LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

UCRL-TR-218541

# Spatial Statistical Procedures to Validate Input Data in Energy Models

G. Johannesson, J.S.Stewart, C. Barr

January 31, 2006

This work was performed under the auspices of the U.S. Department of Energy by University of California, Lawrence Livermore National Laboratory under Contract W-7405-Eng-48.

# Spatial Statistical Procedures to Validate Input Data in Energy Models

Gardar Johannesson, Jeffrey Stewart, Chris Barr

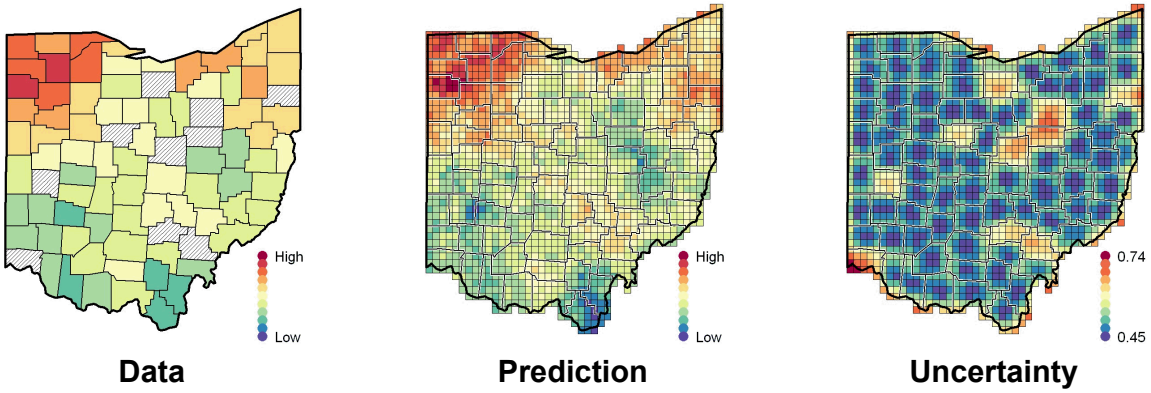
Lawrence Livermore National Laboratory

Acknowledgement to

Liz Brady Sabeff, Ray George, Donna Heimiller, Anelia Milbrandt

National Renewable Energy Laboratory

**Validation of Predictions Using Sparse Spatial Data**




**Data** **Prediction** **Uncertainty**

# Contents

# Abstract

Energy modeling and analysis often relies on data collected for other purposes such as census counts, atmospheric and air quality observations, economic trends, and other primarily non-energy–related uses. Systematic collection of empirical data solely for regional, national, and global energy modeling has not been established as in the above-mentioned fields. Empirical and modeled data relevant to energy modeling is reported and available at various spatial and temporal scales that might or might not be those needed and used by the energy modeling community. The incorrect representation of spatial and temporal components of these data sets can result in energy models producing misleading conclusions, especially in cases of newly evolving technologies with spatial and temporal operating characteristics different from the dominant fossil and nuclear technologies that powered the energy economy over the last two hundred years. Increased private and government research and development and public interest in alternative technologies that have a benign effect on the climate and the environment have spurred interest in wind, solar, hydrogen, and other alternative energy sources and energy carriers. Many of these technologies require much finer spatial and temporal detail to determine optimal engineering designs, resource availability, and market potential. This paper presents exploratory and modeling techniques in spatial statistics that can improve the usefulness of empirical and modeled data sets that do not initially meet the spatial and/or temporal requirements of energy models. In particular, we focus on (1) aggregation and disaggregation of spatial data, (2) predicting missing data, and (3) merging spatial data sets. In addition, we introduce relevant statistical software models commonly used in the field for various sizes and types of data sets.


# 1. Introduction

Energy modelers often try to predict the impact policies or new technologies may have on the energy economy and its associated environmental consequences. Some of these technologies have very different modeling requirements than those that powered the advanced energy economies of the twentieth century. According to Energy Information Agency estimates, domestic coal, natural gas, and nuclear made up 88% of U.S.

electricity production in 2003. These resources have similar energy modeling requirements in that they tend to be plants that are dispatchable, and the fuel supply is easy to estimate over various periods. Thus, estimating their average power production (capacity factor), capital, and operating cost fit neatly within traditional cost models and economic forecast models. They generally have the same spatial and temporal modeling requirements because they are dispatchable and because fuel supplies can be estimated; thus, they can easily match supply, demand, and prices. Some of the technologies that are of interest for the twenty-first century—including wind turbines, solar photovoltaic, biomass, and electrical storage—have much more complex spatial and temporal requirements than the dispatchable nuclear- and fossil-based technologies. The timing and magnitude of their power production do not follow the same patterns of dispatchable generation, and, in the case of biomass, it is more difficult to estimate fuel supplies. With growing interest in Renewable Portfolio Standards in several states and with federal support for renewable energy research, energy models will have to capture appropriate spatial and temporal elements of these new technologies to begin to estimate their penetration and benefits to an energy system.

The National Energy Modeling System (NEMS) is used as the official U.S. energy model for the Department of Energy. Baseline forecasts are developed with NEMS and published annually in the *Annual Energy Outlook*. The regionalization of NEMS was based on information available at the time it was developed in 1993 and on the energy system modeling requirements at the time of its development. For example, the demand modules (residential, commercial, industrial, and transportation) use the 9 Census divisions; the Electricity Market Module uses 15 supply regions based on the North American Electric Reliability Council (NERC) regions; the Oil and Gas Supply Module uses 7 onshore and 3 offshore supply regions based on geologic breakdowns; and the Petroleum Market Module uses 3 regions based on combinations of the 5 Petroleum Administration for Defense Districts. NEMS divides the year into 11 typical time periods, including night, day, evening for winter, spring-fall, and summer. It has also included summer and winter peak hours as the tenth and eleventh time periods. The average demand during these discrete periods is used to build a typical load duration curve. Input data for NEMS and other models run at similar scales require aggregation and

disaggregation of data sets originally collected at various scales. The spatial and temporal scales used by NEMS and other models adequately represent the energy system of the twentieth century and early twenty-first century. However, the increasing public interest in alternative generation has made intermittent technologies potentially more important than they were in the past. This will require more careful treatment of input data to avoid misrepresentation.

## 2. Why Intermittent Technologies Cannot Use the Same Scales as Coal, Nuclear, and Natural Gas

Intermittent generators can have a capacity factor pattern that ranges from zero to a maximum of 1 on any day throughout the year. Unlike dispatchable generators that are under the control of operators, intermittent technologies are driven by the forces of nature such as wind speeds, solar radiation, or seasonal biomass productivity. Lamont and Wu (2005) conducted a series of data comparisons of wind resources and energy demand over a full year to compare with commonly used methods of aggregation. Figure 1 represents the normalized electricity demand for the State of California for each hour of the year and is representative of the true pattern of demand. It shows that the range of energy demand throughout the year is from a low of 0.42 to the highest demand hour of 1 of normalized energy demand.

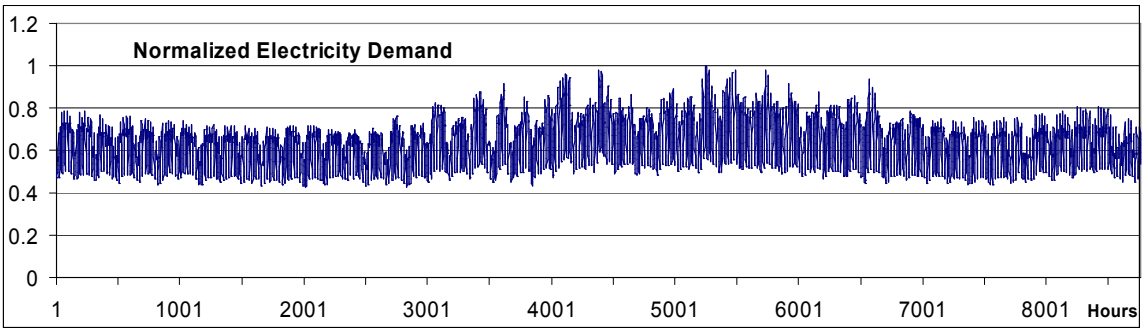



*Figure 1. State of California normalized hourly electricity demand for one year.*

Figure 2 shows the normalized actual wind production factors for one of the large wind sites in California, the Altamont Pass, for the same year. The variation spans between 0 and 1. The load-following ability of wind at this location does not match the electricity peak demand patterns, showing more randomness than the fairly smooth electricity

demand pattern.

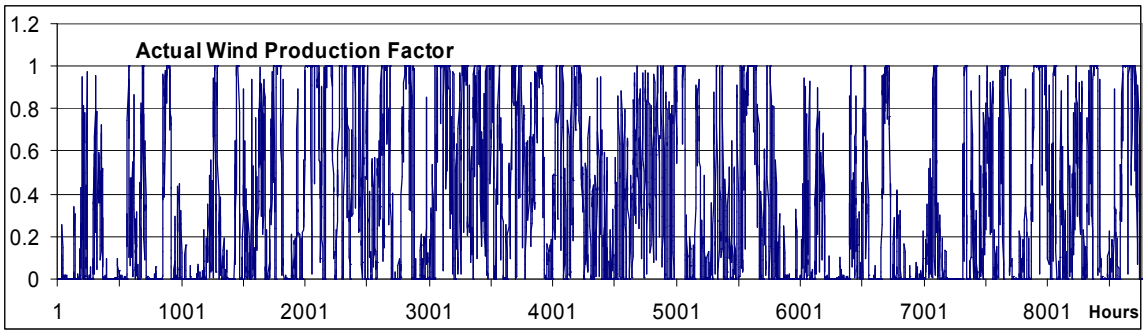
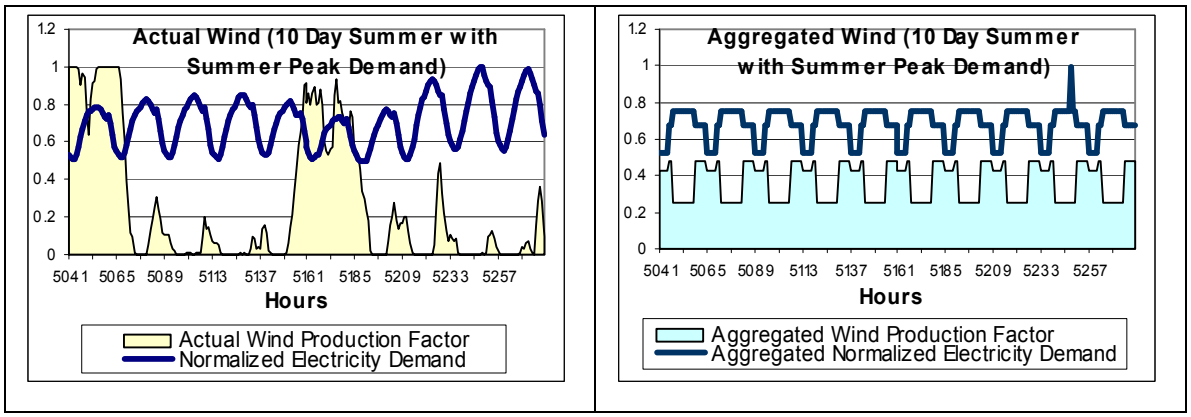



*Figure 2. Altamont, California, normalized wind production factors.*

Aggregation without test for bias can introduce errors into the final analysis. The two graphs in Figure 3 represent the potential problems of aggregation. Figure 3a and 3b represent the same 10-day period taken from the data set that produced Figures 1 and 2. The time period was reduced simply to improve the resolution of the graphs. Figure 3a represents the actual normalized wind from the Altamont wind site and demand data from the state of California. Figure 3a shows periods of high wind generation and periods of moderate to no wind generation. The electricity demand data show a smooth pattern of rising and falling demand through the 10-day period. Energy models reflecting this pattern would show results with zero generation coming from wind at times when there was demand for electricity. In a scenario where wind generators were the only producers, the system shows a pattern of producing more electricity than required and at other times not producing any energy during positive demand periods. Storage technologies would show a different result. Figure 3b on the right represents the aggregated data for the same period. The wind data have been represented by a continuous production pattern and by eliminating the periods of high production and zero production. The electricity demand pattern more closely resembles the actual electricity demand pattern in Figure 3a. One noticeable distinction is that the aggregated data displayed in the NEMS time steps mentioned above only displays one peak period that reaches 100%, while the actual data show three periods that approach 100%. Figure 3b shows a smooth pattern of wind resources that never reaches zero or exceeds demand. Aggregation in this example shows that the important information is lost and leads to the conclusion that the wind resources will always be used because they are always generating some power.




