Taft Summer Research Fellowship Cover Sheet

Checklist

- Signed Cover Sheet - DONE
- Project Description (800-1200 words) - DONE
- Supplement explanation, if necessary - DONE
- 2 page CV - DONE

Review Taft website for full application guidelines.

Date of Application: 1/31/13

Name, Department, Rank:
Emily Lei Kang
Tenure-Track Assistant Professor
Department of Mathematical Sciences

Time Period:
Summer 2013

Location:
Cincinnati, Ohio

Title of Project:
Simulation and Analysis of Massive Spatial Data

Requested Research Supplement (if any): $1,150 of traveling for collaboration
(besides $8,000 for summer salary)

Probable Results of a Grant (such as external funding, publications, and presentations):
Funding from the Jet Propulsion Laboratory (JPL) at California Institute of Technology (Caltech)
Related work will also be presented at scientific conferences and published in academic journals as peer-reviewed articles..

Other Funding Applied For or Received for This Project (list source and amounts requested and awarded):
Subcontract from JPL for Spring 2013. Amount=$17,000 (including cost to buy out one semester course and related traveling expenses for the PI to visit collaborators at JPL)

*** Note:** this grant is for non-teaching quarters without pay and leave quarters without pay. *****

Taft Grants Received in the Last Five Years:

1. Type and Dates: Domestic Conference Travel Grant, 08/2012
   Amount:
   $700

   Project Title:
   Filtering Partially Observed Multiscale Systems with Heterogeneous-Multiscale-Methods-Based Reduced Models

   Resulting Publications and Presentations:
   - Contributed presentation at the Joint Statistical Meetings (JSM), 2013 at San Diego, CA
   - Kang, E. L. and Harlim, J. (2012). Filtering partially observed multiscale systems with Heterogeneous-Multiscale-Methods-based reduced climate models. Monthly Weather Review, 140, 860-873
   - Kang, E. L., Harlim, J., and Majda, A. J. Regression models with memory for the linear response of turbulent dynamical systems, submitted.

2. Type and Dates: International Conference Travel Grant, 01/2013
   Amount:
   $600.25
   **The PI was originally awarded $2,697from Taft but only used $600.25, since the PI was awarded traveling grants from the Association for Women in Mathematics (AWM)-National Science Foundation (NSF).*

   Project Title:
   Bayesian Analysis of High-resolution Regional Climate Model Output over North America

   Resulting Publications and Presentations:
   - Invited presentation at the International Society for Bayesian Analysis (ISBA) Regional Meeting & International Workshop/Conference on Bayesian Theory and Applications (IWCBTA) at Varanasi, India

   - Ojiambo, P. S. and Kang, E. L. Modeling spatial frailties in survival analysis of cucurbit downy mildew epidemics, submitted.

Signature:

*** Note:** this grant is for non-teaching quarters without pay and leave quarters without pay. *****

Project Details

1 Focus and Significance

With the revolution in information and communication technologies such as the Geographical Information Systems (GIS), scientists and researchers in a variety of disciplines today have access to massive amounts of data collected with spatial and temporal information. However, classical statistical approaches for analyzing spatial and spatio-temporal datasets often break down due to expensive matrix inversion operations, whose computational complexity increases in cubic order with the size of the datasets. In this project, I will develop new statistical approaches for analyzing massive spatial and spatio-temporal datasets and provide novel solutions across a wide-range of applied problems. Specifically, a new method for modeling the spatial covariance function is developed which combines the advantages of state-of-the-art methods but also allows feasible computation with massive datasets. This project will also develop a full suite of computationally efficient statistical methods for making optimal spatial (and potentially spatio-temporal) predictions from massive, noisy, and incomplete data obtained from one or multiple sources, while borrowing strength across multiple scales in space and time and multiple data sources. A simulation testbed infrastructure will be designed to simulate synthetic geophysical fields of atmospheric carbon dioxide (CO_2), which will be used to study prediction skills of the proposed model and also lay down the tools for future quantitative comparisons between different methods. This research will improve the interpretability and usability of remote-sensing data retrieved from instruments aboard satellites launched by the National Aeronautics and Space Administration (NASA) through the applications of hierarchical spatial and spatio-temporal statistical models. The proposed method also will directly carry over to the areas of uncertainty quantification for climate modeling and analysis of infectious disease epidemiology.

The proposed research fits my research interests and will lay down the foundation of my continuing collaboration with Dr. Amy Braverman from the Jet Propulsion Laboratory (JPL), California Institute of Technology. I am also expecting to work with faculty from Geography and Environmental Health at University of Cincinnati so that the proposed method can be applied in diverse fields of study.

