Appendix: INLA for GMRFs (e.g. GLMMs, Spatial Models) with Spatial Examples of Leukemia Cases and Heavy Metal Concentrations

Frances Lin

A1. Spatial Examples Using The `R-INLA` Package

Two data sets are considered. The `NY8` data set is areal data. The `meuse` data set is point-referenced (or geostatistical) data. Since the focus of this part of the project is on the use of `R-INLA` package, the details of the areal and geostatistical models used for each data set are only briefly introduced here.

A1.1. A Spatial Areal Example of Leukemia Incident Cases

The `NY8` data set contains the number of incident leukemia cases per census tract in an eight-country region of upstate New York from 1978-1982 (Waller & Gotway, 2004; Bivand et al., 2008). The `NY8` data set can be accessed from the **R** package `DClusterm`, and it is a `SpatialPolygonsDataFrame` object.

For this data set, a total of 5 models (fixed effects, random effects (iid), ICAR, BYM and Leroux et al.) are fitted, and results include criteria for model selection (marginal log-likelihood, DIC and WAIC).

Since the number of incident leukemia cases is count, Poisson GLMs with fixed effects and random effects are fitted. Since there is spatial dependence in the data, spatial models (GLMs with spatial random effects) such as ICAR (Intrinsic Conditional autoregressive), BYM and Leroux et al. model are also considered.

Without going into details, recall that the GLMs have the following form

$$Y = X\beta + Z\alpha + \varepsilon,$$

where β is a vector of fixed effects with design matrix X, α is a vector of random effects with design matrix Z, and ε is an error term, where it is assumed that $\varepsilon_i \sim N(0, \sigma^2), i = 1, ..., n$. The vector of random effects α is modeled as MVN (it is assumed that)

$$\alpha \sim N(0, \sigma_\alpha^2 \Sigma),$$

where the covariance matrix Σ is defined such that it induces higher correlation with adjacent areas.

There are several ways to include spatial dependence in Σ, and in spatial areal model especially, it is more common to model the precision matrix Q directly, where $Q = \Sigma^{-1}$. In ICAR (Intrinsic CAR), $\Sigma^{-1} = diag(n_i) - W$, where n_i is the number of neighbors of area i. In Leroux et al.'s model (mixture

of matrices), $\Sigma^{-1} = ((1-\lambda)I_n + \lambda M), \lambda \in (0,1)$, where M is precision of intrinsic CAR specification. The BYM (Besag, York and Mollié) model includes two latent random effects: an ICAR latent effect and a Gaussian iid latent effect.

Results show that for spatially dependent data, spatial models generally perform better than GLM with fixed or random (iid) effects. It is also no surprise that the baseline model (fixed effects model) appears to be the poorest fit of all.

Table 1: criteria for model selection

	Marg_logLik	DIC	WAIC
fixed	-514.4	1016	1017
iid	-512.1	979.2	983.6
ICAR	-718.9	968.3	972.2
BYM	-458.9	967.4	971.9
Leroux	-508.3	967.3	971.3

A1.2. A Spatial Geostatistical Example of Heavy Metal Concentrations

The `meuse` data set contains locations, topsoil heavy metal concentrations and a number of soil and landscape variables at the observed locations in a flood plain of the river Meuse. The `meuse` data set can be accessed from **R** package `gstat`, and it is a `data.frame` object.

For this data set, an universal kriging model and a continuous spatial process with a Matérn covariance function using the INLA-SPDE approach (referred to as the SPDE model) are fitted, and results include summary results and predicted plots of kriging vs SPDE.

The steps for fitting an universal kriging model include: (1) convert the `data.frame` objaction to a `SpatialPointsDataFrame` object, (2) calculate the empirical/sample variogram and fit a spherical variogram model to the sample variogram, and (3) fit the universal kriging model using the fitted variogram. The steps for fitting the SPDE model are more involved. These steps include: (1) define the boundary and create a mash to approximate the continuous GF (Gaussian field) as a discrete GMRF (Gaussian Markov random field), (2) make the latent model/create a SPDE model on the mesh, (3) make a projection matrix A to map the GF from the observed points to the triangulation vertices, (4) organize the data to be in a particular format for estimation and prediction and join stacks of data, (5) fit the model.

Results show that the universal kriging and the SPDE model provide similar estimates. For the fitted variogram model, sill (or variance) = 0.2053, nugget (i.e., variance when distance = 0) = 0.07643, and range (i.e. the distance after which the variogram levels off) = 728.7. For the SPDE model, mean variance = 0.2351 with $Q1 = 0.183$ and $Q3 = 0.2754$, and max range = 955.8 (This value is a bit higher than expected).