(a) (b)

*Figure 3. (a) 10-day actual wind production and normalized electricity demand. (b) 10-day aggregated wind production and normalized electricity demand.*

The ability to aggregate data depends on the level of bias being introduced. That will depend on other factors in the analysis, such as whether time-of-day pricing or fixed pricing is used and whether the wind production is higher during the peak hours and lower during the off-peak hours. If all-important factors are not understood, bias can be introduced, causing errors in the final analysis.

To determine capacity and cost factors, intermittent technologies often require information and data at a relatively fine spatial and temporal scale because of their (uncontrollable) natural variability. However, such information might not necessarily exist at the scales needed but be available at (courser) scales and be gathered for an altogether different purpose. The question is then how one proceeds to extract the relative information *needed* from the data *available* and then quantify the uncertainty in the produced result.

Statistical modeling techniques have been used successfully in numerous applications to estimate and predict variables not directly observed and to produce associated uncertainty measures. This applies particularly to the area of temporal, spatial, and spatiotemporal statistical modeling, where one generally expects nearby observations (in space and/or time) to be more alike than those observed further apart. For example, in the area of spatial statistics, it is central to characterize and quantify the extent of the spatial correlation in the available data. This provides useful information to decide on the resolution and scale one might want to use to represent or map the data and is also crucial

for accurate spatial prediction and interpolation that is based on the observed data.

An optimal prediction map produced by a spatial model is rarely the ultimate and final goal of an analysis but rather serves as a tool to answer "questions," which might have been the reason the map was produced in the first place. An example question is, What is the size of the area where the variable being mapped exceeds a given threshold? One possible answer is given by "plugging" the map into a procedure that produces an answer. Depending on the question being asked, in many cases such plug-in approaches might provide an answer of sufficient accuracy for any practical use. However, one has to keep in mind that such plug-in approaches fail to take into account the uncertainty associated with the map used.

Statistical models do not only provide optimal predictions or estimates, but they also produce uncertainty measures and, in some cases, whole distributions. Such uncertainty information can be valuable in assessing whether the available data is sufficient to produce, for example in a spatial setting, a map of the quality needed for reliable analysis. Alternatively and preferably, the uncertainty information can be propagated through the post-analysis procedures applied to yield uncertainty bounds on the answers produced. In the wind example mentioned above, if a probability distribution over the mean wind power were produced, an energy modeler would be able to improve an analysis by incorporating the uncertainty into the aggregated time steps. Employing a probability distribution over the mean wind power would improve the accuracy of the estimated value of the wind generator.

We will introduce several examples of spatial statistical modeling that should prove useful to researchers interested in testing their data assumptions for energy analysis.[1] The examples that follow are designed to test the data's suitability to answer specific questions. Before determining the specific spatial statistical procedures to use, the energy modelers must have a good understanding of the question they are trying to answer, including the level of spatial and temporal resolution desired. The spatial modeling teams can then begin to select the appropriate statistical procedures for

[1] There are many statistical models and tests that can be used, and we barely scratch the surface in this paper.

extracting the best representation of the information desired by the energy modelers.

Section 3 introduces two real energy-related data sets of spatial nature that partly motivated this paper: state-specific residue biomass data and nationwide electricity load data. The two data sets were provided by the researchers acknowledged in this paper as typical data sets they use to provide data for energy modelers. Section 4 introduces spatial statistical modeling with a focus on models that are relevant to the problems raised in Section 3 and gives an application to the residue biomass data and the electricity load data. Section 5 follows up with a discussion and conclusion.

# 3. Energy Related Spatial Data

We now introduce the two energy-related data sets with spatial features that, along with others, motivated this paper. In particular, we give a description of what is observed and reported in each data set and then what is sought, that is, what questions need to be answered. We then follow with a general description and notation for spatial data that we use throughout this paper.

## *Estimating Residue Biomass Spatial Distribution (Minnesota Example)*

The National Renewable Energy Laboratory (NREL) biomass assessment is based on statistical data reported by county to the United States Department of Agriculture (USDA), the USDA Forest Service, the Environmental Protection Agency (EPA), and other public and private organizations (Milbrandt, 2005).

The following biomass is included: forest, major crops, primary mill, secondary mill, urban residue, methane emissions from landfills, and animal manure. In this paper, biomass and residue (waste biomass material) will have the same meaning of energy source from plants or animal waste.

The spatial resolution of the current county-based residue raw data might, in many cases, be insufficient to answer accurately some energy-related questions, for example, transportation distances and transmission line proximity to the site of interest for cost analyses. Typically, an analyst would either associate the county-specific data

with a single spatial location within each county or assume that the residue biomass is evenly distributed within each county. In reality, the residue can resemble any number of patterns that could affect the costs and viability of power generation. Possible data to support better spatial distribution within each county include land use (crop and forest residue), animal farm locations (animal manure), population (urban waste), and mill and landfill locations. For example, Figure 4 shows land use data sets that could be used to provide a finer spatial presentation of crop and forest residue biomass.

The question is then whether it is possible to produce a realistic finer-resolution residue biomass map than can be produced at the county level. If so, what is the gain of using those finer-resolution maps instead of assuming that the biomass is evenly distributed within each county?

(a) Residue Biomass (b) Residue Density (c) Land Use

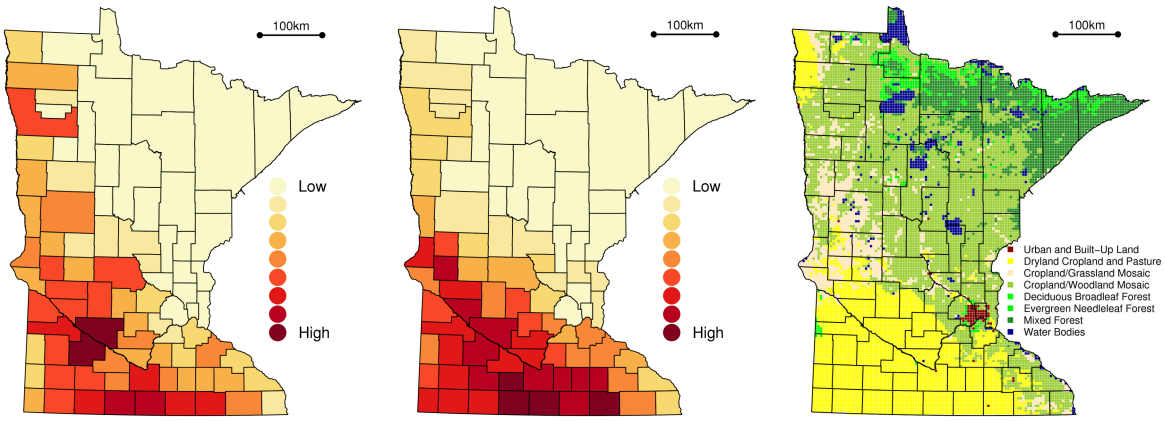



*Figure 4. Minnesota annual crop residue biomass (a) and density (b) by county, along with land-use pattern (c).*

## *National Electric Load and Demand Data by Control Areas*

Electricity load and demand data are sporadically available at utility, market area, or control-area levels. For national energy modeling, coverage at the control-area level is easiest to acquire for consistent reporting across the area of interest. However, control areas represent relatively large areas that are not consistent with the modeling regions of interest (North American Electric Reliability Council ((NERC)), State Boundaries, Utility Territories Census, NEMS or other), and yet the load data need to be allocated appropriately to those modeling regions for accurate representation of input data for

various energy models. In the following sections, we will give some examples of spatial modeling techniques to address the questions of how to aggregate data sets, use explanatory variables to improve the modeled output, and test for spatial variation.

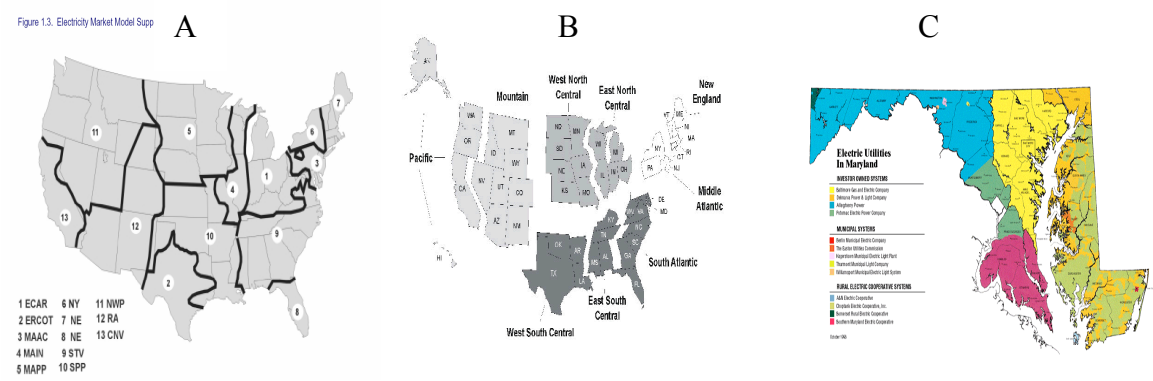



*Figure 5. Examples of various region scales for which data are collected (left to right: NERC regions, Census divisions, Utility District). Source: A. The National Energy Modeling System http://www.eia.doe.gov/oiaf/aeo/overview/figure_10.html. Source .B The National Energy Modeling System http://www.eia.doe.gov/oiaf/aeo/overview/figure_1.htm. Source C http://www.choosemaryland.org/datacenter/utilities/terr_elec.asp*

One way electricity load data has been aggregated and disaggregated is by assuming that electricity load is directly proportional to population. This assumption can work well in situations where other possible explanatory variables such as income level and commercial and industrial activities do not have much influence or are evenly represented in each spatial area of interest. Population data are collected at a very fine spatial resolution, and are easily available from the United States Census Bureau; industrial facility information is less readily available and does not have one single data provider responsible for disseminating it in formats useful for energy analysis. Other factors that may or may not be significant influences on the electricity load distribution include climate, cost of electricity, and housing trends.