2 Background, Theoretical Framework, and Methods

2.1 Preliminary Results

To tackle the challenges of massive spatial data, two approaches have been suggested recently: One approach is *approximation*-based, including covariance tapering (Furrer et al., 2006; Kaufman et al., 2008) and predictive process (Banerjee et al., 2008), which *approximates* the $n \times n$ covariance matrix of n data with a computationally convenient matrix to obtain

approximately optimal predictions. The other approach, fixed rank kriging with the spatial random effects (SRE) model (Cressie and Johannesson, 2008), is *model*-based by using a set of spatial basis functions. In my preliminary studies, I have found that both these two approaches have their drawbacks. In particular, the adequacy of approximation cannot be quantified accurately for the approximation-based approach. Therefore, its applications in spatio-temporal setting and in modeling multiple-process systems is hampered, since the not-appropriately-gauged approximation errors can propagate or even become amplified through time and interactions between different processes. For the model-based approach, the SRE model, I have found in my preliminary simulation studies that it lacks the ability of capture small spatial dependence, although it allows *exact* computation of optimal predictions.

2.2 Statistical Models for Massive Spatial Datasets

This section presents the theoretical framework and the methods proposed in this project. Let $\{Y(\mathbf{s}) : \mathbf{s} \in D \subset \mathbb{R}^d\}$ be an underlying real-valued spatial stochastic process. We are interested in making inferences on the Y-process based on data that have measurement errors incorporated. Consider the process $Z(\cdot)$ of actual and potential observations,

$$Z(\mathbf{s}) = Y(\mathbf{s}) + \varepsilon(\mathbf{s}),\ \mathbf{s} \in D, \tag{1}$$

where $\{\varepsilon(\mathbf{s})\}$ is a spatial white-noise process with mean zero and $\text{var}(\varepsilon(\mathbf{s})) = \sigma^2 > 0$, representing measurement errors. In practice, the process $Z(\cdot)$ is known only at a finite number of spatial locations, $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$, that is, n observation locations.

We seek statistically optimal prediction of $Y(\mathbf{s}_0)$, $\mathbf{s}_0 \in D$, regardless of whether $\mathbf{s}_0$ is or is not an observation location. Inspection of the formulas of the optimal predictor for $Y(\mathbf{s}_0)$ and the associate prediction error (called "kriging" in Cressie, 1993) shows that $\mathbf{\Sigma}^{-1}$ is an essential component, where $\mathbf{\Sigma}$ is the $n \times n$ covariance matrix of the n observations $(Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n))$. Since the computational complexity of $\mathbf{\Sigma}^{-1}$ is $O(n^3)$, classical methods will computationally break down when n is large (e.g., several thousands).

To alleviate this difficulty, I propose a new approach to model the spatial covariance function of Y-process, $C(\mathbf{u}, \mathbf{v}) \equiv \text{cov}(Y(\mathbf{u}), Y(\mathbf{v}))$:

$$C(\mathbf{u}, \mathbf{v}) = C_l(\mathbf{u}, \mathbf{v}) + C_s(\mathbf{u}, \mathbf{v}),\ \mathbf{u}, \mathbf{v} \in D, \tag{2}$$

where the first component $C_l(\mathbf{u}, \mathbf{v}) = \mathbf{S}(\mathbf{u})'\mathbf{K}\mathbf{S}(\mathbf{v})$ is to capture the *large* scale spatial dependence through a set of r (not necessarily orthogonal) known basis functions, $\mathbf{S}(\mathbf{u}) \equiv (S_1(\mathbf{u}), \ldots, S_r(\mathbf{u}))'$; $\mathbf{K}$ is an $r \times r$ positive-definite matrix. The other component in (2), $C_s(\mathbf{u}, \mathbf{v})$, is to capture the *small* scale spatial dependence, which is parameterized with a compactly supported nonstationary spatial covariance function so that covariance for distant pairs of observations is set to zero. In this way, the $n \times n$ covariance matrix $\mathbf{\Sigma}$ can be written as the summation of two parts $\mathbf{SKS}'$ and $\mathbf{C}^*$, with $n \times r$ matrix $\mathbf{S}$ from r basis functions and $n \times n$ *sparse* matrix $\mathbf{C}^*$ allowing efficient computation with sparse matrix techniques. Then,

by using the Sherman-Morrison-Woodbury formulae (Henderson and Searle, 1981), $\mathbf{\Sigma}^{-1}$ can be computed efficiently and *exactly*; so do the optimal predictor of $Y(\mathbf{s}_0)$ and its associated prediction error. It can also be shown that the corresponding computational complexity is substantially reduced from $O(n^3)$ to $O(n)$.

