Regionalization of Multiscale Spatial Processes using a Criterion for Spatial Aggregation Error

Jonathan R. Bradley¹, Christopher K. Wikle², Scott H. Holan²

Abstract

The modifiable areal unit problem and the ecological fallacy are known problems that occur when modeling multiscale spatial processes. We investigate how these forms of spatial aggregation error can guide a regionalization over a spatial domain of interest. By "regionalization" we mean a specification of geographies that define the spatial support for areal data. This topic has been studied vigorously by geographers, but has been given less attention by spatial statisticians. Thus, we propose a criterion for spatial aggregation error (CAGE), which we minimize to obtain an optimal regionalization. To define CAGE we draw a connection between spatial aggregation error and a new multiscale representation of the Karhunen-Loéve (K-L) expansion. This relationship between CAGE and the multiscale K-L expansion leads to illuminating theoretical developments including: connections between spatial aggregation error, squared prediction error, spatial variance, and a novel extension of Obled-Creutin eigenfunctions. The effectiveness of our approach is demonstrated through an analysis of two datasets, one using the American Community Survey and one related to environmental ocean winds.

Keywords: American Community Survey; Empirical orthogonal functions; MAUP; Reduced rank; Spatial basis functions; Survey data

¹(to whom correspondence should be addressed) Department of Statistics, University of Missouri, 146 Middlebush Hall, Columbia, MO 65211, bradleyjr@missouri.edu

²Department of Statistics, University of Missouri, 146 Middlebush Hall, Columbia, MO 65211-6100

1 Introduction

There has long been interest in non-statistical methods for specifying geographies to summarize spatial data (e.g., Openshaw (1977), Murtagh (1992), Martin (2002), Guo (2008), and Logan (2011)). In general, this is known as "regionalization," and it is an important (and sometimes required) task for many applications. For example, the American Community Survey (ACS) is an ongoing survey administered by the US Census Bureau that produces estimates of important US demographic variables. The ACS provides public-use data referenced over areal units (e.g., median household income over US counties). Similar to the decennial census, many of these geographic regions are required (e.g., states, counties, etc.), however, other regions are consistently being evaluated and changed (e.g., combined statistical areas, metropolitan divisions, metropolitan statistical areas, etc.) in a sub-optimal manner based on population controls (e.g., Blank et al. (2011)). This suggests that there is a clear need for regionalization methodology. Thus, we develop a principled statistical methodology for evaluating spatial aggregation error and optimal statistical regionalization.

Regionalization is a topic that has been considered primarily by geographers. The current state-of-the-art is the deterministic "max-p algorithm" (Duque et al., 2012; Spielman and Logan, 2013; Folch and Spielman, 2014; Spielman and Logan, 2015). In general, the max-p algorithm is a greedy search algorithm (using any desired criterion) that groups data defined on n_A areal units into $p \leq n_A$ contiguous regions. The max-p algorithm offers a solution, but there are many known pitfalls to this approach. The most significant issue from the perspective of multiscale spatial inference is that the regions obtained by this approach are not protected from the *ecological fallacy* (Robinson, 1950). Hence, proper inferential conclusions must be limited to a single (often difficult to interpret) spatial support.

We interpret the ecological fallacy as a type of spatial aggregation error, which will be critical to our approach for regionalization. In particular, the ecological fallacy refers to the situation where conclusions at the point-level spatial support differ from conclusions at an aggregate-level spatial support. Similarly, *ecological inference* is explicitly defined as inference on individual behavior drawn from aggregate data (also sometimes referred to as downscaling). This topic has experienced growing interest within a variety of subject matter disciplines. For example, see King (1997) for the sociological data setting; Darby et al. (2001), and the references therein, for applications in epidemiology; and Mearns et al. (2014), and the references therein, for the climatology setting. Following the terminology of Kolaczyk and Huang (2001), a similar problem is known as *image segmentation*, which involves optimally dividing an image into smaller regions (e.g., see Kolaczyk and Nowak (2004), Kolaczyk et al. (2005), and Ferreira et al. (2011)). For reviews of ecological inference and image segmentation see Wakefield (2004), Waller and Gotway (2004), and Ferreira and Lee (2007).

The *modifiable areal unit problem* (MAUP) is another type of spatial aggregation error. Waller and Gotway (2004) consider the MAUP to be the geographic manifestation of the ecological fallacy. That is, the MAUP refers to situations where conclusions on one aggregate spatial support differ from conclusions on another distinct aggregate spatial support, whereas, the ecological fallacy concerns conflicting conclusions at point-level and aggregate-level supports. The MAUP has a rich history, originally considered by Gehike and Biehl (1934), and later by Openshaw and Taylor (1979). Recently, the MAUP has become a topic covered in standard textbooks including Cressie (1993), Waller and Gotway (2004), Cressie and Wikle (2011), and Banerjee et al. (2015), among others.

The aforementioned forms of spatial aggregation error are closely related to the *spatial change* of *support* (COS) problem, which refers to conducting statistical inference on a support that differs from the spatial support of the data (e.g., Waller and Gotway (2004), Cressie and Wikle (2011), and Banerjee et al. (2015)). Methods for spatial COS allow one to choose any support on which to perform statistical inference. However, different choices for the spatial support result in different magnitudes of spatial aggregation error. Nevertheless, the inherent flexibility to use any desired

spatial support for inference has made spatial COS a popular area of research in both multiscale spatial analysis and other subject matter disciplines. For example, see Wikle and Berliner (2005) for the environmental data setting; Mugglin et al. (1998) for the public health setting; Bradley et al. (2015b) for the survey data setting; and Waller and Gotway (2004) and Trevisani and Gelfand (2013) for a review. To capitalize on the flexibility of spatial COS methods, we adopt a multiscale spatial perspective to quantify spatial aggregation error and to develop a method for regionalization.

The known presence of spatial aggregation error suggests an approach for an optimal regionalization. Specifically, our primary inferential question is the following: can we choose a spatial support that minimizes spatial aggregation error? To motivate this perspective, consider an example dataset obtained from the ACS. In Figures 1(a) and 1(b), we plot 5-year period estimates of median household income by county and state, respectively, for 2013. Upon comparison, Figures 1(a) and 1(b) show that the state-level ACS estimates suffer from noticeable spatial aggregation error. For example, Figure 1(b) suggests that households in Virginia have moderately high income, yet Figure 1(a) shows that only households in counties near Richmond have high income. Similarly, Figure 1(b) suggests that households in New York state have a moderately high income while Figure 1(a) shows that only households in counties near Manhattan have high income. These examples, and many others that are quite obvious upon study of these figures, provide evidence that states are not an appropriate (i.e., optimal) spatial support to summarize median household income, political reasons notwithstanding.

In what follows, we formalize this intuition and develop a criterion to quantify spatial aggregation error and an associated method for regionalization. Our approach is to quantify spatial aggregation error using what we call the *criterion for spatial aggregation error* (CAGE). Hence, an optimal spatial support is obtained by minimizing CAGE. The primary theoretical tool used to develop this criterion is the Karhunen-Loève (K-L) expansion (Karhunen, 1947; Loève, 1978), which is a well-known representation of a point-referenced process as the weighted sum of spa-

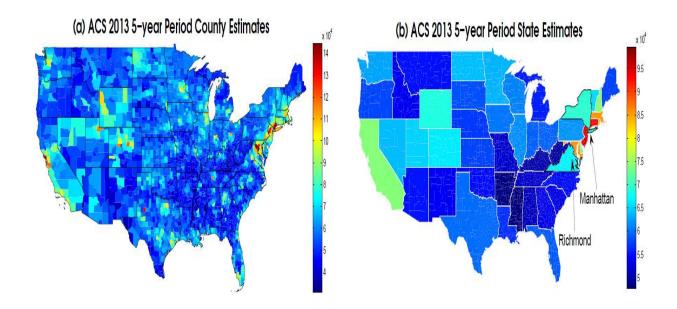


Figure 1: ACS 5-year period estimates of median household income for 2013. In (a), we plot the ACS estimates by counties, and in (b) we plot the ACS estimates by state. We superimpose the state boundaries as a reference in both panels. Notice that the color-scales are different for each panel. In (b), the borders of the states are highlighted in white except for New York and Virginia, whose borders are highlighted in black. Also, Richmond Virginia and Manhattan are indicated with arrows in (b).

tially varying eigenfunctions, where the weights are random. In more precise terms, we develop CAGE through a powerful technical result, which dictates that spatial aggregation error does not occur when the eigenfunctions of a spatial random process are constant between spatial scales. Thus, CAGE is a measure of between spatial scale homogeneity of eigenfunctions within a novel multiscale representation of the K-L expansion.

To date, there has been no such criterion that quantifies spatial aggregation error in this manner. The spatial statistics literature places an emphasis on prediction error (e.g., Cressie (1993)), and thus, such an aggregation-based approach for uncertainty quantification offers an exciting new perspective for spatial statistics. Therefore, to develop this perspective we provide technical results relating CAGE to prediction error and spatial variance.

After having defined CAGE, we can choose a regionalization in a manner that mitigates spatial aggregation error. In particular, we propose an efficient search algorithm (with CAGE as the selection criterion) to specify a regionalization over the spatial domain of interest. This search algorithm involves two stages. In the first stage, a naive algorithm, say *k*-means (e.g., Hartigan and Wong (1979)) is used to determine a collection of spatial supports from which to select. Then, in the second stage CAGE is used to select a single spatial support from among the collection of spatial supports determined in the first stage of the search algorithm. This two-stage approach is extremely efficient because it uses an easy-to-compute deterministic algorithm to direct the path of spatial supports from which to choose. As such, it can be incorporated efficiently within a Bayesian framework using a Markov chain Monte Carlo (MCMC) implementation of a latent spatial model, which facilitates uncertainty quantification.

Finally, to apply our search algorithm in practice, we provide a specification for the multiscale eigenfunctions. Thus, we introduce a general class of eigenfunctions that leads to a consistent class of multiscale spatial processes. To do this, we utilize the often overlooked, but remarkable framework of Obled and Creutin (1986). Obled and Creutin (1986) show that any class of geostatistical basis functions can be re-weighted so that they are eigenfunctions within a (single-scaled) K-L expansion. This notion of what we call *generating basis functions* (GBFs), is central to our development of multiscale eigenfunctions. As interest in spatial and spatio-temporal processes has turned to "big data" problems with large numbers of prediction and/or data locations, the modeling focus has shifted to this basis function perspective incorporating complete, over-complete, and reduced-rank expansions (Bradley et al., 2015a). Thus, the use of GBFs greatly increases the generality and utility of our approach. Furthermore, the use of GBFs is a necessity for our approach to regionalization because they allow us to perform spatial COS without assuming some form of between scale homogeneity.

The remainder of this paper is organized as follows. In Section 2, we introduce the multiscale K-L expansion and CAGE. Next, in Section 3 we describe how to use CAGE in practice, which

includes details on truncating the multiscale K-L expansion and the introduction of the two stage regionalization algorithm. Section 4 provides derivations of a consistent class of multiscale eigenfunctions to use within the CAGE framework. Then, in Section 5 a demonstration is given using the motivating dataset of ACS 5-year period estimates of median household income from Figure 1. In addition to demonstrating the regionalization algorithm for ACS period estimates, this application also highlights an important use of optimal regionalization, namely, aggregation for the purpose of dimension reduction. Finally, Section 6 contains a concluding discussion. We provide additional Supplemental Materials including: the proofs of technical results, simulation studies, and an additional application using a dataset consisting of Mediterranean wind measurements (a subset of the data used in Milliff et al. (2011)). The Mediterranean wind example is used to illustrate that the two-stage regionalization algorithm is flexible enough to handle multiscale spatial data.

2 Quantifying Aggregation Error

Here, we provide requisite extensions of the K-L expansion to the multiscale setting (Section 2.1). These results are then used to formally define CAGE (Section 2.2).