Table 2: sill, nugget and range from fitted variogram

model	psill	range	kappa	ang1	ang2	ang3	anis1	anis2
Nug	0.07643	0	0	0	0	0	1	1
Sph	0.2053	728.7	0.5	0	0	0	1	1

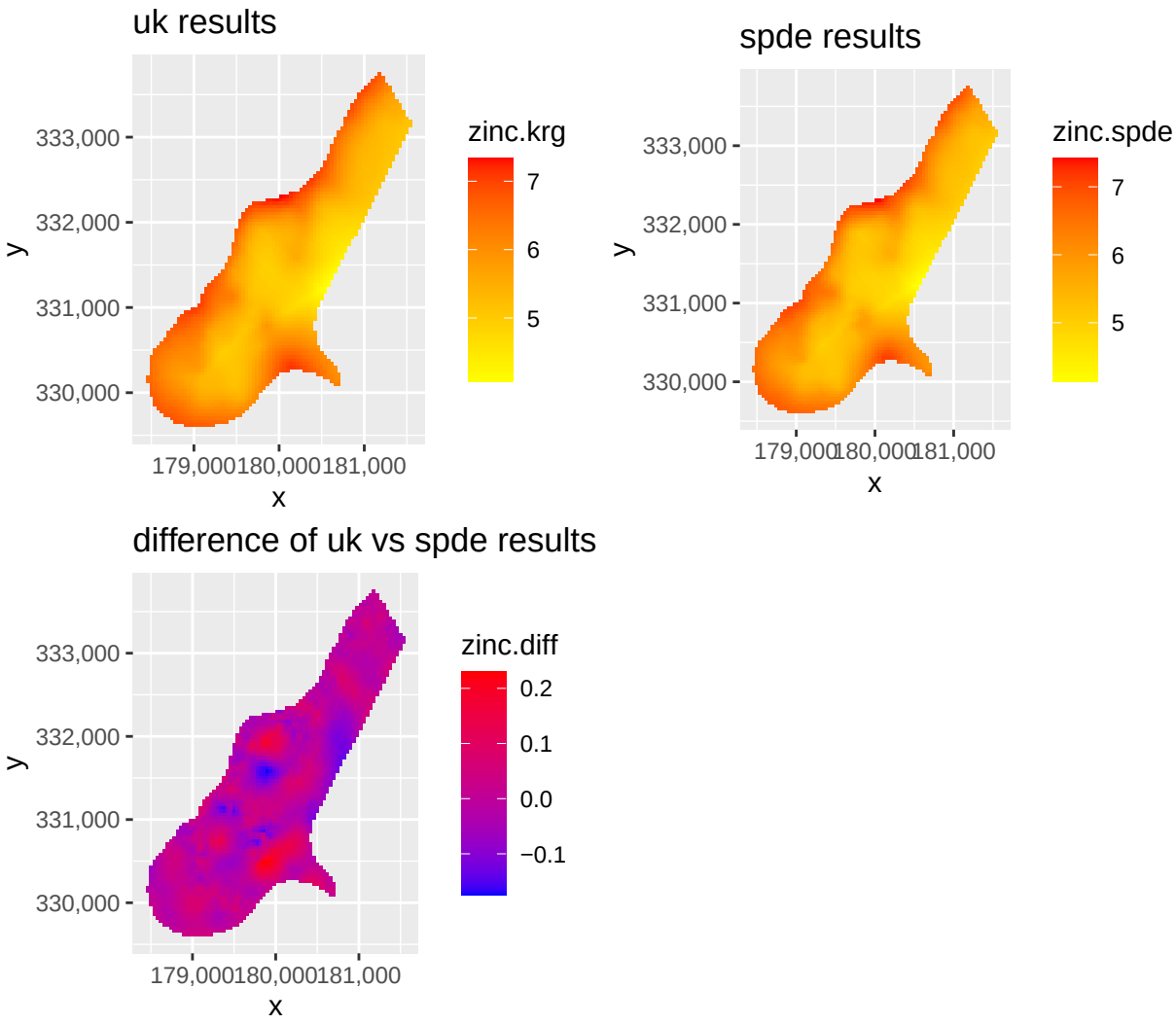
Table 3: variance from SPDE

mean	sd	quant0.025	quant0.25	quant0.5	quant0.75	quant0.975
0.2352	0.0723	0.1245	0.1831	0.2245	0.2754	0.4062

Table 4: range from SPDE

mean	sd	quant0.025	quant0.25	quant0.5	quant0.75	quant0.975
570.2	165.3	311.2	451	547.9	664.4	955.7

Both plots show higher estimated means of log-concentrations of zinc at locations closer to the Meuse river. The differences may be due to the ways how the models were defined and how their model components were specified.


uk results
333,000
332,000
331,000
330,000
y
179,000
180,000
181,000
x
zinc.krg
7
6
5
spde results
333,000
332,000
331,000
330,000
y
179,000
180,000
181,000
x
zinc.spde
7
6
5
difference of uk vs spde results
333,000
332,000
331,000
330,000
y
179,000
180,000
181,000
x
zinc.diff
0.2
0.1
0.0
−0.1

Reference

Gómez-Rubio, V. (2019). R-bloggers. Spatial Data Analysis with INLA. https://www.r-bloggers.com/2019/11/spatial-data-analysis-with-inla/.

Gómez-Rubio, V. (2020). 7.3 Geostatistics. *Bayesian inference with INLA.* CRC Press.

Moraga, P. (2019). 8 Geostatistical data. *Geospatial health data: Modeling and visualization with R-INLA and shiny.* CRC Press.