Nearest Neighbor Process (NNGP)

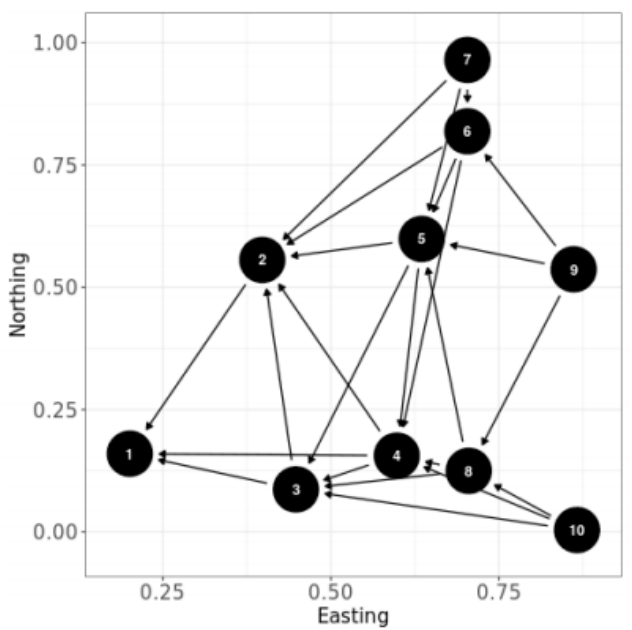
Morgan Schulte, Yulun Xu, Brice Turner

Motivation for NNGP

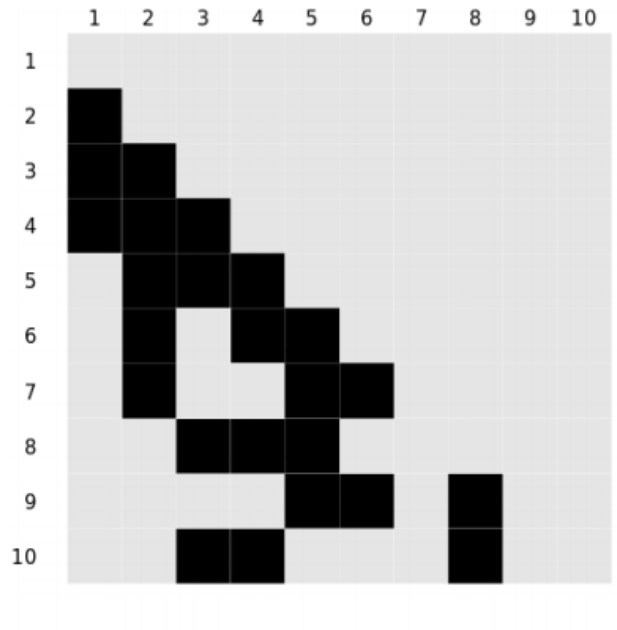
- Issue: Model fitting with a large n requires n3 operations and storage of n2.
- Solution: Introduce sparsity, have a small number of non-zero values
- Vecchia (1988) used m nearest-neighbors for approximating likelihoods
- Datta et al (2016) introduce sparsity in models using specified neighbor sets, and then use these sets to extend finite-dimensional models to a valid spatial process over uncountable sets.
- NNGP is a legitimate proper prior for random fields and is applicable to any class of distributions that support spatial stochastic process, such as Bayesian Kriging.

Motivation for NNGP

- Benefits
  - The storage and computational requirements are linear to n
  - Very scalable! To millions of high-dimensional locations
- Limitations
  - Uses stationary and isotropic covariance functions
  - Kriging may no longer be accurate if stationarity assumption is not met