2.1 The Multiscale Karhunen-Loéve Expansion

Consider a real-valued spatial process that is realized at (possibly) both point-level and aggregate-level spatial supports. That is, the values in the sets $\{Y_s(\mathbf{s}): \mathbf{s} \in D_s\}$ and $\{Y_A(A): A \in D_A\}$ can be realized, where Y_s is a continuous spatial random process defined on D_s , $D_s \subset \mathbb{R}^d$, and Y_A is a spatial random process defined on areal support D_A with $D_A \equiv \{A_i: i=1,...,n_A\}$ and $A_i \subset \mathbb{R}^d$. The set A_i is an areal unit (e.g., a county, state, or census tract) and may be overlapping, contained in, or superimposed over another distinct areal unit $A_i \in D_A$ for $i \neq i$.

The corresponding multiscale spatial process can be written as

$$Y(\mathbf{u}) = \begin{cases} Y_s(\mathbf{u}) & \text{if } \mathbf{u} \in D_s \\ Y_A(\mathbf{u}) & \text{if } \mathbf{u} \in D_A; \ \mathbf{u} \in D_s \cup D_A. \end{cases}$$
 (1)

We interpret $Y_A(\cdot)$ as being computed from the point-level process $\{Y_s(\cdot)\}$. In particular, as is standard in spatial statistics (e.g., Cressie (1993), p. 284), we assume

$$Y_A(A) \equiv \frac{1}{|A|} \int_A Y_s(\mathbf{s}) d\mathbf{s}; \ A \in D_A,$$
 (2)

where |A| represents the cardinality of the set A. Consequently, placing a statistical model on Y_s implicitly places a statistical model on Y_A and Y through (1) and (2). We explore this dependency between (1) and (2) using the well-known K-L expansion (e.g., Cressie and Wikle (2011), p. 156),

$$Y_s(\mathbf{s}) = \sum_{j=1}^{\infty} \phi_j(\mathbf{s}) \alpha_j; \ \mathbf{s} \in D_s,$$
 (3)

where, without loss of generality, $\{Y_s(\cdot)\}$ is assumed to be mean-zero, the random variables in the set $\{\alpha_j : j = 1, 2, ...\}$ are uncorrelated with associated variances $\{\lambda_j : j = 1, 2, ...\}$ (called eigenvalues), the orthonormal real-valued functions $\{\phi_j(\mathbf{s}) : j = 1, 2, ...\}$ (called eigenfunctions) have domain D_s , and satisfy a Fredholm integral equation for a given valid covariance function. (Note that the conditions needed for the K-L expansion are given in the statement of Proposition 1.)

The use of the K-L expansion greatly increases the generality of our approach, since Mercer's theorem dictates that point-level covariance functions can be decomposed according to the K-L expansion (Mercer, 1909) under a very general set of assumptions (Ferreira and Menegatto, 2009). This leads us to define a multiscale K-L expansion, which we formalize through Proposition 1 below.

Proposition 1: Let $(\Omega, \mathcal{F}, \mathcal{P})$ be a probability space, where Ω is a sample space, \mathcal{F} is a sigmaalgebra on Ω , and \mathcal{P} is a finite Borel measure. Let $Y_s(s)$ be defined by the mapping $Y_s: D_s \times \Omega \to \mathbb{R}$, such that $Y_s(s)$ is measurable for every $s \in D_s$, and $D_s \subset \mathbb{R}^d$ is a topological Hausdorff space. Assume that $C(s, \mathbf{u}) \equiv \text{cov}\{Y_s(s), Y_s(\mathbf{u})\}$ is a valid covariance function that exists for each $s, \mathbf{u} \in D_s$. Let $L^2(\Omega)$ denote the Hilbert space of real-valued square integrable random variables.

i. Then, for each $A \subset D_s$ we have that

$$Y_A(A) = \sum_{i=1}^{\infty} \phi_{A,j}(A)\alpha_j, \tag{4}$$

in $L^2(\Omega)$, where for each positive integer j, $\phi_{A,j}(A) \equiv \int_A \phi_j(s) ds/|A|$, the random variables in the set $\{\alpha_j : j = 1, 2, ...\}$ are uncorrelated with associated variances $\{\lambda_j : j = 1, 2, ...\}$ (called eigenvalues), the orthonormal real-valued functions $\{\phi_j(s) : j = 1, 2, ...\}$ (called eigenfunctions) have domain D_s , and satisfy the Fredholm integral equation for $C(s, \mathbf{u})$.

ii. Then for any $A \subset D_s$ and $B \subset D_s$ we have that

$$\operatorname{cov}\left\{Y_{A}(A), Y_{A}(B)\right\} = \lim_{n \to \infty} \sum_{i=1}^{n} \phi_{A,i}(A) \phi_{A,i}(B) \lambda_{i}. \tag{5}$$

The proof of this proposition can be found in the Supplemental Materials.

Remark 1: We call the expression in (6) the multiscale K-L expansion since Proposition 1.i extends the K-L expansion in (6) to a similar infinite-dimensional process that is a function of *any* $A \subset D_s$. Similarly, the expression in (6) can be seen as an extension of Mercer's theorem to the multiscale spatial setting.

Remark 2: In practice, the latent multiscale spatial process of interest Y is not observed perfectly. Instead, we observe the n-dimensional data vector given by $\mathbf{Z} \equiv (Z(\mathbf{u}) : \mathbf{u} \in D_s^O \cup D_A^O)'$, where the

observed locations are denoted by $D_s^O \equiv \{\mathbf{s}_i^O : i = 1, ..., n_s^O\} \subset D_s$ and $D_A^O \equiv \{A_j : j = 1, ..., n_A^O\} \subset D_A$, and $n = n_s^O + n_A^O$. We assume that the stochastic processes $Z : D_s \times \Omega \to \mathbb{R}$ and Y are generated based on the generic probability space $(\Omega, \mathscr{F}, \mathscr{P})$ such that the conditional probability density function of $Y(\mathbf{u})|\mathbf{Z}$ exists for each $\mathbf{u} \in D_s \cup D_A$.

Remark 3: For purposes of implementation it is helpful to define a set $D_B \equiv \{B_j : j = 1,...,n_B\}$ with $B_j \cap B_\ell = \emptyset$ for $j \neq \ell$ and $B_j \subset D_s$ for each j. Here, D_B represents the finest resolution spatial support on which one is willing to perform inference. Then, after observing data $Z(\cdot)$, statistical inference is performed using sample draws from the distribution of $\mathbf{Y}_B | \mathbf{Z}$, where the n_B -dimensional process vector is given by $\mathbf{Y}_B \equiv (Y_A(B) : B \in D_B)'$.

2.2 The Criterion for Spatial Aggregation Error (CAGE)

There is an implicit conceptual challenge involved with quantifying spatial aggregation error. As Gotway and Waller (2011) discuss, the consequences of spatial aggregation error extend beyond between-scale differences of the values of a single statistic (e.g., correlation coefficient, mean, etc.). Thus, we say that spatial aggregation error occurs when there are between-scale differences for *any* generic statistic. The multiscale K-L expansion in (6) provides insight on a formalization of this concept, which we state in Proposition 2.

Proposition 2: Assume that the conditions of Proposition 1 hold. Let f be a measurable real-valued function with domain \mathbb{R}^{n_A} that is discontinuous only on a set with measure zero. Let λ_k be strictly greater than zero for each k = 1, 2, Define a generic point-level support $\{x_j : j = 1, ..., n_A\}$, such that $x_j \in B_j \subset A_j \in D_A$ for $j = 1, ..., n_A$, $Y_s^{(A)} \equiv (Y_s(x_j) : j = 1, ..., n_A)'$, $Y_B^{(A)} \equiv (Y_A(B_j) : j = 1, ..., n_A)'$, and $Y_A \equiv (Y_A(A) : A \in D_A)'$. Then the following statements hold for $Y(\cdot)$ in (1):

- i. $\phi_k(\mathbf{x}_j) = \phi_{A,k}(A_j)$ for $j = 1, ..., n_A$ and every positive integer k, if and only if $f(\mathbf{Y}_s^{(A)}) = f(\mathbf{Y}_A)$ almost surely.
- ii. $\phi_k(B_j) = \phi_k(A_j)$ for $j = 1, ..., n_A$ and every positive integer k, if and only if $f(\mathbf{Y}_B^{(A)}) = f(\mathbf{Y}_A)$ almost surely.
- iii. If $\phi_k(\mathbf{x}_j) = \phi_k(A_j)$ for every positive integer k, and every $\mathbf{x}_j \in B_j$ and j, then $f(\mathbf{Y}_B^{(A)}) = f(\mathbf{Y}_A)$ almost surely.

Remark 4: Proposition 2 provides a condition so that there is no ecological fallacy between $\mathbf{Y}_s^{(A)}$ and \mathbf{Y}_A , and no MAUP between $\mathbf{Y}_B^{(A)}$ and \mathbf{Y}_A . By "no ecological fallacy" and "no MAUP," we mean that for any real-valued, measurable, (almost) continuous statistic f, $f(\mathbf{Y}_s^{(A)}) = f(\mathbf{Y}_A)$ and $f(\mathbf{Y}_B^{(A)}) = f(\mathbf{Y}_A)$ almost surely. This ensures that conclusions using the summary statistic f stay the same regardless of the scale of Y. In general terms, Propositions 2.i and 2.ii show that "no spatial aggregation error" is equivalent to between-scale homogeneity of eigenfunctions within a multiscale K-L expansion. Furthermore, Propositions 2.i and 2.iii provide a relationship between the ecological fallacy and the MAUP; namely, if there is uniformly no ecological fallacy for any of the sets in $\{B_j\}$ (i.e., $\phi_s(\mathbf{x}_j) = \phi(A_j)$ for every $\mathbf{x}_j \in B_j$ and j), then there is no MAUP.

Proposition 2 guarantees that spatial aggregation error does not occur when the point-level eigenfunctions are constant over each region in D_A . This leads naturally to a criterion that measures departures from the absence of spatial aggregation error. Specifically, we define CAGE as follows:

$$CAGE(A) = E \left[\int_{A} \frac{\sum_{j=1}^{\infty} \left\{ \phi_{j}(\mathbf{s}) - \phi_{A,j}(A) \right\}^{2} \lambda_{j}}{|A|} d\mathbf{s} | \mathbf{Z} \right], \tag{6}$$

where A is a generic areal unit (i.e., $A \subset D_s$), and the expectation is taken with respect to the conditional distribution given the data. The logic behind (6) is straightforward: if CAGE(A) is equal to zero there is no loss of information when aggregating D_s to D_A , and if CAGE(A) is close to (far

from) zero then we lose a small (large) amount of point-level information when aggregating to A. Hence, maps of $\{CAGE(A_i): i=1,...,n_A\}$ can be used to assess whether statistical inference on Y_A is reasonable relative to the point level process.

In some settings the latent process cannot realistically be defined at the point level. For example, the median (over counties) household income in Figure 1 cannot be interpreted on D_s (see Banerjee et al. (2015) for a discussion and more examples). Hence, for these settings the multiscale K-L expansion is used for spatial change of support, and the lowest spatial resolution on which Y is defined is D_B . We use the following discretized CAGE (abbreviated as "DCAGE") in these settings:

$$DCAGE(C) \equiv E\left[\sum_{h \in H} \frac{\sum_{j=1}^{\infty} \left\{\phi_{A,j}(B_j) - \phi_{A,j}(C)\right\}^2 \lambda_j}{|C|} | \mathbf{Z}\right],\tag{7}$$

where $C = \bigcup_{h \in H} B_h$, $H \subset \{1,...,n_B\}$, and $B_h \in D_B$ for each $h \in H$. Proposition 2.ii implies the following logic for (7): if DCAGE(C) is equal to zero there is no loss of information when aggregating D_B to higher spatial resolutions, and if DCAGE(C) is close to (far from) zero then we lose a small (large) amount of lower resolution information when aggregating D_B to higher spatial resolutions (see Remark 3).

