Generalised Spatial Fusion Model Framework for Multivariate Analysis of Point and Areal Data

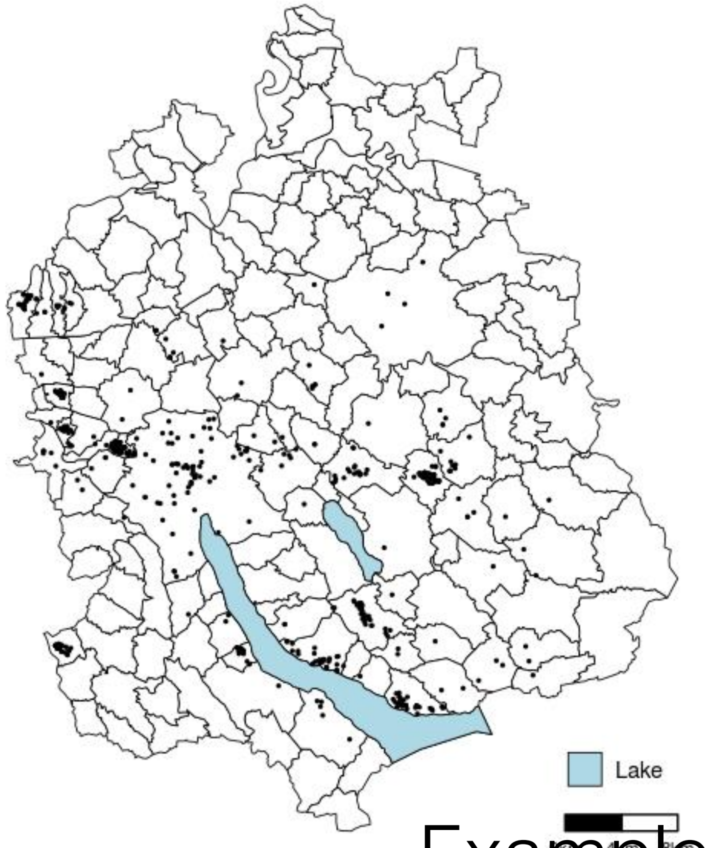
Craig Wang[a], Prof. Milo Puhan[b], Prof. Reinhard Furrer[a,c]

[a] Applied Statistics Group, Department of Mathematics, University of Zurich, Switzerland

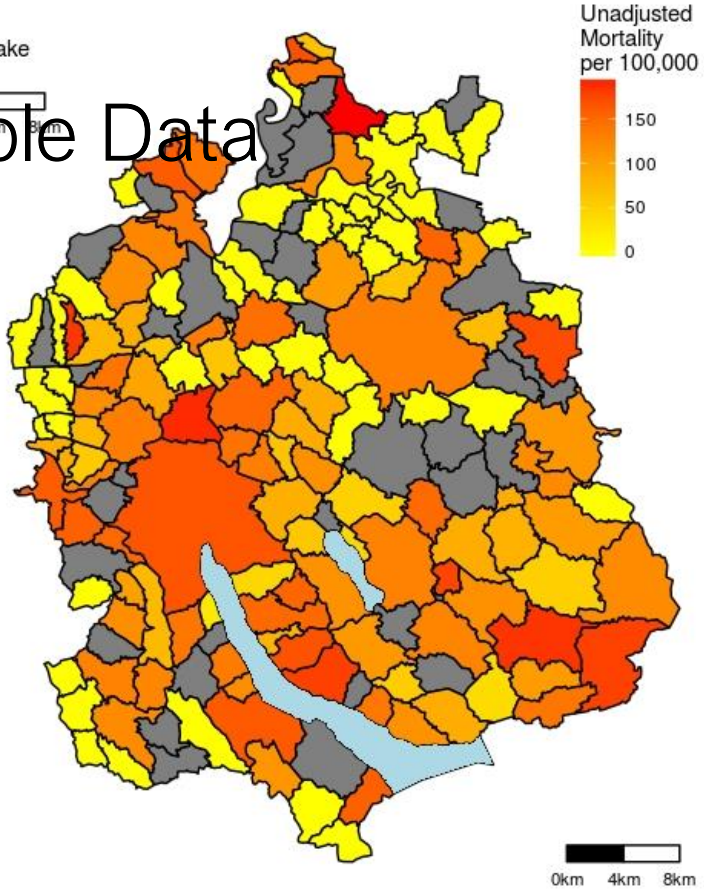
[b] Epidemiology, Biostatistics and Prevention Institute, University of Zurich, Switzerland

