Spatial Statistics 23 (2018) 72–90

Contents lists available at ScienceDirect

Spatial Statistics

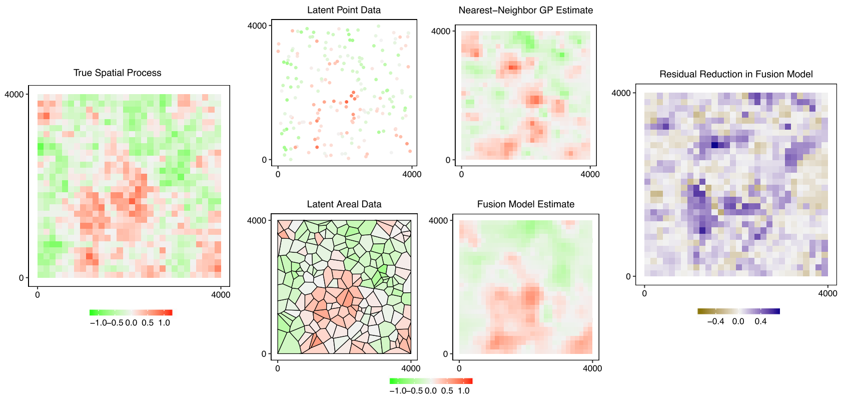
journal homepage: www.elsevier.com/locate/spasta

Generalized spatial fusion model framework for joint analysis of point and areal data

Craig Wang [a], Milo A. Puhan [b], Reinhard Furrer [a,c,*], for the SNC Study Group

[a] Department of Mathematics, University of Zurich, Zurich, Switzerland
[b] Epidemiology, Biostatistics and Prevention Institute, University of Zurich, Zurich, Switzerland
[c] Department of Computational Science, University of Zurich, Zurich, Switzerland

GRAPHICAL ABSTRACT



HIGHLIGHTS

- We propose a new generalized spatial fusion model framework.
- Our framework allows jointly analyzing non-Gaussian point and areal data.
- A simulation study shows fusion models improve prediction performance.

[*] Correspondence to: Department of Mathematics, University of Zurich, Winterthurerstrasse 190, CH-8057, Zurich Switzerland.

E-mail address: reinhard.furrer@math.uzh.ch (R. Furrer).

https://doi.org/10.1016/j.spasta.2017.11.006

- We identify situations where fusion models can lead to large improvements.
- A Gaussian–Poisson fusion model is applied to an epidemiological dataset.

ARTICLE INFO

Article history:
Received 6 October 2017
Accepted 22 November 2017
Available online 5 December 2017

Keywords:
Data fusion
Bayesian hierarchical model
Nearest neighbor Gaussian process
Change-of-support
Spatial epidemiology

ABSTRACT

The availability of geo-referenced data increased dramatically in recent years, motivating the use of spatial statistics in a variety of research fields, including epidemiology, environmental science, remote sensing, and economics. Combining data measured at both point and areal support can improve parameter estimation and increase prediction accuracy. We propose a new generalized spatial fusion model framework for jointly analyzing point and areal data. Assuming a common latent spatial process, we take a Bayesian hierarchical approach to model both types of data without distributional constraints. The models are implemented with nearest neighbor Gaussian process in Stan modeling language to increase computational efficiency and flexibility. Our simulation study shows that generalized fusion models under this framework model the latent process better than spatial process models. We identify scenarios where fusion models can offer large improvements. We then apply the framework to epidemiological data to identify the spatial risk pattern of respiratory diseases and lung cancer in Canton of Zurich, Switzerland.

1. Introduction

An increase in geo-referenced data spurred the use of spatial statistics in a variety of research areas, including epidemiology, environmental science, remote sensing, and economics (Gelfand et al., 2005; Shi and Cressie, 2007; Lawson et al., 2016; Paci et al., 2017). Often, researchers analyze a single spatial dataset, but a single source of spatial data may not be the best choice for parameter inference due to problems such as missing data and selection bias. For example, in remote sensing, cloud cover can interfere with regional observations. In disease mapping, data collection methods can cause selection bias in certain populations. In addition, there can be modeling difficulties, small sample size or weak spatial correlation may make it hard to estimate parameters (Irvine et al., 2007). In these situations, using multiple data sources can offer an advantage. Data may be collected at different spatial support for a variety of reasons, including budget constraints and privacy considerations. These data can be combined with the assumption of a common underlying spatial process in spatial fusion models. For example, air pollution modeling can be done based on measurements from monitoring stations, or numerical model output from computer simulations. We can assume that the same pollution process influences both measurements taken at the station, and the results of the simulation. Another example can be found in epidemiology, where, for privacy reasons, aggregated case counts at the district level are much more common than individual case locations of a disease. We can assume the same disease risk pattern drives the occurrence of individual cases and aggregated counts. As more database hosts and organizations collaborate, it is becoming easier to link datasets, providing more opportunities to carry out fusion tasks.

The approach of jointly analyzing multiple data sources that have different spatial support is called data fusion or data assimilation (Banerjee et al., 2014), or Bayesian melding (Fuentes and Raftery, 2005; Liu et al., 2011) in different literatures. They take a slightly different approach to modeling, but the basic idea is the same: combining point and areal data in a single statistical model. Fuentes and Raftery (2005) proposed one of the first fusion models, which predicts the spatial distribution of air pollution level. The model specifies both point-referenced measurements from monitoring stations and the regional air pollution model. Block averaging, based on systematic sampling, relates the different spatial supports to approximate stochastic integrals. The model was fitted without any covariates; its computational disadvantage makes it suitable only for small sample sizes. Bourgeois et al. (2012) proposed a hierarchical model combining three different types of Poisson-based areal data on weed measurements. The model assumed a latent log Gaussian Cox process to represent weed density, and used sampling points to approximate stochastic integrals. The model handles non-Gaussian distributed data but still lacks computational speed for larger datasets. Sahu et al. (2010) and McMillan et al. (2010) proposed alternative measurement error models, where the underlying spatial process is defined by a conditionally autoregressive (CAR) model. Computational time dropped dramatically, especially for large numbers of areal observations. The underlying CAR assumption is justified, given that there are more than 10^4 gridded areal observations in their dataset. The resulting spatial prediction at areal level is almost indistinguishable from a continuous process. For applications where areal shapes are large and irregular, or when data in many areas are missing, such approach can become problematic.

Cowles et al. (2009) introduced a more general fusion model that accounts for covariates, non-spatial random effects and allows for different measurement errors from point and areal data. They also proposed a four-stage slice sampling-based Markov chain Monte Carlo (MCMC) algorithm to obtain posterior samples, which is more efficient than naively implementing the fusion model in OpenBUGS (Lunn et al., 2009). Instead of using MCMC methods for Bayesian inference, Moraga et al. (2017) proposed an Integrated Nested Laplace Approximation (INLA) approach. INLA's main advantage is that it does not require MCMC sampling algorithms for fusion models, which often generate highly dependent posterior samples. Moraga et al. (2017) pointed out that their fusion model only applies to Gaussian data, and their model requires the same responses and covariates for both spatial supports. Instead of fusing point and areal data, Berrocal et al. (2010) took a downscaling approach to model air quality data. They used numerical model output as a covariate for observations at monitoring stations, and showed it made better predictions than Fuentes and Raftery (2005)'s fusion model and ordinary kriging. The downscaling approach is not suitable when one is interested in the common latent process, or when the association between point and areal observations is not immediately clear. Finally, there are non-Bayesian approaches, such as spatial statistical data fusion, based on fixed rank kriging (Nguyen et al., 2012), area-and-point kriging (Goovaerts, 2010), and geographically weighted regression (Murakami and Tsutsumi, 2015). Here, we only focus only on Bayesian models.

Cowles et al. (2009) and Moraga et al. (2017) showed that the prediction performance of fusion models based on both point and areal data can be better than models that use a single data source, but applications continue to focus on air quality modeling. The fusion models have sometimes been applied in epidemiology (Huang et al., 2015; Lee et al., 2017), but they were only used for mapping air pollution in the first stage of analysis, and the result is used for adjusting disease data in the second stage. Existing literature on fusion models has two limits. First, air pollution data applications aim to model the distribution of some air quality measure, which is often directly available as observations at different spatial support; only Sahu et al. (2010) assumed a common underlying atmospheric driver. Second, air quality data or some transformation of it are assumed to follow only Gaussian distributions, a common assumption of all existing fusion models that limited the scope of potential applications. Diggle et al. (1998) described cases in which assuming Gaussian distribution in spatial models is inappropriate. The model proposed by Bourgeois et al. (2012) did not suffer those two limitation but it was application specific, fusing only Poisson-distributed areal data.

We extend existing fusion models and propose a generalized spatial fusion model framework that incorporates data from both point and areal support by assuming a common latent spatial process. In contrast to existing approaches, we do not restrict either point or areal observations to follow Gaussian distributions. Our framework increases the computational efficiency of fusion models by using nearest neighbor Gaussian process (NNGP) (Datta et al., 2016) and implementing it in Stan programming language (Carpenter et al., 2017). We show, via a simulation study, that a Gaussian–Poisson fusion model under our framework outperforms spatial process models. In Section 2, we introduce the motivating dataset and perform some exploratory analysis. In Section 3, we describe the formulation and implementation of the framework. In Section 4, we conduct a simulation study to compare different models. In Section 5, we apply a fusion model under our new framework to an epidemiological dataset. We end the paper with a discussion, followed by Appendices including our model source files and other additional information.

