surveillance models have been used

1. medical malpractice and failures to deliver appropriate standards of health care (ex. Marshall et al. (2004));
2. outbreaks of communicable diseases (ex. WHO/Europe (2006a));
3. discover unusual trends in non-communicable diseases (ex. Li et al. (2012)).

In these contexts the data are characterised by a spatial and temporal dimension and therefore appropriate surveillance methods that are able to capture spatial and temporal patterns need to be employed.

Disease surveillance for COPD

COPD[1] is one of the most common chronic respiratory diseases (CRDs), that affect the airways and other structures of the lung (WHO, 2012).

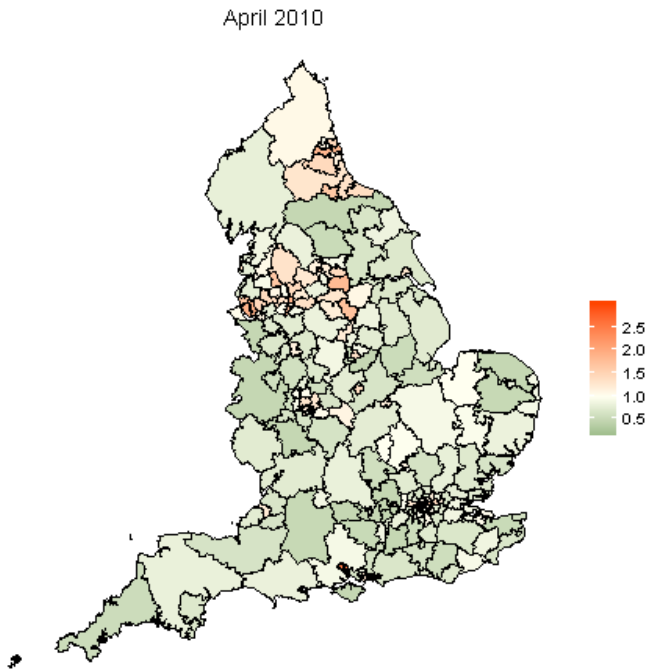
- 64 million people currently suffering from the disease;
- Responsible for 6% of deaths worldwide; projection to become the 3rd leading cause of death by 2030 (WHO, 2014);
- 5th cause of death in UK.

[1] Chronic Obstructive Pulmonary Disease

Maps of Crude Rates

Spatial resolution: 211 Clinical Commissioning Groups (CCGs)
Temporal resolution: monthly data, April 2010 - March 2011.

Figure: Hospital Episode Statistics (HES) SMRs



Spatio-temporal mixture model

- Extend the spatio-temporal hierarchical framework to a mixture model with two components
  - accounts for spatial and temporal correlation;
  - able to detect areas with trend different from the national one (unusual).

$$O_{it} \sim \text{Poisson}(\lambda_{it}\ E_{it})$$

areas $i = 1, \ldots, 211$; time points $t = 1, \ldots, 12$.

$$\log(\lambda_{it}) = p_{it}\ \log\big(\ \lambda_{it}^{\text{C}}\ \big) + (1 - p_{it})\ \log\big(\ \lambda_{it}^{\text{AS}}\ \big)$$

Common Trend

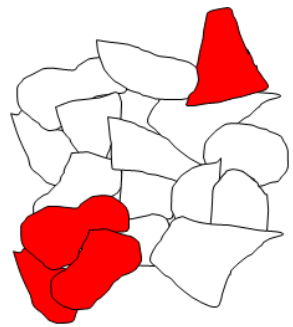
- overall intercept
- spatial component
- temporal component

Area-Specific Trend

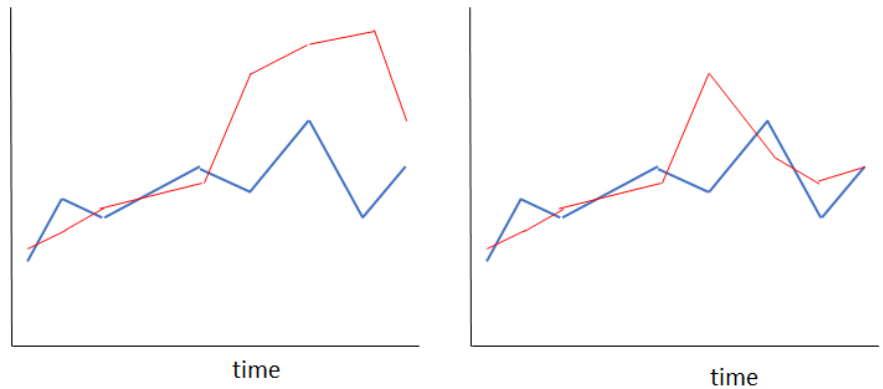
- area-specific intercept
- area-specific temporal component

Modelling the probability for the two components

- Each area is characterised by a (vector of) probability $\boldsymbol{p}_i = (p_{i1}, \ldots, p_{iT})$ of being assigned to the common trend component
- p_{it} can be modelled as a combination of several effect:
  - local and global dependency in space (clusters or isolated areas)

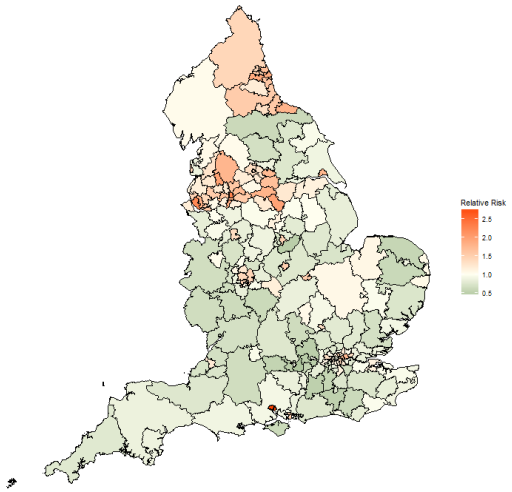


  - local and global dependency in time (similar trend or isolated time points)