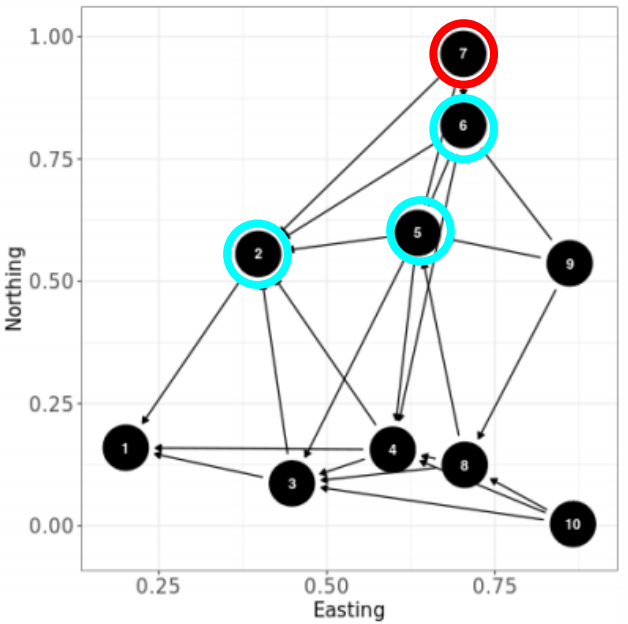


(a)

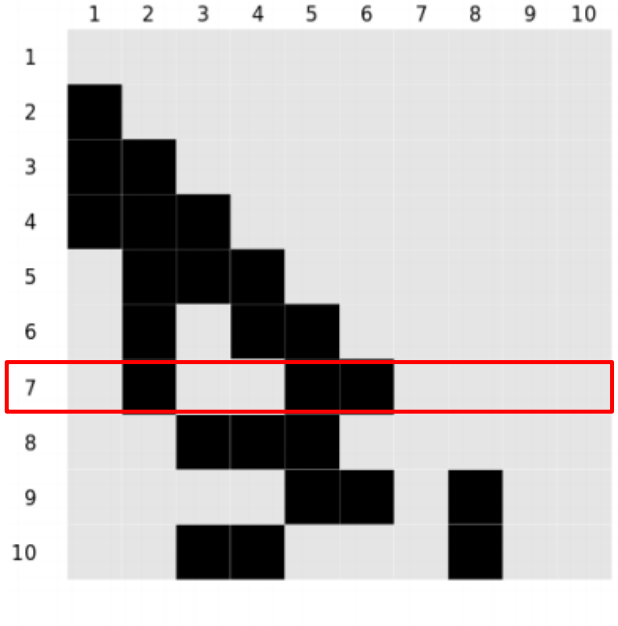


(b)

Figure 1: (a) Illustration of a conditional graph using three neighbors for ten observations with ordering along the easting axis. (b) The sparse lower-triangular matrix (black elements are non-zero) corresponding to the graph in (a) that records the neighbor index (columns) for each observation index (rows).



(a)



(b)

Figure 1: (a) Illustration of a conditional graph using three neighbors for ten observations with ordering along the easting axis. (b) The sparse lower-triangular matrix (black elements are non-zero) corresponding to the graph in (a) that records the neighbor index (columns) for each observation index (rows).

$$Y \sim N\left(X\beta, \tilde{\Sigma}(\phi)\right), \text{ where}$$

$\tilde{\Sigma}$ is the NNGP covariance matrix.

By letting $\alpha = \sigma_\varepsilon^2 / \sigma_w^2$, then the model can be expressed as

$$Y \sim N\left(X\beta, \sigma_w^2 \tilde{R}(\phi, \alpha)\right), \text{ where}$$

$\tilde{R}$ is the NNGP matrix derived from $C(\phi) + \alpha I$,
$C(\phi)$ is the correlation matrix of the exponential Gaussian Process.

Fixing α and ϕ gives a conjugate normal-inverse Gamma posterior distribution for β and σ_w^2.

Conjugate NNGP

Heaton et al 2019

Implementation details

-Software: R

-R Package: - spNNGP

- Function:

    - spNNGP (full MCMC-based inference)

    - spConjNNGP (chosen)

    Advantages: 1) MCMC-free inference

    2) faster to compete big-spatial data

Main steps

a) Fit the linear model of (Y ~ longitude + latitude) to get the residuals

a) Check the variogram to get the sigma square value

a) Set the theta.alpha matrix with the obtained sigma square value

a) Make the model with conjugate NNGP method

a) Make the prediction model by the training set and using the previously obtained phi and alpha values

Arguments of spConjNNGP

-n.neighbors: number of neighbors used in the NNGP.

-score.rule: "crps" (continuous ranked probability score) will be selected here

-theta.alpha: matrix with columns named phi, alpha,

-sigma.sq.IG: a vector of length two that holds the shape and scale respectively

-k.fold: specifies the number of k folds for cross-validation

-cov.model: keyword that specifies the covariance function (“exponential", "matern"...)