[c] Department of Computational Science, University of Zurich, Switzerland

Motivations

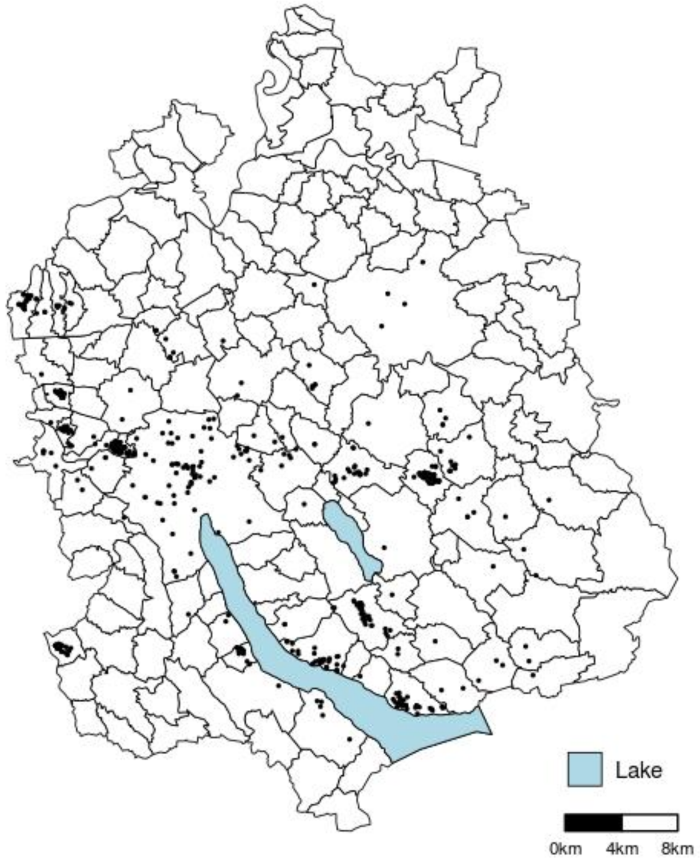


Example Data

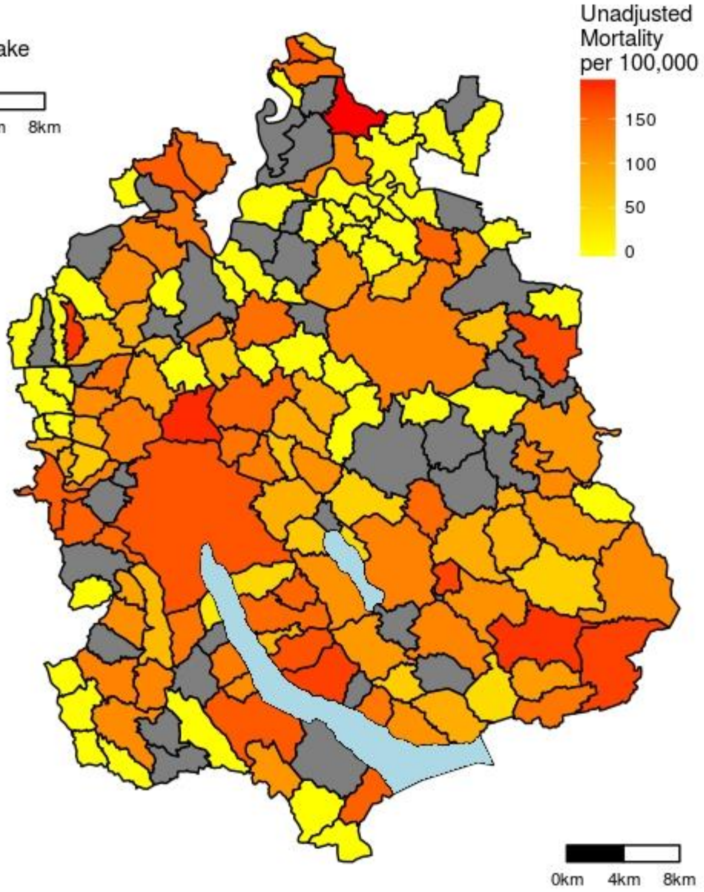


Motivations

- Scarce observations



Few data points

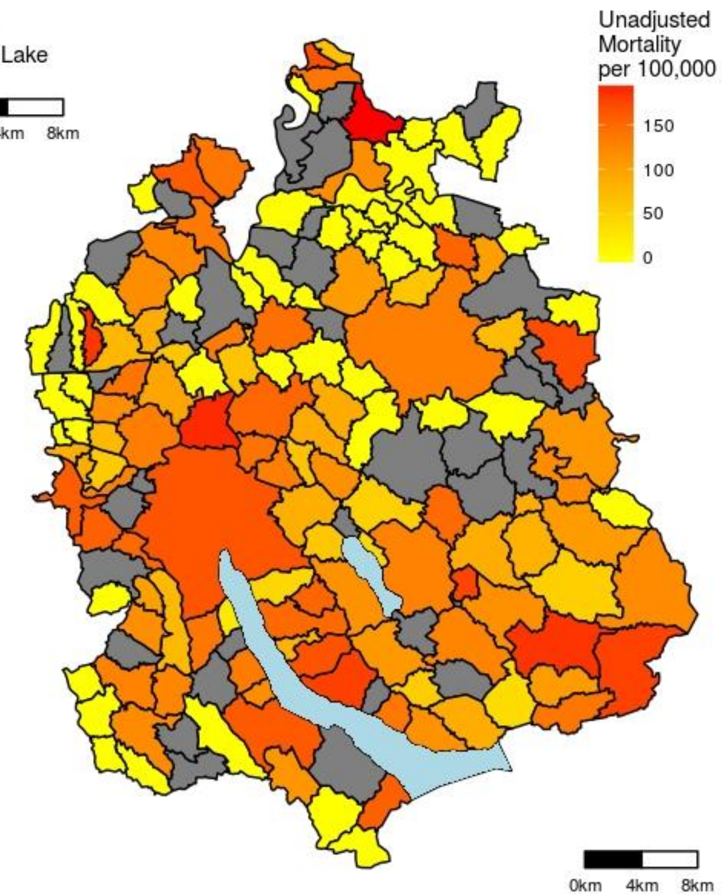


Missing areas

Motivations

- Scarce observations
- Multiple data sources
- Different resolutions

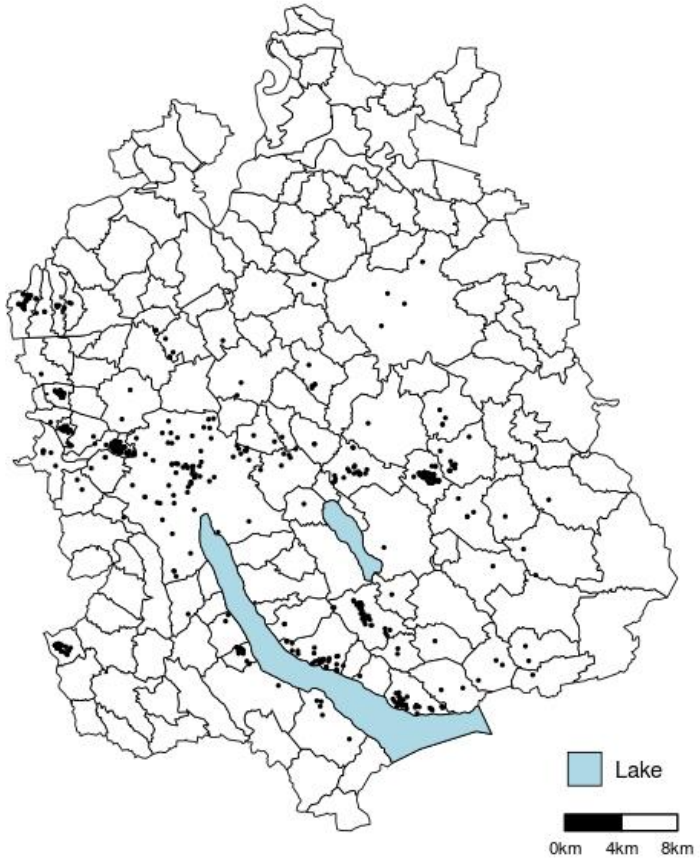