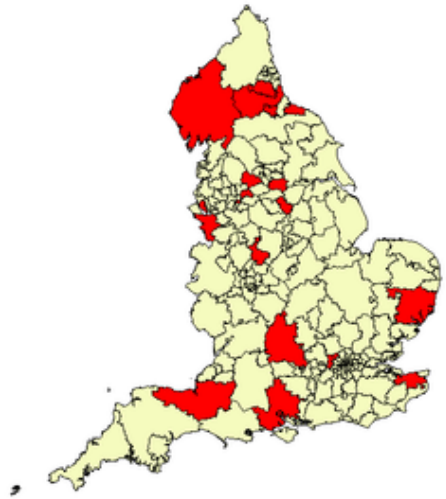


Results: Spatial & Temporal Patterns

Spatial Posterior Mean



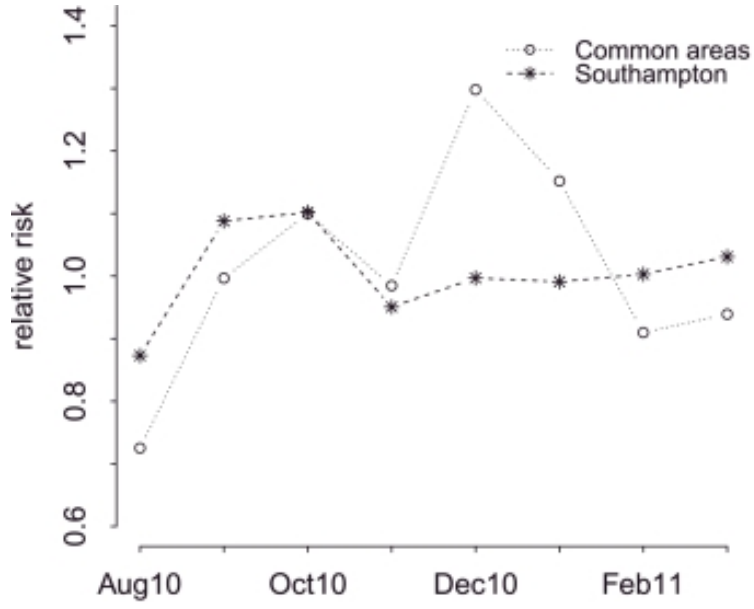
Unusual Areas

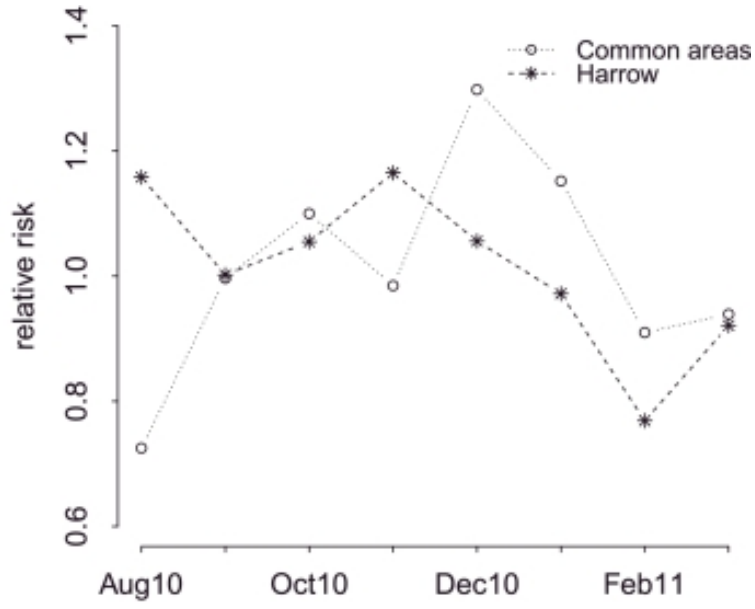


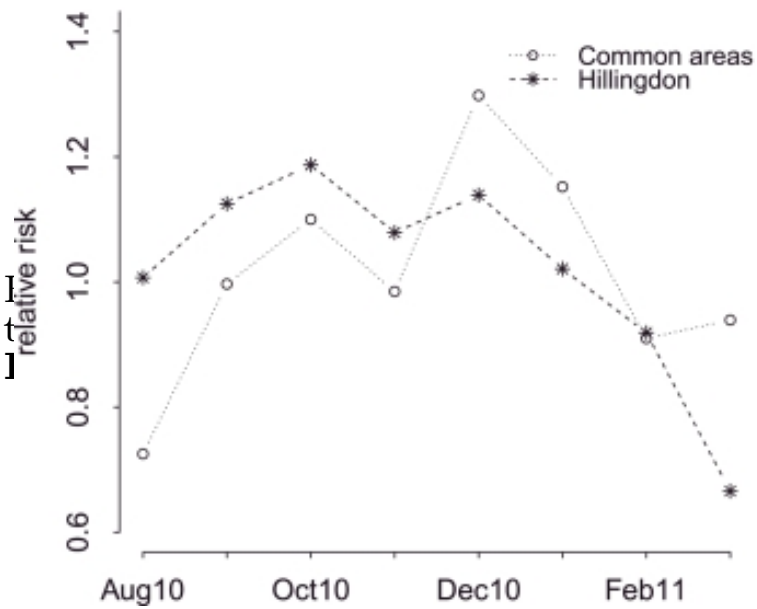
- Higher risk in the north of England across the time period
- Unusual areas: Mostly isolated but some clustered in the North of England

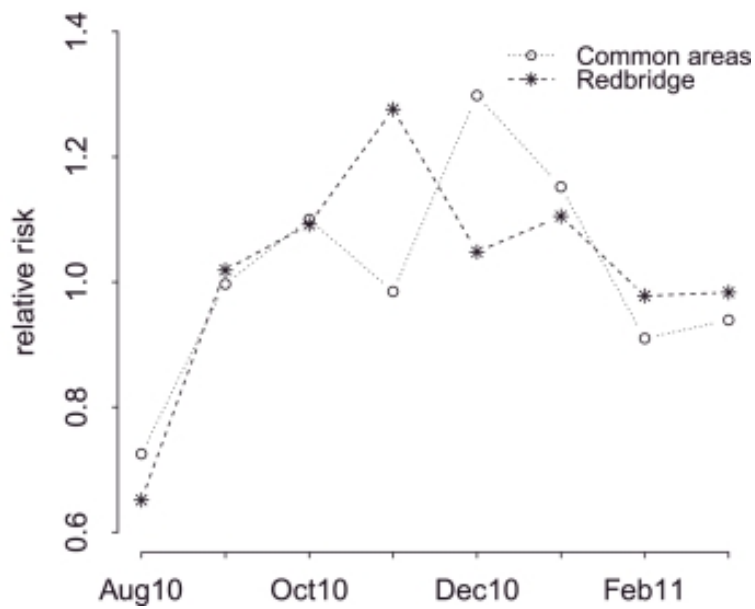
Boulieri A, Hansell A, Blangiardo M, "Investigating trends in asthma and COPD through multiple data sources: A small area study", **Spatial and SpatioTemporal Epidemiology**, 2016, 19: 28-36.

Results: Spatial & Temporal Patterns









Next Steps

- Extend this framework to model several health outcomes.
- Create a user-friendly tool for people to disseminate this method.
- Similar framework to be used to investigate trends in life expectancy (PHE funded).

Boulieri A, Bennett J, Blangiardo M, "A Bayesian mixture modelling approach for disease surveillance", submitted to **Biostatistics**

Risk Assessment

Health effects of polluted air: some estimates

- IARC (2013): Outdoor air pollution is carcinogenic to humans;
- WHO/Europe (2006b): Outdoor PM causes a reduction in life expectancy of average population by approximately a year in Europe;
- WHO (2016): Outdoor $PM_{2.5}$ causes more than 3 million deaths per year worldwide; 92% of the world's population lives in places where air quality levels exceed WHO limits.

Long-term effects vs Short-term

- Long-term considers averages of air pollution concentration and counts of outcomes (at individual level or) over small areas:
  - typically air pollution data from exposure models (deterministic or more recently statistical);
  - important to account for area level confounders (e.g. population and area characteristics);
  - statistical approach is usually ecological regression.

Long-term effects vs Short-term

- Long-term considers averages of air pollution concentration and counts of outcomes (at individual level or) over small areas:
  - typically air pollution data from exposure models (deterministic or more recently statistical);
  - important to account for area level confounders (e.g. population and area characteristics);
  - statistical approach is usually ecological regression.
- Short-term considers the temporal variation in air pollution and evaluate the association with the temporal variation in health:
  - typically air pollution data from monitoring stations (one or more);
  - important to account for seasonality and meteorological variables;
  - statistical modelling usually in a time-series framework.

Risk Assessment: residual confounding

Air pollution and health in London

- CHD[2] admissions (Hospital Episode Statistics - HES), ICD codes I20 to I25.
- Analysis conducted at the middle super output area in London (MSOA - approx 6,500 individuals).
- Concentration of PM_{10} modelled through a land use regression (X).
- Area level confounders: ethnicity and social deprivation ($\boldsymbol{C}$).
- Log linear model with spatially structured/unstructured random effects.

$$\begin{aligned} O_i &\sim \text{Poisson}(\lambda_i \text{E}_i) \\ \log(\lambda_i) &= \alpha + \mathbf{X}_i^\top \boldsymbol{\beta_X} + \boldsymbol{C}_i^\top \beta_C + u_i + v_i \end{aligned}$$

[2] Coronary Heart Disease

Air pollution and CHD hospital admissions in London: Results

Data for 2001, PM_{10} in quintiles

PM_{10}	Posterior Mean	95%CI
<24.2	1.00	ref
24.2-25.3	1.03	0.96-1.10
25.3-26.2	0.94	0.88-1.01
26.2-27.5	0.88	0.81-0.94
>27.5	0.75	0.70-0.81

Several models investigated (spatial only, random effect only, spatial + random effect).