Figure 6 (left) shows average summer-peak electricity load as reported for 2001 by control areas for part of the United States. There are 116 control regions that vary considerably in size and spatial coverage. Figure 6 (right) better demonstrates this by showing the spatial coverage of control regions in an area centered on Illinois (county borders are outlined). We also note that some areas within the U.S. are not assigned to any control region (shown as white areas in Figure 6). We begin the next section with an

introduction to spatial statistics.

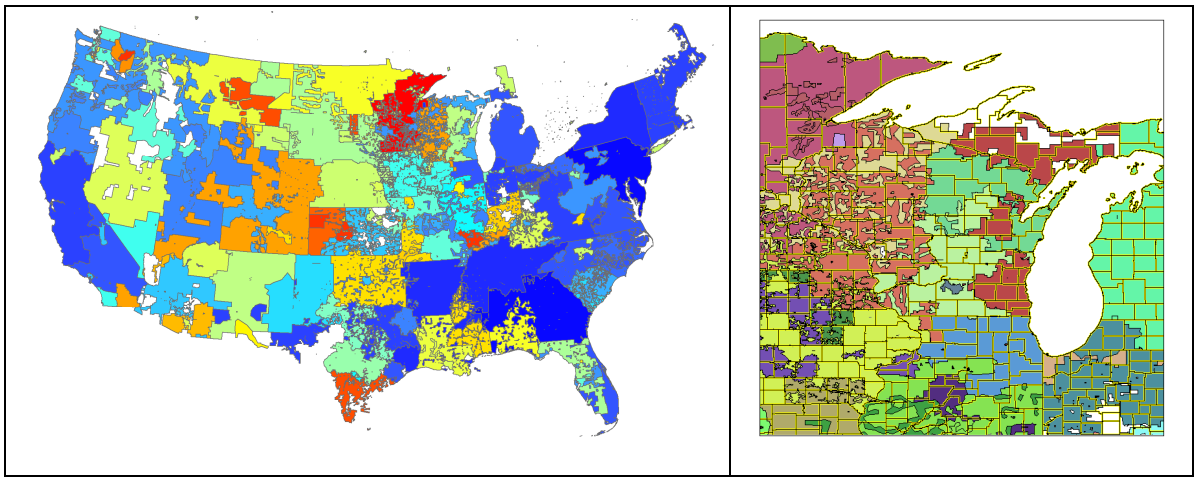


*Figure 6. Left, 2001 average summer peak electricity load (demand) by control regions (red = low, blue = high). Right, an example of the variation in the spatial coverage of control regions (colored areas; each control region has a different color) with respect to counties (black lines on yellow background).*

## *Introduction to Spatial Data and Notation*

Let $D$ be our spatial domain of interest. In the case of the biomass data, it is the state of Minnesota, and in the case of the electricity load data, it is the whole of the U.S. Given a spatial domain, we mainly focus on two primary types of spatial data that are observed within $D$: point-referenced data and areal data.

For point-referenced data, each observation (datum) is associated with a point location (Figure 7a). An example would be the observed wind speed at a given time at a given site. We denote such data by

$$Z_i = Z(s_i) = \text{the } i\text{-th observation at point-location } s_i \text{ in } D,\ i = 1, \ldots n$$

where $Z_i = Z(s_i)$ is recorded for each of the $n$ spatial point locations $s_1, \ldots s_n$. $Z(\cdot)$ is often referred to as the *data process*.

For areal data, each observation is associated with an areal unit (a pixel, a cell, a zone; see Figure 7b). An example would be both the biomass data and the electricity load data. We denote areal data by

$$Z_i = Z(D_i) = \text{the } i\text{-th observation at areal unit } D_i \text{ in } D,\ i = 1, \ldots n$$

In many instances, areal data is the result of spatial aggregation of the biomass data and the load data. For these two data sets, the underlying spatial data process 'lives' at a finer spatial scale but is observed at an aggregated scale. We will come back to this later.

There is a natural extension of spatial data to space-time data, where the temporal index can either be a point (a snapshot in time) or an interval (a time period). In this setup, $Z_i = Z(s_i, t_i)$ denotes, for example, the $i$-th observation that is taken at point-location $s_i$ at time $t_i$.

(a) Point-Referenced Data (b) Areal Data

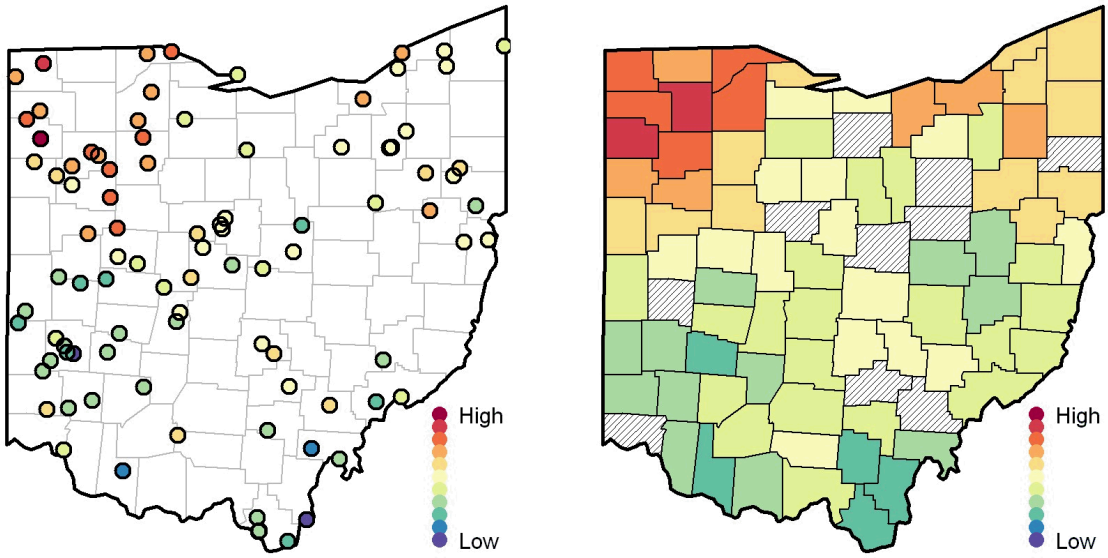



*Figure 7. Example of point-referenced data (left) and areal data (right).*

# 4. Spatial Analysis and Statistical Modeling

Here we briefly introduce classical statistical modeling of spatial data, including exploratory analysis of spatial data. We focus on the underlying assumptions of these models (in somewhat simplified form) and what they are capable of providing, but putting less emphasis on the technical details of their inner workings. See Cressie (1993), Banerjee, Carlin, and Gelfand (2004), and Schabenberger and Gotway (2005) for further details and extensions.

Before giving an overview of classical statistical spatial models, recall the problems associated with the use of the reported biomass residue data and the electricity load data; both are reported on a spatial scale that might not be fine enough or appropriate to answer the questions they are intended to answer for energy analysis (for example, supply, capacity, and energy-economic–related questions). The approach we take to address these problems is to develop a spatial statistical model for the data (biomass,

load, etc.), fit the model to the available (aggregated) data, and finally, predict at the spatial scale needed for further analysis (for example, supply and cost analysis). The statistical spatial model provides optimal predictions and also provides associated uncertainty measure, which can be useful for further analysis.

## *Point-Referenced Spatial Models (Geostatistics)*

We now give a brief overview of classical statistical spatial models, often referred to as geostatistical models, for point-referenced data.

Consider the point-referenced data vector $Z = (Z(s_1), \ldots Z(s_n))^T$, where $x^T$ denotes the transpose of a vector $x$. The data are thought to be realizations from a *spatial process* $Z(s)$, where $s$ is any location within the spatial domain $D$, but the process is only observed at the given locations, $s_1, \ldots s_n$. Our main goal is to predict the *Z-process* at any unobserved location (point prediction) or predict the *average* of the $Z$-process over a given region (block prediction) and provide measure-of-prediction accuracy.

The characterization and modeling of the spatial correlation in the $Z$-process plays a central role in statistical spatial modeling. Assume, for the moment, that the $Z$-process has mean equal to $\mu$ and variance equal to $\sigma^2$; that is

$$\mathrm{E}[Z(s)] = \mu \text{ and } \mathrm{Var}[Z(s)] = \mathrm{E}[(Z(s) - \mu)^2] = \sigma^2 \text{ for any } s \text{ in } D$$

The $Z$-process is not assumed to be uncorrelated but is assumed to have a (spatial) *covariance* structure

$$\mathrm{Cov}[Z(s), Z(s + h)] = \mathrm{E}[(Z(s) - \mu)(Z(s + h) - \mu)] = \sigma^2 K(s, s + h)$$

where $K$ is the spatial correlation (in the $Z$-process) between two locations. The geostatistical model is often written in terms of the two components

$$Z(s) = \mu + \delta(s)$$

where $\mu$ is the mean of the process and $\delta(s)$ is a zero mean spatial process with variance-covariance structure given by $\sigma^2 K(s, s + h)$. Hence, the first component captures the large-scale feature, and the second component captures the small-scale variation around the mean; we shall expand on this later. For the observed data vector $Z$, the geostatistical model can be written in matrix notation as

$$Z = 1\mu + \delta \qquad (1)$$

where 1 is a vector of 1's of length $n$, and $\delta = (\delta(s_1), \ldots \delta(s_n))^T$ has mean 0 and variance-covariance matrix $\sigma^2 K$, where the correlation matrix $K = (K(s_i, s_j)$.

Given the mean parameter $\mu$, the variance $\sigma^2$, and the spatial correlation function $K$, one can produce optimal prediction, referred to as *kriging*, of the $Z$-process at any location $s_0$ in $D$, based on the observed data and without making any further assumption about the $Z$-process (see, for example, Cressie, 1993, Chapter 3). However, the assumption of normality is often added, with the small-scale variation ($\delta$) assumed to be normal (Gaussian) distributed. With this added assumption, the prediction of $Z(s_0)$, given the data, $Z$, can be summarized by a probability distribution, in fact by the normal distribution in this case,

$$\Pr(Z(s_0) \mid Z) = \text{Normal}(\text{mean} = M(s_0; Z), \text{variance} = V(s_0; Z))$$

where $\Pr(Z(s_0) \mid Z)$ denotes the probability distribution of $Z(s_0)$ given the observed data, $Z$. It is worth noting that the data-dependent mean of the predictive distribution above, $M(s_0; Z)$, is identical with the optimal prediction value derived without assuming normality. Similarly, the variance, $V(s_0; Z)$, is the same as the uncertainty associated with the optimal prediction. In addition to predicting at a point, the geostatistical model can easily predict $Z(D_0)$, the average value of the $Z$-process over the area $D_0$ within $D$.

It is rarely the case that the mean and the variance-covariance structure of the $Z$-process are known. However, there are well-established methods to estimate those when they are unknown. The extra uncertainty introduced in the estimation procedures is propagated to the predictive distribution and manifests itself in wider prediction uncertainty. (We note that there are cases where the estimation uncertainty is only partly propagated to the prediction uncertainty.)

