Modeling Spatiotemporal Rates of Change in Interval-Censored Airborne Exposure Data from the Deepwater Horizon (BP) Oil Spill

As a second-year PhD student in Biostatistics specializing in methods for spatially-indexed data, collaborative research will be a key element of my future work. Currently, my methodological interests are in spatiotemporal modeling, that is, modeling data that are collected both spatially and over time while accounting for autocorrelation in the responses resulting from either source. In order to truly excel in this field, however, it will be necessary not only to be a developer of novel statistical methods, well-versed in the associated theory, but also to be well-trained in efficient computing, and knowledgeable of the underlying science. Given the time commitment required to achieve these goals as a biostatistician and complete my related dissertation work, I believe a Doctoral Dissertation Fellowship from the University of Minnesota Graduate School would offer ideal and invaluable support for my education.

While my Masters research and other student research assistantships led to several coauthorships in refereed journals, my first serious foray into biostatistical methods research occurred this past fall, when I completed the first paper from my dissertation. In this paper, along with my advisors, Dr. Brad Carlin and Dr. Sudipto Banerjee, I developed a Bayesian statistical approach based on areally-referenced temporal processes to model monthly asthma hospitalization rates in the counties of California (Quick et al, 2011). Compared to standard methods, Bayesian methods are better suited for complex modeling without excessive simplifying assumptions. While either approach would allow us to adjust for and judge the significance of various risk factors, my methodology also permits inference on temporal gradients (instantaneous rates of change), expanding upon previous work in the continuous space-only setting (Banerjee et al, 2003) and its discrete space-only counterpart (Lu and Carlin, 2005).

In a statistical model, residuals are the components of the response variables that remain after accounting for the predictor variables, or covariates. In my first paper, the statistical model was developed under a framework where observations were collected from a finite set of geographical regions over a continuous period of time. So, generally speaking, the temporal gradients allow us to learn how the residuals change over time. This is important, because, invariably, there will be important covariates that are not included in the model, and maps or time-series plots of the residuals can provide researchers with clues as to what these covariates may be. In the case of the asthma data, there was a noticeable decrease in hospitalization rates during the warmer summer months, and the temporal gradients revealed a gradual change over the years in precisely when the transition into and out of the summer occurred. Researchers familiar with risk factors for asthma may believe this is attributable to climate change, suggesting that weather-related covariates such as average monthly temperature should be included, while others may point to the implementation of a new public health policy. Whatever the case may be, a goal for this methodology is to provide researchers working with spatiotemporal data an additional statistical tool to more fully analyze their data and draw substantive conclusions.

Having said all this, my earlier work was somewhat narrowly focused on technical issues in biostatistical methods and computing, limiting its impact. To fully achieve my goals for my dissertation, my methods must be flexible enough that a wide-range of researchers can analyze
their data without unrealistic simplifying model assumptions, while also producing results that are easily interpretable to non-statisticians. Not only that, but my methods must also be able to be implemented on a standard computer in a reasonable amount of time; otherwise, they will be left by the wayside in favor of alternative approaches. Thus, for my work to extend beyond the pages of the statistical literature and impact the theory and practice of environmental science and its consequence for health outcomes, my methods must also be informed by an interdisciplinary team of experts, including epidemiologists, environmental scientists and computer scientists.

The motivation for this work is a dataset of airborne exposure data collected on the workers engaged in the clean-up following the Deepwater Horizon (BP) oil spill in the Gulf of Mexico in April, 2010. These data present a number of challenges that need to be addressed before their analysis can proceed. First and foremost, the data are point-level (i.e., each measurement takes place at a precise, known spatial location), rather than being areally aggregated (e.g., county-level data). While this permits inference on spatiotemporal gradients, allowing us to describe rates of change in both space and time, this will require more sophisticated methods to account for the more complex correlation structure. Whereas before we defined spatial correlation in terms of which areal regions neighbored one another (i.e., shared a border), here the correlation between two spatial coordinates can be more precisely related to the inverse distance between them. An alternative approach would be to aggregate the observations using a grid and use our earlier approach developed for the asthma data, but doing so would result in an undesirable loss of information.

In order to ensure that the statistical methodology developed to analyze these data remains focused on approaches and models relevant to aerosol science and accessible to researchers not expert in the complex statistical theory behind it, I have recruited Dr. Gurumurthy Ramachandran from the Division of Environmental Health Sciences to join my dissertation committee. An expert in the assessment of occupational and ambient exposures, Dr. Ramachandran is also quite familiar with existing Bayesian spatial methods through his previous work with my co-advisor, Dr. Banerjee, making him an excellent candidate to help me balance theory with application. Preliminary meetings with Dr. Ramachandran have already helped focus the work methodologically. For example, he alerted us to the large amount of observations below the detectable limit, suggesting our methods must be equipped to handle interval-censored data. Furthermore, his knowledge of airborne particle dispersion as it pertains to wind currents in the gulf will be crucial in the development of our spatiotemporal correlation structure (e.g., exposure levels at the site of the spill should have greater influence on future levels downwind than they should on levels upwind).

A challenge of using Bayesian approaches is the associated computational burden, as all unknown model parameters must be estimated iteratively via simulation. That is, unknown model parameters are assigned initial values, and then updated, one by one, using a Bayesian sampling algorithm, which is run until the variability of our estimates stabilizes. As the complexity of the statistical theory increases, so too does the associated computational complexity. In our asthma hospitalization study, we had 18 years of monthly observations from the 58 counties of California, causing us to have roughly 12,000 model parameters to estimate, posing relatively minor computational issues. Now, with over 150,000 personal exposure measurements from the workers and hundreds of thousands of area concentration measurements,
even storing and accessing the data, much less modeling it, will require highly efficient use of computational resources. For instance, the amount of data and model parameters may be too large to fit in a standard computer’s random access memory (RAM). Because of this, I am also fortunate to have the participation of Dr. Shashi Shekhar from the Department of Computer Science and Engineering on my dissertation committee. Dr. Shekhar will work with me to devise ways to both manage and analyze the enormous BP dataset. Not only is Dr. Shekhar an expert in computer science, but his primary area of expertise is in large spatial databases and the computational issues surrounding them. With his help, I hope to develop methods which will allow us to break up the data into manageable pieces that can be modeled separately, making the algorithm faster and more attractive for use by other researchers.

Due to the complex nature of this project, I feel that having a Doctoral Dissertation Fellowship would be invaluable. DDF support will free me from teaching and other, more routine biostatistical research support obligations, leaving sufficient time to fully commit to my dissertation research in biostatistical, computer, and environmental science. This will speed its completion, dramatically enhance its quality and depth, and permit me the time to submit it for publication and begin the associated revision process while I am still in graduate school. The payoff of this experience for me would thus be at least two-fold: not only would it accelerate my dissertation work, allowing me to focus on completing my degree by May 2013, but it would also help prepare me for a career as a well-rounded research biostatistician. After graduation, I hope to work in an academic or government research facility, building on the foundation in spatiotemporal modeling that I’ve established here at the University of Minnesota. In my work as a research biostatistician, I’ll collaborate with experts in a variety of fields in order to better develop and implement statistical methodologies to help solve complex interdisciplinary problems. In addition to advancing statistical methods, I will also welcome the challenge of impacting the way other researchers from other areas view these methodologies, whether by publishing leading methods papers in biostatistical journals, influential applied work in environmental science and public health journals, or by developing and distributing more efficient algorithms with computer scientists. Because of these goals—both for my dissertation and my professional development—as well as my ability to use this support to help achieve them, I feel I am a worthy candidate for Doctoral Dissertation Fellowship. Thank you for your consideration.

Key References