2. Motivating dataset and exploratory analysis

The motivating dataset for our case study is the LuftiBus project (Zürich, 2017), a health promotion campaign of the Zurich Lung Association. LuftiBus was initiated in Switzerland in 1993, to raise awareness about lung diseases and their associated risk factors. A LuftiBus vehicle, equipped with spirometers and other measurement devices, drives around the greater Zurich area and other regions of Switzerland to collect lung function measurements and health information from local residents. LuftiBus data was the basis for published population-based reference values for lung functions and exercise capacity (Kuster et al., 2008; Strassmann et al., 2013). Recently, the records of LuftiBus and census-based Swiss National Cohort (SNC) (The SNC Study Group, 2017) were deterministically linked by date of birth, residential postcode, and sex. The LuftiBus-SNC dataset contains 56,223 people with demographic, health and environmental variables.

We will investigate the spatial pattern of respiratory disease and lung cancer risk in our target population after adjusting for age and gender. In disease mapping, this kind of analysis is usually based on aggregated areal data from each municipality (Chammartin et al., 2016), however such approach can be prone to ecological bias. Additionally municipal boundaries are artificial, we argue that a continuous spatial pattern is a better assumption. Our dataset is unique because it records each participant's residential location, which enables us to model a continuous spatial pattern.

Patients with respiratory diseases such as chronic obstructive pulmonary disease (COPD) can be diagnosed by measuring lung functions. A measure of how much air a person can exhale during a forced breathe is forced expiratory volume in one second (FEV1). FEV1 can be used to diagnose disease and predict mortality related to respiratory functions (Menezes et al., 2014). Before we fit a model, we apply exclusion criteria (see Appendix A) to process the data, which reduces the risk of data recording error and allow us to focus on our population of interest. For purposes of illustration, we restrict our analysis to measurements from Canton of Zurich, made in 2010. After exclusion, we have 2315 people remaining in our analysis. We start by fitting a multiple linear regression with FEV1 as the response variable, age and gender (female is encoded as 1 and male is encoded as 0) as covariates, and investigate spatial structure in the residuals.

Fig. 1 maps 171 municipal boundaries in Canton of Zurich and the residential locations of the 2315 people in the LuftiBus-SNC dataset. Fig. 2 shows the empirical semi-variogram from the residuals of the multiple linear regression. We observe a high nugget-to-partial sill ratio of 5 ($= 0.23/0.046$); the effective range is about 2.2 km, based on the fitted exponential variogram model. The 95% pointwise variogram envelope is computed from the variogram estimates based on 1000 Monte Carlo permutations of residuals on the locations (Ribeiro Jr. and Diggle, 2016). Data with semi-variance estimates outside the envelope are considered to exhibit spatial structure. The high nugget-to-partial sill ratio indicates weak spatial correlation, but several estimates fall a short distance outside the envelope, indicating some spatial structure. Irvine et al. (2007)'s simulation studies showed that spatial parameters are difficult to estimate when spatial correlation is weak, therefore there is a need to consider additional spatial information. In Section 5, we use aggregated cause-specific mortality data as areal observations. By assuming the same latent spatial process for risk of decreased FEV1 and increased mortality, we jointly analyze those two variables with a fusion model. Because there is no appropriate fusion model approach to include Poisson distributed mortality, we propose a new framework unfettered by this distributional constraint.

3. Generalized spatial fusion model framework

3.1. Model formulation

We propose a generalized spatial fusion model framework that jointly analyzes point and areal data. Without loss of generality, we consider two responses: $Y(\boldsymbol{s})$ and $Q(\boldsymbol{a})$. The point-referenced

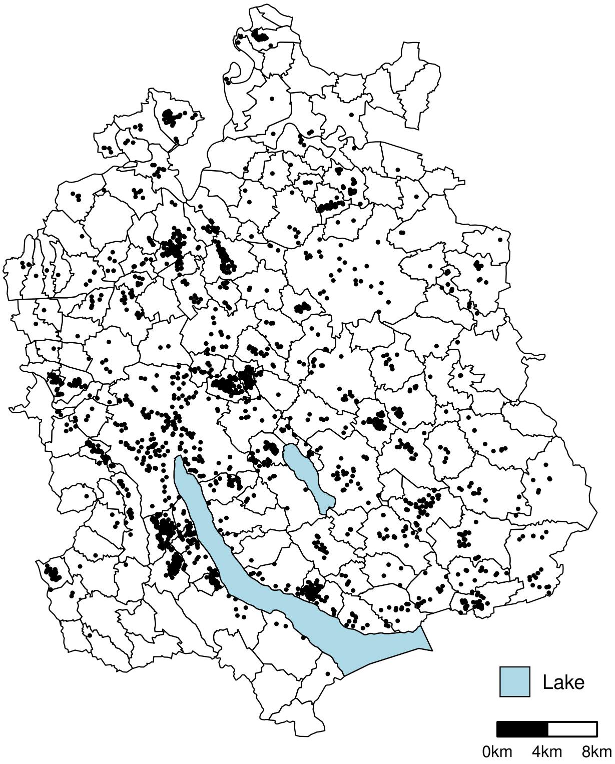


Fig. 1. Residential locations of 2315 LuftiBus candidates in the processed dataset.

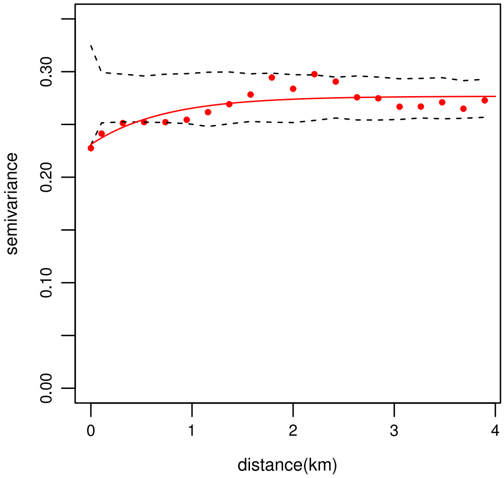


Fig. 2. The empirical semi-variogram of residuals from the multiple linear regression is shown in red dots; the solid line represents the fitted exponential variogram model. The dashed lines are the 95% pointwise variogram envelope, which are based on the 2.5th and 97.5th percentile of semi-variance estimates from Monte Carlo simulations.

response variable $Y(\boldsymbol{s})$ is observed at site $\boldsymbol{s} \in D \subseteq \Re^2$, with $\mathcal{S} = \{\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_{n_s}\}$. The areal response variable $Q(\boldsymbol{a})$ is observed in area $\boldsymbol{a} \in D$, with $\boldsymbol{a} \in \mathcal{A} = \{\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_{n_a}\}$. We assume a latent Gaussian process (GP) $w(\boldsymbol{u})$ at site $\boldsymbol{u} \in D$, which is associated with both responses; it has zero mean and a covariance function $C(\cdot, \cdot; \boldsymbol{\theta})$, i.e., $w(\boldsymbol{u}) \sim GP(0, C(\cdot, \cdot; \boldsymbol{\theta}))$. We denote $\boldsymbol{w}_{\mathcal{S}} = \left(w(\boldsymbol{s}_1), w(\boldsymbol{s}_2), \ldots, w(\boldsymbol{s}_{n_s})\right)^T$ and $\boldsymbol{w}_{\mathcal{A}} = \left(w(\boldsymbol{a}_1), w(\boldsymbol{a}_2), \ldots, w(\boldsymbol{a}_{n_a})\right)^T$, where $w(\boldsymbol{a}_i) = |\boldsymbol{a}_i|^{-1} \int_{\boldsymbol{u}\in\boldsymbol{a}_i} w(\boldsymbol{u})d\boldsymbol{u}$ is the aggregated process for area $\boldsymbol{a}_i$. We address the change of spatial support by approximating the stochastic integrals, with

$$w(\boldsymbol{a}_i) \approx \frac{1}{L} \sum_{j=1, \boldsymbol{s}'_{ij}\in\boldsymbol{a}_i}^{L} w(\boldsymbol{s}'_{ij}), \tag{1}$$

where $w(\boldsymbol{s}'_{ij})$ is the jth sampling points within area $\boldsymbol{a}_i$, and L is the number of sampling points in each area. This approximation was also used by Fuentes and Raftery (2005), Berrocal et al. (2010) and Liu et al. (2011). They showed that a small L is a good trade-off between computational efficiency and model accuracy. We further denote the set of all sampling points as $\mathcal{S}'$. The set of locations in the latent process $\mathcal{U}$ consists of the observed locations and the locations at sampling points, i.e., $\mathcal{U} = \mathcal{S} \cup \mathcal{S}'$.

We assume $1 \times p_s$ and $1 \times p_a$ vectors of spatially-referenced covariates $\boldsymbol{X}_s^T(\boldsymbol{s})$ and $\boldsymbol{X}_a^T(\boldsymbol{a})$. The model can be written as

$$\begin{aligned} f\left(\mathbb{E}\left[Y(\boldsymbol{s})|w(\boldsymbol{s})\right]\right) &= \boldsymbol{X}_s^T(\boldsymbol{s})\boldsymbol{\beta} + w(\boldsymbol{s}), \\ g\left(\mathbb{E}\left[Q(\boldsymbol{a})|w(\boldsymbol{a})\right]\right) &= \boldsymbol{X}_a^T(\boldsymbol{a})\boldsymbol{\alpha} + w(\boldsymbol{a}), \end{aligned} \tag{2}$$

where $f(\cdot)$ and $g(\cdot)$ are suitable link functions that depend on the distributions of $Y(\boldsymbol{s})$ and $Q(\boldsymbol{a})$; $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are the corresponding coefficients for the covariates. Additional error terms can be added, depending on the distribution. In the spirit of Fuentes and Raftery (2005) and Liu et al. (2011), we let the first n_s elements in $w_{\mathcal{U}}$ coincide with $w_{\mathcal{S}}$, and the $(n_s + 1)$th to $(n_s + n_a L)$th elements be the process at sampling points. The latent process is transformed into the average of sampling points within each area as in Eq. (1) by design matrix K with $\boldsymbol{w}_{\mathcal{A}} \approx K\boldsymbol{w}_{\mathcal{U}}$, where K is specified as

$$K = \underbrace{\begin{bmatrix} 0 & \ldots & 0 & k_a & \ldots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \ldots & 0 & 0 & \ldots & k_a \end{bmatrix}}_{n_a \times (n_s + n_a L)} \tag{3}$$

with each k_a a row vector of length L with elements $1/L$.