When the large-scale trend ($\mu$) of the $Z$-process is (partly) unknown, it is typically approximated by a linear regression function

$$\mu(s) = o(s) + f_1(s)\,\beta_1 + \ldots + f_p(s)\,\beta_p = f(s)^T\beta$$

where $o(s)$ is a known offset (trend), the $f(s) = (f_1(s), \ldots f_p(s))^T$ are $p$ known functions of the location $s$, and the $\beta = (\beta_1, \ldots \beta_p)^T$ are unknown regression parameters to be estimated. For example, if $s$ is a 2D location, $s = (x, y)$, a linear large-scale trend in the location coordinates is given by $\mu(s) = \beta_1 + \beta_2 x + \beta_3 y$. When the small-scale variation

($\delta$) has an unknown correlation structure, it is typically modeled with a parametric model that only (assuming isotropic process) depends on the distance between locations:

$$K(s, s+h) = K_0(h; \phi)$$

where $K_0$ is a parametric correlation function with parameter vector $\phi$. An example of such isotropic spatial correlation function is the exponential correlation function

$$K_0(h; \phi) = \exp(-h/\phi)$$

where $\phi > 0$. Further, the variance ($\sigma^2$) of the process might not be constant throughout the spatial domain $D$ and as such allowed to vary, spatially, similarly to the large-scale trend. In general, there is a large flexibility and much literature for explaining how one can model the large-scale trend, the small-scale spatial variation, and correlation.

In addition to allowing the large-scale trend $\mu(s)$ to vary with spatial location, one can use external input variables to help explain the variation in the $Z$-process. For example, if $x(s)$ is a vector of input variables that are available at any location $s$ and known to be (potentially) related to the $Z$-process, the large-scale trend can be extended to incorporate $x(s)$ via

$$\mu(s) = o(s) + f(s)^T\beta + x(s)^T\eta$$

where $\eta$ is a vector of regression parameters to be estimated.

In many cases (most, some would say), the observed data is corrupted by error; however, we seek to predict the underlying error-free process. To make this distinction clearer, we write

$$\begin{aligned} &\text{Data Model: } Z(s_i) = Y(s_i) + \varepsilon(s_i) \\ &\text{Process Model: } Y(s_i) = o(s_i) + f(s_i)^T\beta + x(s_i)^T\eta + \delta(s_i) \end{aligned} \qquad (2)$$

where $Y(s_i)$ is now the error-free process we want to predict and $\varepsilon(s_i)$ are (independent) observation errors. Note that if the above model does not include the small-scale spatial process term $\delta(s_i)$, the model would be a simple linear regression. As such, the model (2) can be thought of as extending the classical linear regression model to take advantage of spatial association.

We finally note that the observed data do not necessarily need to be associated with points. One can estimate a point-referenced spatial model using aggregated data,

$Z(D_1)$, ..., $Z(D_n)$, often referred to as the *change-of-support problem* (see, for example, Banerjee et al., 2004, Chapter 6, for an overview). This simply follows from the fact that a spatial covariance function that is defined between any two points also yields a spatial covariance function between a point and a region (block) and, more generally, between any two blocks (even if they overlap). Similarly, other terms associated with the large-scale trend can be aggregated to the aggregation blocks in question. Hence, given the aggregated data, one can still predict the error-free $Y$-process at any given point, and in particular, predict the $Y$-process on a given set of grid-points (or pixels) for mapping. From a practical point of view, one often defines in advance a set of spatial grid points (or pixels) that covers the spatial domain $D$ of interest, and the $Y$-process is predicted at those grid-points. This grid can then be aggregated up, in different ways, to yield the regions $D_1$, .., $D_n$ on which the data is observed. In this case, model (2) can be written as

$$\begin{aligned} &\text{Data Model} \quad Z(D_i) = \frac{1}{|D_i|}\sum_{j=1,\ldots,N} A_{ij} Y(B_j) + \varepsilon(D_i) \\ &\text{Process Model: } Y(B_i) = o(B_i) + f(B_i)^T\beta + x(B_i)^T\eta + \delta(B_i) \end{aligned} \tag{3}$$

where $B_j$ , $j = 1, \ldots N$, is the $j$-th pixel and $A_{ij} = |D_i \cap B_j|$ is the area of the intersection between $D_i$ and $B_j$. Often the pixels $B_j$'s are small enough compared to the $D_i$'s so that they can be well represented by grid points $s_1$, ..., $s_N$. As such, $A_{ij} = 1$ or 0 (assuming |Bj| = |B| = 1, to simplify), depending on whether the $j$-th grid point is or is not within the $i$-th region.

## *Areal-Data Spatial Models (Lattice Models) and Misaligned Data*

Assume we have observed areal data $Z(D_1)$, ... $Z(D_n)$, and our interest is only to predict at those same areal units, that is, smooth the data. The basic ingredients for the areal data models are very similar to that of the point-referenced model (2) and its block-aggregated extension (3):

$$\begin{aligned} &\text{Data Model: } Z(D_i) = Y(D_i) + \varepsilon(D_i) \\ &\text{Process Model: } Y(D_i) = o(D_i) + f(D_i)^T\beta + x(D_i)^T\eta + \delta(D_i) \end{aligned} \tag{4}$$

The main difference lies in how the spatial correlation associated with the small-scale variation term $\delta(D_i)$ is modeled; it does not have to be derived from a smooth underlying point process that is aggregated up to the areal units in question. The correlation structure

for the point-referenced model is based on a valid distance measure between two points. This is not often easily transferred over to areal units, although, there are times when one might want to represent the spatial associate among areal units using, for example, the location of their centroids. However, an alternative is to base the correlation of $\delta(D_i)$ on a *neighbor* structure (Figure 8a and 8b). For example, two areal units might be labeled as neighbors if they are adjacent (adjacent counties, pixels, etc.) or their centroids are within a given distance.

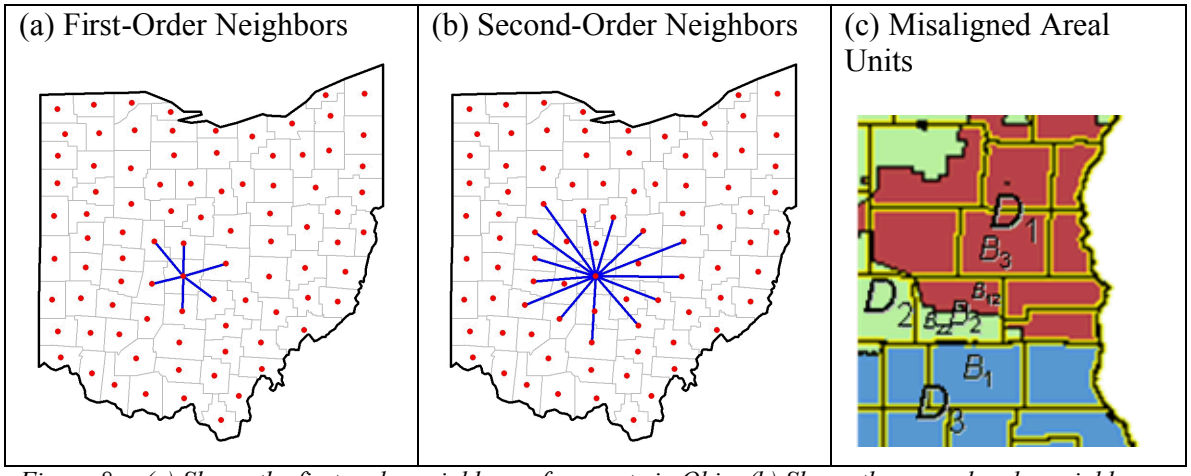



*Figure 8. (a) Shows the first-order neighbors of a county in Ohio. (b) Shows the second-order neighbors. (c) Shows misaligned areal units with observations available at the $D_i$ units, external data available at $B_j$ units, while the process is modeled at the $B_{ij}$ units.*

Two areal-data spatial models have become popular: the conditionally autoregressive (CAR) models and simultaneous autoregressive (SAR) models (see, for example, Banerjee et al., 2004, Chapter 3). Both models yield a parametric model for the *inverse* of the spatial correlation structure; that is, they yield a model for $\text{Cov}[\delta]^{-1}$, where $\text{Cov}[\delta]$ is the spatial covariance matrix of small-scale variation vector $(\delta(D_1), \ldots \delta(D_n))^T$. In addition, both models assume that the small-scale variation is distributed as a normal (Gaussian) distribution.

The CAR model is specified through a set of conditional distributions

$$\delta_i \mid \delta_{-i} \sim N\left(\rho \sum_{j=1,\ldots,n} w_{ij}\delta_j, \tau_i^2\right), \quad i = 1,\ldots n \qquad (5)$$

that specified the distribution of $\delta_i = \delta(D_i)$, conditional on all the remaining $\delta$'s, $\delta_{-i} = \{ \delta_j$

: $j = 1, .., i - 1, i + 1, n$ } as a normal distribution with mean equal to a weighted average of its neighbor, with neighborhood weights given by $w_{ij}$, with $w_{ii} = 0$. There are certain conditions that the weights need to satisfy so that (5) yields a valid spatial covariance model (for example, Banerjee et al., 2004, p. 78). The SAR model is specified in a similar way, and we refer to Cressie (1993, Chapter 6) and Banerjee et al. (2004, Chapter 3) for further details.

There are a number of cases in which one observes data at one set of areal units, $Z(D_1)$, ... $Z(D_n)$, but it might be more convenient to model the process at altogether different areal units, $B_1$, ... $B_N$; for example, due to availability of external data reported on the $B_1$, ...$B_N$ units. This is a case of *misaligned data* (for example, Banerjee et al., 2004, Chapter 6), also referred to as the *modifiable areal unit problem*. It might be the case that the $B_j$'s are nested with the $D_i$'s (for example, counties within states), but it is more often that some of the $B_j$'s contribute to one or more $D_i$'s, as in Figure 8c, which shows how some counties contribute to more than one electricity control region. In general, the spatial model can be written as follows:

$$\begin{aligned} &\text{Data Model} \quad Z(D_i) = \sum_{j=1,\ldots,N} A_{ij} Y(B_{ij}) + \varepsilon(D_i) \\ &\text{Process Model: } Y(B_{ij}) = o(B_{ij}) + f(B_{ij})^T \beta + x(B_{ij})^T \eta + \delta(B_{ij}) \end{aligned} \tag{6}$$

where $B_{ij} = D_i \cap B_j$ and $A_{ij} = 1$ if $D_i$ and $B_j$ do overlap, but 0 otherwise.

Estimation of CAR or SAR parameters is carried out using maximum likelihood-based methods.

## *Spatial Exploratory Data Analysis*

Exploratory data analysis is a very important ingredient to any statistical analysis. In spatial settings, classical exploratory data analysis can be used to infer large-scale trends and potential use of external input variables. The *empirical semivariogram* is a powerful tool for exploring spatial correlation in point data and even areal data that is observed on somewhat 'regular' units (for example, on units where the 'distance' between units can be represented by the distance between centroids). For point data, the classical variogram is defined as (Cressie, 1993, Chapter 2)

$$2\hat{\gamma}(h) = \frac{1}{N(h)} \sum_{(i,j)\in N(h)} (Z(s_i) - Z(s_j))^2 \qquad (7)$$

where $N(h) = \{ (i, j) : |s_j - s_i| = h, i < j \}$ is the (unique) set of points separated by a distance $h$, and $|N(h)|$ is the number of such pairs. Hence, the empirical variogram is the average difference-squared between data separated by a distance $h$. In practice, one forms 'bins' of data pairs that are approximately separated by a given set of distance lags and plots (4) versus the distance $h$. If data nearby are alike, the variogram should be close to zero for small $h$. If the data has a constant mean and covariance given by $\text{Cov}[Z(s), Z(s+h)] = \sigma^2 K(h)$, then

$$\text{Var}[Z(s) - Z(s+h)] = \text{Var}(Z(s)) + \text{Var}(Z(s)) - 2\text{Cov}[Z(s), Z(s+h)] = 2\sigma^2 (1 - K(h))$$

Note that (4) is an empirical estimate of this variance of the difference. The *theoretical* variogram is defined as $2\gamma(h) = \text{Var}[Z(s) - Z(s+h)]$, where $\gamma(h)$ is referred to as the *semivariogram*. Note that $\gamma(h) = \sigma^2(1 - K(h))$, and its empirical counterpart in (4) therefore provides information about the variation, $\sigma^2$, and the spatial correlation function $K(h)$. There are more robust estimates of $\gamma(h)$ than the classical estimator in (4), in addition to semivariograms that explore directional variation in the spatial correlation (anisotropy); see Cressie (1993, Chapter 2).