-n.omp.threads: number of threads to use for SMP parallel processing

Example of spConjNNGP

```
m.c1 <- spConjNNGP(Y1~X1-1, coords=X1, n.neighbors = 15,
                   k.fold = 5, score.rule = "crps",
                   n.omp.threads = 10,
                   theta.alpha = theta.alpha1, sigma.sq.IG = sigma.sq.IG1,
                   cov.model = "exponential")
```

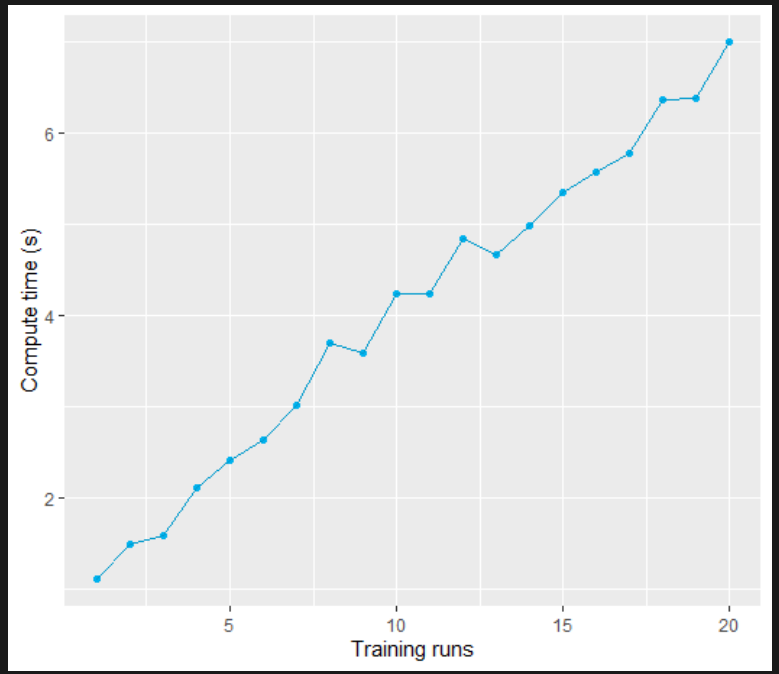
Results

- Compare NNGP to MLE in terms of time running and prediction performance across various sizes of the training set.

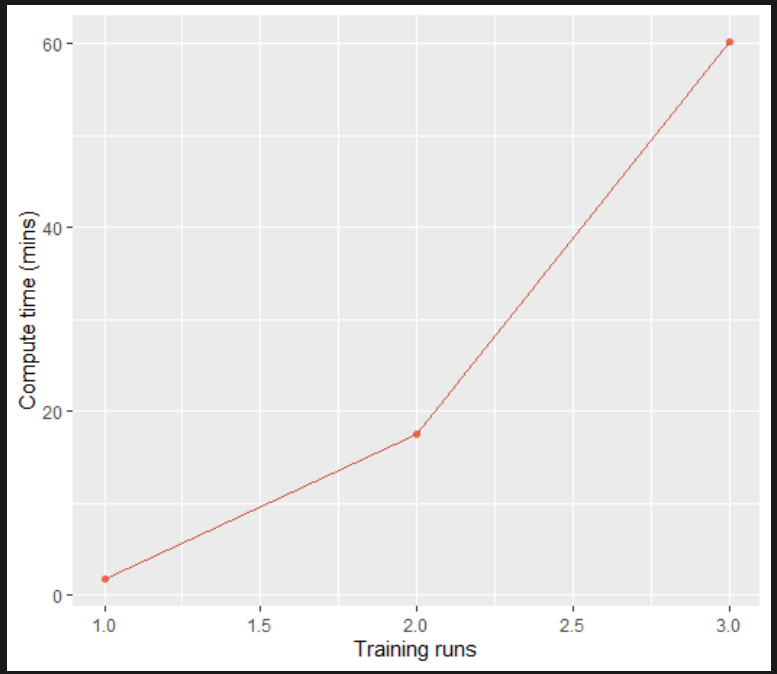
- MLE - A statistical concept to find the parameter to maximize the likelihood of a particular dataset