Fig. 3 is a graphical model representation of the fusion model framework. Unlike previous fusion models, in our model $Y(\boldsymbol{s})$ and $Q(\boldsymbol{a})$ do not need to follow the same distributions. For example, in Section 5, $Y(\boldsymbol{s})$ represents the lung function measurement FEV1 that follows a Gaussian distribution, while $Q(\boldsymbol{a})$ represents the number of cause-specific mortality, which follows a Poisson distribution.

3.2. Model implementation

It is computationally expensive to fit full Gaussian process models for large spatial datasets. The number of floating point operations (flops) is $O(n^3)$ for n locations in evaluating the exact log-likelihood. The model assumption of a continuous latent process becomes infeasible for a fusion model with many locations, but advancements in low rank models (Banerjee et al., 2008; Stein, 2008) and sparse methods (Furrer et al., 2006; Rue et al., 2009; Datta et al., 2016) have significantly reduced computation time without compromising performance. In our implementation, we let the latent spatial process $w(\boldsymbol{u})$ follow an NNGP Datta et al. (2016). NNGP uses conditional densities based on neighboring locations to construct joint density of a full Gaussian process, to avoid dealing with large covariance matrices in a full GP model. A nearest-neighbor Gaussian process for $w_{\mathcal{U}}$ can be constructed by firstly selecting a fixed set of reference locations $\mathcal{R} = \{\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_{n_r}\} \in D$ that may not coincide with the locations in $\mathcal{U}$, then the nearest-neighbor density of $w_{\mathcal{U}}$ can be derived from

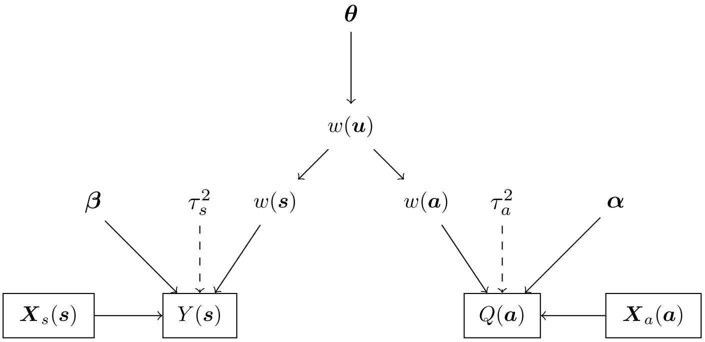


Fig. 3. A graphical model representation of the fusion model framework. Dashed arrows represent optional components that depend on the distribution of response variables. Boxes represent data components.

$p(w_{\mathcal{U}}, w_{\mathcal{R}}) = p(w_{\mathcal{R}}) \times p(w_{\mathcal{U}}|w_{\mathcal{R}})$, with $w_{\mathcal{R}} = \left(w(\boldsymbol{r}_1), w(\boldsymbol{r}_2), \ldots, w(\boldsymbol{r}_{n_r})\right)^T$. For each of the densities, we can express it using conditionals on nearest neighbors, i.e.,

$$\begin{aligned} p(w_{\mathcal{R}}) &= p(w(\boldsymbol{r}_1)) \prod_{i=2}^{n_r} p\Big(w(\boldsymbol{r}_i) \mid w\big(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_{i-1}\big)\Big) \\ &\approx \prod_{i=1}^{n_r} p\Big(w(\boldsymbol{r}_i) \mid w\big(N(\boldsymbol{r}_i)\big)\Big), \end{aligned} \tag{4}$$

where $N(\boldsymbol{r}_i)$ is the set of m nearest-neighbors from $\{\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_{i-1}\}$ for $\boldsymbol{r}_i$, and

$$p(w_{\mathcal{U}}|w_{\mathcal{R}}) \approx \prod_{i=1}^{n_s+n_aL} p\Big(w(\boldsymbol{u}_i) \mid w\big(N(\boldsymbol{u}_i)\big)\Big), \tag{5}$$

where $N(\boldsymbol{u}_i)$ is the set of m nearest-neighbors from $\mathcal{R}$ for $\boldsymbol{u}_i$. As Datta et al. (2016) pointed out, a convenient choice for the reference set is $\mathcal{R} = \mathcal{U}$. The computation in likelihood evaluation of NNGP is then simplified further since $p(w_{\mathcal{U}}|w_{\mathcal{R}}) = 1$. We are left to evaluate the conditional densities in Eq. (4) with $w_{\mathcal{R}} = w_{\mathcal{U}}$. Utilizing NNGP, the computational complexity for modeling $w_{\mathcal{U}}$ is reduced to $O(nm^3)$, where m is usually very small ($\sim$10).

To make our framework more flexible, we implemented our fusion models in Stan (Carpenter et al., 2017), via the R interface rstan (Stan Development Team, 2016). Stan uses the No-U-Turn sampler (Homan and Gelman, 2014) to obtain posterior samples, an improved version of Hamiltonian Monte Carlo (HMC) that adaptively adjust its tuning parameters. Hierarchical modeling of GP with non-Gaussian distributed responses involve high dimensional spatial random effects, which contains highly correlated samples and leads to slow mixing of MCMC algorithms. HMC enables the chains to move much faster especially in high dimensional target distributions by borrowing a concept from physics. It assigns a momentum to each parameter, and updates all the parameters and their momentum at each HMC iteration. Thus we did not have to design custom samplers for the specific fusion models, as Sahu et al. (2010) and Cowles et al. (2009) did. Using Stan also relaxed the constraint on assigning conjugate priors that is typically used in Gibbs samplers. Finally, rstan syntax is easy to understand, which makes it more accessible to researchers. (See Appendix B for the source files.)

4. Simulation study

In this section, we conduct a simulation study to compare the performance of models in different scenarios. Since we assume point and areal observations follow different distributions, the other existing fusion models cannot be fitted for such situations. Hence, we consider our fusion model and other spatial process models. All of the simulation results are obtained in R version 3.3 (R Core Team, 2017), on a Linux server with 256 GB of RAM and two Intel Xeon 6-core 2.50 GHz processors.

4.1. Simulation setup

We are interested in modeling the latent spatial process within a $[0, 4000] \times [0, 4000]$ domain. We start by generating a zero-mean stationary Gaussian process $w(\boldsymbol{u})$ on a fine grid with covariance matrix $C(\cdot, \cdot; \boldsymbol{\theta})$. To generate areal observations, we divide the domain into n_a number of Voronoi cells with uniformly distributed centroids in the domain, and compute $w(\boldsymbol{a})$ for each area. We sub-sample n_s locations from the fine grid to obtain $w(\boldsymbol{s})$ at observed locations. We then generate a covariate $X_{s,1}(\boldsymbol{s}) \sim N(0, 1^2)$ for each location and a covariate $X_{a,1}(\boldsymbol{a}) \sim N(0, 1^2)$ for each area. Adding intercepts, we obtain $\boldsymbol{X}_s^T(\boldsymbol{s}) = [1\, X_{s,1}(\boldsymbol{s})]$ and $\boldsymbol{X}_a^T(\boldsymbol{a}) = [1\, X_{a,1}(\boldsymbol{a})]$. The response variables are then generated according to

$$\begin{aligned} Y(\boldsymbol{s})|w(\boldsymbol{s}) &\sim N\left(\boldsymbol{X}_s^T(\boldsymbol{s})\boldsymbol{\beta} + w(\boldsymbol{s}), \tau^2\right), \\ Q(\boldsymbol{a})|w(\boldsymbol{a}) &\sim \text{Poisson}\left(\exp\left(\boldsymbol{X}_a^T(\boldsymbol{a})\boldsymbol{\alpha} + w(\boldsymbol{a})\right)\right), \end{aligned} \tag{6}$$

where τ^2 is the nugget. In the simulation, we use exponential covariance function from the Matérn family, $C(\boldsymbol{s}_i, \boldsymbol{s}_j; \sigma^2, \phi) = \sigma^2 \exp(-\|\boldsymbol{s}_i - \boldsymbol{s}_j\|/\phi)$, where $\|\boldsymbol{s}_i - \boldsymbol{s}_j\|$ is the Euclidean distance between $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$, σ^2 is the partial sill that represents spatial covariance, and ϕ is the decay parameter that controls spatial range.