The negative association between air pollution and CHD hospital admissions points towards residual confounding.

Using external information to deal with residual confounding

The integration of data sources (registries, cohorts/surveys) has been investigated to deal with confounding issues at the individual level:

Using external information to deal with residual confounding

The integration of data sources (registries, cohorts/surveys) has been investigated to deal with confounding issues at the individual level:

- Jackson et al. (2009), Molitor et al. (2009) used Bayesian graphical models to build sub-models for each source of data that are then linked together in a coherent global analysis.
  $\rightsquigarrow$ Multiple imputation in a Bayesian framework, computationally intensive, works on a limited set of confounders.

Using external information to deal with residual confounding

The integration of data sources (registries, cohorts/surveys) has been investigated to deal with confounding issues at the individual level:

- Jackson et al. (2009), Molitor et al. (2009) used Bayesian graphical models to build sub-models for each source of data that are then linked together in a coherent global analysis.
  $\rightsquigarrow$ Multiple imputation in a Bayesian framework, computationally intensive, works on a limited set of confounders.
- McCandless et al. (2012) proposed to use the propensity score to link data sources at the individual level - application on water chlorination and birth weight using Millennium Cohort Study and Hospital Episode Statistics.

Using external information to deal with residual confounding

The integration of data sources (registries, cohorts/surveys) has been investigated to deal with confounding issues at the individual level:

- Jackson et al. (2009), Molitor et al. (2009) used Bayesian graphical models to build sub-models for each source of data that are then linked together in a coherent global analysis.
  $\rightsquigarrow$ Multiple imputation in a Bayesian framework, computationally intensive, works on a limited set of confounders.
- McCandless et al. (2012) proposed to use the propensity score to link data sources at the individual level - application on water chlorination and birth weight using Millennium Cohort Study and Hospital Episode Statistics.

We develop a framework based on propensity score like indices for ecological (small area) studies.

Notation

- Data measured only at *geographical areas*, $i = 1, \ldots, N$
  - Outcome in analysis O.
  - Exposure of interest X;
  - Set of Q covariates $\boldsymbol{C}$;

Notation

- Data measured only at *geographical areas*, $i = 1, \ldots, N$
  - Outcome in analysis O.
  - Exposure of interest X;
  - Set of Q covariates $\boldsymbol{C}$;
- Data unmeasured at area-level, $i = 1, \ldots, N$
  - Set of K missing confounders $\boldsymbol{M}$.

Notation

- Data measured only at *geographical areas*, $i = 1, \ldots, N$
  - Outcome in analysis O.
  - Exposure of interest X;
  - Set of Q covariates $\boldsymbol{C}$;
- Data unmeasured at area-level, $i = 1, \ldots, N$
  - Set of K missing confounders $\boldsymbol{M}$.
- Data measured at individual-level within several geographical areas $i = 1, \ldots, S \in N$ (e.g. areas covered by a survey)
  - Set of K confounders $\boldsymbol{m}$, on $j = 1, \ldots, J$ subjects.

Notation

- Data measured only at *geographical areas*, $i = 1, \ldots, N$
  - Outcome in analysis O.
  - Exposure of interest X;
  - Set of Q covariates $\boldsymbol{C}$;
- Data unmeasured at area-level, $i = 1, \ldots, N$
  - Set of K missing confounders $\boldsymbol{M}$.
- Data measured at individual-level within several geographical areas $i = 1, \ldots, S \in N$ (e.g. areas covered by a survey)
  - Set of K confounders $\boldsymbol{m}$, on $j = 1, \ldots, J$ subjects.

Idea

- Use a hierarchical model that up-scales spatially-referenced individual data, $\boldsymbol{m}$, at ecological level, to provide latent (unmeasured) confounding factors, $\boldsymbol{M}$, for inclusion in the ecological health-effect model.
- However, how to deal with areas where individual-level data are missing?

Why is propensity score useful in areal-referenced studies?

$$O_i \sim \text{Poisson}(E_i \lambda_i)$$

Areas with survey data

$$\log(\lambda_i) = \alpha + X_i^\top \beta_X + \boldsymbol{C}_i^\top \beta_C + \boldsymbol{M}_i^\top \beta_M + u_i + v_i$$

Areas without survey data

$$\log(\lambda_i) = \alpha + X_i^\top \beta_X + \boldsymbol{C}_i^\top \beta_C + ? + u_i + v_i$$

- Without imputing the missing data only a partial spatial analysis can be carried out.
  $\rightsquigarrow$ Not useful in a public health/surveillance perspective.

- Imputing several confounders raises methodological challenges and is computationally intensive.
  $\rightsquigarrow$ Propensity Score to provide dimension reduction.

Proposed approach

(1) Up-scaling the $\boldsymbol{m}$ to $\boldsymbol{M}$ ($i \in S$):

$$\begin{aligned} m_{ikj} &\sim \text{Bern}(M_{ik}) \\ g_k(M_{ik}) &= \eta_k + \xi_{ik} \qquad k = 1, \ldots, K; \end{aligned}$$

- ξ_{ik} are multivariate random effects;
- η_k is confounder-specific fixed intercept and has a vague prior.

Proposed approach

(1) Up-scaling the $\boldsymbol{m}$ to $\boldsymbol{M}$ ($i \in S$):

$$\begin{aligned} m_{ikj} &\sim \text{Bern}(M_{ik}) \\ g_k(M_{ik}) &= \eta_k + \xi_{ik} \qquad k = 1, \ldots, K; \end{aligned}$$

- ξ_{ik} are multivariate random effects;
- η_k is confounder-specific fixed intercept and has a vague prior.

(2) Ecological PS estimation (EPS) for non-missing areas ($i \in S$).

- For a binary exposure, we follow McCandless et al. (2012), that proposed the estimation of a partial PS for modelling missing confounders:

$$\text{logit}(\Pr(X_i = 1 | \boldsymbol{C}_i, \boldsymbol{M}_i)) = \delta_0 + \sum_q \delta_q C_{qi} + \underbrace{\sum_k \delta_k M_{ik}}_{\text{EPS}}$$

- M_{ik} comes from the multilevel model.
- Only one quantity has to be imputed where missing at the small area level.

Proposed approach (cont.)

(3) Imputation model for missing areas:

$$\text{EPS}_i \sim \text{Normal}(\eta + f(\boldsymbol{C}_i) + \zeta_i, \sigma^2_{EPS})$$

- ζ_i incorporates the spatial structure in the model.

Proposed approach (cont.)