To date there has been no attempt to quantify the magnitude of spatial aggregation error using criteria like (6) and (7). In the geostatistical setting, emphasis is usually placed on minimizing the squared prediction error (Cressie, 1993). From this point-of-view, it is worthwhile to note that there are connections between the squared prediction error, spatial variance, and CAGE in (6), which we formally state in Proposition 3 below.

Proposition 3: Assume that the conditions of Proposition 1 hold. Also, assume that the stochastic process $Z: D_s \times \Omega \to \mathbb{R}$ is generated based on a generic probability space $(\Omega, \mathcal{F}, \mathcal{P})$ such that the conditional probability density function of $Y(\boldsymbol{u})|\mathbf{Z}$ exists for each $\boldsymbol{u} \in D_s \cup D_A$, where Z is defined

in Remark 2. Then, CAGE in (6) has the following alternative expressions:

$$CAGE(A) = E\left[\int_{A} \frac{\left\{Y_{s}(s) - Y_{A}(A)\right\}^{2}}{|A|} ds |\mathbf{Z}\right]$$
(8)

$$CAGE(A) = E\left[\int_{A} \frac{\operatorname{var}\left\{Y_{s}(s)\right\}}{|A|} ds - \operatorname{var}\left\{Y_{A}(A)\right\} | \mathbf{Z}\right]$$
(9)

$$CAGE(A) = E\left[\int_{A} \frac{\left\{Y_{s}(s) - \widehat{Y}_{A}(A)\right\}^{2}}{|A|} ds |\mathbf{Z}\right] - E\left[\left\{\widehat{Y}_{A}(A) - Y_{A}(A)\right\}^{2} |\mathbf{Z}\right],\tag{10}$$

where A is a generic areal unit (i.e., $A \subset D_s$), and $\widehat{Y}_A(A) \equiv E(Y_A(A)|\mathbf{Z})$.

Remark 5: Each expression in Proposition 3 provides interesting motivation for CAGE. For example, (6) was motivated by Proposition 2 (i.e., by measuring the departure from the absence of spatial aggregation error), however, one could argue to use (6) from a practical perspective. That is, intuition suggests that it is reasonable to make finer scale inference using the aggregate process if $Y_s(\mathbf{s})$ is consistently "close" to $Y_A(A)$. However, it is important to note that our use of the K-L expansion is important because it allows us to perform spatial change of support to obtain Y_A without assumptions of between-scale homogeneity. Additionally, the expression in (6) is especially interesting from a historical perspective, since many of the early references on spatial aggregation error focused on second order statistics (Robinson, 1950). Here, we see that between-scale differences of variances have a connection (through Propositions 1, 2, and 3) to between-scale differences of any statistic.

Remark 6: The "ANOVA-type" decomposition in (6) offers a different perspective in which to interpret (6). The first term on the right-hand-side of (6) (from left to right) represents a within-areal unit prediction error. Specifically, the first term represents the prediction error between the point-level process Y_s and the aggregate-level estimator \widehat{Y}_A . The second term in (6) shows that a

minimax-type approach is used for between areal unit error. That is, we minimize the squared prediction error to obtain \widehat{Y}_A , but penalize for choosing A so that Y_A is close to \widehat{Y}_A . One could conceive of a version of Proposition 3 that provides similar identities for the DCAGE in (7). In Supplemental Materials, we provide the statement and proof of this technical result.

3 Statistical Methodology for Regionalization

In practice, higher order components, of the infinite sum in (6), correspond to a decreasing percentage of variation. Thus, it is standard practice to truncate the K-L expansion, and assume that the residual is negligible (e.g., see Obled and Creutin (1986), and Cressie and Wikle (2011) p. 267). In this section, we extend the results from Section 2 to accommodate this common assumption. In particular, for our applications we truncate the multiscale K-L expansion (Section 3.1), which leads to another version of CAGE (Section 3.2). With these details in place, we can describe how to use CAGE for regionalization (Section 3.3).

3.1 The Truncated Multiscale Karhunen-Loéve Expansion

A common simplification of the K-L expansion is to truncate the infinite sum in (6) and assume that

$$Y_s(\mathbf{s}; \boldsymbol{\phi}_s) = \sum_{j=1}^r \phi_{s,j}(\mathbf{s}) \alpha_j \equiv \boldsymbol{\phi}_s(\mathbf{s})' \boldsymbol{\alpha}; \ \mathbf{s} \in D_s,$$
(11)

where r is a fixed and "known" integer, the r-dimensional vector of eigenfunctions is given by $\phi_s(\cdot) \equiv (\phi_{s,1}(\cdot),...,\phi_{s,r}(\cdot))'$, and the associated r-dimensional random vector is $\boldsymbol{\alpha} \equiv (\alpha_1,...,\alpha_r)'$. It is important to note that $Y_s(\mathbf{s}; \boldsymbol{\phi}_s) \neq Y_s(\mathbf{s})$ in general due to the truncation in (11).

Now, (2) and (11) provide an immediate expression for Y_A , namely,

$$Y_A(A; \boldsymbol{\phi}_s) = \sum_{j=1}^r \frac{1}{|A|} \left\{ \int_A \phi_{s,j}(\mathbf{s}) d\mathbf{s} \right\} \alpha_j \equiv \boldsymbol{\phi}(A; \boldsymbol{\phi}_s)' \boldsymbol{\alpha}; \ A \in D_A,$$
 (12)

where $\phi(A; \phi_s) \equiv \left(\frac{1}{|A|} \int_A \phi_{s,j}(\mathbf{s}) d\mathbf{s} : j = 1,...,r\right)'$. Then, (1), (11), and (12) imply the following expression for the truncated K-L expansion of the multiscale spatial process Y,

$$Y(\mathbf{u}; \boldsymbol{\phi}_s) = \begin{cases} \boldsymbol{\phi}_s(\mathbf{u})' \boldsymbol{\alpha} & \text{if } \mathbf{u} \in D_s \\ \boldsymbol{\phi}(\mathbf{u}; \boldsymbol{\phi}_s)' \boldsymbol{\alpha} & \text{if } \mathbf{u} \in D_A; \ \mathbf{s} \in D_s \cup D_A, \end{cases}$$
(13)

where it is important to note that the r-dimensional random vector α is the same for both supports. Validity of the implied covariance function for Y follows immediately from the quadratic form (see Supplemental Materials for more details).

The distributional assumptions governing Propositions 1—3 were very general (see Remark 2). For the truncated multiscale K-L expansion we incorporate additional distributional assumptions. In particular, we assume the following:

$$Z(\mathbf{u})|Y(\cdot), \boldsymbol{\theta}_D \stackrel{\text{ind}}{\sim} \text{Normal}\left\{Y(\mathbf{u}), \sigma_Z^2(\mathbf{u})\right\}; \ \mathbf{u} \in D_s \cup D_A,$$
 (14)

where $\sigma_Z^2(\mathbf{u}) > 0$, and

$$Y(\mathbf{u}) = \mu + Y(\mathbf{u}; \boldsymbol{\phi}_s) + \delta(\mathbf{u}; \boldsymbol{\xi}); \ \mathbf{u} \in D_s \cup D_A, \tag{15}$$

is the unknown process of interest. In principal, one could easily adopt the generalized linear mixed effects model framework and replace the normal distribution in (14) with the appropriate probability density function from the exponential class of distributions. For example, if $Z(\cdot)$ is count-valued than one might let $Z(\mathbf{u})|Y(\mathbf{u}), \boldsymbol{\theta}_D$ be distributed as Poisson with the log link.

The unknown real value μ is interpreted as a constant "trend term." Additionally, in (15) we assume that α is an r-dimensional random vector with mean zero and covariance matrix $\Lambda \equiv \operatorname{diag}(\lambda_1,...,\lambda_r)$. The specification of ϕ_s , the distribution of α , and associated prior distributions for ϕ_s and Λ , are stated in Section 5. It is important to note that it is typically straightforward to take an empirical Bayesian approach by directly estimating ϕ_s and Λ instead of placing prior distributions on these unknown quantities.

The δ process represents "fine-scale variability." We adopt the models for δ used in Wikle and Berliner (2005) and Bradley et al. (2015b). That is, let $\xi \equiv (\xi_j : j = 1, ..., n_B)'$ consist of i.i.d. random variables with mean zero and variance σ_{ξ}^2 , and let

$$\delta(\mathbf{s};\boldsymbol{\xi}) = \boldsymbol{\xi}_i,\tag{16}$$

for any $\mathbf{s} \in D_s$ such that \mathbf{s} is in the j-th areal unit in D_B . Thus, $\delta(B_j; \boldsymbol{\xi}) = (1/|B_j|) \int_{B_j} \delta(\mathbf{s}; \boldsymbol{\xi}) d\mathbf{s} = \xi_j$ for $B_j \in D_B$. In general, (16) implies that the fine-scale variability term is constant within each of the $j = 1, ..., n_B$ areal units in D_B (with the respective value ξ_j). The specification of the distribution of $\boldsymbol{\xi}$ and a prior for σ_{ξ}^2 shall also be given in Section 5.

3.2 CAGE for the Truncated Karhunen-Loéve Expansion

It is not immediate that Proposition 2 (which motivated CAGE) holds for the process Y in (15). Thus, we provide an extension of Proposition 2 that develops the spatial aggregation error properties of Y in (15). We formally state this result in Proposition 4.

Proposition 4: Let f be any real-valued function with domain \mathbb{R}^{n_A} , and λ_k be strictly greater than zero for each k=1,...,r. Recall that a regionalization of D_B is given by $D_C=\{C_\ell:\ell=1,...,n_C\}$ with $C_j\cap C_\ell=\emptyset$ for $j\neq \ell$, $C_\ell=\cup_{h\in H}B_h$, $H\subset\{1,...,n_B\}$, and $B_h\in D_B$ for $\ell=1,...,n_C\leq n_B$. Define a generic point-level support $\{x_j:j=1,...,n_C\}$, such that $x_j\in B_j\in D_B$, where $B_j\subset C_j$

and $j = 1,...,n_C$. Let $Y_s^{(C)} \equiv (Y_s(x_j) : j = 1,...,n_C)', Y_B^{(C)} \equiv (Y_A(B_j) : j = 1,...,n_C)'$, and $Y_C \equiv (Y_A(C) : A \in D_C)'$. Then the following statements hold for Y in (15):

i.
$$\phi_s(\mathbf{x}_j) = \phi(C_j; \phi_s)$$
 for $j = 1, ..., n_C$, if and only if $f(\mathbf{Y}_s^{(C)}) = f(\mathbf{Y}_C)$ almost surely.

ii.
$$\phi(B_j; \phi_s) = \phi(C_j; \phi_s)$$
 for $j = 1, ..., n_A$, if and only if $f(Y_B^{(C)}) = f(Y_C)$ almost surely.

iii. If
$$\phi_s(\mathbf{x}_j) = \phi(C_j; \phi_s)$$
 for every $\mathbf{x}_j \in B_j$ and j , then $f(\mathbf{Y}_B^{(C)}) = f(\mathbf{Y}_C)$ almost surely.

Remark 7: For the process Y in (15) to have no spatial aggregation error on D_C we (again) require between scale homogeneity of the eigenfunctions. There are two key differences between Propositions 2 and 4. The first difference is that Proposition 4 can be seen as an extension of Proposition 2 from the multiscale K-L expansion in (6) to the truncated process Y in (15). The second difference is that Proposition 4 can be seen as a discretized version of Proposition 2. That is, Proposition 2 allows B_j to be any subset of A_j , and Proposition 4 requires B_j to be defined on the (discrete) areal support D_B .