We simulate a total of 40 scenarios with a combination of different sample sizes and parameters, where $n_s = \{50, 100, 200, 500, 1000\}$, $n_a = \{50, 100, 200, 300\}$, and $(\sigma^2, \tau^2, \phi) = \{(0.1, 1.0, 300), (0.5, 0.5, 300)\}$. The coefficients are $\boldsymbol{\beta} = (1, 5)^T$, $\boldsymbol{\alpha} = (1, 2)^T$. We use the same random seed to generate the data such that the latent spatial process is consistent for the same set of spatial parameters. We can thus make a fair comparison when sample sizes vary. We consider the following three models: (i) a full Gaussian process (full GP) model based on point observations, as implemented in spBayes package (Finley et al., 2015); (ii) an NNGP model based on point observations, implemented in Stan; and, (iii) a fusion model under our new framework, implemented in Stan. For all models, the intercepts and coefficients are assigned with independent $N(0, 5^2)$ priors. The variance parameters σ^2 and τ^2 are assigned with inverse Gamma priors $IG(2, 1)$, which has a mean of one and infinite variance. For spatial decay, model (i) has a uniform prior $U(0.002, 0.01)$ on $1/\phi$, while model (ii) and (iii) have a weakly informative normal prior $N(300, 100^2)$ truncated at zero on ϕ. The priors in the fusion model can have arbitrary distributions with appropriate support, since Stan implementations do not rely on conjugacy between priors and likelihoods. We use the Stan implementation of model (ii) because using the alternative spNNGP package (Finley et al., 2017) implementation would prevent us from using the same priors as our fusion model, to make a fair comparison. To balance the trade-off between computational time and model accuracy, we use $m = 5$ nearest neighbors and $L = 5$ sampling points. The sampling points consist of the centroid plus four randomly selected locations for each area. For model (i) we run a single chain of 22,500 iterations with 2500 warm-up samples, and thin by a factor of 10. For model (ii) and (iii), we run 4 chains of 4000 iterations with 2000 warm-up samples, without thinning. Multiple chain convergence is checked with potential scale reduction factors (Brooks and Gelman, 1998).

Additional 900 sites are chosen under each scenario to evaluate predictive performance. The prediction sites are located on a 30×30 grid that uniformly covers the sampling domain. To compare model performance, we compute the width of 95% credible intervals and their coverage probabilities, and root mean squared prediction errors (RMSPE), defined as

$$\text{RMSPE} = \left(\frac{1}{n}\sum_{i=1}^{900}\left(w(\boldsymbol{u}_i) - \hat{w}(\boldsymbol{u}_i)\right)^2\right)^{1/2}. \tag{7}$$

4.2. Simulation results

We first look at the predictive performance of the models under different scenarios, as a function of sample size. Then we look closely into four specific scenarios from the simulation study to investigate performance on parameter estimation. See Appendix C for computation times.

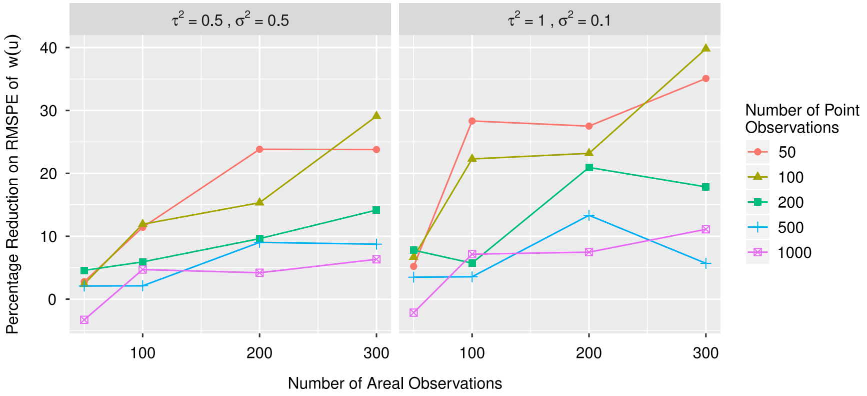


Fig. 4. Percentage reduction of RMSPE on latent spatial process $w(\boldsymbol{u})$ for different combinations of sample size. Left panel shows $\tau^2 = 0.5$ and $\sigma^2 = 0.5$ ($\tau^2/\sigma^2 = 1$). Right panel shows $\tau^2 = 1.0$ and $\sigma^2 = 0.1$ ($\tau^2/\sigma^2 = 10$).

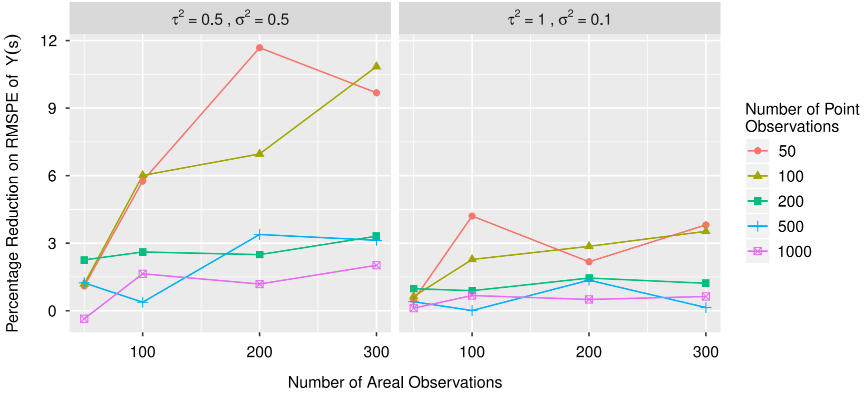


Fig. 5. Percentage reduction of RMSPE on point response $Y(\boldsymbol{s})$ for different combinations of sample size. Left panel shows $\tau^2 = 0.5$ and $\sigma^2 = 0.5$ ($\tau^2/\sigma^2 = 1$). Right panel shows $\tau^2 = 1.0$ and $\sigma^2 = 0.1$ ($\tau^2/\sigma^2 = 10$).

4.2.1. Predictive performance

Figs. 4 and 5 summarize the percentage reduction in RMSPE of the fusion model over NNGP; this can also be interpreted as percentage improvement in prediction. (The figures are in different scales.) The RMSPE of the full GP model is similar to NNGP, so we do not include it in the figures. We see an overall increase in the reduction of RMSPE as the number of areal observation increases. $w(\boldsymbol{u})$ improves slightly when the nugget-to-partial sill ratio is lower and the number of areal observations is small. In this situation, parameter estimation mainly relies on point data. Improvement is generally greater when the nugget-to-partial sill ratio is higher, with $\tau^2 = 1.0$ and $\sigma^2 = 0.1$. In both cases, improvement increases with the number of areal observations. The improvement of RMSPE for point response $Y(\boldsymbol{s})$ is less than that for $w(\boldsymbol{u})$ because much of the variance is modeled by the coefficients and the measurement error τ^2. There is little room for improvement by modeling the spatial component. The effect of latent spatial process on the point response in our simulation is small, so even though we saw great improvement in predicting $w(\boldsymbol{u})$, the effect on $Y(\boldsymbol{s})$ was small. The modeling of $Y(\boldsymbol{s})$ improves only when the nugget-to-partial sill ratio is low.

Overall, the fusion model improves the prediction of latent process $w(\boldsymbol{u})$ more noticeably than $Y(\boldsymbol{s})$. The amount of improvement depends on the relative sample size of point and areal observations,

Table 1
Four of the scenarios from the simulation study. They have common parameter values of $\boldsymbol{\beta} = (1, 5)^T$, $\boldsymbol{\alpha} = (1, 2)^T$, and $\phi = 300$.

Parameter	Scenario 1	Scenario 2	Scenario 3	Scenario 4
n_s	500	500	1000	1000
n_a	200	200	200	200
σ^2	0.5	0.1	0.5	0.1
τ^2	0.5	1.0	0.5	1.0

and on the nugget-to-partial sill ratio or the strength of spatial correlation. See Appendix D for a visualization of the estimated latent process from different models.

4.2.2. Parameter estimation

For parameter estimation, we focus on four different scenarios from the 40 combinations. Table 1 summarizes the parameters; Table 2 presents the results. For the parameter estimates, we show the posterior medians and 95% equal-tailed credible intervals (CI). We also include RMSPE, width of 95% credible intervals, and coverage probabilities for both $w(\boldsymbol{u})$ and $Y(\boldsymbol{s})$. All three models provide reasonable parameter estimates for coefficients $\boldsymbol{\beta}$, but only the fusion model simultaneously provided the estimate for the coefficients of areal covariate $\boldsymbol{\alpha}$. The models tend to overestimate the partial sill σ^2 in Scenario 2 and 4, where the true value is 0.1, in part because we assigned the prior $IG(2, 1)$ to the model. The prior allocates very small probability mass to low values and it has a mean of one, nevertheless the fusion model returned the most accurate estimates. To obtain more reasonable estimates, a different prior can be assigned to the model or σ^2 and τ^2 can be re-parameterized as a ratio (Cowles et al., 2009). We briefly investigate the influence of priors on estimating σ^2 in Appendix E. The spatial decay parameter ϕ is only weakly identified, hence we are not surprised it varies in the estimation across the scenarios. The overall estimation of the spatial parameter is improved for Scenario 1 and 2 in the fusion model, when there is relatively little information from the point observations. The width of the posterior credible interval for $w(\boldsymbol{u})$ in the fusion models is smaller in all scenarios, and especially in Scenarios 1 and 2, which indicates predictions can be achieved with lower uncertainty when some areal observations are taken into account.