Source 1: Point data



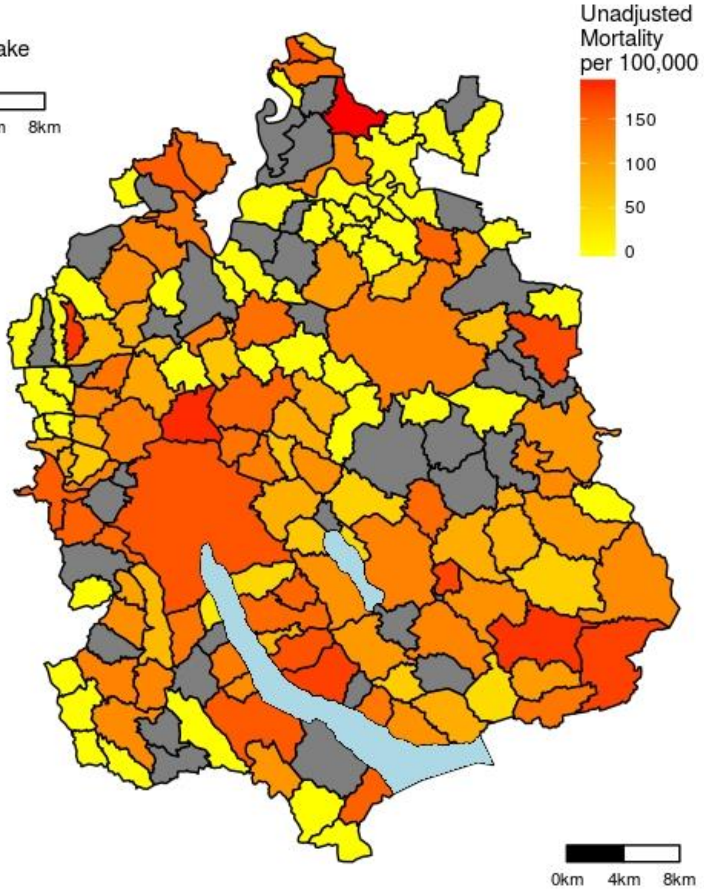
Source 2: Areal data

Motivations

- Scarce observations
- Multiple data sources
- Different resolutions
- Different distributions



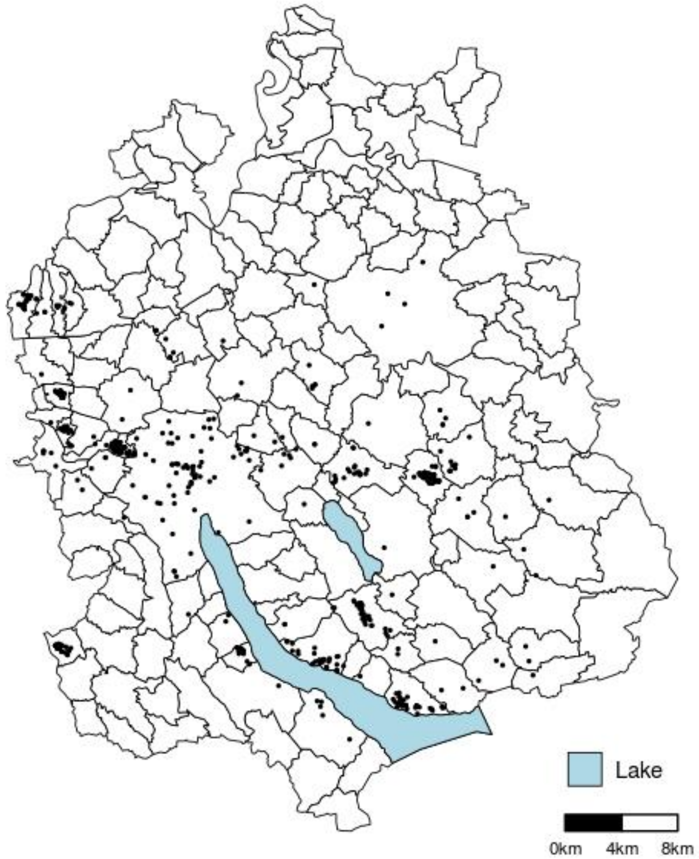
Lung function ~ Normal



Mortality ~ Poisson

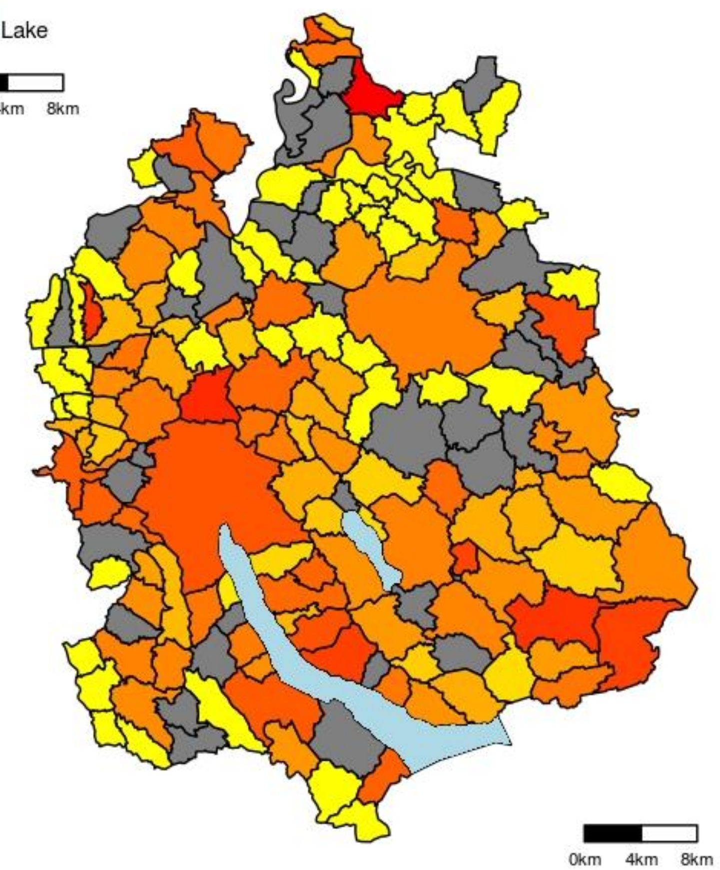
Motivations

- Scarce observations
- Multiple data sources
- Different resolutions
- Different distributions