Remark 8: The choice to set $r < \infty$ is intimately related to the concept of spatial aggregation error. It is well known that predictors based on spatial basis functions with r-large display more fine-level details than predictors based on spatial basis functions with r-small (Stein, 2013; Bradley et al., 2014a). Thus, if r is chosen to be "too small" then predictions of Y_s will have less variability over D_s (i.e., be more constant), and consequently the differences between Y_s and Y_A (or CAGE; see Proposition 3.i) will be smaller than they should be. We strongly recommend performing an in-depth sensitivity analysis to choose r when using CAGE. To investigate the consequences of choosing r "too small" we provide a small sensitivity study in the Supplemental Materials. Additionally, in the Supplemental Materials we provide a sensitivity analysis for the choice of r for the application in Section 5.

Similar to Proposition 2, we have that Proposition 4 guarantees that spatial aggregation error does not occur for the spatial process in (15) when *a finite number* of point-level eigenfunctions are constant over each region in D_A . This leads naturally to a definition of CAGE for the spatial process in (15):

$$CAGE(A) \equiv E\left[\int_{A} \frac{\{\boldsymbol{\phi}_{s}(\mathbf{s}) - \boldsymbol{\phi}(A; \boldsymbol{\phi}_{s})\}' \boldsymbol{\Lambda} \{\boldsymbol{\phi}_{s}(\mathbf{s}) - \boldsymbol{\phi}(A; \boldsymbol{\phi}_{s})\}}{|A|} d\mathbf{s} | \mathbf{Z}\right]$$
(17)

DCAGE(C)
$$\equiv E\left[\sum_{h\in H} \frac{\{\boldsymbol{\phi}(B_h; \boldsymbol{\phi}_s) - \boldsymbol{\phi}(C; \boldsymbol{\phi}_s)\}' \boldsymbol{\Lambda} \{\boldsymbol{\phi}(B_h; \boldsymbol{\phi}_s) - \boldsymbol{\phi}(C; \boldsymbol{\phi}_s)\}}{|C|} |\mathbf{Z}\right],$$
 (18)

where A is a generic areal unit (i.e., $A \subset D_s$), $\mathbf{\Lambda} \equiv \operatorname{diag}(\lambda_i : i = 1, ..., r)$, $C = \bigcup_{h \in H} B_h$, $H \subset \{1, ..., n_B\}$, $B_h \in D_B$ for each $h \in H$, and the expectation is taken with respect to the posterior distribution derived from (14) and (15). Notice that (17) and (18) are the truncated versions of CAGE and DCAGE in (6) and (7), respectively. In a similar manner a truncated version of Proposition 3 exists. We state and prove this result in Supplemental Materials.

3.3 A Two-Stage Regionalization Algorithm

The CAGE(A) measure allows us to evaluate whether or not the generic areal unit A has poor spatial aggregation properties. However, it is not immediately clear how it can be used to specify an optimal spatial support. We now describe the use of CAGE to explicitly obtain an optimal regionalization. Recall that D_B is the finest level aggregate support on which we wish to predict. In general, our approach is to consider many different regionalizations (combinations) of elements of D_B and select from among them the support that produces the smallest average CAGE. By "regionalizations of D_B " we mean a generic set $D_C \equiv \{C_\ell : \ell = 1, ..., n_\ell\}$, where $C_j \cap C_\ell = \emptyset$ for $j \neq \ell$ and for each ℓ , $C_\ell = \bigcup_{h \in H} B_h$, $H \subset \{1, ..., n_B\}$, and $B_h \in D_B$.

A greedy search algorithm that seeks the minimum of the average CAGE (i.e., $\sum_{\ell=1}^{n_\ell} \text{CAGE}(C_\ell)/n_\ell$) poses a considerable computational challenge (see Spielman and Logan (2013) for related discus-

sion). To address this computational issue we use a two stage search algorithm. In the first stage, a naive clustering algorithm is applied to each of the M samples of \mathbf{Y}_B from $[\mathbf{Y}_B|\mathbf{Z}]$, denoted $\mathbf{Y}_B^{[m]}$, for m=1,...,M. For example, we could apply a k-means algorithm to $\mathbf{Y}_B^{[m]}$ to define a set $D_C^{(k)}(\mathbf{Y}_B^{[m]}) \equiv \{C_\ell^{[m]}: \ell=1,...,k\}$, where $C_\ell^{[m]}$ is the ℓ -th cluster returned by the k-means algorithm. The superscript "(k)" denotes the number of areal units in $D_C^{(k)}$, and we keep track of the dependence of the m-th replicate $\mathbf{Y}_B^{[m]}$. In this article, we consider using the k-means algorithm. We set the input of the k-means algorithm to be the centroids of the areal units in D_B and $\mathbf{Y}_B^{[m]}$. In the Supplemental Materials we also consider *structural hierarchical clustering* (SHC) (Marsland, 2009) in place of k-means. The choice of clustering algorithm depends on the application. In settings where computation is of particular interest k-means is preferable over structural hierarchical clustering. However, structural hierarchical clustering allows one to incorporate neighborhood information to obtain contiguous areal units, which is a preferred regionalization in some applications.

The first stage of our algorithm defines a collection of "candidate" spatial supports

$$\mathscr{C} = \{ D_C^{(k)}(\mathbf{Y}_R^{[m]}) : k = g_L, ..., g_U; m = 1, ..., M \}.$$
(19)

Here, $g_L(g_U)$ represents the smallest (largest) number of areal units one is willing to consider, and both g_L and g_U must be pre-specified. Notice that there are a total of $M \times (g_U - g_L + 1)$ spatial supports in \mathscr{C} , which is considerably fewer than the total number of possible candidate spatial supports to chose from.

In the second stage of the search algorithm we find the best (i.e., smallest average CAGE) subset of \mathscr{C} . To do this, we compute

$$D_C^{op} = \underset{D_C^{(k)}(\mathbf{Y}_B^{[m]}) \in \mathscr{C}}{\arg \min} \left[\frac{1}{k} \sum_{\ell=1}^k \text{CAGE}(C_\ell^{[m]}) \right], \tag{20}$$

where $D_C^{op} \equiv \{C_j^{op}: j=1,...,n_C^{op}\}$ and $C_k^{op} \subset \mathbb{R}^d$ for $k=1,...,n_C^{op}$. It should be noted that D_C^{op} , by

definition, is optimal since it is obtained by minimizing error. However, one might obtain a smaller value for the average CAGE by optimizing over a different set than \mathscr{C} . Furthermore, one has to determine for their application whether or not it is appropriate to use CAGE or DCAGE in (20); that is, in the case where the process is not interpretable on D_s then one should replace CAGE in (20) with DCAGE. A step-by-step presentation of the regionalization procedure is provided in the Supplemental Materials.

4 A Class of Multiscale Eigenfunctions

Propositions 2 and 4 show that between scale differences in the eigenfunctions indicate that spatial aggregation error is present. Thus, the importance of the eigenfunctions for quantifying spatial aggregation error suggests that it should be parameterized. This will allow us to estimate eigenfunctions, and hence, CAGE can be informed by the data. Below, we discuss the construction of what we call Obled-Creutin (O-C) eigenfunctions as a weighted combination of generic GBFs. We then discuss the properties of these basis functions.

4.1 Obled-Creutin Eigenfunctions

It has become common to express spatial random processes in terms of a basis expansion on random effects. As such, there are many possible choices for basis functions (Wikle, 2010; Bradley et al., 2014a). The insight provided by Obled and Creutin (1986) is that one can use *any* of these classes of point-level spatial basis functions to build an eigenfunction. We define an Obled-Creutin (O-C) eigenfunction as any real-valued function on D_s that takes the following form:

$$\phi_k^{\text{OC}}(\mathbf{s}; \mathbf{F}) \equiv \sum_{i=1}^r \psi_i(\mathbf{s}) F_{ik}; \ \mathbf{s} \in D_s, k = 1, ..., r,$$
(21)

where \mathbf{F} is an $r \times r$ matrix with (i,k)-th element given by the real value weight F_{ik} , and the r-dimensional vector $\boldsymbol{\psi}(\cdot) \equiv \{\psi_1(\cdot),...,\psi_r(\cdot)\}'$, with $\psi_i(\cdot):D_s \to \mathbb{R}$ for i=1,...,r, corresponds to the aforementioned GBF basis vectors. One can organize the O-C eigenfunctions into the r-dimensional vector, $\boldsymbol{\phi}_s^{\mathrm{OC}}(\cdot;\mathbf{F}) \equiv (\phi_1^{\mathrm{OC}}(\cdot;\mathbf{F}),...,\phi_r^{\mathrm{OC}}(\cdot;\mathbf{F}))'$, which we call an Obled-Creutin (O-C) vector.

It is not necessarily true that $Y(\cdot; \phi_s^{OC})$ in (11) leads to a multiscale truncated K-L expansion. In Proposition 5 below, we specify the condition such that $Y(\cdot; \phi_s^{OC})$ admits a multiscale truncated K-L expansion.

Proposition 5: Let $Y\left\{\cdot; \boldsymbol{\phi}_s^{OC}(\cdot; \boldsymbol{F})\right\}$ be the multiscale spatial process defined in (13), where $\lambda_j \geq 0$ and > 0 for at least one j=1,...,r. Here, $\psi_1(\cdot),...,\psi_r(\cdot)$ are r real-valued functions with domain D_s . Additionally, let \boldsymbol{F} be an invertible $r \times r$ real-valued matrix. If $\boldsymbol{F}'\boldsymbol{WF} = \boldsymbol{I}$ then $Y\left\{\cdot; \boldsymbol{\phi}_s^{OC}(\cdot; \boldsymbol{F})\right\}$ admits a multiscale truncated K-L expansion, where \boldsymbol{I} is an $r \times r$ identity matrix and we define the (i,j)-th element of the $r \times r$ matrix \boldsymbol{W} as $W_{ij} \equiv \int_{D_s} \psi_i(s) \psi_j(s) ds$.

Remark 9: Proposition 5 is crucial for implementing the two-stage regionalization algorithm. That is, with a given GBF (i.e., radial basis functions, Fourier basis functions, wavelets, etc.) one can construct eigenfunctions, which can then be used within the two-stage regionalization algorithm from Section 3.3. There are many choices of GBFs available in the literature (e.g., Bradley et al. (2015a)), and in Section 5 we use the local bisquare functions from Cressie and Johannesson (2008). In the Supplemental Materials, we also consider using Wendland basis functions (Wendland, 1998).

4.2 Specification of the O-C Weight Matrix, F

We capitalize on the fact that the $r \times r$ matrix \mathbf{F} is unknown. Estimating \mathbf{F} will allow the data to inform the value of CAGE. However, Proposition 5 suggests that one needs to specify \mathbf{F} with care; specifically, we require $\mathbf{F}'\mathbf{WF} = \mathbf{I}$ to ensure that $Y_s(\cdot; \boldsymbol{\phi}_s^{\text{OC}})$ is a multiscale truncated K-L expansion. We achieve this by introducing a novel class of \mathbf{F} matrices. This contribution is formally stated in Proposition 6.

Proposition 6: For a given r-dimensional vector of basis functions $\boldsymbol{\psi}$ let \boldsymbol{W} be positive definite. Let \boldsymbol{G} be an $r \times r$ real-valued orthogonal matrix. Then,

$$F(G) \equiv P_{\mathbf{W}} \Lambda_{\mathbf{W}}^{-1/2} G, \tag{22}$$

satisfies F(G)'WF(G) = I, where $P_W \Lambda_W^{-1/2}$ is the Cholesky square root of the matrix W^{-1} .