5. Case study: LuftiBus-SNC dataset

To investigate the spatial risk pattern for respiratory diseases and lung cancer in Canton of Zurich, we use our new framework to fit a spatial fusion model. The lung function measurement FEV1 $Y(\boldsymbol{s})$ is recorded at the individual level in the LuftiBus-SNC dataset, and is assumed to be Gaussian distributed. We analyze it with Poisson distributed cause-specific (respiratory diseases and lung cancer) mortality $Q(\boldsymbol{a})$ from the SNC data at the municipality level. To maintain consistency with the selected LuftiBus-SNC dataset, we only consider mortality from 2010. Expected mortality $E(\boldsymbol{a})$ for each municipality is computed based on the reference mortality from the same population, which is grouped by gender and 5-year age categories. If we assume that the latent spatial process of risk is associated with both decreased FEV1 and increased cause-specific mortality, the first level of the hierarchical model can be formulated as

$$\begin{aligned} Y(\boldsymbol{s})|w(\boldsymbol{s}) &\sim N\left(\beta_0 + \text{age}(\boldsymbol{s}) \times \beta_1 + \text{gender}(\boldsymbol{s}) \times \beta_2 - w(\boldsymbol{s}), \tau^2\right), \\ Q(\boldsymbol{a})|w(\boldsymbol{a}) &\sim \text{Poisson}\Big(\exp\big(\log E(\boldsymbol{a}) + w(\boldsymbol{a})\big)\Big). \end{aligned} \tag{8}$$

We also assume an exponential covariance function for the process. We choose priors $\sigma^2 \sim IG(2, 0.1)$, $\tau^2 \sim IG(2, 1)$, and $\phi \sim N(700, 100^2)$ in meters, based on the exploratory analysis in Section 2. The coefficients $\boldsymbol{\beta}$ are given independent normal priors $N(0, 5^2)$. We run the model with 4 chains of 4000 iterations, including 2000 warm-up iterations. All of the potential scale reduction factors are under 1.1, so approximate convergence is reached. The trace plots are shown in Appendix F. Parameter estimates are summarized in Table 3. Age and gender have a significant effect on lung function FEV1, consistent with the reference equations (Kerstjens et al., 1997; Kuster et al., 2008).

Table 2
Parameter estimation and model performance summary from the simulation study. The parameter estimates are based on the median, 2.5th and 97.5th percentiles from the posterior distributions. RMSPE indicates root mean squared prediction error. CI width and coverage are calculated based on 95% credible intervals of $w(\boldsymbol{u})$ and $Y(\boldsymbol{s})$.

	Scenario 1			Scenario 2		
	Full GP	NNGP	Fusion	Full GP	NNGP	Fusion
β_0	0.84 (0.60,1.10)	0.86 (0.61,1.12)	0.90 (0.65,1.17)	0.92 (0.80,1.05)	0.93 (0.76,1.09)	0.94 (0.77,1.12)
β_1	5.03 (4.96,5.11)	5.04 (4.96,5.11)	5.02 (4.95,5.09)	5.04 (4.96,5.13)	5.04 (4.96,5.13)	5.01 (4.95,5.12)
α_0	–	–	0.83 (0.57,1.12)	–	–	0.98 (0.80,1.17)
α_1	–	–	2.03 (1.92,2.14)	–	–	1.97 (1.88,2.06)
σ^2	0.48 (0.31,0.70)	0.48 (0.32,0.71)	0.53 (0.39,0.73)	0.26 (0.13,0.48)	0.21 (0.12,0.53)	0.16 (0.10,0.25)
τ^2	0.48 (0.34,0.62)	0.50 (0.38,0.61)	0.49 (0.40,0.59)	0.81 (0.60,0.99)	0.86 (0.54,1.02)	0.93 (0.81,1.06)
ϕ	294 (174,482)	345 (230,503)	398 (289,542)	127 (100,305)	262 (52,460)	380 (232,553)
$w(\boldsymbol{u})$						
RMSPE	0.5250	0.5272	0.4796	0.3124	0.3045	0.2639
CI width	2.01	1.93	1.78	1.89	1.69	1.17
Coverage	94.3	93.7	92.8	99.8	99.9	97.8
$Y(\boldsymbol{s})$						
RMSPE	0.8771	0.8829	0.8530	1.045	1.045	1.031
CI width	3.347	3.341	3.242	3.98	3.98	3.95
Coverage	95.4	95.4	95.8	95.0	95.4	95.1
	Scenario 3			Scenario 4		
	Full GP	NNGP	Fusion	Full GP	NNGP	Fusion
β_0	0.89 (0.62,0.16)	0.94 (0.69,1.19)	0.95 (0.69,1.23)	0.95 (0.80,1.10)	0.97 (0.81,1.13)	0.97 (0.82,1.14)
β_1	4.97 (4.92,5.02)	4.97 (4.95,5.02)	4.98 (4.93,5.03)	4.95 (4.89,5.02)	4.95 (4.89,5.02)	4.96 (4.94,5.02)
α_0	–	–	0.92 (0.65,1.22)	–	–	0.95 (0.78,1.13)
α_1	–	–	1.97 (1.88,2.06)	–	–	1.99 (1.91,2.07)
σ^2	0.51 (0.38,0.70)	0.54 (0.40,0.73)	0.47 (0.40,0.70)	0.18 (0.12,0.29)	0.18 (0.12,0.28)	0.15 (0.10,0.23)
τ^2	0.50 (0.42,0.57)	0.49 (0.43,0.57)	0.51 (0.44,0.57)	0.91 (0.82,1.01)	0.93 (0.83,1.03)	0.94 (0.86,1.04)
ϕ	362 (251,472)	373 (269,513)	397 (298,528)	284 (140,467)	356 (199,537)	397 (257,567)
$w(\boldsymbol{u})$						
RMSPE	0.4522	0.4478	0.4290	0.2622	0.2596	0.2402
CI width	1.75	1.76	1.66	1.36	1.29	1.08
Coverage	93.3	93.4	93.8	99.0	98.6	97.1
$Y(\boldsymbol{s})$						
RMSPE	0.8326	0.8343	0.8244	1.036	1.037	1.032
CI width	3.23	3.24	3.21	3.97	3.98	3.95
Coverage	94.4	94.6	94.4	94.9	94.9	94.4

Table 3
Parameter estimates from the case study, with posterior median and 95% equal-tailed CIs.

Parameter	Posterior median	95% CI
β_0	5.640	(5.528, 5.757)
β_1	-0.037	(-0.039, -0.035)
β_2	-0.919	(-0.964, -0.876)
σ^2	0.017	(0.010, 0.026)
τ^2	0.261	(0.245, 0.278)
ϕ	607	(431, 805)

The nugget-to-partial sill ratio is high, indicating weak spatial correlation. This is consistent with the estimates from the empirical semi-variogram. Effective range is about 1.8 km.

We select 10,000 gridded prediction locations within Canton of Zurich to obtain a continuous spatial risk map. The posterior median estimates, and 95% CI width at the locations are shown in Fig. 6. The risk map represents the unexplained spatial component in the model, after taking age and gender into account. On the left, red regions indicate higher value of the estimated latent spatial process,

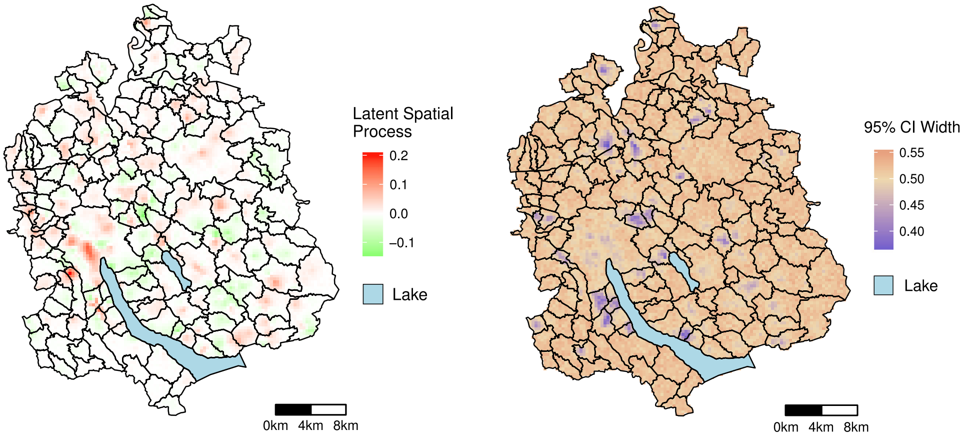


Fig. 6. Posterior median estimates and 95% CI width of the spatial risk process for respiratory disease and lung cancer in Canton of Zurich, after adjusting for age and gender. (For interpretation of the references to colour in this figure, the reader is referred to the web version of this article.)

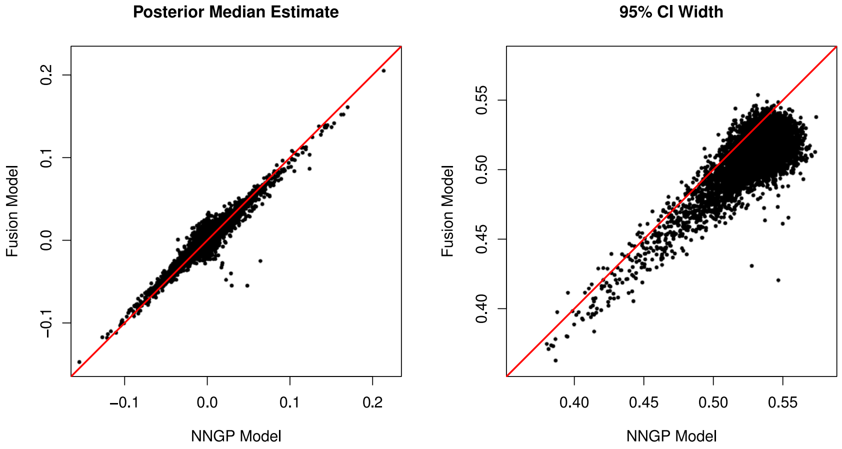


Fig. 7. Comparison of the latent spatial process estimates from the NNGP and the fusion model, with posterior median estimates and 95% CI width on 10'000 gridded prediction locations. $y = x$ diagonal lines are red.

corresponding to higher risk. North-west regions have slightly higher risk, while south-east regions have lower risk. There are a few clusters of higher risk near the city center, but none of those risk estimates are significantly different from zero, based on 95% CIs.