Generalizations Based on the model of the spatial covariance function in (2), I will also develop a new approach for fusing massive spatial data from multiple instruments, measuring the same underlying spatial process or various but related spatial processes. Meanwhile, I will extend this spatial model to the spatio-temporal setting and lay down a general framework for simulation and analysis of massive spatio-temporal data.

Simulation Studies I will develop a detailed mathematical formulation of the *large* and the *small* scale spatial dependence structures, and use it to simulate synthetic geophysical fields of the atmospheric carbon dioxide (CO_2): First, the simulated values will be generated at high spatial resolution, resulting massive data. Meanwhile, the simulated values must obey certain constraints imposed by the requirement that they aggregate properly to coarser resolutions in the sense that they match physical model predictions made at those coarse scales. I will also derive algorithms for estimating model parameters and implement them within the simulation testbed infrastructure. In addition, comparisons between the proposed approach and the state-of-the-art approaches will be carried out in simulation studies with spatial dependence of various scales assumed, exploring the sensitivity and flexibility of different methods. New computing technologies such as parallel computing and divide-and-conquer will also be investigated in the aforementioned simulation studies.

References

Banerjee, S., Gelfand, A. E., Finley, A. O. and Sang, H. (2008) Gaussian predictive process models for large spatial datasets. *Journal of the Royal Statistical Society, Series B*, **70**, 825–848.

Cressie, N. (1993) *Statistics for Spatial Data*, revised edn. New York: Wiley.

Cressie, N. and Johannesson, G. (2008) Fixed Rank Kriging for very large spatial data sets. *Journal of the Royal Statistical Society, Series B*, **70**, 209–226.

Furrer, R., Genton, M. G. and Nychka, D. (2006) Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics*, **15**, 502–523. Gelfand, A. E., Zhu, L. and Carlin, B. P. (2001) On the change of support problem for spatio-temporal data. *Biostatistics*, **2**, 31–45.

Henderson, H. V. and Searle, S. R. (1981) On deriving the inverse of a sum of matrices. *SIAM Rev.*, **23**, 53–60.

Kaufman, C., Schervish, M. and Nychka, D. (2008) Covariance tapering for likelihood-based estimation in large spatial datasets. *Journal of the American Statistical Association*, **103**, 1556–1569.

Budge with explanation of costs
Total Amount Requested: $9,150
This includes:
--$8,000 as salary for Summer 2013
--$1,150 as traveling expenses for the PI to visit her collaborator at JPL for 2 days
-- $650 Airfare + $380 Lodging + $120 Meals

Funding is requested for summer salary support so that the PI can devote the summer to the proposed research project and will enable her to visit Dr. Amy Braverman at the Jet Propulsion Laboratory (JPL) for collaboration.

*** Note:** this grant is for non-teaching quarters without pay and leave quarters without pay. *****

EMILY LEI KANG

Department of Mathematical Sciences
4199 French Hall West
University of Cincinnati
Cincinnati OH 45221-0025

Email: kangel@ucmail.uc.edu
Tel: 513-556-5138
Fax: 513-556-3417
http://homepages.uc.edu/~kangel/

EDUCATION

Ph. D. Statistics, Department of Statistics, The Ohio State University Statistics, 12/13/2009
Thesis title: *Reduced-dimension hierarchical statistical models for spatial and spatio-temporal data* Advisor: Noel Cressie
M.S. Statistics, Department of Statistics, The Ohio State University, 06/2006
B.S. Applied Mathematics, Department of Mathematics, Tianjin University, China, 07/2004
B.A. Finance, Department of Finance, Nankai University, China, 07/2004

RESEARCH INTERESTS

- Spatial statistics and spatio-temporal statistics
- Bayesian methodology and hierarchical modeling
- Computation strategies in multiscale systems
- Data assimilation in atmospheric sciences
- Applications of statistical models in public health, environmental and climate sciences

ACADEMIC and PROFESSIONAL APPOINTMENTS

09/2011 – present	**University of Cincinnati**
	Assistant Professor, Department of Mathematical Sciences
2009 – 2011	**Statistical and Applied Sciences Institute (SAMSI)**
	Postdoctoral Fellow, Program on Space-Time Analysis
2009 – 2011	**North Carolina State University**
	Postdoctoral Fellow, Department of Mathematics
06/2008 – 09/2008	**National Center for Atmospheric Research**
	Research Fellow, Geophysical Statistics Project (GSP)
06/2007 – 09/2007	**Commonwealth Scientific and Industrial Research Organization**
	Research Intern, Division of Mathematics, Informatics and Statistics
2006 – 2009	**The Ohio State University**
	Research Assistant, Department of Statistics