(3) Imputation model for missing areas:

$$\text{EPS}_i \sim \text{Normal}(\eta + f(\boldsymbol{C}_i) + \zeta_i, \sigma^2_{EPS})$$

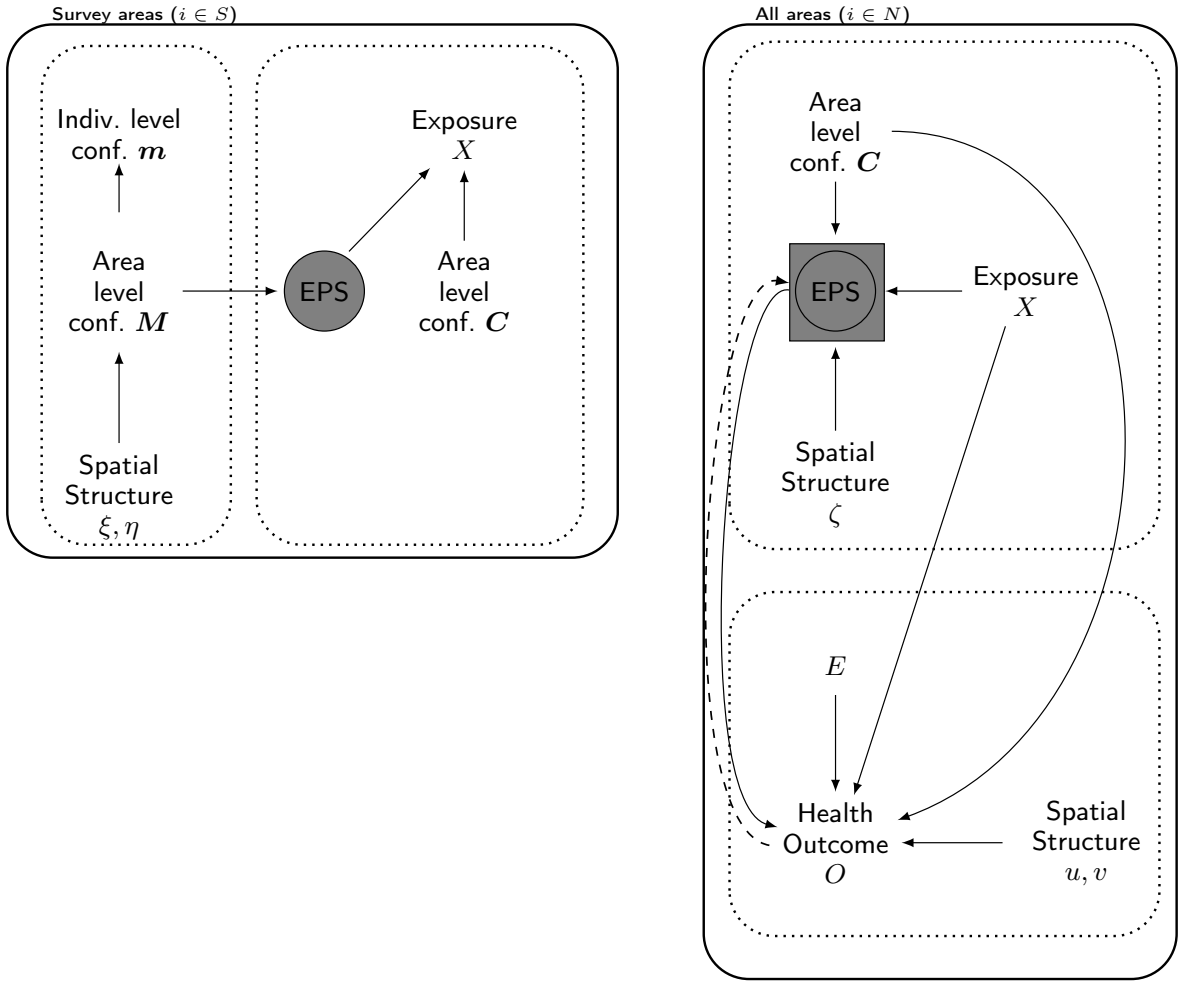
- ζ_i incorporates the spatial structure in the model.

(4) Health-effect model adjustment:

$$O_i \sim \text{Poisson}(\text{E}_i\lambda_i), \quad i = 1, \ldots, N$$
$$\log(\lambda_i) = \beta_0 + X_i^\top \boldsymbol{\beta}_X + \mathbf{C}_i^\top \boldsymbol{\beta}_C + h(\text{EPS}_i) + u_i + v_i$$

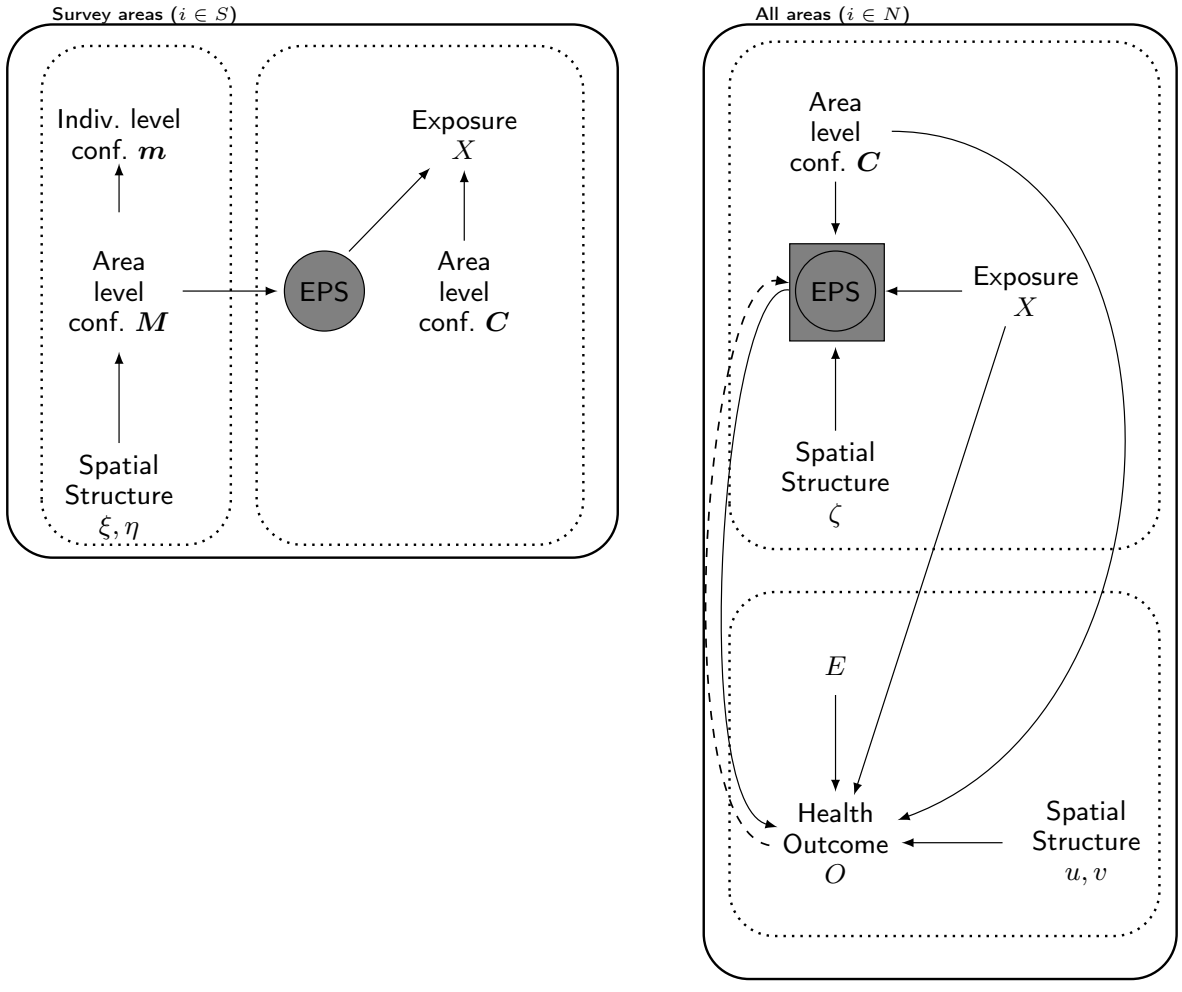
- Data-driven function $h()$ to link EPS to the outcome-exposure analysis in order to minimise the model specification bias.