Finally, we compare the estimated spatial pattern with the results from the NNGP model in Fig. 7. The posterior median estimates at those prediction locations are similar for both models. We obtain a slightly smoother map from the fusion model, since the estimates are above the diagonal line at small values and below the line at large values. There are a few locations where the fusion model estimated a much lower value than the NNGP model. At those locations, the point observations are sparsely distributed, so the information mainly come from areal mortality observations. The width of 95% CIs is lower in the fusion model, indicating the estimated map has a smaller uncertainty.

6. Discussion

We proposed a generalized spatial fusion model framework that jointly analyzes both point and areal data. The framework offers several advantages over existing alternatives. First, it does not require point and areal observations be Gaussian distributed. This makes it suitable for a much wider range of applications, particularly in epidemiology, where areal observations are usually Poisson-distributed counts. Second, the framework is relatively easy to be adapted. It is implemented in the Stan modeling language, which offers computational gain and intuitive syntax. Users can avoid constructing custom samplers for MCMC, so the models can be modified easily. We use NNGP to model the latent process $w(\boldsymbol{u})$, such that the framework is viable for larger datasets because the computation grows linearly with the number of observations.

The simulation study showed that the fusion model outperforms full GP and NNGP models. Improvement is great when the number of point observations is small relative to the number of areal observations, and when the nugget-to-partial sill ratio is high. However, when we look at the prediction of $Y(\boldsymbol{s})$, the improvement diminishes if the nugget-to-partial sill ratio is high, this is because adding areal data does not contribute much additional information on estimating the coefficients for point data. If interest is in modeling $Y(\boldsymbol{s})$ directly as a spatial process without covariates, we expect the fusion model to also greatly improve the prediction of $Y(\boldsymbol{s})$. In the simulation study, we used $IG(2, 1)$ priors for σ^2 and τ^2, where 95% of the probability mass lies between 0.18 and 4.16. In Scenario 1 and 3 of the simulation study, there is a clear preference for low σ^2, since the posterior distribution of all three models departs from the prior. When a more informative prior is assigned to the models, such as $IG(2, 0.1)$ for σ^2, the difference between model performance diminishes because the relative contribution of the prior is greater (see Appendix E). The simulation results demonstrated that the fusion model can reduce bias in parameter estimation, even if the prior information is misleading. We chose the true spatial decay ϕ to be 300, which corresponds to an effective range of 900 based on the exponential covariance function. This range is generally large compared to area sizes in our simulation study, allowing the areal data to help inferring ϕ as well as the latent process. When the spatial range is small compared to area sizes, the areal data can borrow information about ϕ from the point data and still help inferring the latent process, albeit to a lesser extent.

In the case study, we considered only age and gender information in the models. Assuming a common latent spatial process, the resulting map represents the unexplained spatial risk component for respiratory diseases and lung cancer. This map can be used as a model-based exploratory tool to guide further research in epidemiology. For example, it can be used jointly with cluster analysis to identify potential risk factors, such as smoking prevalence, social–economic grouping, or environmental pollution.

Convergence problems can arise in some situations. When the nugget-to-partial sill ratio is high, it is particularly difficult to identify spatial parameters. This causes a problem not only in Bayesian models but also in likelihood methods (Irvine et al., 2007). In practice, the spatial range parameter can be fixed, or informative priors around the maximum likelihood estimate can be assigned. We have experimented with different parameterizations and prior combinations, and our current implementation converged nicely in all of the simulated scenarios, as it did in the case study. Some modifications may be necessary in completely different settings, when, for example, the nugget-to-partial sill ratio is very low.

We made choices about implementing our framework. We used NNGP because it offers a computational advantage. If the number of observations is relatively small, the latent spatial process can be modeled with a full Gaussian process instead of an NNGP. An alternative approach is to use a Gaussian Markov random field representation of $w_{\mathcal{U}}$ and exploit the resulting sparse precision matrix. When we approximated stochastic integrals, we followed Fuentes and Raftery (2005), Berrocal et al. (2010) and Liu et al. (2011) in our choice of four sampling points. This is appropriate if the area size is relatively small, and the spatial process in each area is not very heterogeneous. If domains have uneven area sizes or drastic changes in the spatial process, one can set the number of sampling points to be proportional to the area size to reflect the spatial process inside each area. A more flexible strategy for selecting sampling points can lead to further improvements. Further, in modeling the areal response $Q(\boldsymbol{a})$, we used a disease mapping approach to link its mean with aggregated spatial process $w(\boldsymbol{a})$. The ecological

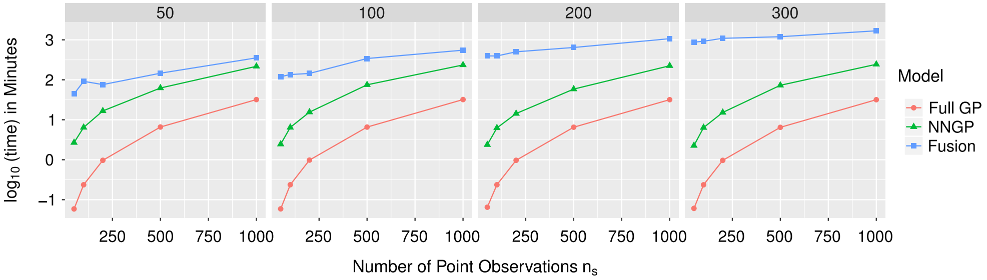


Fig. A.1. Computation time for different combination of sample sizes. Each column represents a different number of areal observations n_a. The computation time for full GP grows linearly with n_s^3, NNGP grows linearly with n_s, while the fusion model grows linearly with $n_s + n_a L$.

bias (Greenland, 1992) introduced by this approach can be addressed by firstly transforming the process depending on the link function, then do the aggregation.

The generalized spatial fusion framework we proposed should make fusion models more applicable when spatial information is available on both point and areal support. For instance, it can be used for proportional mortality analysis that consider two or more diseases, where the mortality of a single disease follows a binomial or multinomial distribution (Lawson, 2013). It can also be used to map species abundance where observations at both point and areal level follow Poisson distributions. Many real world data have multiple responses. For instance, we used FEV1 as the lung function measure, but that is one among many spirometry measurements. To explore the relationship between multiple responses and model their dependencies, an extension to multivariate spatial fusion models is required. Spatio-temporal fusion models can also be proposed in the future to incorporate temporal dependencies.

We proposed a generalized spatial fusion model framework, and implemented it with flexibility and adequate computational efficiency. The framework can handle point and areal data that are not Gaussian distributed. We hope to encourage the usage of spatial fusion models on a much wider range of practical problems when suitable data are available.

Acknowledgments

We thank the Swiss Federal Statistical Office for providing mortality and census data and for the support which made the Swiss National Cohort and the case study possible. The work was supported by the Swiss National Science Foundation (grant nos. 3347CO-108806, 33CS30_134273, 33CS30_148415 and 175529). The members of the Swiss National Cohort Study Group are Matthias Egger (Chairman of the Executive Board), Adrian Spoerri and Marcel Zwahlen (all Bern), Milo Puhan (Chairman of the Scientific Board), Matthias Bopp (both Zurich), Nino Künzli (Basel), Michel Oris (Geneva) and Murielle Bochud (Lausanne).

Appendix A. Exclusion criteria for LuftiBus-SNC data

To reduce data recording errors and to focus on the population of interest, we apply the following exclusion criteria to the participants from LuftiBus-SNC data:

- living outside of Canton of Zurich at the time of measurement,
- without linkage with census 2000 and mortality records,
- BMI outside the range of 15–50,
- and age under 40 years old.