Remark 10: For a given set of spatial basis functions $\{\psi_i\}$ we suggest verifying that **W** is positive definite. Then from (13), (21), and (22) one can write Y_s as

$$Y_{s}\left[\cdot;\boldsymbol{\phi}_{s}^{\text{OC}}\left\{\cdot;\mathbf{F}(\mathbf{G})\right\}\right] = \boldsymbol{\phi}_{s}^{\text{OC}}\left\{\cdot;\mathbf{F}(\mathbf{G})\right\}'\boldsymbol{\alpha} = \boldsymbol{\psi}(\cdot)'\mathbf{F}(\mathbf{G})\boldsymbol{\alpha} = \boldsymbol{\psi}(\cdot)'\mathbf{P}_{W}\boldsymbol{\Lambda}_{W}^{-1/2}\mathbf{G}\boldsymbol{\alpha},\tag{23}$$

where α has mean-zero and $r \times r$ covariance matrix Λ . If a closed form expression for W is not available then numerical integration or direct Monte Carlo sampling can easily be applied to approximate W. In the case of the latter, one can randomly generate n_w points $\{\mathbf{s}_k : k = 1, ..., n_w\} \subset D_s$ using a uniform distribution on D_s , and approximate W_{im} with $(1/n_w) \sum_{k=1}^{n_w} |D_s| \psi_i(\mathbf{s}_k) \psi_m(\mathbf{s}_k)$

In our Bayesian implementation given in Section 5, we use the following equivalent reparameterized expression of $Y_s \left[\cdot; \phi_s^{OC} \left\{ \cdot; \mathbf{F}(\mathbf{G}) \right\} \right]$ derived from the representation of Y_s in (23):

$$Y_s \left[\cdot; \boldsymbol{\phi}_s^{\text{OC}} \left\{ \cdot; \mathbf{F}(\mathbf{G}) \right\} \right] = \boldsymbol{\psi}^*(\mathbf{u})' \boldsymbol{\eta}; \ \mathbf{u} \in D_s \cup D_A, \tag{24}$$

where $\psi^*(\mathbf{s})' \equiv \psi(\mathbf{s})' \mathbf{P}_{\mathbf{W}} \mathbf{\Lambda}_{\mathbf{W}}^{-1/2}$ for $\mathbf{s} \in D_s$, $\psi^*(A)' \equiv \frac{1}{|A|} \int_A \psi(\mathbf{s})' d\mathbf{s} \ \mathbf{P}_{\mathbf{W}} \mathbf{\Lambda}_{\mathbf{W}}^{-1/2}$ for $A \in D_A$, and η ($\equiv \mathbf{G} \boldsymbol{\alpha}$) has mean zero and $r \times r$ covariance matrix $\mathbf{Q} \equiv \mathbf{G} \mathbf{\Lambda} \mathbf{G}'$. Additionally, we assume that \mathbf{Q} consists of random parameters that can be sampled. (For each application we undergo independent sensitivity analyses to select a prior distribution. For details behind the prior specification, and for related empirical results, see Supplemental Materials.) Then, it is straightforward to obtain samples of \mathbf{Q} and $\boldsymbol{\eta}$, respectively, via a MCMC algorithm. Note that if a closed form expression for $\frac{1}{|A|} \int_A \psi(\mathbf{s})' d\mathbf{s}$ is not available then numerical integration or direct Monte Carlo sampling can easily be applied to obtain an approximation. In the case of the latter, one can randomly generate n_w points $\{\mathbf{s}_k : k = 1, ..., n_w\} \subset A \subset D_s$ using a uniform distribution on A, and approximate $\frac{1}{|A|} \int_A \psi(\mathbf{s})' d\mathbf{s}$ with $(1/n_w) \sum_{k=1}^{n_w} \psi(\mathbf{s}_k)'$. In general, we have found that the value of n_w needs to be large for these approximations to be reasonable (in Section 5 we set $n_w = 20,000$).

Additionally, one can obtain samples of the eigenfunction $\phi_s^{\text{OC}}\left\{\cdot;\mathbf{F}(\mathbf{G}^{[m]})\right\}$ to use within the expression of CAGE in (6). That is, denote the m-th replicate of \mathbf{Q} with $\mathbf{Q}^{[m]}$, and let the corresponding spectral decomposition be written as $\mathbf{Q}^{[m]} = \mathbf{G}^{[m]} \mathbf{\Lambda}_{\mathbf{Q}}^{[m]} \mathbf{G}^{[m]}$. Then, the corresponding m-th replicate of $\phi_s^{\text{OC}}\left\{\cdot;\mathbf{F}(\mathbf{G}^{[m]})\right\}$ is given by

$$\phi_s^{\text{OC}}\left\{\cdot; \mathbf{F}(\mathbf{G}^{[m]})\right\} = \boldsymbol{\psi}^*(\cdot)'\mathbf{G}^{[m]}; \ m = 1, ..., M.$$
 (25)

We shall henceforth use the representation of $Y_s\left[\cdot; \boldsymbol{\phi}_s^{\text{OC}}\left\{\cdot; \mathbf{F}(\mathbf{G})\right\}\right]$ in (24), and the O-C eigenfunction $\boldsymbol{\phi}_s^{\text{OC}}\left\{\cdot; \mathbf{F}(\mathbf{G}^{[m]})\right\}$ in (25).

5 Application: Median Household Income from the American Community Survey

We revisit the ACS 5-year period estimates of median household income for 2013 presented in Figure 1. This data can be downloaded at http://factfinder2.census.gov/. This is an important

example because there has been a growing interest in regionalizing data from ACS (Spielman and Logan, 2013, 2015).

For this example, $D_s^O = \emptyset$, and $D_A^O = D_A$ consists of the n = 3,109 counties in the continental US. Since US counties are the finest spatial resolution of the dataset in Figure 1, we set $D_B = D_A$. Let $[Z(\cdot)|Y(\cdot)]$ be a normal probability density function with mean $Y(\cdot)$ and known variance $\sigma_Z(\cdot) > 0$, which are computed from margin of error estimates that are publicly available. Here, $Z(\cdot)$ is the log median household income, and we let $Y(\cdot)$ be distributed according to (15).

Both α and ξ are assumed to be Gaussian, and we perform regionalization using $\phi_s^{OC}(\cdot; \psi)$, where $\psi(\cdot) \equiv (\psi_j(\cdot) : j = 1, ..., 75)'$ is a 75-dimensional vector consists of local bisquare functions (Cressie and Johannesson, 2008):

$$\psi_{j}(\mathbf{s}) \equiv \begin{cases} \{1 - (||\mathbf{s} - \mathbf{c}_{j}||/w)^{2}\}^{2} & \text{if } ||\mathbf{s} - \mathbf{c}_{j}|| \leq w \\ 0 & \text{otherwise; } \mathbf{s} \in D_{s}, \end{cases}$$
(26)

with j=1,...,75 equally spaced knots \mathbf{c}_j , and where w is 1.5 times the smallest distance between two different knots. The placement of knots was achieved using a space filling design (Nychka and Saltzman, 1998). We performed empirical studies that explore the relationship between r and n_C^{op} (see discussion in Remark 8). These investigations suggest that r=75 is appropriate for this example. (From our experience, our method is rather robust to the placement and number of knots, and the empirical results guiding this experience are provided in the Supplemental Materials.) We considered many different choices of prior distributions for the $r \times r$ covariance matrix \mathbf{Q} , and through independent sensitivity studies we found that the so-called MI prior (Bradley et al., 2015c) appeared to be the most appropriate choice for this example (see the Supplemental Materials for more details). The k-means algorithm is used to define $\mathscr C$ in (19), and we let $g_L=175$ and $g_U=195$. Since the latent field is not interpretable on D_s , we use DCAGE within the expression of D_C^{op} in (20). The variances of $\{\varepsilon(A_i): i=1,...,n\}$ are estimated a priori by ACS, and hence, are

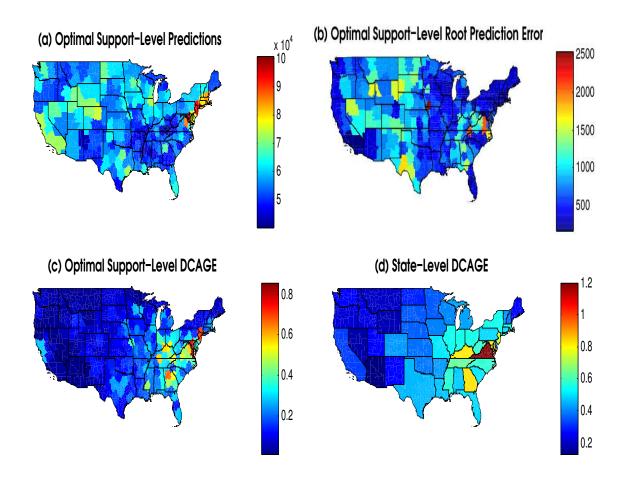


Figure 2: In (a), we present maps (for the contiguous US) of predicted median household income (US dollars) defined on the optimal spatial support (i.e., D_C^{op}) consisting of 185 areal units. Recall, we consider areal units 175 through 195, and the value chosen using DCAGE is 185. We superimpose the state boundaries as a reference to compare to Figure 1(b). In (b) and (c), we present maps of the posterior standard deviations and DCAGE. In (d), we plot DCAGE by states.

assumed known.

In Figure 2(a) and 2(b), we present the predictions and corresponding prediction error of median household income on the optimal spatial support D_C^{op} (and add state boundaries as a reference). In Figure 2(b), the predictions appear fairly precise with largest prediction error occurring in regions near Virginia, which have posterior standard deviation around 2,500 (which is roughly 5% of the mean median household income). The problems with spatial aggregation error indicated by Figures 1(a) and 1(b) described in the Introduction are no longer present in D_C^{op} , which consists of 185 areal units. For example, counties near Richmond constitute a distinct region. Also, the state

of New York is divided into multiple distinct regions: areas near and in Manhattan, western New York, and upstate New York are all separated. However, it is worth noting that in Figure 2(c) the square root DCAGE values are comparatively larger around the state of Virginia.

The DCAGE can also be used for uncertainty quantification. That is, state-level representatives may not be interested in the optimal regionalization produced by the two stage search algorithm, and instead, be interested in the median income over states. The DCAGE can be used to identify which states have poor spatial aggregation error properties. In Figure 2(d), we plot DCAGE over states (i.e., treat states as fixed areal units), which has an average DCAGE of 0.24. This value is larger than the average DCAGE corresponding to the optimal solution, which is 0.19. Notice that the DCAGE corresponding to Virginia (and states near Virginia) are relatively high, while other states in the Midwest and West coast have comparatively smaller values of DCAGE. This would suggest that one should be concerned about assuming that statistics over Virginia can interpreted at lower spatial resolutions.

6 Discussion

The ecological fallacy and MAUP have become popular pedagogical tools for discussion in geography and spatial statistics (Robinson, 1950; Openshaw and Taylor, 1979; Cressie, 1993; Cressie and Wikle, 2011; Banerjee et al., 2015). However, very little has been done to characterize and mitigate these forms of spatial aggregation error from a statistical perspective. Thus, we provide a measure to formally characterize such error and a principled way to obtain an optimal (in terms of spatial aggregation error) regionalization defined over the generic continuous domain $D_s \subset \mathbb{R}^d$. Regionalization has traditionally been solved using techniques outside the realm of statistics (Duque et al., 2012; Spielman and Logan, 2013; Folch and Spielman, 2014; Spielman and Logan, 2015), and our work offers a new perspective that respects the uncertainty of spatial random processes. Consequently, our methodology can significantly impact federal statistics, survey methodology,

geography, spatial statistics, and remote sensing/data acquisition settings.

The heart of our methodology lies in the criterion for spatial aggregation error (CAGE), which we minimize to obtain our optimal regionalization. The methodological development of CAGE is intricate and involves a novel multiscale Karhunen-Loève (K-L) expansion. The introduction of a multiscale K-L expansion provides an approach to spatial COS that is not based on assumptions of between scale homogeneity. Furthermore, the multiscale K-L expansion leads to a powerful technical result that shows that any statistic does not suffer from spatial aggregation error as long as the multiscale eigenfunctions are homogeneous across scales. Thus, CAGE represents a measure of between scale homogeneity of eigenfunctions within a multiscale K-L expansion. There are many additional motivating features of CAGE, including connections to prediction error and across scale homogeneity of variances.