To sum up


Survey areas ($i \in S$)
Indiv. level conf. $\boldsymbol{m}$
Area level conf. $\boldsymbol{M}$
Spatial Structure ξ, η
EPS
Exposure X
Area level conf. $\boldsymbol{C}$
All areas ($i \in N$)
Area level conf. $\boldsymbol{C}$
EPS
Exposure X
Spatial Structure ζ
E
Health Outcome O
Spatial Structure u, v

To sum up

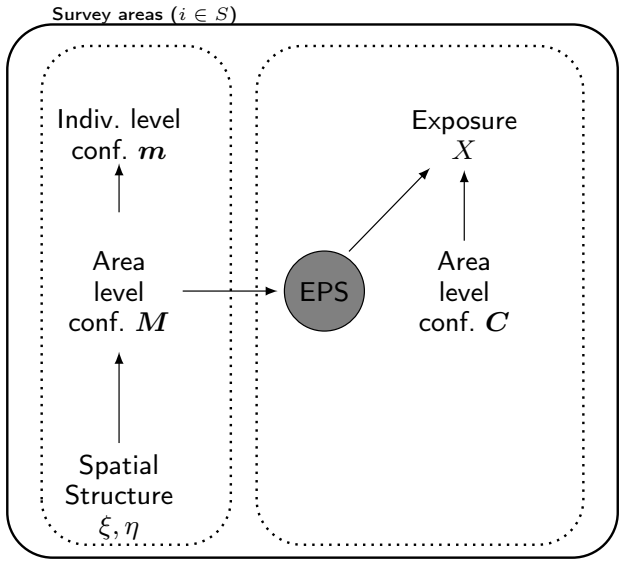
Uncertainty on the parameters is propagated throughout the model