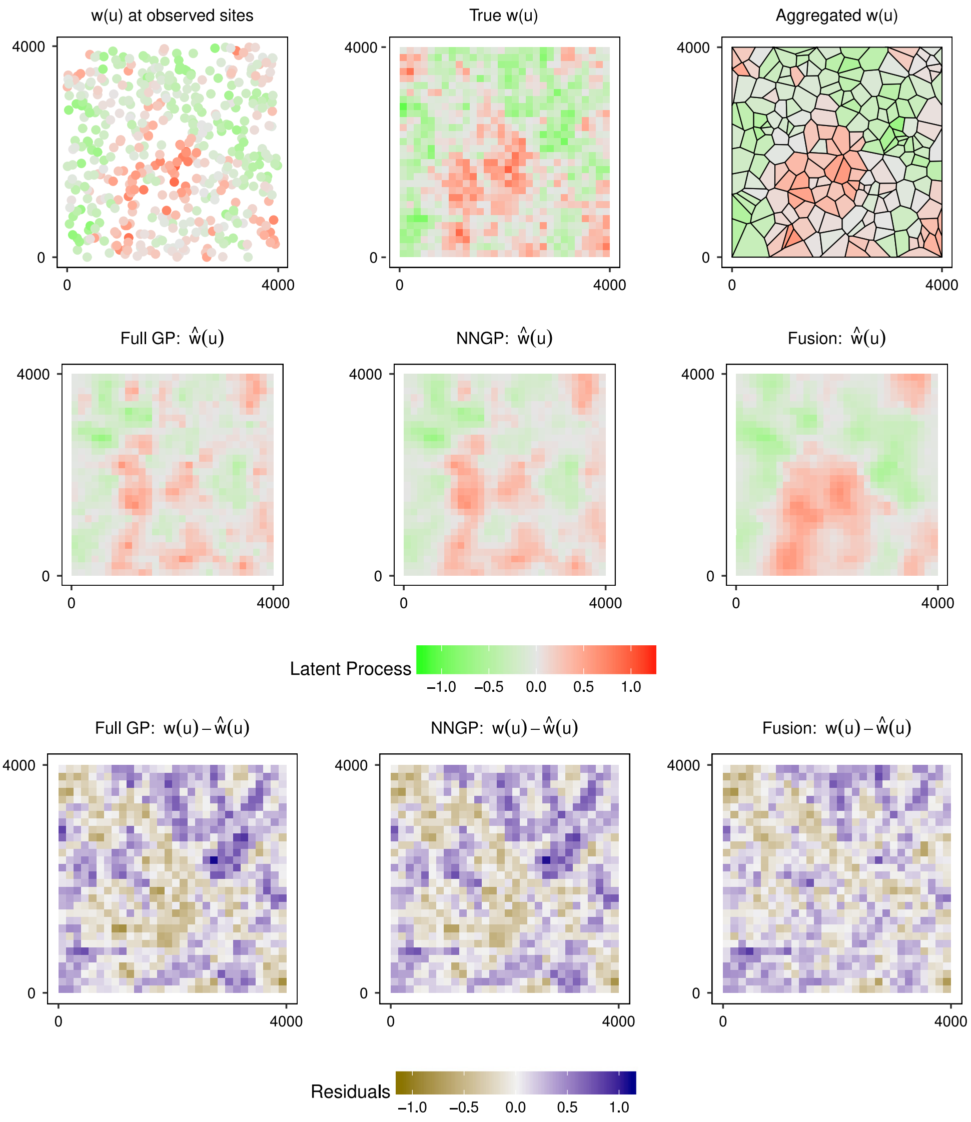


Fig. A.2. Modeling results from Scenario 2 in Table 1. First row shows a realization of the process at observed locations, interpolated process, and aggregated process. Second row shows the estimated latent process from the models. Third row shows the difference between estimated process and true process; gray regions represent the most accurate predictions. The results from the fusion model most closely resemble the true latent spatial process.

Appendix B. Source files

A zip file is available at: www.math.uzh.ch/furrer/download/fusion2017_spatstat_appendixB.zip. It contains the following source files.

- README.txt: description of file contents and references
- model_simulation.R: data generation
- model_fit.R: prediction
- fusion.stan: implementation of our fusion model in Stan.

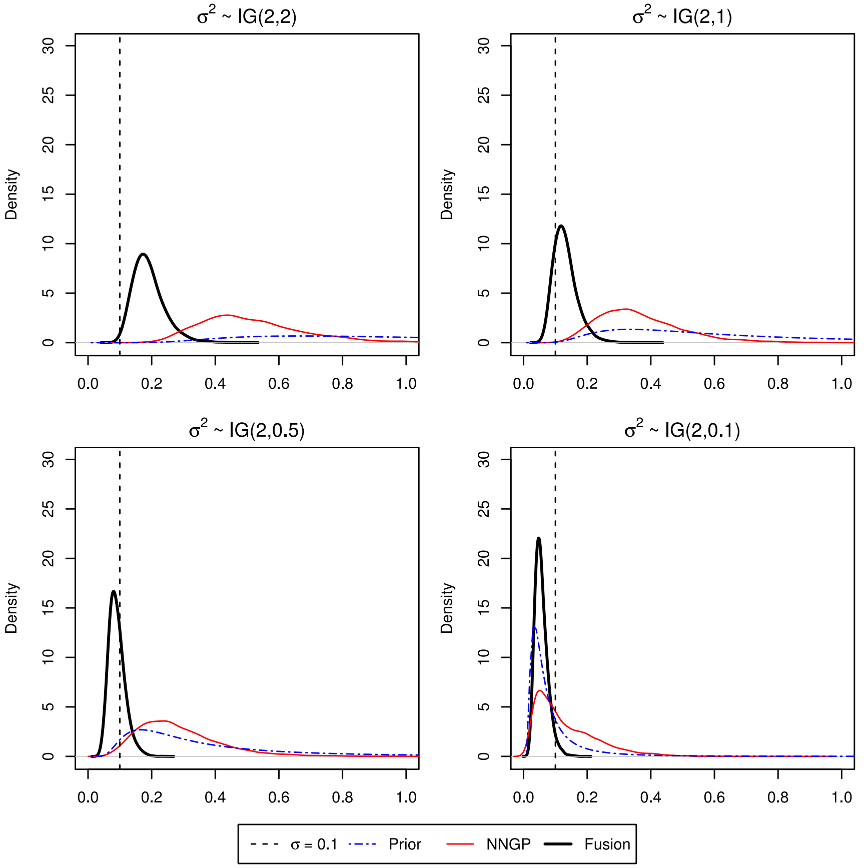


Fig. A.3. The influence of four different priors for the partial sill on the posterior distribution. The prior $IG(2, 1)$ was used in the simulation study.

Appendix C. Computation time with varying sample size

The NNGP model used in our simulation study is implemented in Stan without marginalizing the spatial process $w(\boldsymbol{u})$, hence the computation is generally slower than the marginalized full GP model implemented in the spBayes package (see Fig. A.1). We expect decreases in the computation time if the NNGP model is also marginalized.

We conduct the simulation study with the un-marginalized NNGP in order to aid the comparison with our generalized fusion model, where marginalization of $w(\boldsymbol{u})$ is not possible.

Appendix D. Visualization of latent process estimation

See Fig. A.2.

Appendix E. Influence of priors on the simulation study

In addition to the simulation study, we investigate the influence of priors on σ^2. An inverse Gamma prior $IG(2, 1)$ was chosen for the partial sill in all simulated scenarios, and we observed some over-estimation of σ^2 when the true value is 0.1. In the following, we simulate with $n_s = 200, n_a =$

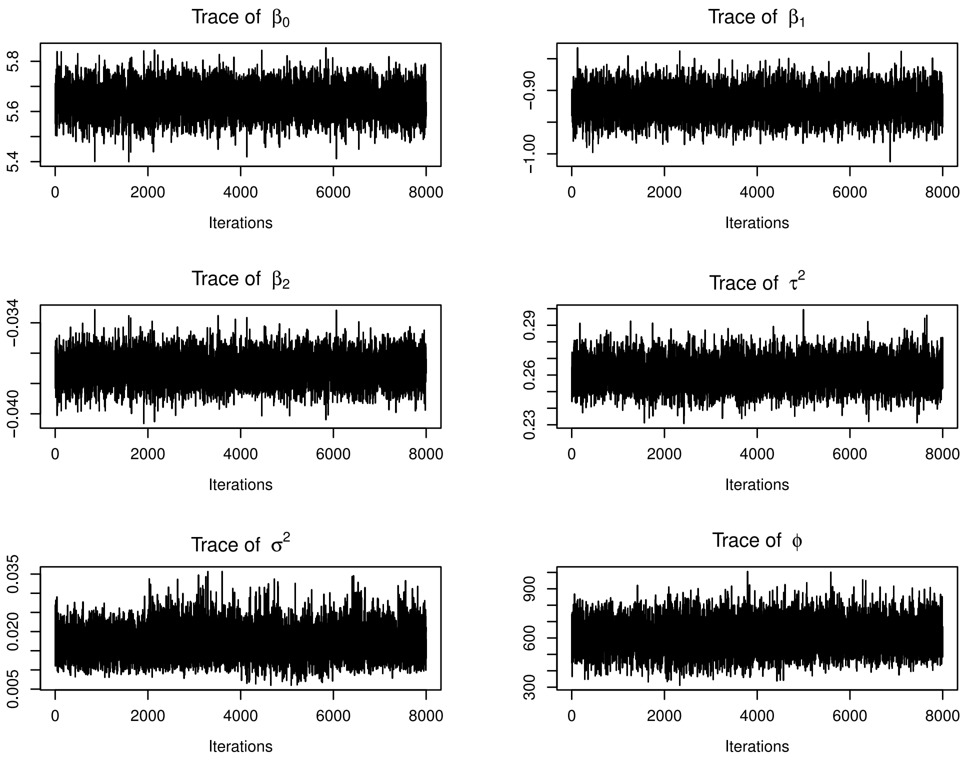


Fig. A.4. Trace plots of the parameters from the case study. The samples from four chains are concatenated.

100, $\sigma^2 = 0.1$, $\tau^2 = 1$, and $\phi = 300$. The coefficients $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are the same as in the simulation study in Section 4.

We employed four different priors for σ^2 to fit our fusion model and the NNGP model separately, while leaving the other priors unchanged. Fig. A.3 shows that, as the mean of priors decrease, the posterior distributions shift towards lower values. The posterior distribution from the fusion model is robust under different prior specifications, and shows much smaller shifts than the NNGP model. With the prior $IG(2, 0.1)$, the posteriors of σ^2 from the two models have a large overlapping region.

Appendix F. Trace plots from the case study

See Fig. A.4.

References

Banerjee, S., Carlin, B.P., Gelfand, A.E., 2014. Hierarchical Modeling and Analysis for Spatial Data. CRC Press, p. 136-139.

Banerjee, S., Gelfand, A.E., Finley, A.O., Sang, H., 2008. Gaussian predictive process models for large spatial data sets. J. R. Stat. Soc. Ser. B Stat. Methodol. 70 (4), 825–848. http://dx.doi.org/10.1111/j.1467-9868.2008.00663.x.

Berrocal, V.J., Gelfand, A.E., Holland, D.M., 2010. A spatio-temporal downscaler for output from numerical models. J. Agric. Biol. Environ. Stat. 15 (2), 176–197. http://dx.doi.org/10.1007/s13253-009-0004-z.