Binary

Counts



Existing Spatial Fusion Models

- Also known as data assimilation (Banerjee et al. 2014) or Bayesian melding (Fuentes and Raftery, 2005)

- Fusion models include Sahu et al. (2010), Berrocal et al. (2010), Goovaerts (2010), Liu et al. (2011), Moraga et al. (2017), and Shi & Kang (2017)

- Some drawbacks
  - Constraint on distribution assumption
  - Model specification for single application
  - Designing of custom Monte Carlo sampler

A Framework for Spatial Fusion Models

- Point data $\boldsymbol{s} = \{s_1, \cdots, s_n\}$

$$f\left(\mathbb{E}\left[Y(\boldsymbol{s})|w(\boldsymbol{s})\right]\right) = \boldsymbol{X}_s^T(\boldsymbol{s})\boldsymbol{\beta}_s + w(\boldsymbol{s})$$

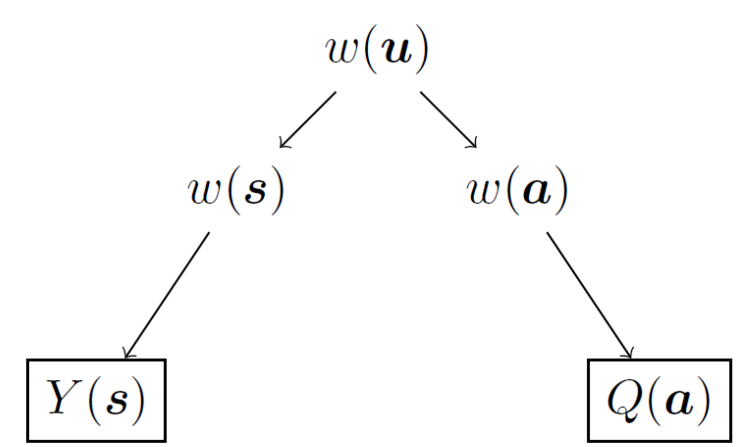
- Areal data $\boldsymbol{a} = \{a_1, \cdots, a_m\}$

$$g\left(\mathbb{E}\left[Q(\boldsymbol{a})|w(\boldsymbol{a})\right]\right) = \boldsymbol{X}_a^T(\boldsymbol{a})\boldsymbol{\beta}_a + w(\boldsymbol{a})$$

- Data fusion via sampling points $\boldsymbol{s}' = \{s'_1, \cdots, s'_l\}$

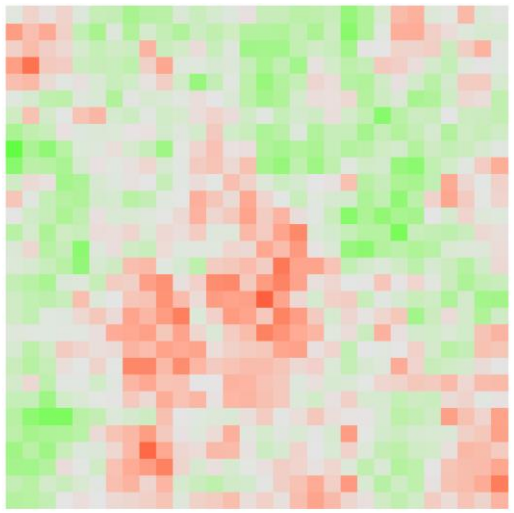
$$w(a_i) = \int_{\boldsymbol{s} \in a_i} w(\boldsymbol{s}) d\boldsymbol{s} \approx \frac{1}{L} \sum_{j=1, s'_j \in a_i}^{L} w(s'_j)$$

$$w(\boldsymbol{u}) = [w(\boldsymbol{s}) \quad w(\boldsymbol{s}')] \sim GP(0, C(\cdot, \cdot; \theta))$$

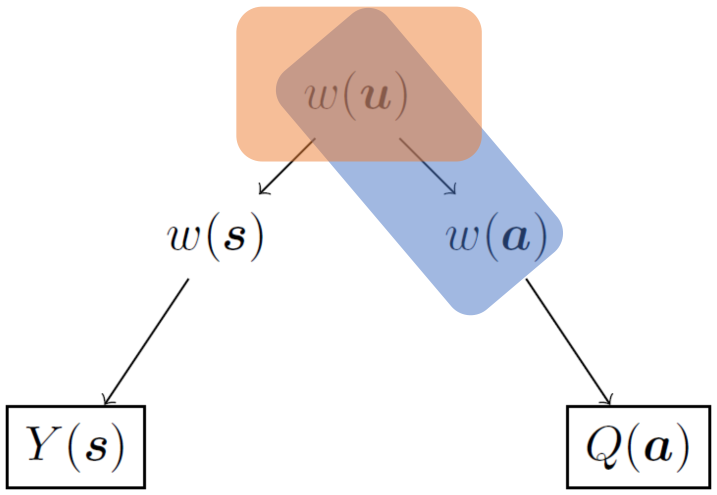


Modelling Approach

- Bayesian hierarchical model
- Common latent spatial process
- Approximation to stochastic integral
- Efficient Computation
  - Nearest neighbour Gaussian Process
  - Stan modelling language



A simulated Gaussian spatial process.