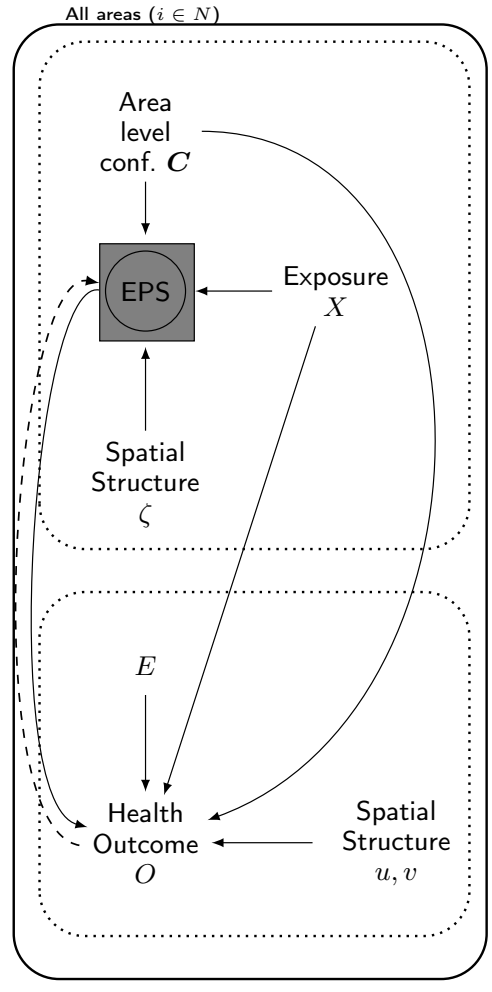


To sum up

Uncertainty on the parameters is propagated throughout the model

→From the estimates of $\mathbf{M}$ into the EPS estimation

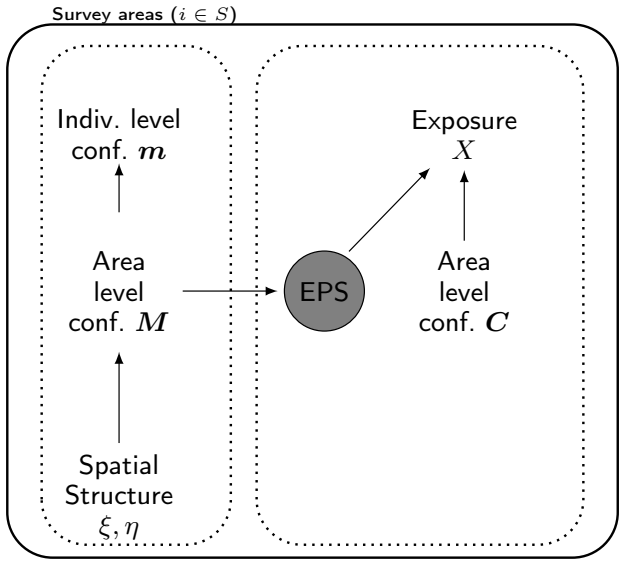


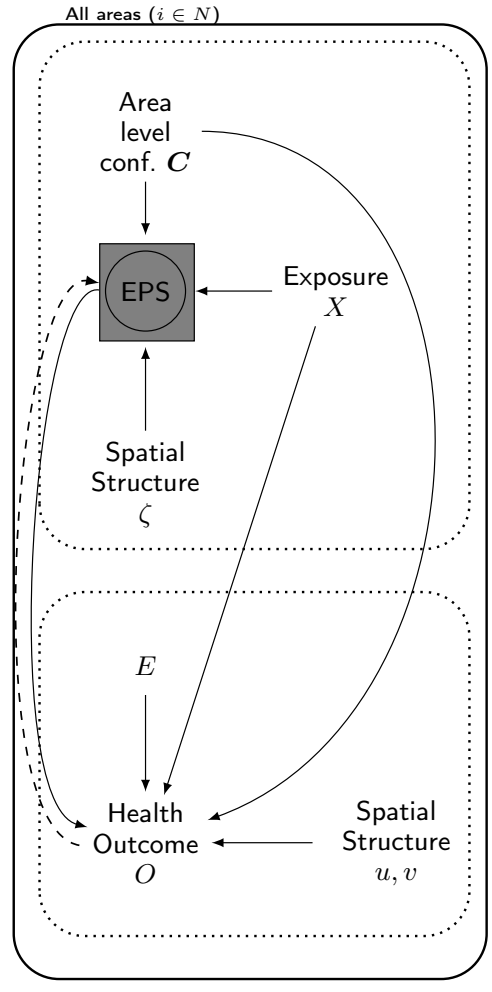


To sum up

Uncertainty on the parameters is propagated throughout the model

→From the estimates of $\mathbf{M}$ into the EPS estimation



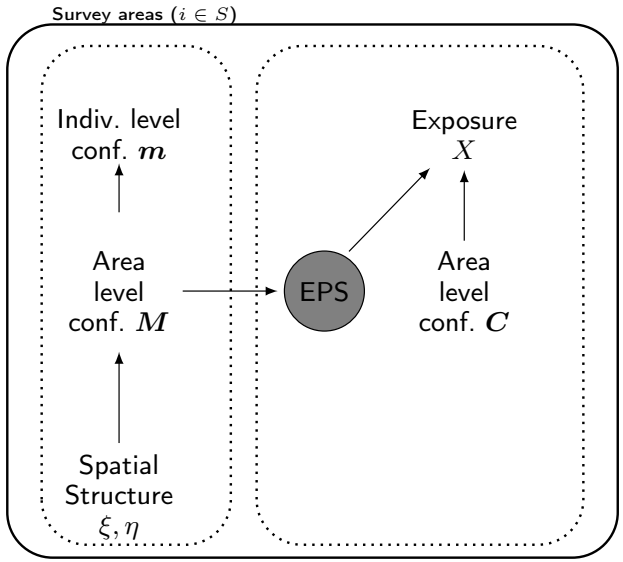


→ From the EPS imputation into the analysis model

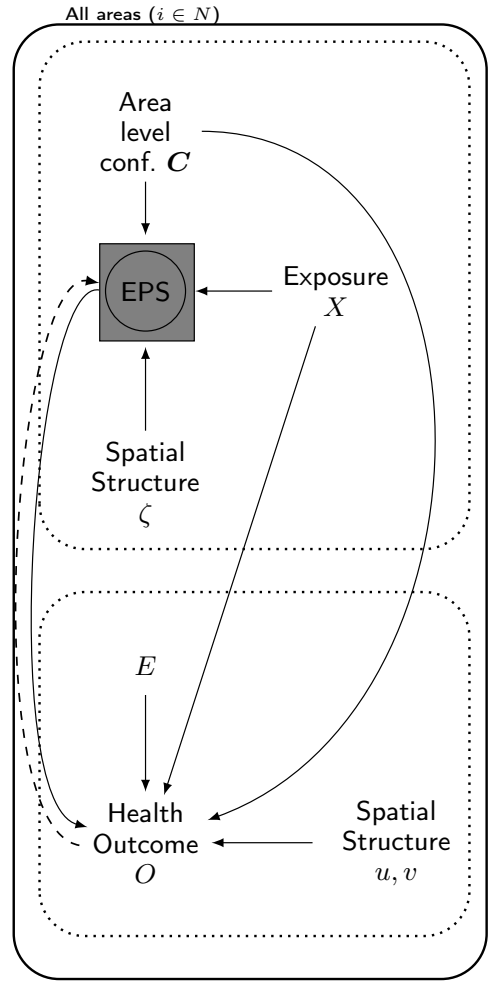
To sum up

Uncertainty on the parameters is propagated throughout the model

→From the estimates of $\mathbf{M}$ into the EPS estimation



Feedback from X to M is prevented

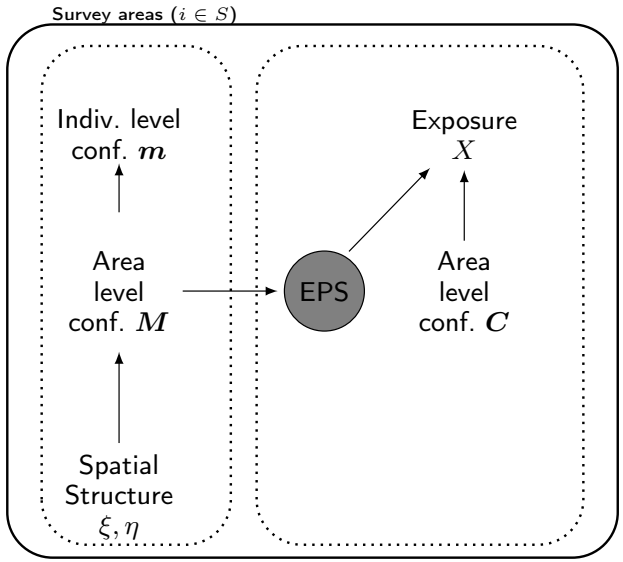


→ From the EPS imputation into the analysis model

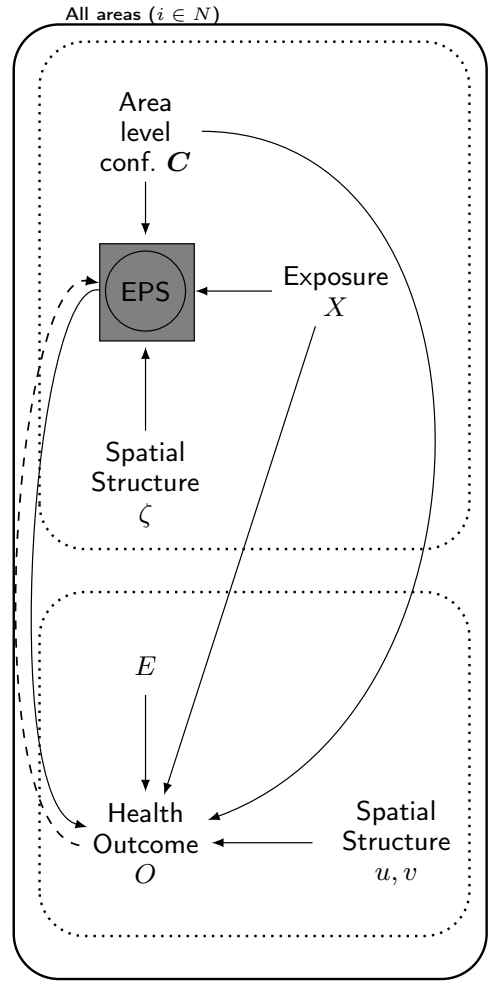
To sum up

Uncertainty on the parameters is propagated throughout the model

→From the estimates of $\mathbf{M}$ into the EPS estimation



Feedback from X to M is prevented



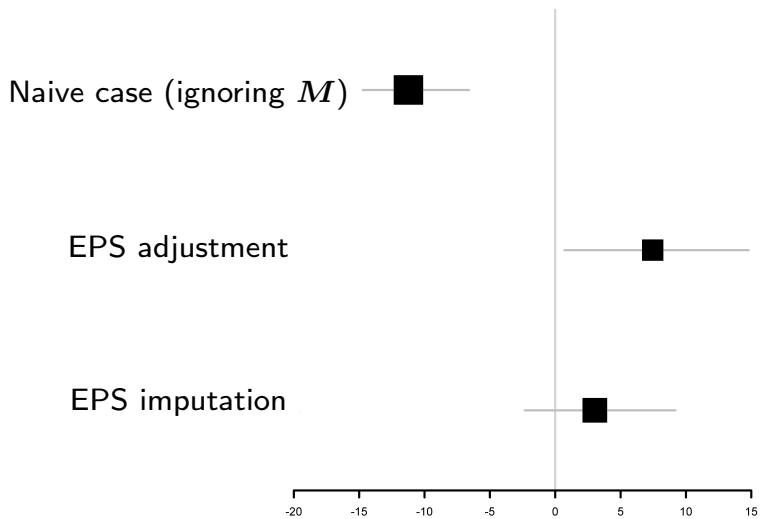
→ From the EPS imputation into the analysis model

Feedback from O to EPS is allowed

London air pollution and CHD hospitalisations: adjusting for EPS

Association between CHD and exposure to PM_{10} dichotomised using a cut-off of $25\mu g/m^3$.

Individual level data obtained from the Health Survey of England 1994-2001. 13 potential confounders were considered covering smoking, drinking, BMI, physical activity, diet.

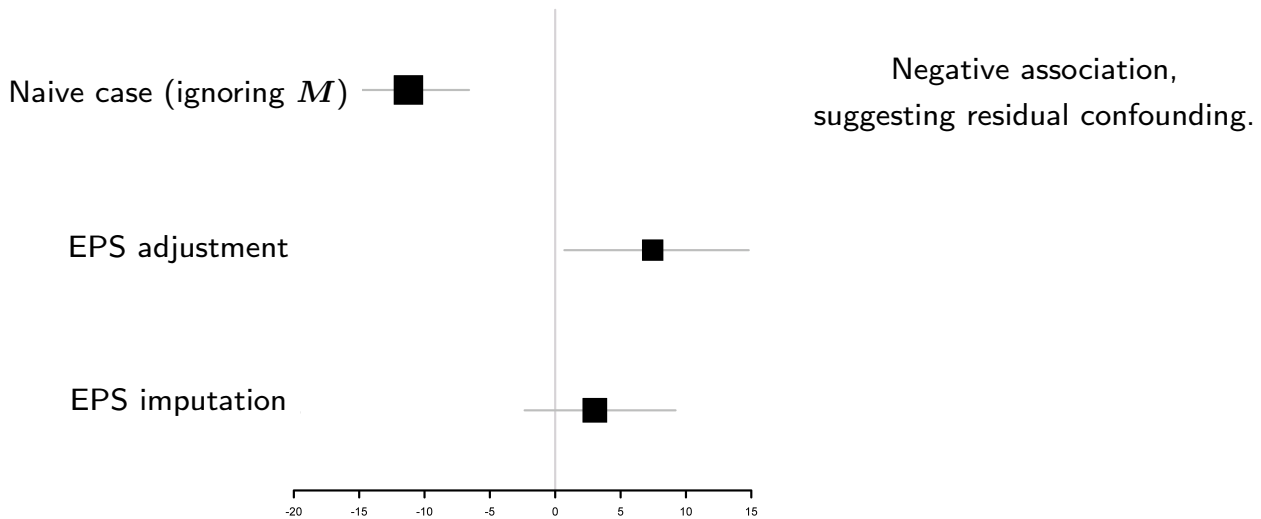


Wang Y, et al. (2017), Using ecological propensity score to adjust for missing confounders in small area studies; **Biostatistics** (accepted).