To apply CAGE we need a parameterization of the multiscale eigenfunctions. This allows the eigenfunctions to be estimated, and hence, the CAGE can be informed by the data. Thus, we provide a new class of Obled-Creutin (O-C) eigenfunctions motivated by the seminal paper of Obled and Creutin (1986). The proposed class of O-C eigenfunctions has broad applicability in the sense that any class of generating basis functions (GBF) can be used to build eigenfunctions.

Finally, CAGE is used within an efficient two-stage regionalization algorithm. In the first stage of the algorithm (for a given number of areal units) a deterministic clustering algorithm is applied to each of the M samples from the posterior distribution of the latent process. This defines M spatial supports to select from. Then, in the second stage, the spatial support with the smallest (average) CAGE is chosen. This approach is extremely efficient, and accounts for the variability of the data by performing the search algorithm within the latent process space.

An illustration of our algorithm was given using American Community Survey (ACS) 5-year period estimates of median household income. Comparisons of the optimal spatial support to the state-level ACS estimates indicate that the optimal regionalization preserves the county-level spatial information. Additionally, the size of this dataset is 3,109, and notably, the optimal spatial

support consists of just 185 areal units. The dramatic decrease of the dimensionality of the problem has important implications for modeling very large spatial datasets.

The application of CAGE to reduce the dimensionality of spatial data is just one of many exciting avenues for future research. For example, the introduction of spatially varying covariates into the statistical model will undoubtedly effect the spatial aggregation error properties. Also, as previously mentioned, model selection considerations, such as the number of basis functions and class of basis functions, may effect the conclusions of the two-stage regionalization algorithm. The truncation of the multiscale K-L expansion is especially important from the point of view of regionalization, since fewer basis functions lead to less variable predictions of the latent process, which then leads to fewer areal units produced by the regionalization algorithm. Another interesting idea for future research would be to construct a prior distribution for the regionalization by using the values of the CAGE to define prior weights.

There are minor modifications to CAGE and the two-stage regionalization algorithm that would be reasonable to consider. For example, Proposition 2 shows that spatial aggregation error does not occur when point-level eigenfunctions are constant over each region in the aggregate-level spatial support. Thus, we use the squared distance between point-level and aggregate-level eigenfunctions to measure departures from the absence of spatial aggregation error. However, other distances besides the squared distance might be used. This is similar to considering other forms of prediction error besides squared error. Also, there are a number of alternative search algorithms that one might consider. For example, one could use CAGE within a forward selection algorithm, or perhaps, one might use Spielman and Logan (2015)'s ACS regionalization (AReg) algorithm within the first stage of the two-stage algorithm. It would be difficult to incorporate AReg into the two-stage algorithm practically, since it is not computationally efficient for high-dimensional spatial datasets. The specifications we use are computationally efficient and are shown to give favorable results.

Acknowledgments

This research was partially supported by the U.S. National Science Foundation (NSF) and the U.S. Census Bureau under NSF grant SES-1132031, funded through the NSF-Census Research Network (NCRN) program. In addition, C.K. Wikle acknowledges the support of NSF grant DMS-1049093 and Office of Naval Research (ONR) grant ONR-N00014-10-0518.

Supplemental Materials: Regionalization of Multiscale Spatial Processes using a Criterion for Spatial Aggregation Error

Jonathan R. Bradley³, Christopher K. Wikle⁴, Scott H. Holan²

Keywords: American Community Survey; Empirical orthogonal functions; MAUP; Reduced rank; Spatial basis functions; Survey data

 $^{^3}$ (to whom correspondence should be addressed) Department of Statistics, University of Missouri, 146 Middlebush Hall, Columbia, MO 65211, bradleyjr@missouri.edu

⁴Department of Statistics, University of Missouri, 146 Middlebush Hall, Columbia, MO 65211-6100

I Introduction

In this supplement to "Regionalization of Multiscale Spatial Processes using a Criterion for Spatial Aggregation Error," by J.R. Bradley, C.K. Wikle, and S.H. Holan, we give additional insight to CAGE and the two-stage regionalization algorithm outside of what was presented in the main text. In particular, we have applied the algorithm to another dataset, performed many different sensitivity analyses, and provided additional material that is meant to aid readers interested in implementing our procedure.

This supplement is organized as follows. In Section II, we provide guidance on the implementation of our algorithm including: a summary of the statistical model used in Section 5, details on prior distribution considerations, a step-by-step outline of estimation and the two-stage regionalization procedure, and additional discussion on model and regionalization specifications. Note, we use Roman numerals for section titles in this Supplement to distinguish from section titles in the main text. In Section III, we provide sensitivity analyses including: a comparison to a current state-of-art method for regionalization within the geography literature from Speilman et al. (2013), a sensitivity analysis to the choice of D_A , and a simulation study investigating the choice of the rank of the truncated multiscale K-L expansion. Next, in Section IV we provide a demonstration of the two-stage regionalization algorithm to a dataset consisting of Mediterranean wind measurements (a subset of the data used in Milliff et al. (2011)), which is used illustrate that the two-stage regionalization algorithm is flexible enough to handle multiscale spatial data. Finally, in Section V we provide the proofs to the technical results from the main-text.

II Additional Details for Implementation

Here, we give guidance on the implementation of our algorithm including: a summary of the statistical model used in Section 5 (Section II.i), details on prior distribution considerations (Sec-

tion II.ii), a step-by-step outline of the estimation and the two-stage regionalization procedure (Section II.iii), and additional discussions on model and regionalization specification (Section II.iv).

II.i Outline of the Statistical Model

The statistical model introduced in Section 3.1 is summarized in Algorithm 1 below. We choose to describe this Bayesian hierarchical model using the data, process, and parameter model terminology from Berliner (1996).

Algorithm 1: Outline of the statistical model introduced in Section 3.1

Data Model: $Z(\mathbf{u})|\mu, \boldsymbol{\eta}, \mathbf{Q}, \boldsymbol{\xi} \stackrel{\text{ind}}{\sim} \text{Normal } \{\mu + \boldsymbol{\psi}^*(\mathbf{u})'\boldsymbol{\eta} + \delta(\mathbf{u}; \boldsymbol{\xi}), \sigma_Z^2(\mathbf{u})\};$

Process Model 1: $\eta | \mathbf{Q} \sim \text{Gaussian}(\mathbf{0}, \mathbf{Q});$

Process Model 2: $\boldsymbol{\xi} | \sigma_{\boldsymbol{\xi}}^2 \sim \text{Gaussian} \left(\mathbf{0}, \sigma_{\boldsymbol{\xi}}^2 \mathbf{I}_{n_B} \right);$

Parameter Model 1: $\mu \sim \text{Normal}\left(0, \sigma_{\mu}^{2}\right)$;

Parameter Model 2: $\sigma_{\xi}^2 \sim \text{IG}\left(\alpha_{\xi}, \beta_{\xi}\right)$;

Parameter Model 3: $\mathbf{Q} \sim [\mathbf{Q}]; \ \mathbf{u} \in D_s \cup D_A$.

Here, the n_B -dimensional random vector $\boldsymbol{\xi} \equiv (\xi_1, ..., \xi_{n_B})'$, $\sigma_{\mu}^2 > 0$, $\alpha_{\xi} > 0$, $\beta_{\xi} > 0$, and we let $[\mathbf{Q}]$ denote a probability density function for the unknown $r \times r$ covariance matrix \mathbf{Q} . We consider many different choices for $[\mathbf{Q}]$, and provide these details in Section II.ii. The value of σ_{μ}^2 is chosen to be large so that the prior distribution on μ is interpreted to be vague, and similarly, we set $\alpha_{\xi} = \beta_{\xi} = 1$ so that the prior distribution on σ_{ξ}^2 is flat.

II.ii Prior Distributions to Consider

As Sorbye and Rue (2014) discuss, the prior distribution (and the associated hyperparameters) on the $r \times r$ covariance matrix \mathbf{Q} affects posterior inference. As such, we consider several different choices for priors on covariance matrices. In particular, we consider three different prior distributions. The first prior distribution we consider is the conjugate inverse Wishart distribution. This is a fairly common choice because it allows for direct sampling of the full-conditional distribution corresponding to \mathbf{Q} , however, in high-dimensions this prior is known to perform poorly (Hodges, 2013).

The second prior distribution we consider is from Bradley et al. (2014b) and Bradley et al. (2015c), where it is assumed that

$$\mathbf{Q} = \frac{1}{\sigma^2} \left[\mathbf{R}_B^{-1} \mathscr{A}^+ \left\{ \mathbf{Q}_B' \left(\mathbf{I} - \mathbf{A} \right) \mathbf{Q}_B \right\} \mathbf{R}_B^{-1} \right]^{-1}, \tag{2.0}$$

where $\mathscr{A}^+(\mathbf{M})$ is the best positive approximate (Higham, 1988) of a square real-valued matrix \mathbf{M} , $\sigma^2 > 0$ is unknown, the $n_B \times r$ matrix $\mathbf{\Psi}_B \equiv (\mathbf{\psi}(B)' : B \in D_B)'$, $\mathbf{\Psi}_B = \mathbf{Q}_B \mathbf{R}_B$ is the QR decomposition, and \mathbf{A} is the $n_B \times n_B$ adjacency matrix corresponding to D_B . Notice that (6) incorporates spatial information, but is not spatially referenced. That is, this prior for \mathbf{Q} is motivated by specifying $\operatorname{cov}(\mathbf{\Psi}_B \mathbf{\eta})$ so that it is "close" to the covariance from an ICAR model on D_B , where $\mathbf{\Psi}_B$ is spatially referenced but $\mathbf{\eta}$ is not. An inverse gamma prior is placed on σ^2 where the hyperparameters are chosen based on the suggestions in Section 3.2 of Sorbye and Rue (2014). Following Bradley et al. (2014b) and Bradley et al. (2015c), we refer to this prior specification as the "MI" prior distribution due to a connection to the Moran's I statistic.

The third prior distribution we consider is the Givens angle prior (Yang and Berger, 1994; Bradley et al., 2015b), where the spectral decomposition is written as $\mathbf{Q} = \mathbf{P}_{\mathbf{Q}} \mathbf{\Lambda}_{\mathbf{Q}} \mathbf{P}_{\mathbf{Q}}$, and the $r \times r$ diagonal matrix $\mathbf{\Lambda}_{\mathbf{Q}}$ has diagonal entries set equal to the eigenvalues of (6). The parameter σ^2 is assumed to follow a flat inverse gamma distribution (i.e., with shape and scale set equal to 1). The

 $r \times r$ orthogonal matrix P_Q is decomposed into a Givens rotator product,

$$\mathbf{P}_{\mathrm{Q}} \equiv (\boldsymbol{O}_{1,2} \times \boldsymbol{O}_{1,3} \times \cdots \times \boldsymbol{O}_{1,r}) \times (\boldsymbol{O}_{2,3} \times \cdots \times \boldsymbol{O}_{2,r}) \times \cdots \times \boldsymbol{O}_{r-1,r},$$

where $O_{i,j}$ is a $r \times r$ identity matrix with the (i,i)-th and (j,j)-th element replaced by $\cos(\theta_{i,j})$ and the (i,j)-th ((j,i)-th) element replaced by $-\sin(\theta_{i,j})$ $(\sin(\theta_{i,j}))$. Here, $\theta_{i,j} \in [-\pi/2, \pi/2]$ is unknown, and let the shifted and rescaled $\theta_{i,j}$ be denoted as $\zeta_{i,j} \equiv 1/2 + \theta_{i,j}/\pi$. Then, it is assumed that

$$logit(\zeta_{i,j}) = a + b \times g_{i,j}(\mathbf{P}_{O}); \ i < j = 1, ..., r,$$
(2.0)

where $\operatorname{logit}(\zeta_{i,j}) \equiv \operatorname{log}\{\zeta_{i,j}/(1-\zeta_{i,j})\}$, $a,b \in \mathbb{R}$, and $g_{i,j}(\mathbf{P}_{Q})$ represents the (i,j)-th Givens angle of \mathbf{P}_{Q} . Finally, a vague Gaussian prior is placed on (a,b)' (i.e., Gaussian with mean zero and variance 1000). For all of our analyses we considered all three prior distributions. These sensitivity analyses suggested that the MI prior lead to the best predictive performance for the application in Section 5, and the inverse Wishart prior led to the best predictive performance in Section V.

