Final Report on "Spatio-Temporal Data Analysis at Scale Using Models Based on Gaussian Processes"

Award Number: DE-FG02-14ER26171/DE-SC0011087

Title: Spatio-Temporal Data Analysis at Scale Using Models Based on Gaussian Processes

Principal Investigator and co-PIs: Michael Stein (stein@galton.uchicago.edu) is the principal investigator for this grant.

Dates of Award: 1/1/2014–12/31/2016

This award is part of a larger project with Mihai Anitescu of Argonne National Laboratory as Project Director. This report describes only the work done as part of the award to Michael Stein at the University of Chicago.

Executive Summary

Gaussian processes are the most commonly used statistical model for spatial and spatio-temporal processes that vary continuously. They are broadly applicable in the physical sciences and engineering and are also frequently used to approximate the output of complex computer models, deterministic or stochastic. We undertook research related to theory, computation, and applications of Gaussian processes as well as some work on estimating extremes of distributions for which a Gaussian process assumption might be inappropriate.

Our theoretical contributions include the development of new classes of spatial-temporal covariance functions with desirable properties and new results showing that certain covariance models lead to predictions with undesirable properties. To understand how Gaussian process models behave when applied to deterministic computer models, we derived what we believe to be the first significant results on the large sample properties of estimators of parameters of Gaussian processes when the actual process is a simple deterministic function. Finally, we investigated some theoretical issues related to maxima of observations with varying upper bounds and found that, depending on the circumstances, standard large sample results for maxima may or may not hold.

Our computational innovations include methods for analyzing large spatial datasets when observations fall on a partially observed grid and methods for estimating parameters of a Gaussian process model from observations taken by a polar-orbiting satellite. In our application of Gaussian process models to deterministic computer experiments, we carried out some matrix computations that would have been infeasible using even extended precision arithmetic by focusing on special cases in which all elements of the matrices under study are rational and using exact arithmetic.

The applications we studied include total column ozone as measured from a polar-orbiting satellite, sea surface temperatures over the Pacific Ocean, and annual temperature extremes at a site in New York City. In each of these applications, our theoretical and computational innovations were directly motivated by the challenges posed by analyzing these and similar types of data.

Comparison of accomplishments with goals and objectives of project

The overarching goal of this project was to "advance the state of the art in Gaussian spatio-temporal process methodology and scope and to demonstrate the benefits of this advance in DOE mission applications at their operational scales." Stein's role in this project was mostly on the methodological side, including modeling, computation and inference for Gaussian processes. We successfully completed research in modeling, computation and theory for Gaussian processes and their application to environmental processes and computer experiments.

Research on computation, which was to be done jointly with Anitescu, was hampered by the departure of Jie Chen, a co-PI on this project, for industry, and by the departure of Zhen Zhang, a postdoc partially supported by this grant, for industry before he completed a substantial computational project. Nevertheless, we did complete some significant research on computational issues related to the analysis of data on partially observed grids and approximate likelihoods for data collected by polar-orbiting satellite, a common mode for collecting environmental data.

On the theoretical side, we made fundamental advances in understanding the relationship between Gaussian process models and the properties of predictions based on these models. In addition, we made what we believe is the first serious effort to understand the behavior of estimates of the parameters of Gaussian process models when they are applied to smooth deterministic functions, circumstances that frequently arise when applying Gaussian process models to computer model output. Finally, the project description recognized that Gaussian process models will not be appropriate for many spatio-temporal processes and proposed to look at alternative models. Inferences about extremes is an important situation in which one would want to be cautious about assuming Gaussianity. Motivated by the problem of modeling maximum annual temperatures at a site in a way that appropriately accounts for seasonality, we examined theoretical properties of maxima of many bounded random variables where the upper bound changes slowly.