London air pollution and CHD hospitalisations: adjusting for EPS

Association between CHD and exposure to PM_{10} dichotomised using a cut-off of $25\mu g/m^3$.

Individual level data obtained from the Health Survey of England 1994-2001. 13 potential confounders were considered covering smoking, drinking, BMI, physical activity, diet.

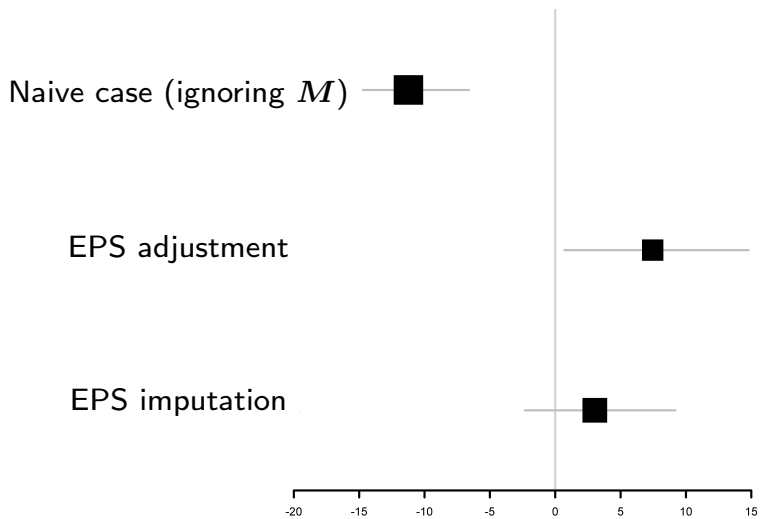


Wang Y, et al. (2017), Using ecological propensity score to adjust for missing confounders in small area studies; **Biostatistics** (accepted).

London air pollution and CHD hospitalisations: adjusting for EPS

Association between CHD and exposure to PM_{10} dichotomised using a cut-off of $25\mu g/m^3$.

Individual level data obtained from the Health Survey of England 1994-2001. 13 potential confounders were considered covering smoking, drinking, BMI, physical activity, diet.



Negative association,
suggesting residual confounding.

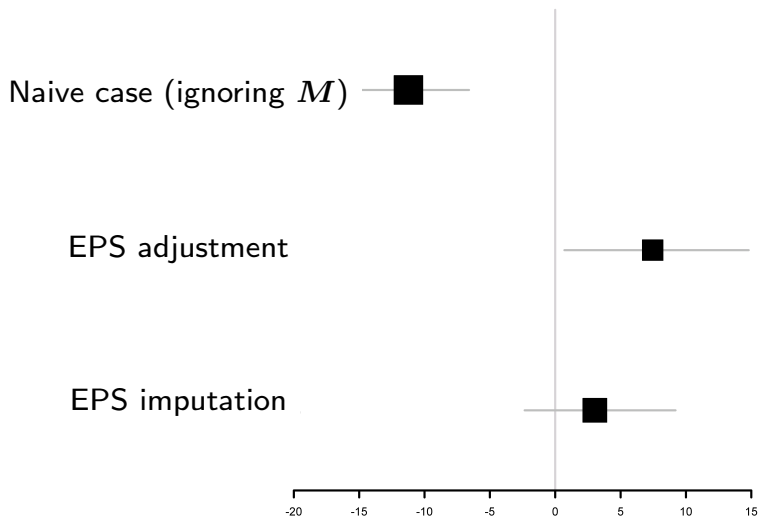
Positive association,
posterior probability 98.4%.

Wang Y, et al. (2017), Using ecological propensity score to adjust for missing confounders in small area studies; **Biostatistics** (accepted).

London air pollution and CHD hospitalisations: adjusting for EPS

Association between CHD and exposure to PM_{10} dichotomised using a cut-off of $25\mu g/m^3$.

Individual level data obtained from the Health Survey of England 1994-2001. 13 potential confounders were considered covering smoking, drinking, BMI, physical activity, diet.



Negative association, suggesting residual confounding.

Positive association, posterior probability 98.4%.

Positive association more noise due to imputation, posterior probability 85.3%.

Wang Y, et al. (2017), Using ecological propensity score to adjust for missing confounders in small area studies; **Biostatistics** (accepted).

Next steps

- Extension of the framework for a multi-categorical or continuous exposure.
- Model to account for nonlinearity between the individual level confounders and the outcome.
- From spatial to spatio-temporal dimension.

Risk Assessment: correlated exposures

From a single to a multi-pollutant approach

- The quantification of the impact of air pollution on population health has been historically undertaken through a **single pollutant approach**.
- This is mainly due to:
  - regulatory strategies of air quality management which have addressed a single pollutant at a time.

From a single to a multi-pollutant approach

- The quantification of the impact of air pollution on population health has been historically undertaken through a **single pollutant approach**.
- This is mainly due to:
  - regulatory strategies of air quality management which have addressed a single pollutant at a time.

However, the air we breathe is a mixture and:

- It is unlikely that all parts of the air pollution mix are equally harmful;
- It is clear that the effect estimates can be affected by correlation, measurement error and exposure misclassification (locally varying vs regional pollutants).

Therefore, we need new/revised statistical methods and approaches for a multi-pollutant approach (e.g. Coull and Park 2015 and Molitor et al. 2016).

Multi-pollutant approach so far

- Air quality indexes (Daily Air Quality Index in the UK)
  → typically used by governments;
  → easy to build and to communicate to the public.

- Bayesian Kernel Machine Regression (Bobb et al. 2015)
  → the pollutants are included in the model through a smooth function represented using a kernel;
  → authors found that Gaussian kernel outperformed linear and ridge regression kernels.

- Dirichlet process mixture model (profile regression, Pirani et al. 2015)
  → days are clustered based on their concentration profiles;
  → concentration and health outcome are modelled jointly.

Another way forward

We propose the use of a hierarchical Bayesian time-series approach which is formed by two linked components:

- A **pollutant component** which estimates the true 'latent' concentration values;
- A **health component** which links the estimated concentration to the health outcome;
- The two components are jointly modelled;
- The modelling framework allows to estimate a health effect for each pollutant.

Data description

- Daily measurements of five regulated pollutants and the number of particles present in any given volume of air (PCN) available from a monitoring site in North Kensington for 2011-2012.
- Daily count of mortality for cardio-vascular disease (ICD-10, Chapter I) available for the same period.

	Number of Days	Percentiles 10th	25th	50th	75th	90th	IQR
Mortality	731	28	32	37	42	47	10
Meteorological data:							
Temperature ($°C$)	731	5.1	8.0	11.7	15.5	18.1	7.4
Relative Humidity (%)	731	61.6	69.6	78.0	84.2	88.5	14.5
Pollutants:							
CO (mg/m^3)	715	0.1	0.2	0.2	0.3	0.4	0.1
NO_2 ($\mu g/m^3$)	706	18.2	23.2	33.3	46.9	57.9	23.6
O_3 ($\mu g/m^3$)	695	11.4	24.3	39.1	51.1	64.9	26.8
SO_2 ($\mu g/m^3$)	717	0.0	0.4	1.8	2.6	3.6	2.2
$PM_{2.5}$ ($\mu g/m^3$)	730	5.0	6.0	9.0	14.0	25.0	8.0
PCN (p/mm^3)	636	7.8	9.7	12.1	14.9	17.9	5.2

H2M: POLLUTANT COMPONENT

- We specify X_{pt} as the measured (standardised) concentration level of pollutant p ($p = 1, ..., P = 6$) on day t ($t = 1, ..., T = 731$) from the monitoring site:

$$\begin{aligned} X_{pt} &\sim N(\mu_{pt}, \sigma_p^2) \\ \mu_{pt} &= \gamma_{0p} + \sum_j \gamma_{jp} z_{jt} + \theta_{pt} \end{aligned}$$

True (latent) concentration model

- $\mathbf{z}_t$ - meteorological variables:
  $\Rightarrow$ Evidence of non-linear relationship with the pollutant levels - inclusion of linear and quadratic terms.