TEACHING

University of Cincinnati: Applied Spatial Statistics (Fall 2012), Probability and Statistics I (Fall 2012, Autumn 2011), Probability and Statistics II (Winter 2012), Applied Regression (Winter 2012)
North Carolina State University: Introduction to Statistics and Distribution Theory (Summer 2010)

SECTED AWARDS and GRANTS

2013	**Subcontract ($17,000)**, the Jet Propulsion Laboratory (JPL), California Institute of Technology (buyout of one semester course and related traveling expenses)
2013	**Taft International Conference Travel Grant**
2012-2013	**Travel Award ($1,950)**, the Association for Women in Mathematics (AWM)-National Science Foundation (NSF)

*** Note:** this grant is for non-teaching quarters without pay and leave quarters without pay. *****

2012-2013 **Mini-Grant Award ($1200)**, Center for Geospatial Information & Environmental Sensor Networks (GIESN), University of Cincinnati

2012 **Taft Domestic Conference Travel Grant**

2011 – 2012 **Faculty Research Grant for Summer Stipend ($8000)**, University Research Council, University of Cincinnati

PEER-REVIEWED PUBLICATIONS

1. **Kang, E. L.,** Cressie, N. (2012). Bayesian hierarchical ANOVA of regional climate change projections from NARCCAP Phase II. *International Journal of Applied Earth Observation and Geoinformation*. doi: 10.1016/j.jag.2011.12.007
2. **Kang, E. L.,** Cressie, N. and Sain, S. R. (2012). Bayesian hierarchical models to combine outputs from the NARCCAP regional climate models. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, doi: 10.1111/j.14679876.2011.01010.x.
3. **Kang, E. L.**, and Harlim, J. (2012). Filtering nonlinear spatio-temporal chaos with autoregressive linear stochastic model. Physica D, in press.
4. **Kang, E. L.** and Harlim, J. (2012). Filtering partially observed multiscale systems with Heterogeneous-Multiscale-Methods-based reduced climate models. *Monthly Weather Review*, 140, 860-873.
5. **Kang, E. L.** and Cressie, N. (2011). Bayesian inference for the spatial random effects model. *Journal of the American Statistical Association*, 106, 975-983.
6. Cressie, N. and **Kang, E. L.** (2010). High-resolution digital soil mapping: Kriging for very large datasets. In *Proximal Soil Sensing*, Eds R. Viscarra-Rossel, A. B. McBratney, and B. Minasny, 49-63.
7. Cressie, N., Shi, T. and **Kang, E. L.** (2010). Fixed rank filtering for spatio-temporal data. *Journal of Computational and Graphical Statistics*, 19, 724-745.
8. **Kang, E. L.**, Cressie, N. and Shi, T. (2010). Using temporal variability to improve spatial mapping with application to satellite data. *Canadian Journal of Statistics*, 38, 271-289.
9. **Kang, E. L.**, Liu, D. and Cressie, N. (2009). Statistical analysis of small-area data based on independence, spatial, non-hierarchical, and hierarchical models. *Computational Statistics and Data Analysis*, 53, 3016-3032.
10. Morton, R., **Kang, E. L.**, and Henderson, B. (2009). Smoothing splines for trend estimation and prediction in time series. *Environmetrics*, 20, 249-259.

SELECTED INVITED PRESENTATIONS

Invited presentation at the International Workshop/Conference on Bayesian Theory and Applications (IWCBTA), Varanasi, India, January 2013
Invited presentation in the Financial Statistics Seminar in the Financial Engineering Research Center of Suzhou University, Suzhou, Jiangsu, China, December 2012
Invited presentation at Department of Statistics, North Carolina State University, March 2012
Invited presentation at Department of Statistics, Purdue University, February 2012
Invited presentation at the Third North American Regional Meeting of The International Environmetrics Society (TIES), La Crosse, Wisconsin, July 2011

REVIEWER FOR: *Biometrics, Biostatistics, Canadian Journal of Statistics, Environmetrics, Journal of Agricultural, Biological, and Environmental Statistics, Journal of Computational and Graphical Statistics, Journal of the American Statistical Association, Journal of Time Series Analysis, Scandinavian Journal of Statistics, Spatial Statistics*

*** Note:** this grant is for non-teaching quarters without pay and leave quarters without pay. *****