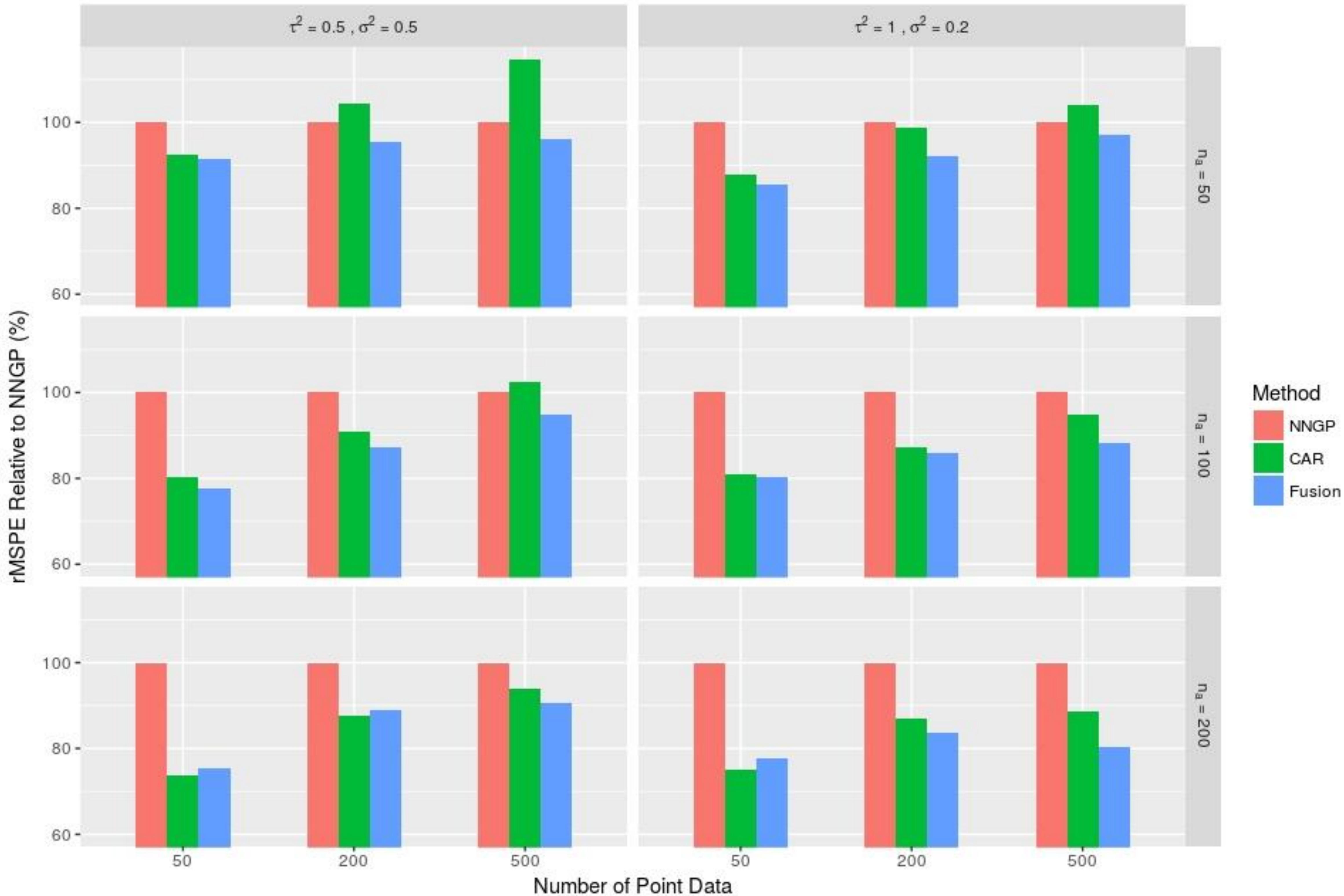


Simulation Study

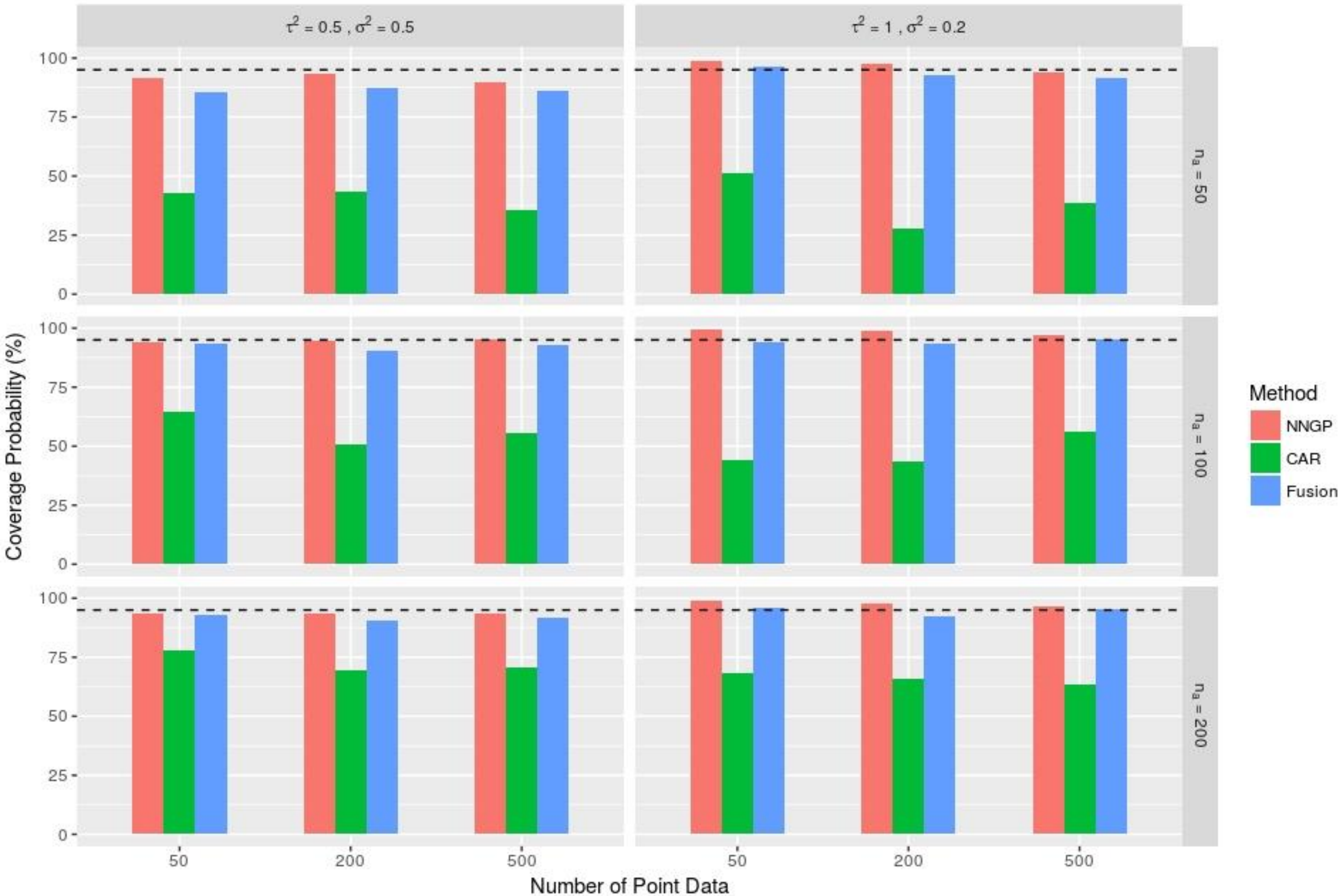
$$Y(\boldsymbol{s})|\boldsymbol{\beta}_s, w(\boldsymbol{s}) \sim \mathrm{Normal}(X_s^\top \boldsymbol{\beta}_s + w(\boldsymbol{s}), \tau^2)$$
$$Q(\boldsymbol{a})|\boldsymbol{\beta}_a, w(\boldsymbol{a}) \sim \mathrm{Poisson}(X_a^\top \boldsymbol{\beta}_a + w(\boldsymbol{a}))$$