The empirical semivariogram is not only computed for the raw observed data but more importantly used to explore the potential spatial variation in the small-scale term $\delta(s)$ of (3) and therefore computed for the residue $Z(s_i) - [o(s_i) + f(s_i)^T\beta + x(s_i)^T\eta]$, given some estimates of $\beta$ and $\eta$ (for example, from a linear regression ignoring the possible spatial correlation).

Similar empirical statistics are available for areal data that can be used to explore the spatial association in neighboring areal units; see Moran's I and Geary's C statistics (Schanbenberger and Gotway, 2005).

# 5. Application

In Section 4, we introduced residue biomass data for Minnesota and national electricity load data by control areas. Both data sets are provided on an aggregated scale

that is considered to be too coarse for further use. We shall now apply spatial exploratory and modeling techniques to these data and predict the underlying (generating) processes of interest at a finer scale. First, we give an example of the impact spatial aggregation can have on inference.

## *Average Wind Data, Aggregation, and Spatial Correlation*

We shall now investigate the impact aggregation has on variation and spatial correlation using an estimate of the annual average wind-power density at 50-m height, as provided by TrueWind Solutions. Figure 9 shows the wind-power density over a region in California (centered on San Francisco East Bay) at its provided 200-m resolution and five aggregated scales. In this example, we are interested in the size of the region that exceeds a certain threshold (i.e., capacity), where we take what is reported at 200 m as the true data.

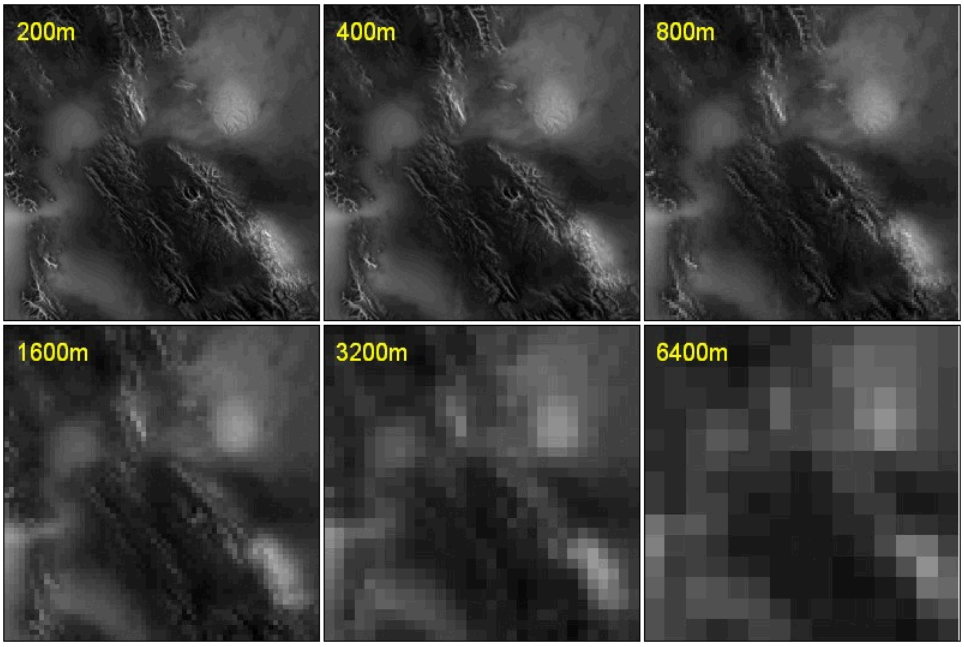
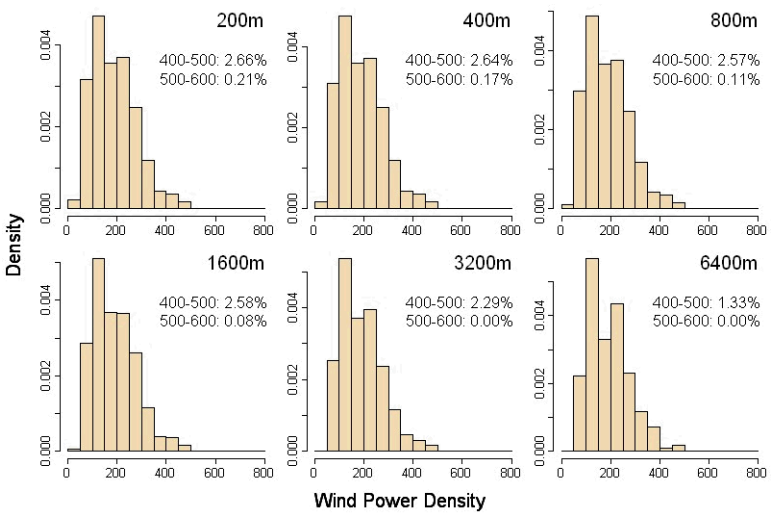



*Figure 9. Average annual wind density at 50-m height at different resolutions (dark is low, white is high).*

Figure 10a shows the histogram of the wind-power density at the six resolutions shown in Figure 9 along with the six empirical semivariograms shown in Figure 10b. The scales 400–500 and 500–600 in Figure 10a correspond with the most favorable wind classes 4 and 5, respectively. The histogram shows the magnitude of data loss with increasing aggregation. This histogram can be used as one of the early analyses to show

energy modelers for a joint discussion on determining the appropriate resolution required by the energy modeling team to use in their analysis. If class 5 is significant, this example shows a loss of almost 52% of the class 5 resource as one moves from 200- to 800-meter resolution. The same aggregation only shows a loss of approximately 3% of the class 4 resource. The energy modeler will need to determine the significance of the loss of information versus the extra computation time required to model data sets at finer scales. Aggregation reduces variation, shrinking the histograms as the resolution gets coarser. However, aggregation has relatively little impact on the range of the spatial correlation (i.e., when the semivariogram starts to level off), which is around 20–30 km. Aggregating down to one-quarter of the correlation range (i.e., about 6000 km) is clearly too course, while 200 m might be excessive. What if the data were only available at the coarsest resolution, and one could therefore not carry out the histogram comparison presented? The spatial statistical properties of the process can be estimated from the coarsest-resolution semivariogram shown and used to produce a finer-resolution map with the same histogram properties as those shown.

(a)



(b)

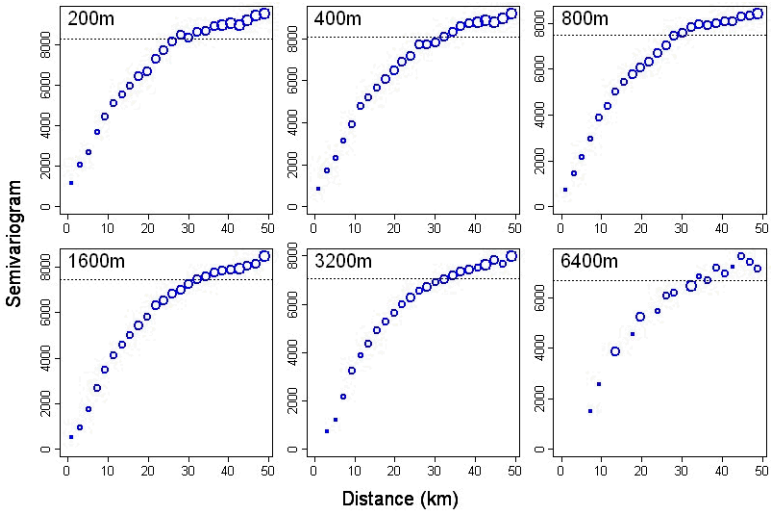



*Figure 10. (a) Histogram and (b) empirical semivariogram for the wind-power density reported at the six different resolutions shown in Figure 9.*

## *Minnesota's Annual Crop Residue Biomass Data*

From Figure 4a we conclude that the crop residue biomass by county has a smooth large-scale variation. The large-scale feature of the process obviously reflects the land use pattern in Figure 4c. Our goal with this analysis is to estimate the amount of crop residue biomass at a given distance from a given location (for example, a coal power plant site). Given this knowledge, one can carry out transportation feasibility studies. We shall simply use geodesic distance ("by air" distance) as a proxy for "by road" distance. Our goal is to compare different methods, and we expect the relative difference between the methods to hold for a more realistic distance measurement.

Two trivial methods are often used to accomplish this in practice. The first one simply assigns the biomass within each county to a given point location, for example the centroid. The second one assumes that the biomass is evenly distributed within each county. Both methods "honor" the data, in the sense that when aggregated back up to the county level, they yield the originally reported biomass for each county. The drawback of the first method is obvious when it comes to estimating the amount of biomass within a given distance from a site, especially at relatively short distances. The second method can be considered slightly better, but its accuracy depends on the validity of the assumption of evenly distributed biomass within each county; we note from Figure 4b that this

assumption yields an artificial "blocky" biodensity pattern along the counties' borders. The main question is whether we can improve on evenly distributing the biomass within each county, or is that simply sufficient for the application(s) at hand?

We now apply two spatial models that predict the biomass at a spatial scale finer than the county level. The first model does not take advantage of the external land-use data, but the second model does. The first step is to decide what spatial scale is sufficient to represent the underlying biodensity process. Figure 11 shows the empirical semivariogram of the biodensity using the distance between county centroids as a course proxy for the "distance" between counties.