II.iii Outline: Estimation and Implementation of Regionalization

In this section we give a brief outline of the two-stage regionalization algorithm. It should be acknowledged that, for any given application, minor modifications to these steps may be needed.

- 1. Define the spatial support D_B , which represents the finest resolution one is willing to predict on. If $D_s^O = \emptyset$ we suggest setting $D_B = D_A$, which is the finest resolution information that is available. When $D_s^O \neq \emptyset$ then one has the freedom to choose any spatial support for D_B , however, one should be mindful of the size and spatial coverage of the locations within D_s^O . Thus, for illustration, when $D_s^O \neq \emptyset$ we suggest setting D_B to a fine resolution grid.
- 2. Obtain M MCMC replicates of $\mathbf{Y}_B \equiv (Y_A(B) : B \in D_B)'$, using the statistical model in Algorithm 1. Specifically, let $\boldsymbol{\eta}^{[m]}$ represent the m-th replicate of $\boldsymbol{\eta}$ and $\boldsymbol{\xi}^{[m]}$ represent the m-th

replicate of $\boldsymbol{\xi}$. Then, the *m*-th replicate of \mathbf{Y}_B can be computed as

$$\mathbf{Y}_{B}^{[m]} = \mathbf{\Psi}_{B} \boldsymbol{\eta}^{[m]} + \boldsymbol{\xi}^{[m]}; \ m = 1, ..., M,$$

where the $n_B \times r$ matrix $\Psi_B \equiv (\psi^*(\mathbf{u})' : \mathbf{u} \in D_B)'$. The Bayesian procedure can easily implemented using a Metropolis with in Gibbs sampling algorithm.

- 3. Use a naive clustering algorithm to obtain \mathscr{C} in (19). We consider two clustering algorithms to define \mathscr{C} , namely, the k-means algorithm, and structural hierarchical clustering. In general, the k-means algorithm takes on as it's argument an $n_B \times f$ real-valued matrix \mathbf{J} , and returns a clustering of the rows of \mathbf{J} . Let \mathbf{L} be a $n_B \times d$ matrix with the j-th row equaling the centroid of the j-th areal unit in D_B . Then, we let f = d + 1 and set $\mathbf{J} = [\mathbf{L}, \mathbf{Y}_B^{[m]}]$. The structural hierarchical clustering approach takes on two arguments $\mathbf{J} = [\mathbf{L}, \mathbf{Y}_B^{[m]}]$ and the adjacency matrix corresponding to D_B .
- 4. Choose the spatial support from \mathscr{C} that minimizes CAGE. That is, compute D_C^{op} according to (20). If Y can not be interpreted on D_s substitute CAGE with DCAGE.
- 5. Produce maps of the values in the sets $\{\widehat{Y}_A(C^{op}): C^{op} \in D_C^{op}\}$, $\{\operatorname{var}(Y_A(C^{op}|\mathbf{Z})): C^{op} \in D_C^{op}\}$, and $\{\operatorname{CAGE}(C^{op}): C^{op} \in D_C^{op}\}$ (or $\{\operatorname{DCAGE}(C^{op}): C^{op} \in D_C^{op}\}$ when appropriate). This allows one to visualize the process and its corresponding prediction and spatial aggregation errors.

II.iv Model and Regionalization Algorithm Specifications

To implement the two-stage regionalization algorithm, we need to specify: the number and placement of knots that define the r-dimensional GBF ψ , and the lower and upper bounds on the number of areal units used within the two-stage regionalization algorithm (i.e., g_L and g_U). We now provide discussion on to make these choices in practice.

Specification of Knots: The choice of knots and r is important for preserving the appropriate fine-scale features of Y_s . If the fine-scale features of Y_s are ignored then the two-stage regionalization algorithm may produce too coarse of a regionalization (see simulation study in Section IV.iii). However, the number of areal units produced by the two-stage regionalization algorithm appears to be robust to r "too large." Recall the number of areal units in D_C^{op} is denoted with n_C^{op} . This interaction between the number of optimal areal units and r suggests an approach for selecting the rank r, which we outline into the following steps:

- (1) Consider a fixed range of values for r (i.e., $r = r_L, ..., r_U$).
- (2) For each $r = r_L, ..., r_U$, use the algorithm outlined in II.iv to find an optimal regionalization and n_C^{op} . There will be a different value of n_C^{op} for each each $r = r_L, ..., r_U$.
- (3) Plot r versus n_C^{op} .
- (4) Choose the value of r to be the point in which n_C^{op} does not change dramatically as r increases.

We follow the suggestion of Ruppert et al. (2003, chap. 13, pp. 255-260) and apply a space filling design algorithm to a set of randomly selected points $\{\mathbf{c}_j: j=1,...,r^*\}$, where we set $r^*=600>r$. The space-filling design can be determined using the FUNFITS function in R (Nychka et al., 1998). Then, we choose r according to steps 1-3 above. For the applications in Section 5 and Section V we found that, respectively, r=75 and r=200 are appropriate.

Specification of g_L and g_U : The widest range of values that we can consider for regionalization is $g_L = 2$ and $g_U = n_B - 1$. To specify less extreme choices for $g_L = 2$ and $g_U = n_B - 1$ we consider running a simplified version of the two-stage regionalization algorithm, and use the results of the "simplified two-stage regionalization algorithm" to inform a tighter range between g_L and g_U .

In particular, we first run the two-stage regionalization algorithm (outlined in Section II.iii) with M=1, $g_L=2$, $g_U=n-1$, and use the k-means algorithm. Then, we choose g_L and g_U to be a tight range centered around n_C^{op} found using this simplified two-stage regionalization algorithm.

III Simulations, Sensitivity Analyses, Comparisons, and Technical Clarifications

Here, we provide many different side-studies including: a simulation study to compare the twostage regionalization algorithm to a current state-of-the-art alternative in the geography literature, Spielman and Logan (2015)'s ACS regionalization (AReg) algorithm (Section III.i); a small sensitivity analysis on the choice of D_A (Section III.ii); and a simulation study investigating the choice of the rank of the spatial basis function expansion (Section III.iii).

III.i Simulation Study: A Comparison to Speilman et al. (2013)

In this section, we establish that our approach performs regionalization extremely well relative to the AReg algorithm available in the geography literature. To do this, we generate synthetic data based on a subset of the ACS 5-year period (from 2009 to 2013) estimates of the percentage of households below the poverty threshold. We generate the spatial field,

$$Z(A) = Y_A(A) + \varepsilon(A); \ A \in D_A, \tag{3.0}$$

where D_A is the set of 351 census tracts surrounding the city of Austin (TX). Let $\{Z(A)\}$ represent the perturbed version of the logit transformed percent below the poverty level ACS survey estimate (denoted by $\{Y_A(A)\}$). (Notice that we use the symmeterizing logit transformation, where, for a given percentage p, logit(p) = p/(1-p).) The set $\{\varepsilon(A) : A \in D_A\}$ consists of independent normal random variables with mean-zero and known variance. The published variances for percent below the poverty level are transformed to the logit scale using the delta method (Oehlert, 1992), and used as the known variances of $\{\varepsilon(A)\}$. In practice, the ACS estimates (i.e., $\{Y_A\}$ for this example) are publicly available and are, hence, observed. Nevertheless, for the purposes of this simulation study we will act as if the ACS estimates are an unobserved spatial field to be estimated from Z.

To obtain D_C^{op} , we model this data using the mixed effects model in Algorithm 1, where $\psi(\cdot) \equiv (\psi_j(\cdot) : j = 1,...,42)'$ is a 42-dimensional vector consists of local bisquare functions (Cressie and Johannesson, 2008):

$$\psi_{j}(\mathbf{s}) \equiv \begin{cases} \{1 - (||\mathbf{s} - \mathbf{c}_{j}||/w)^{2}\}^{2} & \text{if } ||\mathbf{s} - \mathbf{c}_{j}|| \leq w \\ 0 & \text{otherwise; } \mathbf{s} \in D_{s}, \end{cases}$$
(3.0)

with j=1,...,42 equally spaced knots \mathbf{c}_j , and w is 1.5 times the smallest distance between two different knots. Note, that we are not restricted to using local bisquare functions, since our modeling framework is general enough to allow for any desired GBF. For computational convenience, we use the k-means algorithm to define $\mathscr C$ in (19), and let $g_L=2$ and $g_U=100$. The latent process in (6) is not defined on D_s , and thus, we shall use DCAGE within the expression of D_C^{op} in (20). Additionally, we denote the output of AReg with $D_A^{\mathrm{AReg}} \equiv \{A_k^{\mathrm{AReg}}: k=1,...,n_A^{\mathrm{AReg}}\}$, and compute it using software made available at https://github.com/geoss/ACS_Regionalization/blob/master/README.md.

The goal of this simulation study is to compare the error properties of D_C^{op} , and D_A^{AReg} . This is done using the following metrics:

$$\begin{aligned} \text{ReMSPE}(Z_{A}) &\equiv \frac{\sum_{j=1}^{n_{A}^{\text{AReg}}} \frac{1}{|A_{j}^{\text{AReg}}|} \left\{ Y_{A}(A_{j}^{\text{AReg}}) - \widehat{Y}_{A}(A_{j}^{\text{AReg}}) \right\}^{2}}{\sum_{j=1}^{n_{C}^{op}} \frac{1}{|C_{j}^{op}|} \left\{ Y_{A}(C_{j}^{op}) - \widehat{Y}_{A}(C_{j}^{op}) \right\}^{2}} \\ \text{ReCAGE}(Z_{A}) &\equiv \frac{\sum_{j=1}^{351} \sum_{k=1}^{n_{A}^{\text{AReg}}} I(A_{j} \subset A_{k}^{\text{AReg}}) \left[\frac{\left\{ Y_{A}(A_{j}) - Y_{A}(A_{k}^{\text{AReg}}) \right\}^{2}}{|A_{k}^{\text{AReg}}|} \right]}{\sum_{j=1}^{351} \sum_{k=1}^{n_{C}^{op}} I(A_{j} \subset C_{k}^{op}) \left[\frac{\left\{ Y_{A}(A_{j}) - Y_{A}(C_{k}^{op}) \right\}^{2}}{|C_{k}^{op}|} \right]}, \end{aligned}$$

where $I(\cdot)$ is the indicator function. Here, ReMSPE stands for "relative mean squared prediction error" and ReCAGE stands for "relative spatial aggregation error," respectively. Values of ReMSPE that are larger (smaller) than 1.0 indicate that prediction on D_C^{op} has smaller (larger) MSPE than when predicting on D_A^{AReg} . Thus, values of ReMSPE that are larger (smaller) than 1.0 indicate that the two-stage algorithm (AReg) leads to better (worse) predictive performance. Likewise, values of ReCAGE that are larger than 1.0 indicate that the two-stage algorithm is preferable in terms of spatial aggregation error.

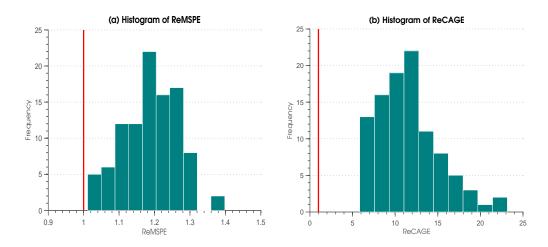


Figure 3: In (a) and (b), we present histograms of ReMSPE and ReCAGE from taken over the 100 replicates of Z defined in (21). The red line indicates the value of 1 in both panels. A value of ReMSPE and ReCAGE greater than 1.0 indicates that the two-stage regionalization algorithm is preferable over AReg.