- 18 different scenarios
  - Different sample size
  - Different spatial signal-to-noise ratio
- Methods
  - nearest neighbor Gaussian process (NNGP) – point data only
  - conditional autoregressive model (CAR) + kriging – areal data only
  - fusion model – both point and areal data
- Evaluation criteria on latent process prediction $w(\boldsymbol{s})$
  - root mean squared prediction error (rMSPE)
  - coverage probability of 95% posterior credible intervals

Simulation Performance - rMSPE



Simulation Performance - Coverage



Case Study: LuftiBus - SNC

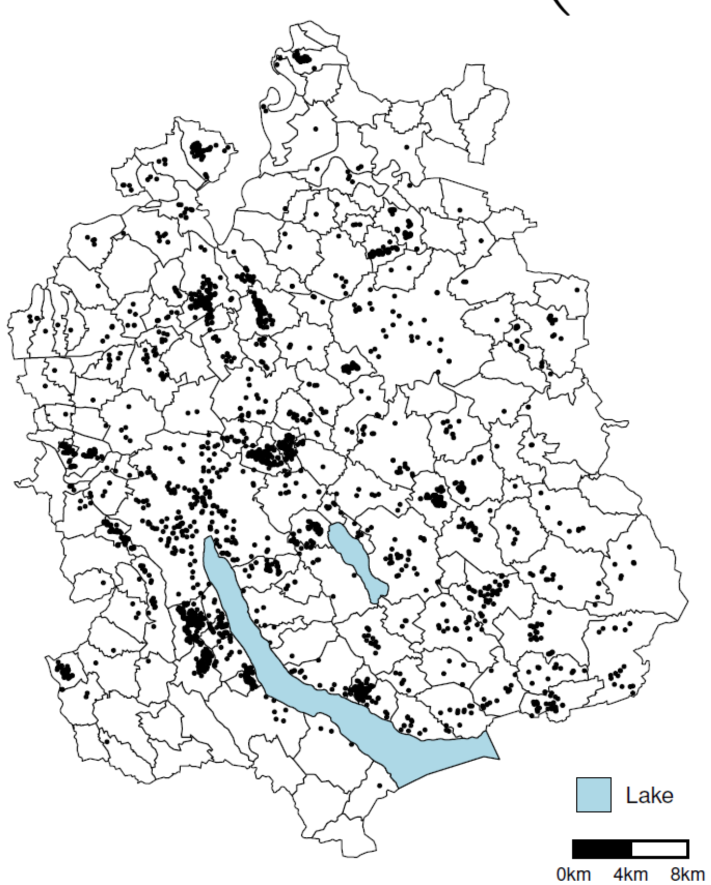
- LuftiBus dataset
  - Time: 2003 to 2012
  - Location: Switzerland
  - Variables: spirometry, demographics
- Swiss National Cohort (SNC)
  - Long-term population-based cohort
  - Variables: mortality, demographics
- Variables of interest
  - Point: Forced Expiratory Volume in one second (FEV1)
  - Areal: Mortality caused by respiratory disease and lung cancer

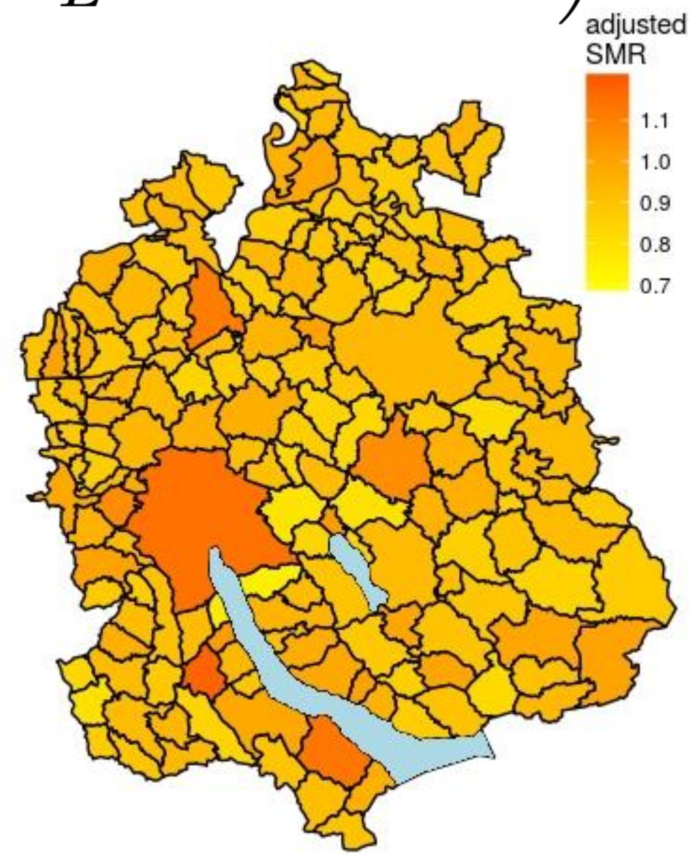
Case Study: Fusion Modelling

What is the spatial distribution of underlying risk in respiratory disease and lung cancer?