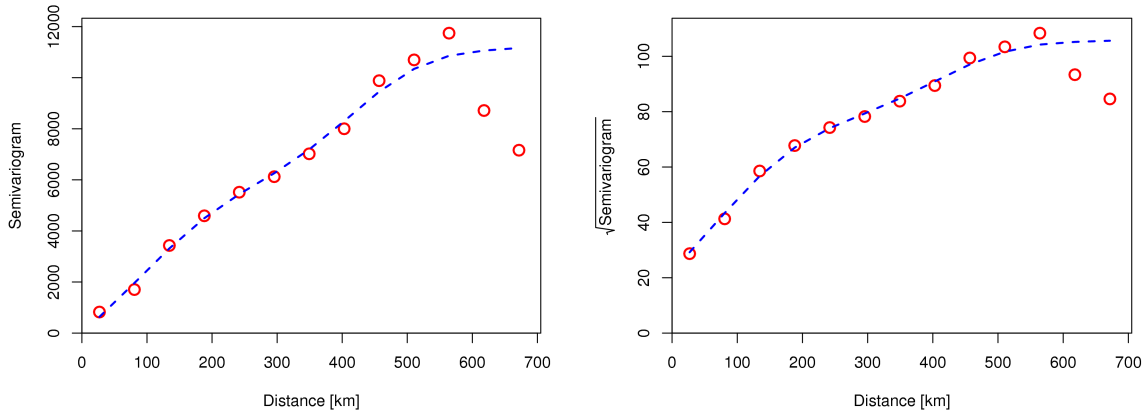



*Figure 11. Semivariogram of county residue biodensity (right) and the square-root of the semivariogram. Distances are based on county centroids, and the density is in tonnes/km$^2$.*

As expected, the semivariogram shows strong spatial correlation due to the smooth variation in the biodensity (Figure 4c). Such smooth variation suggests that we can represent the biodensity at a relatively course scale. The second step is to consider to what the spatial scale is relative, for example computing transportation cost. Note that we aim to map the biomass on a spatial grid and then, for example, when computing the amount of biomass with a given distance from a site, simply count the grid points within that distance. For this analysis, we choose a grid with 4-km grid-point spacing. This yields 13,663 grid points that cover Minnesota. Given further information about transportation cost, one might either increase or decrease the resolution. In terms of notation, let

$s_j$ = the location of the $j$-th grid-point, $j$ = 1, ..., $N$, $N$ =13,663

$Y_i = Y(s_j)$ = the biomass at the $j$-th grid point (what we want to predict)

$Z_i = Z(D_i)$ = the reported biomass at the $i$-th county, $i$ = 1, ..., $n$, $n$ = 87

where $Di$ is the area of the $i$-th county.

In addition, let $Z = (Z_1, ..., Z_n)^T$ and $Y = (Y_1, ..., Y_N)^T$. We adopt the spatial model of (4), which we can write in matrix notation as

$$Z = A\,Y + \varepsilon\,, \text{ with } Y = o + F\,\beta + \mathrm{X}\,\eta + \delta$$

and recall that A is an i-times-$N$ allocation (aggregation) matrix, with the $(i,j)$ element equaling 1 if the $i$-th grid point is in county $j$, but 0 otherwise. We shall assume that $\varepsilon = 0$, zero measurement error, which is in line with the two trivial methods mentioned earlier.

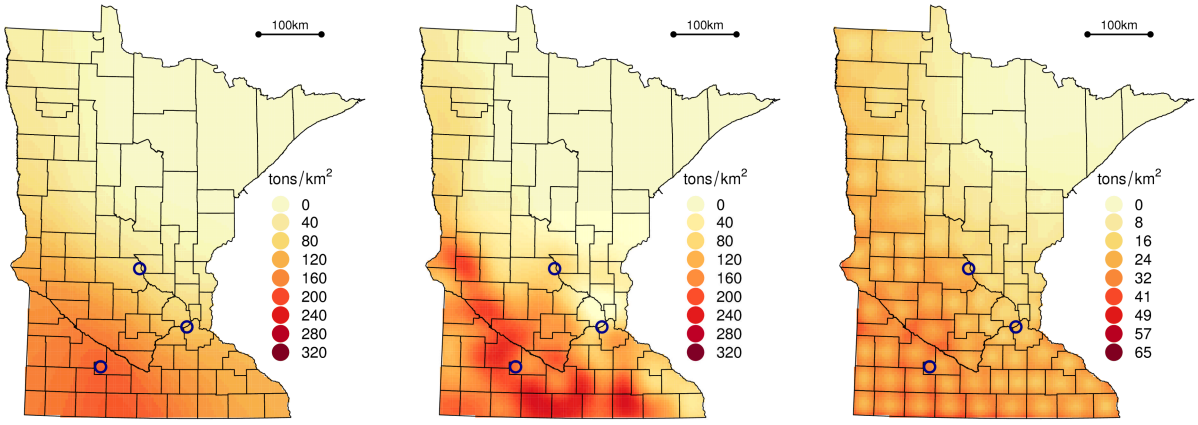
Our first spatial model does not take advantage of the land use data, that is, $X\,\eta = 0$. A spatial kernel smoother (locally weighted average) was applied to the county biodensity data (assigned to the centroid of each county) to extract the smooth spatial trend seen in Figure 4a; this yields the offset $o$ in (3) and is seen in Figure 12a. The remaining trend was simply taken as $f(s)^T\beta = \beta_1$ (an unknown constant to be estimated). Even though we externally extract the large-scale trend, it is still important to include at least a constant large-scale term (unknown), which is estimated along with other parameters of the model and contributes to the prediction uncertainty. An analysis of spread in the residuals $Z(D_i) - o(D_i)$ showed (not surprisingly) an almost linear trend in square-root of $o(D_i)$. We therefore assume that the small-scale variation $\delta(s_i)$ has variance that is proportional to the large-scale trend, $\mathrm{Var}[\delta(s_j)] = \sigma^2\,\mathrm{o}(s_j)$, with $\sigma^2$ to be estimated. A semivariogram analysis of the standardized residuals, $(Z(Di) - o(D_i)) / o(D_i)^{1/2}$, yielded a semivariogram that mirrors the shape of the spherical variogram (Cressie, 1993, p. 61), leveling off at a spatial range around 100 km. We therefore assume a spherical spatial correlation function for the small-scale spatial variation $\delta(s_i)$, given by

$$K(h; r) = 1 - 1.5(h/r) + 0.5(h/r)^3 \text{ if } h < r, \text{ but 0 otherwise} \qquad (8)$$

where $r$ is a spatial correlation range parameter to be estimated. Hence, the parameters to be estimated from the data are $\beta_1$, $\sigma^2$, and $r$. Those parameters were estimated by

maximum likelihood. The resulting optimal prediction at the grid-points are shown in Figure 12b with the associated uncertainty given in Figure 12c. Since the model 'honors' the data, when the prediction map in Figure 12b is aggregated up to the county level, it yields the reported county residue biomass.

(a) Trend (b) Prediction (c) Standard Deviation




.

*Figure 12. Residues density results from a spatial model with a smooth trend:(a) large-scale residue density trend, (b) residue density prediction, (c) the marginal prediction uncertainty*

The second spatial model applied to the data takes advantage of the land-use data. The land-use data are reported at approximately 1-km resolution and are used to estimate the fraction of each land-use type in 4- × 4-km pixels centered at the grid points. We assumed that only crop-related land use yields crop residue, and within Minnesota, there are three major crop-related land-use codes (Cropland/Pasture, Cropland/Grassland, and Cropland/Woodland; see Figure 4c). We therefore take the model at the grid level to be

$$Y(s_j) = \eta_1 x_1(s_j) + \eta_2 x_2(s_j) + \eta_3 x_3(s_j) + \delta(s_j)$$

where $x_1(s_j)$, $x_2(s_j)$, and $x_3(s_j)$ are the fraction of the three crop land-use categories within each 4- × 4-km pixel, and $\delta(s_j)$ is the familiar small-scale variation. Hence, we assume that the residue density at the $j$-th pixel is explained by its land-use pattern plus a spatially-smooth deviation from the global land-use trend (i.e., we expect nearby pixels to deviate in a similar way from the global land-use pattern). The remaining parts of the model are as in the previous one, except with $\mathrm{Var}[\delta(s_j)]$ proportional to the large-scale trend given by the land-use term (not the smooth, moving-average trend). In this case, the

parameters for estimation are $\eta_1, \eta_2, \eta_3, \sigma^2$, and $r$. The results from the model are shown in Figure 13.

(a) Trend (b) Prediction (c) Standard Deviation

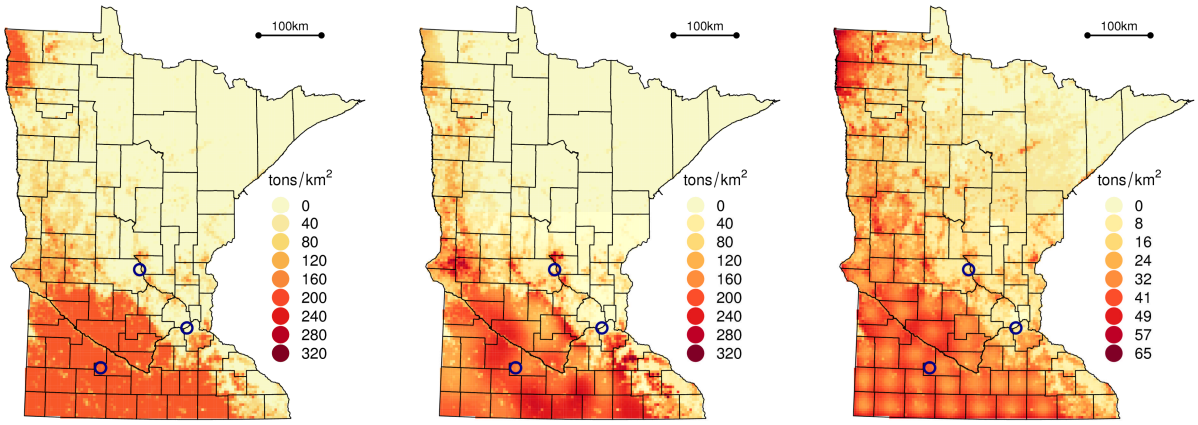



*Figure 13. Residue density results from a spatial model taking advantage of land-use data:(a) land-use-based large-scale residue density trend, (b) residue density prediction, and (c) the marginal prediction uncertainty.*

There is a clear difference between the model results with and without the land-use data, with the land-use results "looking" more realistic. But what is the impact on estimating the available biomass within a given site? Is there a significant difference? And how do these two models stack up to the two 'trivial' methods mentioned earlier: (1) assigning the biomass to the centroids and (2) assuming it is evenly distributed within each county? To investigate that question, we estimated the total biomass within various distances from the three sites shown in Figure 12 and Figure 13. One of these sites is surrounded by homogenous, high-yielding crop land, one is near an urban area, and the third is in an area with variable residue density. A comparison of the amount of residue biomass as a function of distance is presented in Figure 14 for the three locations and four different methods (the two spatial models along with the two trivial methods). There is remarkably little difference between assuming that the biomass is evenly distributed within each county and the two spatial models, with the crude centroid approach showing some deviation from the three (particularly at short distances). We might have been able to tell this at the onset by just looking at the semivariogram presented in Figure 11 and noting that the typical size of a county is well within the range of the spatial correlation shown (remember the wind aggregation example). However, it is just by carrying out the

full spatial analysis and modeling the data that we conclude one can work at a relative coarse resolution, due to the inherent smoothness of the process.