NNPG vs MLE Time Running

NNPG

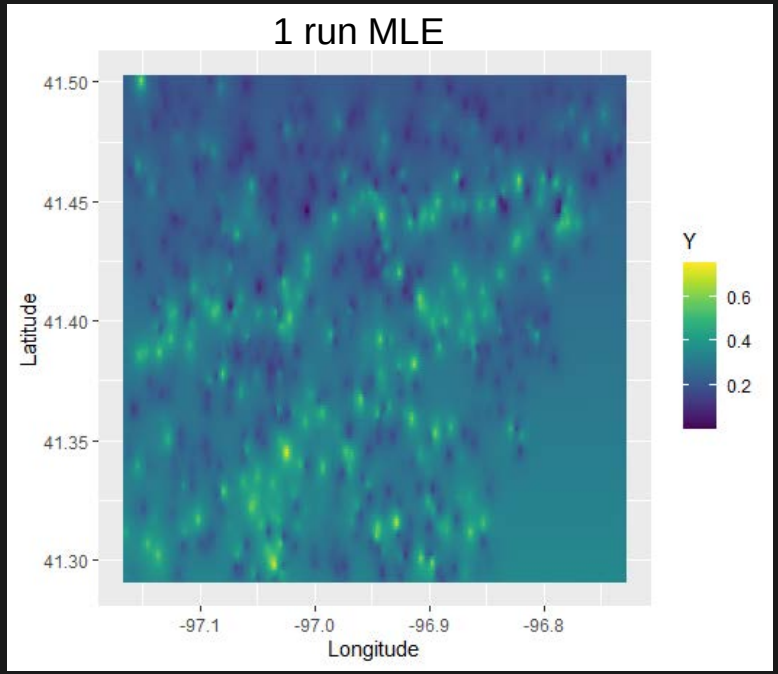


MLE

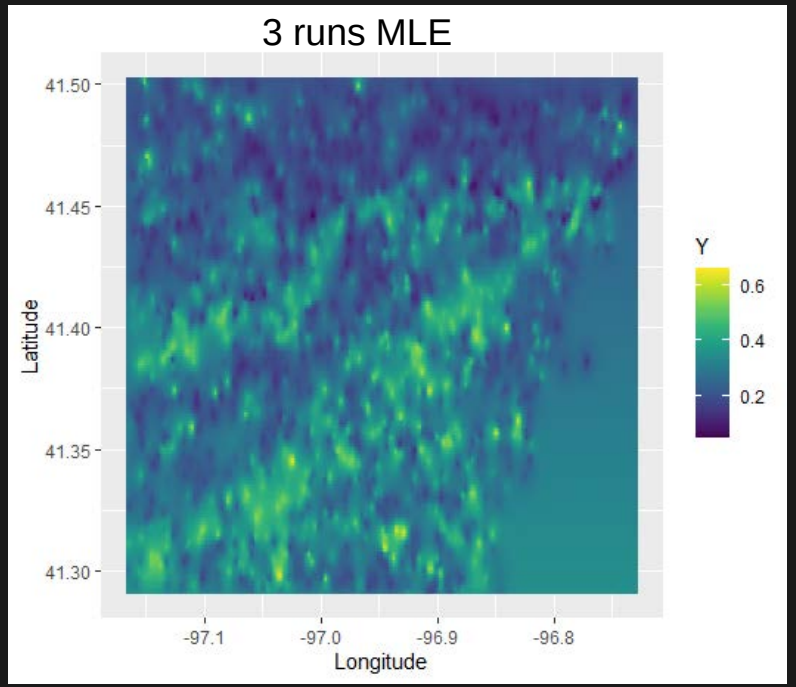


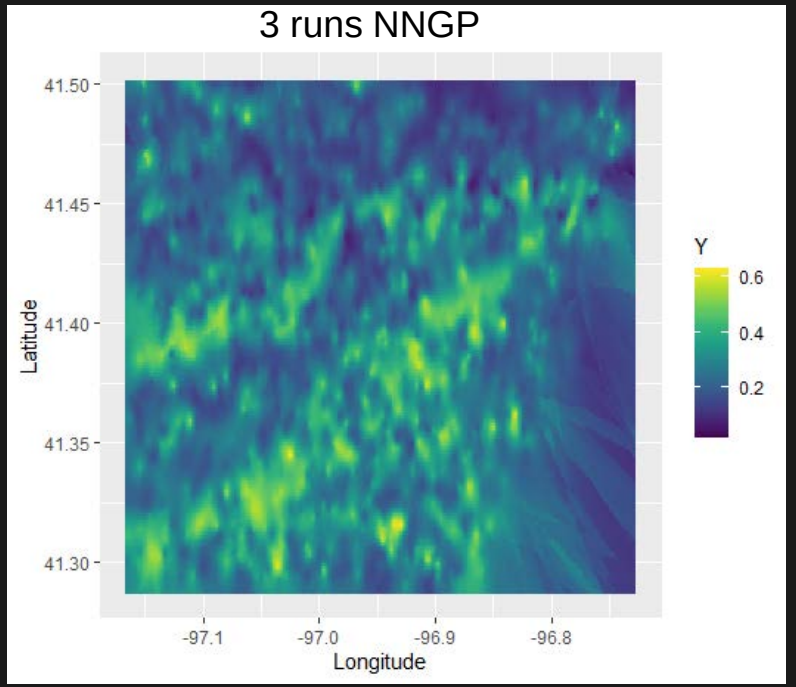
Prediction Performance - MLE

- 1 Run



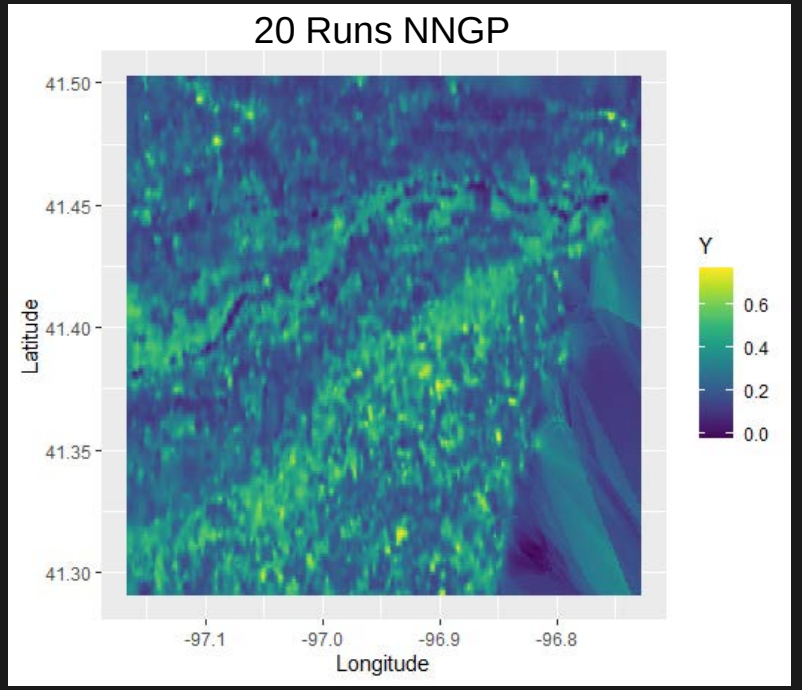
Prediction Performance - MLE 3 Runs vs NNGP 3 Runs





Prediction Performance - NNGP 20 Runs

- 20 Runs



Resources

R package *spNNGP*

Finley, A. O., Datta, A., Banerjee, S. “spNNGP R package for Nearest Neighbor Gaussian Process models

https://arxiv.org/pdf/2001.09111.pdf

Datta, A., Banerjee, S., Finley, A. O., and Gelfand, A. E. (2016a), “Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets,” Journal of the American Statistical Association, 111, 800–812.

Datta, A., Banerjee, S., Finley, A. O., and Gelfand, A. E. (2016b), “On nearest-neighbor Gaussian process models for massive spatial data,” Wiley Interdisciplinary Reviews: Computational Statistics, 8, 162–171.

Heaton MJ, Datta A, Finley AO, Furrer R, Guinness J, Guhaniyogi R, Gerber F, Gramacy RB, Hammerling D, Katzfuss M, et al. (2019). “A case study competition among methods for analyzing large spatial data.” Journal of Agricultural, Biological and Environmental Statistics, 24(3), 398–425.

Vecchia, A. V. (1988), “Estimation and model identification for continuous spatial processes,” Journal of the Royal Statistical Society. Series B (Methodological), 297–312.