Title: A Scalable Partitioned Approach to Model Massive Nonstationary Non-Gaussian Spatial Datasets

Benjamin Seiyon Lee and Jaewoo Park

Abatract: Nonstationary non-Gaussian spatial data are common in many disciplines, including climate science, ecology, epidemiology, and social sciences. Examples include count data on disease incidence and binary satellite data on cloud mask (cloud/no-cloud). Modeling such datasets as stationary spatial processes can be unrealistic since they are collected over large heterogeneous domains (i.e., spatial behavior differs across subregions). Although several approaches have been developed for nonstationary spatial models, these have focused primarily on Gaussian responses. In addition, fitting nonstationary models for large non-Gaussian datasets is computationally prohibitive. To address these challenges, we propose a scalable algorithm for modeling such data by leveraging parallel computing in modern high-performance computing systems. We partition the spatial domain into disjoint subregions and fit locally nonstationary models using a carefully curated set of spatial basis functions. Then, we combine the local processes using a novel neighbor-based weighting scheme. Our approach scales well to massive datasets (e.g., 1 million samples) and can be implemented in nimble, a popular software environment for Bayesian hierarchical modeling. We demonstrate our method to simulated examples and two large real-world datasets pertaining to infectious diseases and remote sensing.

Keywords: spatial partition, basis representation, parallel computing, nonstationary process, non-Gaussian spatial data