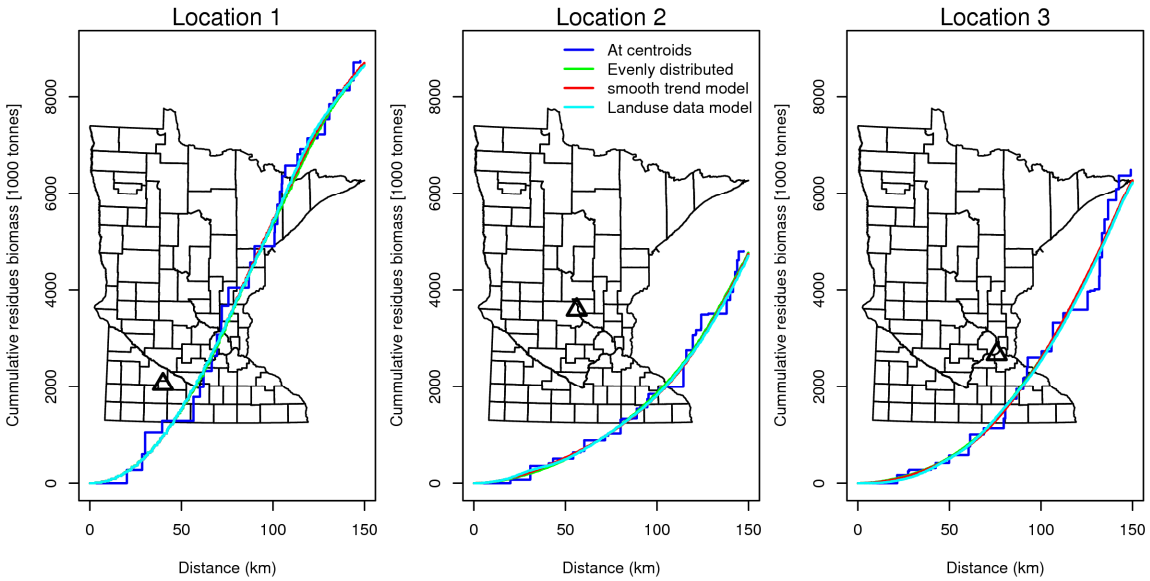



*Figure 14. The cumulative residue biomass within a given (air) distance from three selected locations from four different residue biomass models (legend).*

## *Electricity Load/Demand Data*

The electricity load data is collected at the 116 utility district control areas. It is then aggregated to a new scale and 12 temporal periods for input to the National Energy Modeling System (NEMS) run and managed by the Energy Information Agency. We have selected one of the time periods NEMS needs to forecast energy futures, summer peak average, to illustrate how one can test the confidence of aggregating and disaggregating spatial scales. Any of the 12 time steps would be useful for this purpose.

We seek to predict the load at a finer spatial scale which we, eventually, might want to aggregate up to another coarse-resolution unit (for example, the needed input areal units for a particular model). We have in mind the areal model (6), where the $Z(D_i)$'s, $i = 1, ..., n$ ($n = 116$), are the reported electricity loads and the $Y(B_{ij})$'s are the electricity loads at the units $B_{ij}$ (to be specified), and recall that $D_i$ is the union of the $B_{ij}$'s, $j = 1, ..., N$.

Electricity load is tightly coupled to residential, commercial, and industrial density. Population density has been demonstrated to be a good proxy for energy demand either directly or indirectly. Population data is available down to the census-block level,

while other potential useful input variables are typically gathered and reported at the county level. This suggests that one might model the load at the county level and then aggregate back up to the regions needed. However, about half of the counties are supplied with electricity by more than one control region. To estimate their energy demand we take the model areal units $B_{ij}$ as the intersection of the two, $B_{ij} = D_i \cap B_j$, where $B_j$, $j = 1, ..., N$, are the counties. Fine-resolution population data can be aggregated to the units $B_{ij}$, while other data is typically only reported at the county level (that is, it applies to the union of one or more of the $B_{ij}$'s).

It is very difficult to carry out (spatial) exploratory analysis of the load data (the summer peak average load) in its raw form because of the irregular size and shape of the control units. However, given a possible explanatory variable (e.g., population), one can aggregate the explanatory variable up to the control regions and use classical regression techniques to investigate whether it is potentially correlated with the total load. As expected, such analysis revealed that population is an important factor. Another explanatory variable that was looked at was the estimated industrial water usage in 2000 by county, as estimated by the U.S. Geological Survey (Kenny, J.F., 2004). The industrial water usage is dominated by steel, chemical, smelting, and petroleum activity, all known to be energy-demanding activities and cannot necessarily be inferred from county population size. Another selection criterion we used was that the data set was reported nationally and did not leave area gaps as other data sets collected at the regional level. Industrial water usage was seen to correlate with electricity load when aggregated up to the control regions (even after taking into account population).

We applied a spatial model that uses both the population data and industrial water usage as input variables. The spatial correlation between the small-scale variation terms $\delta(B_{ij})$ was modeled using the distance between their centroids. The process model in (6) is given by

$$Y(B_{ij}) = \eta_0 + (\eta_1 + \eta_2 x_{ij} + \eta_3 y_{ij} + \eta_4 x_{ij} y_{ij}) p_{ij} + \eta_5 w_{ij} + \delta(B_{ij})$$

where $(x_{ij}, y_{ij})$ is the centroid of $B_{ij}$, $p_{ij}$ is the estimated (1998) population in $B_{ij}$ (as based on census data), and $w_{ij}$ is the 2000 estimated industrial water usage in $B_{ij}$ . Note that the impact of the population is allowed to vary (linearly) spatially. The data model assumed

zero measurement error, as in the crop residues biomass case. The spatial correlation among the $\delta(B_{ij})$ is given by the spherical correlation function used previously (8), with variance proportional to the large-scale trend (similar to the biomass analysis). Estimation of model parameters was carried out using maximum likelihood.

Figure 15 shows the predicted average summer-peak load at the county level (i.e., the $B_{ij}$ have been aggregated up to the county level), along with relative prediction standard deviations (standard deviation over prediction). The prediction map reflects the underlying population, with some modification due to industrial water usage. Note that the model is built such that when the prediction map is aggregated to the control regions, it yields the reported average summer-peak loads. The relative standard deviation maps show generally low relative accuracy in areas of low load, but better accuracy in populated areas; the level of accuracy is not surprising keeping in mind the size of the 116 control regions. Figure 16 shows the small-scale variation $\delta(B_{ij})$ at the county level; that is, the predicted load minus the fitted trend. The estimated spatial correlation was relatively week (with a spatial range of 550 km and dropping sharply with distance).

This model can be easily extended to include other (county-level or point-level) explanatory input variables, and the significance of each can be tested. However, given only 116, relatively large control regions, the ability of the model to predict at the county level at a precise level of accuracy is limited. In addition, if the final goal is to aggregate the load up to other, relatively large regions compared to the size of the counties (for example, states), the final impact might be minimal. For example, we compared the current model with a model that does not use the industrial water usage. Figure 17 shows the difference in the predicted load at the county level. There is clearly some difference in the output of the two models, and the computed average relative difference is about 11.5%. However, when the predicted load of these two models is aggregated up to the state level, the average relative difference decreases to about 1.5%. Hence, in addition to the models accuracy (at the county level), the coarseness of the output scale (states) impacts the final accuracy of the predicted load. If the output areas are known in advance, say at the state level, and the predicted loads are not needed at any other spatial scales, the spatial modeling can be done at the intersection of those; that is, for example, the $B_j$ could be taken as the states instead of the counties, and the spatial modeling carried out

on the regions formed by intersecting control regions and states.

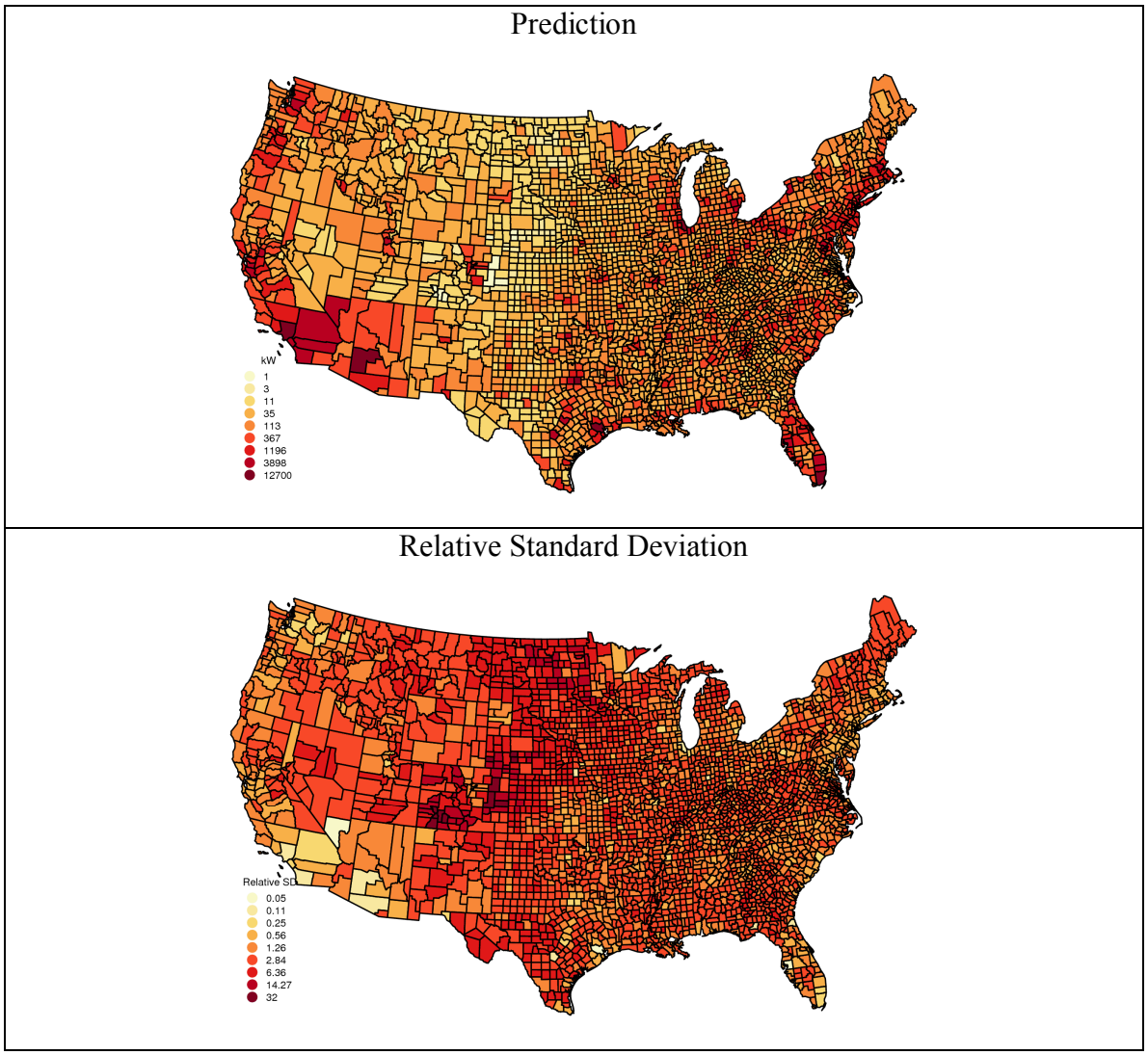



*Figure 15. Predicted average summer-peak load by county (top) and the associated relative prediction standard deviations (standard deviations over predicted values).*