Summary of project activities

Much of the theoretical work we did concerned gaining a better understanding of the relationship between the spectral density for spatial and spatial-temporal stationary Gaussian processes and the properties of these processes, with an eye towards understanding when certain approximations to optimal predictors and likelihood functions will work well. In particular, many approximations in spatial and spatial-temporal statistics are based on the assumption that the distribution of the process at one location is, conditional on some set of nearby observation, close to conditionally independent of all other observations, which is known as the screening effect. As past work of mine showed, it is possible to obtain some rigorous mathematical results showing that some forms of a screening effect hold when the spectral density is well-behaved at high frequencies in a well-defined sense. In particular, some forms of a screening effect will hold if a spectral density f satisfies

$$\lim_{\omega\to\infty}\frac{f(\omega+\omega_0)}{f(\omega)}=1 \tag{1}$$

for all ω_0. Although there were specific examples showing that some spectral densities that do not satisfy (1) also do not show a screening effect, there was no general theory showing when a screening effect does not hold. Stein (2015) provides some first general results showing that a broad class of spectral densities not satisfying (1) also do not satisfy a certain form of the screening effect. This work looks in detail at a fascinating special case for a process on the line with spectral density that behaves like $\{1+\sin(\omega^2)\}/\omega^2$ at high frequencies, which most certainly does not satisfy (1). For this model, it turns out that whether one observes a screening effect depends in subtle ways on what exactly one means by this question. This example also sheds light on the equivalence and orthogonality of Gaussian measures, a fundamental property of Gaussian processes from both probabilistic and statistical perspectives.

Michael Horrell was a doctoral student supported by this grant who graduated in 2015 and now works for Uptake Technologies, LLC as a Data Science Architect. He had previously worked on statistical models for spatial-temporal data at the global scale. While supported by this grant, he worked on connecting two threads of my research. The first is this notion of using spectral densities satisfying (1) and the second is defining spatial-temporal models spectrally in time and in the spatial domain in space. This model formulation is particularly useful (both conceptually and computationally) when data are available at regular time intervals at a modest number of sites, which occurs frequently for environmental monitoring data. Horrell investigated when these "half-spectral" representations satisfy (1) and obtained some nice results on examples of these models that I hope will prove broadly useful in practice. He fitted these new models to a large dataset of total column ozone levels in a latitude band over a month and showed that one in particular fits better than a number of other space-time models recently proposed in the literature. Unlike many works based on polar-orbiting satellite data, we did not use the “Level 3” gridded data product, but instead considered the “Level 2” version of the data, which retains the actual irregular locations and times of the observations. This irregularity of the observation locations and the size of the dataset necessitated the use of approximate methods for evaluating the likelihoods of Gaussian process models and we developed and implemented some effective approximate likelihoods that are appropriate for data taken from a polar-orbiting satellite. This work is described in Horrell and Stein (in press).

With Jon Stroud, now at Georgetown University, I worked on an approach to doing Bayesian computations for stationary spatial processes on a subset of a grid. The approach is based on embedding the observations in a larger, complete grid on which the process can be taken to be periodic. Because it is easy to evaluate the exact likelihood on this larger grid using the fft, one can then use MCMC approaches to average over the process at unobserved locations and obtain effective approximations of the posterior distribution of unknown parameters. This work is described in Stroud, Stein and Lysen (2017).

A major area of application of Gaussian process models is to model the output of deterministic computer experiments as a function of multiple inputs, despite the fact that there is nothing stochastic about the results. Wanting Xu is a fourth-year Ph.D. student who first started working with me during Winter Quarter 2015 on theoretical and numerical issues related to Gaussian processes in computer experiments. Gaussian processes have two important desirable properties in this context: they provide a flexible class of interpolating rules that match the available observations (which is desirable because there is no random error in the computer output) and they provide measures of uncertainties for these interpolations. However, despite their long use, there is very little theory on how well Gaussian process models actually perform in this setting. Xu studied the use of Gaussian process models with a squared exponential covariance function, which is commonly used to model smooth deterministic functions. She proved that the maximum likelihood estimates of the scale parameter of the covariance function can tend to infinity for some simple functions and to zero for others. The specific results were not what I had expected and I think they will surprise others who work in the field. Xu also did various numerical experiments comparing maximum likelihood and cross validation as methods for estimating the parameters of the squared exponential covariance function, many of which use exact arithmetic to avoid the notorious problem of near singularity of covariance matrices based on the squared exponential model. The results are somewhat mixed, but maximum likelihood is often better than cross validation even for deterministic functions that are obviously not realizations of a Gaussian process. This work is described in Xu and Stein (2017).

Zhen Zhang was a postdoc at Chicago for two years and was partially supported on this grant. He worked on a range of computational and modeling issues related to the use of Gaussian processes for spatial data and made quite a bit of progress. Unfortunately, Zhang took a position as a statistician at Dow AgroSciences before completing any papers. It now appears unlikely that any publications will result from this work.