Bourgeois, A., Gaba, S., Munier-Jolain, N., Borgy, B., Monestiez, P., Soubeyrand, S., 2012. Inferring weed spatial distribution from multi-type data. Ecol. Modell. 226 (Supplement C), 92–98. http://dx.doi.org/10.1016/j.ecolmodel.2011.10.010.

Brooks, S.P., Gelman, A., 1998. General methods for monitoring convergence of iterative simulations. J. Comput. Graph. Statist. 7 (4), 434–455. http://dx.doi.org/10.2307/1390675.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., Riddell, A., 2017. Stan: A probabilistic programming language. J. Stat. Softw. 76 (1). http://dx.doi.org/10.18637/jss.v076.i01.

Chammartin, F., Probst-Hensch, N., Utzinger, J., Vounatsou, P., 2016. Mortality atlas of the main causes of death in Switzerland, 2008-2012. Swiss Med. Weekly 146, w14280. http://dx.doi.org/10.4414/smw.2016.14280.

Cowles, M.K., Yan, J., Smith, B., 2009. Reparameterized and marginalized posterior and predictive sampling for complex Bayesian geostatistical models. J. Comput. Graph. Statist. 18 (2), 262–282. http://dx.doi.org/10.1198/jcgs.2009.08012.

Datta, A., Banerjee, S., Finley, A.O., Gelfand, A.E., 2016. Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. J. Amer. Statist. Assoc. 111 (514), 800–812. http://dx.doi.org/10.1080/01621459.2015.1044091.

Diggle, P.J., Tawn, J., Moyeed, R., 1998. Model-based geostatistics. J. R. Stat. Soc. Ser. C. Appl. Stat. 47 (3), 299–350. http://dx.doi.org/10.1111/1467-9876.00113.

Finley, A.O., Banerjee, S., Gelfand, A.E., 2015. spBayes for large univariate and multivariate point-referenced spatio-temporal data models. J. Stat. Softw. 63 (13), 1–28. http://dx.doi.org/10.18637/jss.v063.i13.

Finley, A.O., Datta, A., Banerjee, S., 2017. spNNGP: Spatial Regression Models for Large Datasets using Nearest Neighbor Gaussian Processes, URL: https://CRAN.R-project.org/package=spNNGP R package version 0.1.1.

Fuentes, M., Raftery, A.E., 2005. Model evaluation and spatial interpolation by Bayesian combination of observations with outputs from numerical models. Biometrics 61 (1), 36–45. http://dx.doi.org/10.1111/j.0006-341X.2005.030821.x.

Furrer, R., Genton, M.G., Nychka, D., 2006. Covariance tapering for interpolation of large spatial datasets. J. Comput. Graph. Statist. 15 (3), 502–523. http://dx.doi.org/10.1198/106186006X132178.

Gelfand, A.E., Banerjee, S., Gamerman, D., 2005. Spatial process modelling for univariate and multivariate dynamic spatial data. Environmetrics 16 (5), 465–479. http://dx.doi.org/10.1002/env.715.

Goovaerts, P., 2010. Combining areal and point data in geostatistical interpolation: applications to soil science and medical geography. Math. Geosci. 42 (5), 535–554. http://dx.doi.org/10.1007/s11004-010-9286-5.

Greenland, S., 1992. Divergent biases in ecologic and individual-level studies. Stat. Med. 11 (9), 1209–1223. http://dx.doi.org/10.1002/sim.4780110907.

Homan, M.D., Gelman, A., 2014. The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. J. Mach. Learn. Res. 15 (1), 1593–1623. URL http://dl.acm.org/citation.cfm?id=2627435.2638586.

Huang, G., Lee, D., Scott, M., 2015. An integrated Bayesian model for estimating the long-term health effects of air pollution by fusing modelled and measured pollution data: a case study of nitrogen dioxide concentrations in Scotland. Spat. Spatio-Temporal Epidemiol. 14, 63–74. http://dx.doi.org/10.1016/j.sste.2015.09.002.

Irvine, K.M., Gitelman, A.I., Hoeting, J.A., 2007. Spatial designs and properties of spatial correlation: Effects on covariance estimation. J. Agric. Biol. Environ. Stat. 12 (4), 450–469. http://dx.doi.org/10.1198/108571107X249799.

Kerstjens, H., Rijcken, B., Schouten, J., Postma, D., 1997. Decline of FEV1 by age and smoking status: facts, figures, and fallacies. Thorax 52 (9), 820–827. http://dx.doi.org/10.1136/thx.52.9.820.

Kuster, S.P., Kuster, D., Schindler, C., Rochat, M.K., Braun, J., Held, L., Brändli, O., 2008. Reference equations for lung function screening of healthy never-smoking adults aged 18–80 years. Eur. Respir. J. 31 (4), 860–868. http://dx.doi.org/10.1183/09031936.00091407.

Lawson, A.B., 2013. Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology. CRC Press.

Lawson, A.B., Banerjee, S., Haining, R.P., Ugarte, M.D., 2016. Handbook of Spatial Epidemiology. CRC Press.

Lee, D., Mukhopadhyay, S., Rushworth, A., Sahu, S.K., 2017. A rigorous statistical framework for spatio-temporal pollution prediction and estimation of its long-term impact on health. Biostatistics 18 (2), 370. http://dx.doi.org/10.1093/biostatistics/kxw048.

Liu, Z., Le, N.D., Zidek, J.V., 2011. An empirical assessment of Bayesian melding for mapping ozone pollution. Environmetrics 22 (3), 340–353. http://dx.doi.org/10.1002/env.1054.

Lunn, D., Spiegelhalter, D., Thomas, A., Best, N., 2009. The BUGS project: Evolution, critique and future directions. Stat. Med. 28 (25), 3049–3067. http://dx.doi.org/10.1002/sim.3680.

McMillan, N.J., Holland, D.M., Morara, M., Feng, J., 2010. Combining numerical model output and particulate data using Bayesian spacetime modeling. Environmetrics 21 (1), 48–65. http://dx.doi.org/10.1002/env.984.

Menezes, A.M.B., Pérez-Padilla, R., Wehrmeister, F.C., Lopez-Varela, M.V., Muiño, A., Valdivia, G., Lisboa, C., Jardim, J.R.B., de Oca, M.M., Talamo, C., Bielemann, R., Gazzotti, M., Laurenti, R., Celli, B., Victora, C.G., for the PLATINO team, 2014. FEV1 is a better predictor of mortality than FVC: The PLATINO cohort study. PLoS One 9 (10), 1–10. http://dx.doi.org/10.1371/journal.pone.0109732.

Moraga, P., Cramb, S.M., Mengersen, K.L., Pagano, M., 2017. A geostatistical model for combined analysis of point-level and area-level data using INLA and SPDE. Spat. Stat. http://dx.doi.org/10.1016/j.spasta.2017.04.006.

Murakami, D., Tsutsumi, M., 2015. Area-to-point parameter estimation with geographically weighted regression. J. Geog. Syst. 17 (3), 207–225. http://dx.doi.org/10.1007/s10109-015-0212-8.

Nguyen, H., Cressie, N., Braverman, A., 2012. Spatial statistical data fusion for remote sensing applications. J. Amer. Statist. Assoc. 107 (499), 1004–1018. http://dx.doi.org/10.1080/01621459.2012.694717.

Paci, L., Beamonte, M.A., Gelfand, A.E., Gargallo, P., Salvador, M., 2017. Analysis of residential property sales using space–time point patterns. Spat. Stat. 21, 149–165. http://dx.doi.org/10.1016/j.spasta.2017.06.007.

R Core Team,, 2017. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL https://www.R-project.org/.

Ribeiro Jr., P.J., Diggle, P.J., 2016. geoR: Analysis of Geostatistical Data. URL https://CRAN.R-project.org/package=geoR R package version 1.7-5.2.

Rue, H., Martino, S., Chopin, N., 2009. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. J. R. Stat. Soc. Ser. B Stat. Methodol. 71 (2), 319–392. http://dx.doi.org/10.1111/j.1467-9868.2008.00700.x.

Sahu, S.K., Gelfand, A.E., Holland, D.M., 2010. Fusing point and areal level space-time data with application to wet deposition. J. Roy. Statist. Soc. Ser. C 59 (1), 77–103. http://dx.doi.org/10.1111/j.1467-9876.2009.00685.x.

Shi, T., Cressie, N., 2007. Global statistical analysis of MISR aerosol data: a massive data product from NASA's Terra satellite. Environmetrics 18 (7), 665–680. http://dx.doi.org/10.1002/env.864.

Stan Development Team,, 2016. RStan: the R interface to Stan. URL http://mc-stan.org/ R package version 2.14.1.

Stein, M.L., 2008. A modeling approach for large spatial datasets. J. Korean Stat. Soc. 37 (1), 3–10. http://dx.doi.org/10.1016/j.jkss.2007.09.001.

Strassmann, A., Steurer-Stey, C., Lana, K.D., Zoller, M., Turk, A.J., Suter, P., Puhan, M.A., 2013. Population-based reference values for the 1-min sit-to-stand test. Int. J. Public Health 58 (6), 949–953. http://dx.doi.org/10.1007/s00038-013-0504-z.

The SNC Study Group, 2017. The Swiss National Cohort. URL http://www.swissnationalcohort.ch (Accessed: 13-09-17).

Lunge Zürich, 2017. The LuftiBus Project. URL: http://www.lunge-zuerich.ch/de/projekte/luftibus/ (Accessed: 28-02-17).