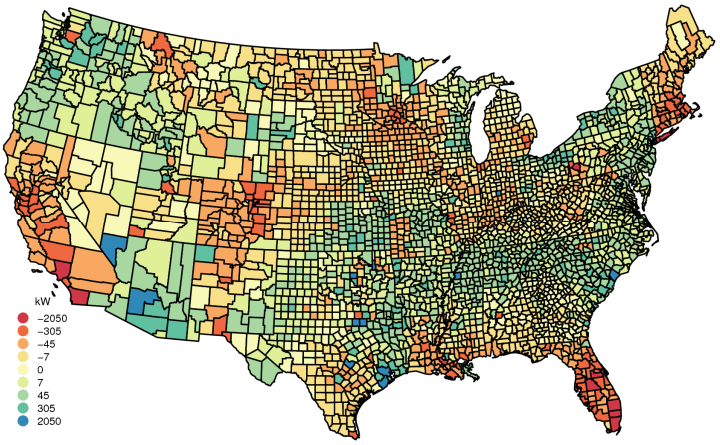



*Figure 16. The spatial variation; predicted load minus trend.*

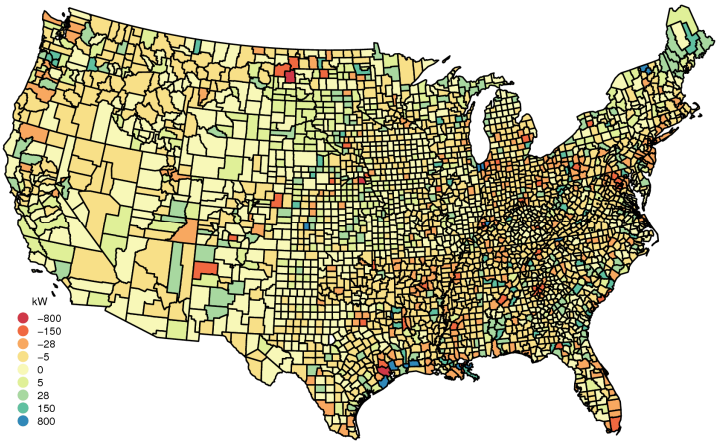



*Figure 17. The difference in county-level load predictions from a model using population and industrial water usage as an input to one using only population as input variable.*

# 6. Discussion

A large body of statistical models, including spatial and spatiotemporal models, can be very useful in assisting energy modelers to extract resource information, demand data, and cost factors needed by merging information available from various, disparate data sources. Several examples were presented in an Energy Efficiency and Renewable Energy Workshop presentation by Johannesson and Stewart in 2005 in support of this paper, including methods for working with point, areal, and misaligned data; modeling the unknown process; and spatial exploratory techniques that are applicable to data requested by energy modelers.

The area of spatial and spatiotemporal modeling of misaligned data from

disparate sources remains a very active research topic in statistical modeling; see Banerjee et al. (2004) for an introduction. Hierarchical Bayesian models have been dominant with sampling-based inference methods. These methods differ from more traditional methods (for example, kriging) in that results are presented as an ensemble of possible outcomes (for example, possible maps given available data) instead of a single (best!) prediction map.

Techniques developed by the spatial modeling community can be applied to the energy modeling area for use in analyzing resource information and new technologies that require more robust methods of estimation. Energy modelers need to have a sense of the importance of spatial and temporal scales required for accurate representation of the energy system under review. This information is important for the spatial modeling team's ability to determine which statistical procedures to apply to the available data. There are numerous ways of merging data sets that exist at different resolutions, and an equal number of tests to validate the results. The type and number of techniques to apply will depend on the level of precision required by the energy modelers to support their analysis. The spatial and energy modeling teams need to exchange information on 1) the questions energy modelers are trying to answer, 2) the spatial and temporal scales desired, 3) the data sets available, and 4) the minimum acceptable accuracy. This exchange will allow the spatial modeling team to determine the best procedures and test to validate the results.

# References


Energy Information Agency, http://www.eia.doe.gov/oiaf/aeo/overview/introduction.html (accessed on December 14, 2005).

Lamont, A. and Wu, T. (2005) Impact of Time Resolution on the Projected Rates of Market Penetration by Intermittent Generation Technologies. UCRL pending, Lawrence Livermore National Laboratory, Livermore, California.

Johannesson, G., Interview with Anelia Milbrandt of National Renewable Energy Laboratory, July 2005.

Banerjee, S., Carlin, B.P., and Gelfand, A.E. (2004) *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall/CRC, Boca Raton, Florida.

Cressie, N.A.C. (1993) *Statistics for Spatial Data*, Revised ed., John Wiley & Sons, New York.

Schabenberger, O. and Gotway, C.A. (2005) *Statistical Methods for Spatial Data Analysis*, Chapman & Hall/CRC, Boca Raton, Florida.

Johannesson, G. and Stewart, J. S., *Geospatial Statistics and Issues in Energy, Modeling*, May 10–11, 2005, GIS/Regionalization Workshop for Energy Efficiency and Renewable Energy, RAND Offices Arlington, Virginia.
http://www.nrel.gov/analysis/workshops/gis_workshop_05.html

Kenny, J.F. (2004) Guidelines for preparation of State water-use estimates: U.S. Geological Survey Techniques and Methods 4-A-4

*The* National Energy Modeling System
*http://www.eia.doe.gov/oiaf/aeo/overview/figure_10.html*.

*The* National *Energy Modeling System*
http://www.eia.doe.gov/oiaf/aeo/overview/figure_1.html

Service Territory Map: Electric
*http://www.choosemaryland.org/datacenter/utilities/terr_elec.asp*

# Appendix: Statistical Software for Spatial Analysis

The field of statistical computing continues to expand at a tremendous rate, and numerous software packages have emerged in response to the significant range and complexity of problems involving spatial and geographic analysis. In many cases, spatial tools have been developed as extensions of existing software environments. Sometimes these add-ons are produced by a firm, and other times industrial or academic statisticians have developed them. The routines which are more commonly used or which were first written have, in some cases, been included in the various base packages for languages. However, for newer and more specialized tools, it is often necessary to purchase, request, or otherwise query for them. Additionally, statistical and analytical tools are increasingly being integrated into GIS software, for both the novice and the advanced user.

This appendix gives a brief introduction to some of the leading software tools available for spatial analysis, with particular focus on the interaction between GIS and statistical modeling. Generally speaking, analytical tools, found in either a statistical analysis package or embedded in GIS software, are tailored to work best with certain types of geospatial data. The data types, or data models, of geospatial information are points, lines and networks, polygons, and fields (or lattice models). While some packages can aid in the analysis of many of these data types, few can excel in providing analysis and visualization for all of them. In addition to the leading software packages, a number of "home grown" software applications have been created for specific geoanalytical needs.

The analysis carried out in this paper used the GIS software called ArcGIS and the statistical software called R. Spatial data were explored and prepared using ArcGIS, while the spatial models where fitted and applied in R.

## *The S Language—R and S-Plus*

R (http://r-project.org) and S-Plus (http://www.splus.com) are both related to a statistical programming environment called S, developed at Bell Labs. R is an open-source implementation of the language, a popular tool in academic research and within research institutions, and S-Plus is a commercial implementation with a stronger focus on

enterprise.

The strength of the S language is its flexibility; it can easily be extended and customized by the end user. It has also gained reputation as a powerful graphical exploratory data analysis tool. Both R and S-Plus come equipped with classical and modern statistical modeling and testing procedures, which can be very useful for any spatial statistical modeling.

Both R and S-Plus have geostatistical (that is, point-referenced data) capabilities, which include tools for exploratory spatial data analysis, trend and variogram analysis, and estimation and prediction procedures. Because of the open-source nature of R, multiple add-in packages provide geostatistical capabilities, including the gstat package (http://www.gstat.org) and the geoR package (http://www.est.ufpr.br/geoR).

Similar capabilities exist within both environments for the treatment of lattice models (areal data), in particular, exploring, fitting, and predicting using SAR and CAR models.

Both R and S-Plus can interface with (import and export) some of the more common GIS data formats, including ESRI's shapefiles. In addition, R has special data classes (such as the sp package) that are designed for various types of spatial data (points, lines, polygons) and operations on them (see http://r-spatial.sourceforge.net).

There are some key differences between the two packages. S-Plus has a rather extensive graphical user interface (GUI) in addition to command-line interface, which some users might find intimidating; however, any advance use of S-Plus requires (eventually) knowledge of the underlying S statistical programming language. Because of the open-source nature of R, a large, active community of research statisticians expands the capabilities of R by providing new, cutting-edge functionality in packages. However, this can be a double-edged sword. The statistical tools within R exist within different packages (with different maintainers), where documentation can range from being very sparse to good. In contrast, in S-Plus, all spatial statistical procedures have been collected in a single, well-documented module (S+SpatialStats) for advanced spatial analysis; it is also an extension to the less-technical ESRI GIS software (S-PLUS® for ArcView GIS).

## *SAS*

SAS (http://www.sas.com), a commercial package, was originally developed for general statistical analysis, and it has had spatial capabilities added recently. SAS has a good reputation in enterprise application, with a good interface to many common data-bases, good handling of large data sets, and a well-established suite of classical statistical procedures.

Currently, SAS has procedures to model point-referenced data (VARIOGRAM and KRIGE2D). However, there are no specific procedures to model areal data (lattice models). On the other hand, SAS has a GIS module with a bridge to ESRI's ArcGIS, which makes it attractive for GIS-related work.

Like S-Plus, advanced use of SAS is command-line based, but SAS also has an extensive GUI.

## *MATLAB*

The MATLAB (http://www.mathworks.com/products/matlab/) matrix manipulation software excels in managing large array sizes. The software is easily extensible and offers free toolboxes to aid in the processing and analysis of field data, namely in detrending data and identifying autocorrelation. In addition, MATLAB can provide tools for fitting variograms and interpolation.

Interactions with common GIS platforms usually occur by passing data through common array forms, such as ASCII grids. Frequently, those who use MATLAB display their results by taking advantage of the superior visualization capabilities of the software.

MATLAB relies mostly on command-line interfaces for operation.

## *Geographical Information Systems*

### ArcGIS

ESRI's ArcGIS (http://www.esri.com), a commercial GIS package, is the most popular and widely used GIS package available. It is available for different levels of users and consists of a number of extensions that have been tailored to specific data, analytic, and visualization needs. These include Spatial Analyst, for the handling of raster data,

Network Analyst, for analyzing linked vector data sets, and Business Analyst, providing tools for commercial needs. ArcGIS also enables web-based visualization and limited analysis through ArcIMS, an Internet Map Server. While a number of free extensions are available for ArcGIS, ESRI recently added a powerful geostatistical analyst toolbox to its suite of powerful spatial data analysis and management tools. The geostatistical toolbox includes exploratory tools, variogram fitting, and kriging predictor. It does not yet have capabilities to handle areal data via block kriging nor CAR/SAR models. However limited, the ArcGIS statistical strength lies more with the interpretation of points, linear features, and polygons than with fields and imagery.

Recently, ESRI added the ability to interface with the Python programming language (http://www.python.org), which is a powerful and flexible interpretation language. The geostatistical procedures are written in Python. Hence, they can be easily modified and extended. ArcGIS can also be programmed and developed in a number of languages, including ARC Macro Language, C and C++, Visual Basic, and Java.

## GRASS

GRASS (http://grass.itc.it) is an open-source Geographic Resources Analysis Support System. It has all the basic GIS capabilities but does not have built-in spatial statistics procedures. However, GRASS has interfaces to spatial statistical tools, including R, and also to gstat and GRASP.

## Other Raster Packages

Two major software packages handle raster (lattice) data: IDRISI Kilimanjaro (http://www.clarklabs.org/IdrisiSoftware.asp?cat=2 ID) and ENVI (http://www.rsinc.com/envi/). Though IDRISI has more vector capability than ENVI, both packages have been increasing the seamless use of both vector and raster data. These packages are primarily used for the analysis, management, and visualization of remote sensing data products. As they are developed in concert with the increased availability of remote sensing data products, these packages will compete more openly with ESRI in the marketplace.