H2M: POLLUTANT COMPONENT - θ

- $\{\theta_{1t}, \ldots, \theta_{Pt}\}$ accounts for the residual temporal effects and for the correlation among pollutants

$$(\theta_{1t}, ..., \theta_{Pt})^\top \sim \text{MVN}\left((\theta_{1,t-\ell}, ..., \theta_{P,t-\ell})^\top, \Sigma_P\right)$$

- $t - \ell$ provides the temporal lag of ℓ days for the t-th day;
- The diagonal of the covariance matrix of the errors Σ_P allows each pollutant to have a different amount of temporal dependence;
- The off-diagonals represent the temporal dependence between the pollutants.

H2M: HEALTH COMPONENT

- The second model component links the **true** concentration μ_{pt} to the health outcome

$$\begin{aligned} O_t &\sim \text{Poisson}(\lambda_t E) \\ \log(\lambda_t) &= \beta_0 + \sum_p \beta_p \mu_{p(t-1)} + \sum_i s(v_{ti}, \psi_i) + \delta_{I_t} + \epsilon_t \end{aligned}$$

- λ_t represents the relative risk of CVD death on day t compared to the average;
- we consider lag $\ell = 1$ (same as Atkinson et al. 2016);
- $\boldsymbol{\beta}$ are the pollutant effects on CVD mortality;
- ϵ_t is an overdispersion parameter;
- $s(v_{ti}, \psi_i)$ is modelled through a low-rank thin plate spline:

$$s(v_{ti}, \psi_i) = \alpha_i v_{ti} + \sum_{k=1}^{K_i} b_{ki} |v_{ti} - \kappa_{ki}|^3$$

Application Results: single vs multi-pollutant model

- Five pollutants (CO, NO_2, O_3, SO_2, $PM_{2.5}$) and particle number concentration (PCN);
- NO_2 and O_3 are the only pollutants to show evidence of an effect on health;
- Consistent with the literature (ex. Williams et al. 2014).

Pollutant		**IQR**	Multi-pollutant model **% Increase (95% CI)**	
CO	(mg/m^3)	0.10	-1.67	(-4.72, 1.65)
NO_2	$(\mu g/m^3)$	23.65	**9.40**	**(3.06, 16.03)**
O_3	$(\mu g/m^3)$	26.85	**3.46**	**(0.18, 6.71)**
SO_2	$(\mu g/m^3)$	2.20	-1.94	(-6.59, 2.80)
PCNT	(p/mm^3)	5.18	-2.89	(-6.36, 1.05)
$PM_{2.5}$	$(\mu g/m^3)$	8.00	-1.24	(-3.45, 0.92)

Blangiardo M, et al., A hierarchical modelling approach to assess multi pollutant effects in time-series studies; **Statistics in Medicine** (under review).

Next steps

- Uncertainty from the pollutant concentration estimates included in estimating the health effects and feedback from outcome is allowed;
  $\Rightarrow$ From the simulation study it seems this helps reduce the bias in the estimates.

- Linear concentration-response - move towards non-linearity?

- Consider 'total' pollutant concentration;
  $\Rightarrow$ Recent trends move from multi-pollutants to multi-components via *source apportionment*.

Conclusions & Future directions

- Need for statistical modelling in environmental epidemiology to deal with emerging challenges
  - Spatial dependency (and temporal) for disease surveillance;
  - Data integration to account for residual confounding;
  - Correlated exposures;
  - Uncertainty propagation.
- New directions
  - HES Opt out - what is the impact for surveillance?
  - Non parametric methods (surveillance for multi-conditions and for correlated exposures);
  - Causality (policy effects - natural experiments).

Acknowledgements

- My team
  - Areti Boulieri
  - Lauren Kanapka
  - Monica Pirani
  - Anna Freni Sterrantino

- Collaborators
  - Gary Fuller (KCL)
  - Sylvia Richardson (MRC-BSU)
  - Alexina Mason (LSHTM)
  - James Bennett (IC)
  - Anna Hansell (IC)

References

Atkinson, R., Analitis, A., Samoli, E., Fuller, G., Green, D., Mudway, I., and Anderson, H. e. a. (2016), "Short-term exposure to traffic-related air pollution and daily mortality in London, UK," *Journal of Exposure Science and Environmental Epidemiology*, 26, 125–132.

Bobb, J., Valeri, L., Henn, B., Christiani, D., Wright, R., Mazumdar, M., Godleski, J., and Coull, B. (2015), "Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures," *Biostatistics*, 16, 493–508.

Coull, B. A. and Park, E. S. (2015), *Development of Statistical Methods for Multipollutant Research*, Boston, MA: Health Effects Institute, Research Report 183 - I & II.

IARC (2013), *IARC monographs on the evaluation of carcinogenic risks to humans. Volume 109. Outdoor air pollution*, Lyon, France: International Agency for Research on Cancer.

Jackson, C. H., Best, N. G., and Richardson, S. (2009), "Bayesian graphical models for regression on multiple data sets with different variables," *Biostatistics*, 10, 335–51.

Li, G., Best, N., Hansell, A., Ahmed, I., and Richardson, S. (2012), "BaySTDetect: detecting unusual temporal patterns in small area data via Bayesian model choice," *Biostatistics*, 13, 695–710.

Marshall, C.and Best, N., Bottle, A., and Aylin, P. (2004), "Statistical Issues in the Prospective Monitoring of Health Outcomes across Multiple Units," *JRSSA*, 167, 541–559.

McCandless, L. C., Richardson, S., and Best, N. (2012), "Adjustment for missing confounders using external validation data and propensity scores," *Journal of the American Statistical Association*, 4, 40–51.

Molitor, J., Coker, E., Jerrett, M., Ritz, B., and Li, A. (2016), *Modeling of Multipollutant Profiles and Spatially Varying Health Effects with Applications to Indicators of Adverse Birth Outcomes*, Boston, MA: Health Effects Institute, Research Report 183 - III.

Molitor, N.-T., Best, N., Jackson, C., and Richardson, S. (2009), "Using Bayesian graphical models to model biases in observational studies and to combine multiple sources of data: application to low birth weight and water disinfection by-products," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172, 615–637.

Pirani, M., Best, N., Blangiardo, M., Liverani, S., Atkinson, R. W., and Fuller, G. W. (2015), "Analysing the health effects of simultaneous exposure to physical and chemical properties of airborne particles," *Environment International*, 79, 56 – 64.

WHO (2016), *Ambient air pollution: A global assessment of exposure and burden of disease*, Geneva, Switzerland.

WHO/Europe (2006a), *Communicable disease surveillance and response systems*, WHO.

— (2006b), *Health Risks of Particulate Matter from Long-range Transboundary Air Pollution*, Copenhagen, Denmark: WHO Regional Office for Europe.

Williams, M., Atkinson, R., Anderson, R., and Kelly, F. (2014), "Associations between daily mortality in London and combined oxidant capacity, ozone and nitrogen dioxide," *Air Quality, Atmosphere and Health*, 7, 407–414.

Thank you