$$\text{FEV1} \sim \text{Normal}\left(\beta_0 + \beta_1 \times \text{gender} + \beta_2 \times \text{age} - w(\boldsymbol{s}), \tau^2\right)$$

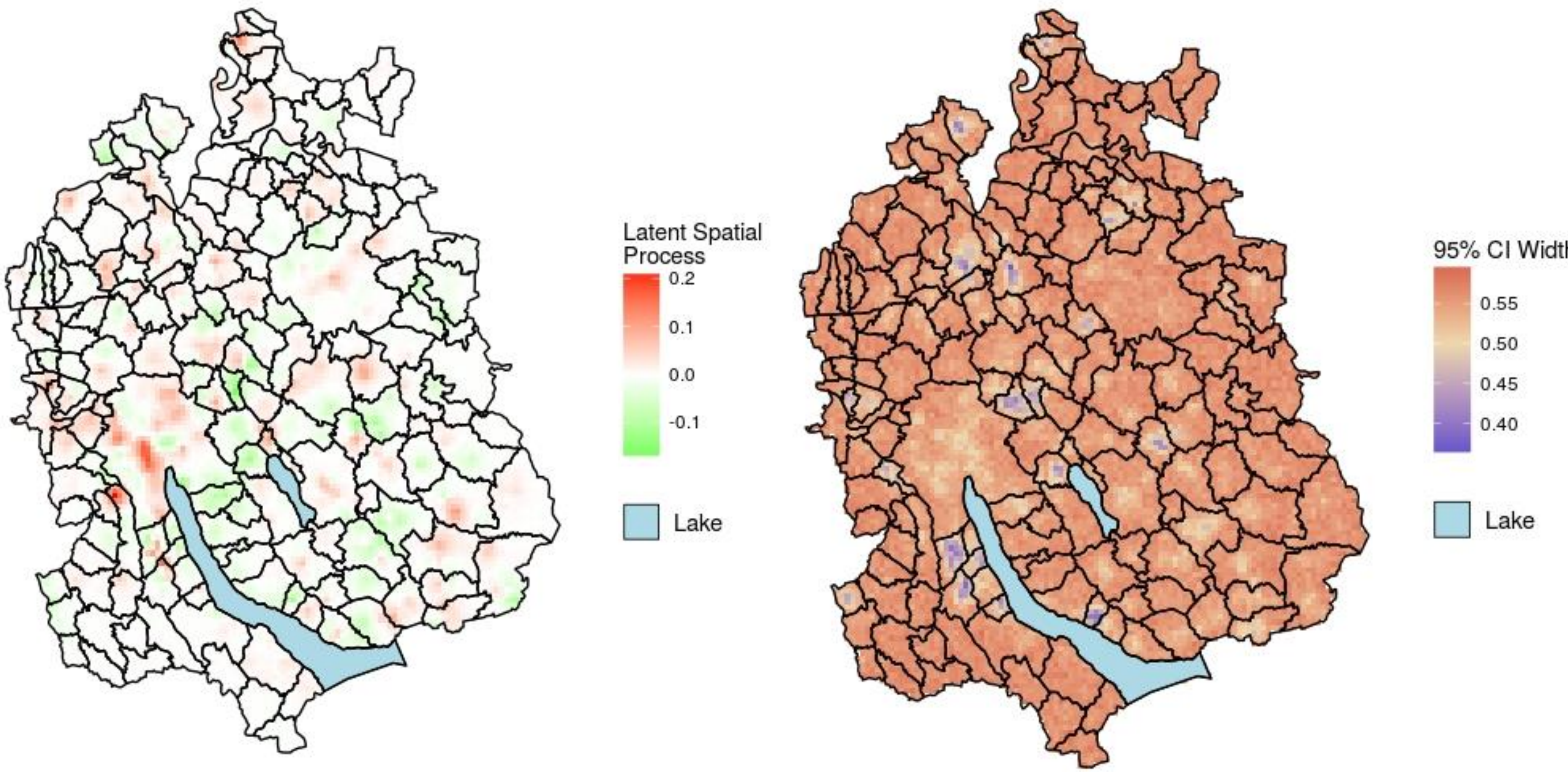
$$\text{Mortality} \sim \text{Poisson}\left(\text{Expected} \times \frac{1}{L}\sum \exp(w(\boldsymbol{s}'))\right)$$





Case Study: Results

Posterior Median of the Latent Spatial Process and its 95% CI Width



Summary

- Introduced generalised spatial fusion framework
  - Multivariate analysis of spatial data with different resolution and distribution
  - Mimicking data generating process, account for ecological bias
  - Can be easily adapted to different problems
- Utilized Stan modelling language and NNGP
- Simulation showed fusion models perform better in latent process prediction
- If you have “fusible” data, try it out!

