Spatial Gaussian Process (GP) models for Massive Geostatistical Datasets

Spatial Gaussian Process (GP) models—subsume kriging—for point-referenced geospatial datasets provide a statistically valid framework for prediction with associated uncertainty for unobserved locations.

Challenge: Spatial process models for analyzing geostatistical data require matrix operations that become prohibitively expensive as the number of observed locations become large. Implementation can rarely use more than about $n=1x10^4$ locations. Hence these valuable statistical tools are not available for mapping efforts (e.g., CMS biomass maps with uncertainty).

New methods: We developed a class of highly scalable Nearest Neighbor Gaussian Process (NNGP) models to deliver model-based inference for large geostatistical datasets. NNGP provides an excellent approximation to GP without onerous matrix operations.

Significance: A rich set of geostatistical models can now be fit to massive datasets (e.g., $n>1x10^7$) and deliver statistically valid probability-based predictions with associated uncertainty. Addition of spatial process components to biomass mapping models help meet model assumptions and improve prediction. NNGP will have far reaching impact in spatial data sciences.

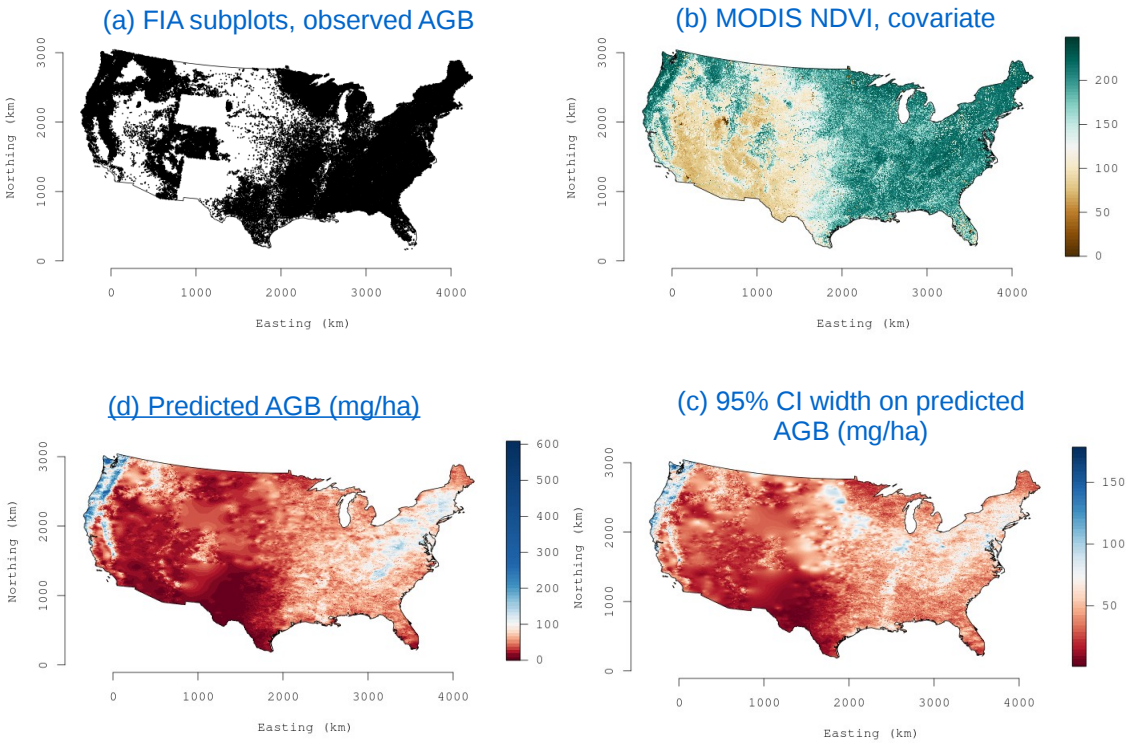

Illustration: Space-varying coefficients model for above ground biomass (AGB) prediction at generic location s

$$AGB(s) = \beta_0(s) + \beta_1(s)NDVI(s) + \varepsilon(s),$$

where $\beta_0(s) \sim NNGP(\beta_0, C(\theta_0))$ and $\beta_1(s) \sim NNGP(\beta_1, C(\theta_1))$

AGB measured at n=114,371 FIA subplots and single MODIS NDVI predictor. Prediction at MODIS resolution. (note, not final biomass map product)

CMS publications from: Hurtt-03, Dubayah-04, Morton-02, Cook-B-01

- Datta, A., S. Banerjee, A.O. Finley, and A.E. Gelfand. (2016) Hierarchical Nearest-Neighbor Gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association*. 10.1080/01621459.2015.1044091
- Datta, A., S. Banerjee, A.O. Finley, N.A.S. Hamm, and M. Schaap. (In press) Non-separable dynamic Nearest Neighbor Gaussian Process models for large spatio-temporal data with application to particulate matter analysis. *Annals of Applied Statistics*. https://arxiv.org/abs/1510.07130
- Finley, A.O., S. Banerjee, A.E. Gelfand. (2015) spBayes for large univariate and multivariate point-referenced spatio-temporal data models. *Journal of Statistical Software*, 63:1-28. www.jstatsoft.org/v63/i13
- Babcock, C., A.O. Finley, J.B. Bradford, R. Birdsey, R. Kolka, and M.G. Ryan. (2015) LiDAR based prediction of forest biomass using hierarchical models with spatially varying coefficients. Remote Sensing of Environment. 169:113-127