Although the funding period for this grant has passed, there are three ongoing projects that received support from this grant and should result in publications. The first is the development of spatial-temporal covariance functions for ocean temperatures as measured by Argo floats, a major international effort to study ocean characteristics, especially below the surface. This work is being carried out by Mikael Kuusela, a very promising postdoc from EPFL in Switzerland who arrived in Chicago last fall. This work is being done in close collaboration with researchers at the Scripps Institute of Oceanography, who have responded very positively to Mikael's initial results. A second project is being undertaken by Wanting Xu on a topic unrelated to her earlier work on Gaussian processes for deterministic computer experiments. This project concerns the analysis of a time series of length 10^6 from an electrical circuit designed to mimic the behavior of a well-known chaotic dynamical system. Although there is a nominally known system of differential equations for this dynamical system, a close examination of the data shows that this model is clearly wrong. We are investigating the use of Gaussian process models and neural networks to obtain better models for the data. We hope that this work will lead to new understanding of how to model data from a low-dimensional chaotic dynamical system whose exact dynamics are unknown and for which measurement errors are small but not ignorable. The third project concerns applications of skeletonization algorithms for factoring large covariance matrices arising in spatial statistics. Work on such algorithms has been a research area in the numerical linear algebra literature for over ten years, but it is only recently that people have recognized the importance of this work in spatial statistics and, as far as I am aware, no statisticians have yet contributed to this literature. Sam Baugh, a student in our highly selective joint BA/MS program, which allows undergraduates to receive both a Bachelor's and a Masters degree after four years of study, is testing recently developed software to see how well it works on realistic problems in spatial statistics. Baugh has a strong background in computer science and, together, we are exploring the effectiveness of these algorithms and developing models and approaches to optimization that make good use of the algorithms.

Products

Funded publications

Horrell, M. T. and Stein, M. L. (in press). Half-Spectral Space-Time Covariance Models. *Spatial Statistics* (http://dx.doi.org/10.1016/j.spasta.2016.12.002).

Stein, M. L. (2015). When Does the Screening Effect Not Hold? *Spatial Statistics*, 11, 65–80.

Stein, M. L. (in press). Should Annual Maximum Temperatures Follow a Generalized Extreme Value Distribution? *Biometrika.* https://doi.org/10.1093/biomet/asw070

Stroud, J. R., Stein, M. L., and Lysen, S. (2017). Bayesian and Maximum Likelihood Estimation for Gaussian Processes on an Incomplete Lattice. *Journal of Computational and Graphical Statistics*, 26, 108-120.

Xu, W. and Stein, M. L. (2017). Maximum Likelihood Estimation for a Smooth Gaussian Random Field Model. *SIAM/ASA Journal on Uncertainty Quantification*, 5, 138-175 (doi: 10.1137/15M105358X).

Networks or collaborations fostered

There have been strong personal and intellectual interactions between the people funded through this award and STATMOS, Research Network for Statistical Methods for Atmospheric and Oceanic Sciences, an NSF-funded network of 19 institutions in academia and government. In particular, postdocs Zhen Zhang and Mikael Kuusela were jointly funded by this award and STATMOS. Kuusela is carrying out his research in close collaboration with researchers at Scripps Institute of Oceanography, which is a STATMOS member. I expect this collaboration to continue throughout this academic year and next, which Kuusela will be spending at SAMSI, The Statistical and Applied Mathematical Sciences Institute, an NSF-supported center, in an arrangement that was made when Kuusela was hired last year. SAMSI has a program in 2017-2018 on Mathematical and Statistical Methods for Climate and Earth Systems, which will enable Kuusela to continue his research in statistical oceanography.

All of my students and postdocs participate in my weekly group meetings, which are regularly attended by people outside of my group, including postdocs from Argonne National Laboratory and, this past year, by Mary Silber, an applied mathematician who recently joined the Chicago faculty, and her students. In addition, Kuusela is an active participant in separate group meetings I hold with Elisabeth Moyer in Geophysical Sciences for a climate science group that is part of RDCEP, The Center for Robust Decisionmaking on Climate and Energy Policy, another NSF-supported center. These group meetings include two postdocs who work with me and Prof. Moyer, one a statistican and one a climate scientist. Kuusela particularly benefits from interactions with the climate scientist, who has a strong background in oceanography.