References

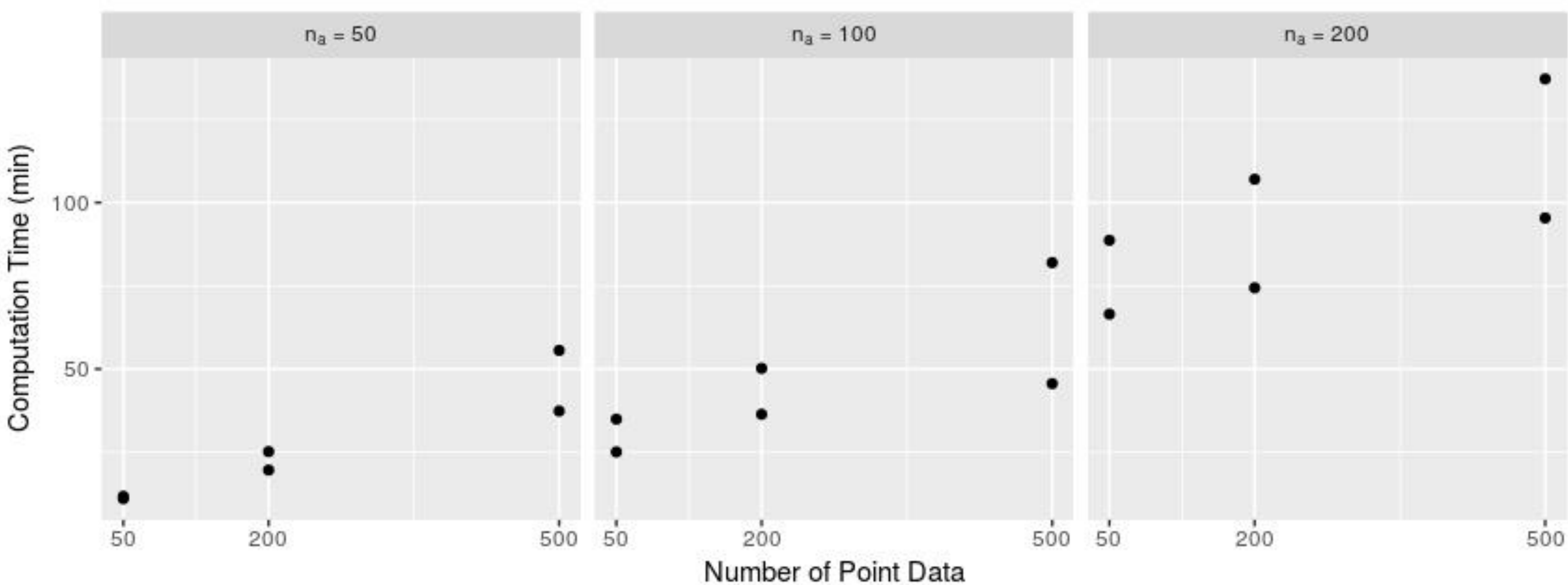
- Banerjee, Carlin & Gelfand, (2014). Hierarchical Modeling and Analysis for Spatial Data. CRC.
- Berrocal, Gelfand & Holland, (2010), *A spatio-temporal downscaler for output from numerical models*, Journal of Agricultural, Biological, and Environmental Statistics.
- Datta, Banerjee, Finley & Gelfand, (2016), *Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets*, Journal of the American Statistical Association.
- Fuentes & Raftery, (2005), *Model evaluation and spatial interpolation by Bayesian combination of observations with outputs from numerical models*,' Biometrics.
- Goovaerts, (2010). *Combining areal and point data in geostatistical interpolation: Applications to soil science and medical geography*. Mathematical Geosciences.
- Liu, Le & Zidek, (2011), *An empirical assessment of Bayesian melding for mapping ozone pollution,* Environmetrics.
- Moraga, Cramb, Mengersen & Pagano, (2017), *A geostatistical model for combined analysis of point-level and area-level data using INLA and SPDE*, Spatial Statistics.
- Sahu, Gelfand & Holland, (2010), *Fusing point and areal level space-time data with application to wet deposition*, Journal of the Royal Statistical Society Series C.
- Shi & Kang, (2017), *Spatial data fusion for large non-Gaussian remote sensing datasets*, Stat.
- Wang, Puhan & Furrer (2018), *Generalized spatial fusion model framework for joint analysis of point and areal data*, Spatial Statistics.

Backup Slides

- Accounting for ecological bias

$$\exp(w(a_i)) = |a_i|^{-1} \int_{\boldsymbol{s} \in a_i} \exp(w(\boldsymbol{s})) d\boldsymbol{s} \approx \frac{1}{L} \sum_{j=1, s'_{ij} \in a_i}^{L} \exp(w(s'_{ij}))$$

- Computation time for fusion model



Stan Model Code

```
parameters{
  vector[pn] beta; // coefficients
  real<lower = 0> sigma_sq;
  real<lower = 0> tau_sq;
  real<lower = 0> phi;
  vector[n+aL] w;
}

transformed parameters{
  vector[n] A0w;
  vector[a] A1w;
  A0w = w[1:n];
  A1w = log(A1 * exp(w));
}


model{
  beta ~ normal(0, 5);
  sigma_sq ~ inv_gamma(2, 0.1);
  tau_sq ~ inv_gamma(2, 1);
  phi ~ normal(700,100);
  w ~ nngp_w(sigma_sq, phi, neardist, neardistM, nearind, n+aL, M);
  Y ~ normal(Xn * beta - A0w, sqrt(tau_sq));
  Q ~ poisson_log(offset + A1w);
}
```

Nearest neighbor Gaussian process

- Datta et al. (2016) proposed a low rank process that achieves $O(nk^3)$ where k is the number of nearest neighbors
- Instead of specifying $w(\boldsymbol{s})$ with a $\boldsymbol{n} \times \boldsymbol{n}$ covariance matrix,

$$w(\boldsymbol{s}) \sim N\left(C_{\boldsymbol{s},N(\boldsymbol{s})}C_{N(\boldsymbol{s})}^{-1}w_{N(\boldsymbol{s})}, C(\boldsymbol{s},\boldsymbol{s}) - C_{\boldsymbol{s},N(\boldsymbol{s})}C_{N(\boldsymbol{s})}^{-1}C_{N(\boldsymbol{s}),\boldsymbol{s}}\right)$$

where $C_{\boldsymbol{s},N(\boldsymbol{s})}$ is the $\boldsymbol{1} \times \boldsymbol{k}$ cross-covariance matrix between location $\boldsymbol{s}$ and each neighbors, and $C_{N(\boldsymbol{s})}$ is the $\boldsymbol{k} \times \boldsymbol{k}$ covariance matrix of $w_{N(\boldsymbol{s})}$

- Computational complexity comparison with n = 10,000
  - Spatial regression with full Gaussian process $O(10^{12})$
  - Gaussian predictive process with 500 knots $O(2.5 \times 10^9)$
  - Nearest neighbor Gaussian process with 10 neighbors $O(10^7)$