We simulate 100 replicates of Z in (6), and compute ReMSPE and ReCAGE for each of the 100 replicates. For both metrics our proposed algorithm consistently outperforms AReg. In fact, in each of the 100 replications of Z we obtain an ReMSPE > 1.0, and a ReCAGE > 1.0, where ReMSPE ranges from 1.0112 to 1.3979 and ReCAGE ranges from 5.8408 to 23.1620, respectively (see Figure 1 for a histogram over the 100 replications of Z). It is somewhat expected that ReCAGE suggests that the two-stage regionalization algorithm is preferable over AReg because from Proposition 3, CAGE is directly related to the squared difference between the lower spatial resolution

process and the aggregate-level estimator. However, it is rather interesting that ReMSPE suggests that the two stage algorithm is also preferrable in terms of squared prediction error, since AReg is motivated by reducing sampling error. This may be due to the fact that AReg does not take into account survey error (i.e., $\{\varepsilon(A)\}$), while the two-stage regionalization algorithm accounts for this error by performing its search in latent space.

III.ii Sensitivity to D_A

Notice that the two-stage search algorithm takes on D_s and D_A (the spatial domains of interest) as an input. Thus, one might be interested in the sensitivity of our approach to the spatial domain of interest. For example, in Figure 2(a) we plot the optimal areal units (i.e., D_C^{op}), found in Section 5, over California, Oregon, Nevada, and Arizona. Now, suppose we let D_A consist of the 126 counties in California, Oregon, Nevada, and Arizona, and we re-run the two stage search algorithm on this restricted domain (i.e., D_A no longer consists of all counties in the mainland of US, but consists only of counties in California, Oregon, Nevada, and Arizona). The D_C^{op} found under this restriction is given in Figure 2(b).

There are 12 areal units in D_C^{op} without restricting D_A , and 11 when one restricts D_A . Upon comparison of Figures 2(a) to 2(b) we see that the general pattern of the two-stage search algorithm is robust to this change in D_A , however, the final answer does change. We note that since the initialization of the k-means algorithm is random, the candidate set of areal units are not necessarily the same each time one runs the two-stage search algorithm.

III.iii Simulation Study: Selection of the Rank of the Truncated Multiscale K-L Expansion

In this section, we use simulation to investigate the impacts of misspecifying the rank of the truncated multiscale K-L expansion. In particular, we choose a simulation model with r = 100 random



(b) D_A equal to CA, OR, NV, AR only



Figure 4: In (a), we plot the optimal areal units (i.e., D_C^{op}), found in Section 5, over the state of California. In (b), we plot the D_C^{op} found by restricting D_A to consist only of counties in California, Oregon, Nevada, and Arizona. Each distinct color identifies a different areal unit, and the relative difference between each color is arbitrary. The state boundaries are superimposed as a reference.

effects, and we perform regionalization with r misspecified and r correctly specified. The regionalization with r correctly specified is treated as the "correct" regionalization, which we compare to.

Let the latent process of interest Y_s be generated as follows:

$$Y_s(\mathbf{s}) = \mu + Y_s(\mathbf{s}; \boldsymbol{\phi}_s^{\text{OC}}) + \delta(\mathbf{s}; \boldsymbol{\xi}); \, \mathbf{s} \in D_s,$$
 (3.-1)

where $D_s \equiv \{\mathbf{s} = (s_1, s_2)' : s_1, s_2 = [0.05, 0.1, 0.15, ..., 1] \times [0.05, 0.1, 0.15, ..., 1]\}$, recall $Y_s(\mathbf{s}; \boldsymbol{\phi}_s^{\text{OC}})$ is defined in (11), and let $\boldsymbol{\phi}_s^{\text{OC}}$ be based generated from 100 equally spaced (over D_s) local bisquare basis functions. The corresponding dataset is generated as follows:

$$Z_s(\mathbf{s}) = Y_s(\mathbf{s}) + \varepsilon_s(\mathbf{s}); \mathbf{s} \in D_s^O \subset D_s$$

 $Z_A(\mathbf{s}) = Y_A(A) + \varepsilon_A(A); A \in D_A,$ (3.-1)

where we randomly select 50% of the observations from D_s to define D_s^O , and D_A consists of the 10×10 grid cells that cover $[0,1] \times [0,1]$. We let $\varepsilon_s(\cdot)$ be a mean zero white-noise process with

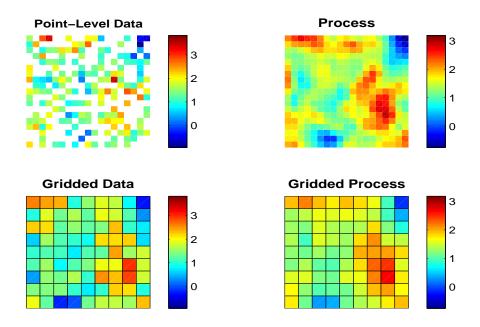


Figure 5: Example simulated data and process. These maps are produced using (3.-1) and (3.0). The top left panel contains simulated data on D_s (with 50% of the field being covered). The top right panel contains the simulated process on D_s . The bottom left panel contains the aggregate data process (i.e., Z_A), which has complete spatial coverage over D_A . The bottom right panel displays Y_A .

variance $\sigma_{\varepsilon}^2 = 0.1820$ (so that the signal-to-noise ratio (=5) is large). Likewise, $\{\varepsilon_A(A) : A \in D_A\}$ consists of i.i.d. independent mean zero random variables with variance 0.1820, and is independent of the spatial random process $\varepsilon_s(\cdot)$. An example of the data and the process is given in Figure 3.

Consider performing regionalization using the outline in Section II.iii, to the data in Figure 3 with r = 9,100, and 256. For illustration let $D_B = D_A$, and set $g_L = 2$ and $g_U = 99$ (the largest possible range). Here, r = 9 represents the case where r is too small, r = 100 represents the case where r is correct, and r = 256 represents the case when r is too large. When r is too small we obtain fewer areal units (6 areal units) than when r is correct (13 areal units); however, the optimal regionalization algorithm is robust to the case where r is too large, which produced 15 areal units. This conforms to intuition as it is well known that predictors based on spatial basis functions with

r-large display more fine-level details than predictors based on spatial basis functions with r-small (Bradley et al., 2011; Stein, 2013; Bradley et al., 2014a). Thus, one would expect that if r is chosen to be "too small" then predictions of Y_s will have less variability over D_s (i.e., be more constant), and consequently lead to coarser regionalizations.

These conclusions are similar over multiple replications; in Figure 4 we provide histograms of n_C^{op} obtained from the two-stage regionalization algorithm over 50 independent replications of $\{Z_s\}$ and $\{Z_A\}$. Notice, however, that the variability associated with r too large is much higher than when r is too small and when r is correct. The p-value of a sign test comparing n_C^{op} when r = 9 (r = 256), to n_C^{op} when r = 100 is 0.0494 (0.5716), which suggests that when r is too small (too large) we obtain coarser (similar) results than when r is correct.

The fact that there is no significant change in the number of areal units when r is too large also conforms to intuition; since there are enough spatial random effects to capture fine-scale behavior, and the remaining random effects are negligible. This interaction between the number of optimal areal units and r suggest an approach for choosing r (i.e., Steps 1–3 in Section II.iv). For the ACS application in Section 5, we consider r = 25,50,75,100,125, and 150. Likewise for the Mediterranean wind example we consider r = 25,50,75,100,125, and 150. In Figure 5, we plot n_C^{op} versus r (i.e., Step 3 from Section II.iv). Here, we see that for the applications in Section 5 and Section V we found that, respectively, r = 75 and r = 200 are appropriate.

III.iv Technical Clarifications: Positive Definiteness of the Multiscale K-L Expansion

A covariance function $\operatorname{cov}\{Y_s(\mathbf{s}),Y_s(\mathbf{u})\}\$ is positive definite if (Cressie, 1993, p. 68),

$$\sum_{i=1}^{m} \sum_{j=1}^{m} b_i b_j \operatorname{cov}\left\{Y_s(\mathbf{s}_i), Y_s(\mathbf{s}_j)\right\} \ge 0$$
(3.-1)

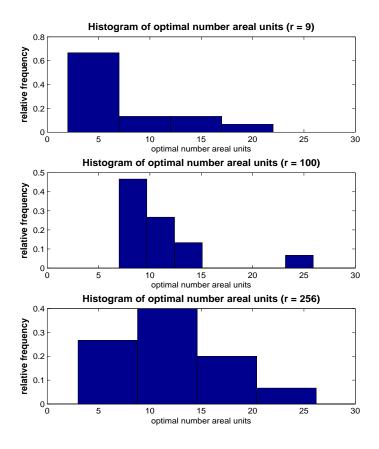


Figure 6: Histograms of n_C^{op} over 50 independent replications of $\{Z_s\}$ and $\{Z_A\}$. The value of r used to fit Algorithm 1 is indicated in the title of the panel.

for any finite number of spatial locations $\{\mathbf{s}_i: i=1,...,m\}$ and any set of real numbers $\{b_i: i=1,...,m\}$. That is, the covariance function, associated with the spatial random process Y_s , is positive definite if a weighted average of covariances implied by any set $\{Y_s(\mathbf{s}_i): i=1,...,m\}$ has nonnegative variance, where $\{b_i: i=1,...,m\}$ are the generic weights. The validity of the covariance of Y_s in (11) follows immediately from the definition of positive definiteness, and the quadratic form of

$$\operatorname{cov}\left(\mathbf{Y}^{(m)}\right) = \mathbf{\Psi}^{(m)} \mathbf{\Lambda} \mathbf{\Psi}^{(m)\prime},$$

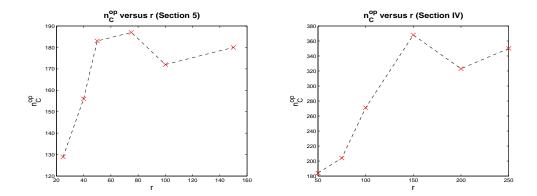


Figure 7: The plot of n_C^{op} versus r as described in Section II.iv. In the left panel we plot n_C^{op} versus r for the ACS example presented in Section 5, and in the right panel we plot n_C^{op} versus r for the wind example in Section IV. The values of r considered in the ACS example in Section 5 were 25, 40, 50, 75, 100, and 150. The values of r considered in the wind example in Section IV were 50, 75, 100, 150, 200, and 250.

where Λ is defined below Equation (15) of the main text,

$$\mathbf{Y}^{(m)} \equiv \left\{ Y_s(\mathbf{s}_1; \boldsymbol{\phi}_s), ..., Y_s(\mathbf{s}_m; \boldsymbol{\phi}_s) \right\}',$$

and

$$\mathbf{\Psi}^{(m)} \equiv \{\boldsymbol{\phi}_s(\mathbf{s}_1),...,\boldsymbol{\phi}_s(\mathbf{s}_m)\}'.$$

That is, let $\mathbf{b} = (b_1, ..., b_m)'$, and notice that

$$\sum_{i=1}^{m} \sum_{j=1}^{m} b_i b_j \operatorname{cov} \left\{ Y_s(\mathbf{s}_i), Y_s(\mathbf{s}_j) \right\} = \operatorname{cov} \left(\mathbf{b}' \mathbf{Y}^{(m)} \right) = \mathbf{b}' \mathbf{\Psi}^{(m)} \mathbf{\Lambda} \mathbf{\Psi}^{(m)'} \mathbf{b} \ge 0,$$

and hence, (6) holds for the covariance associated with Y_s in (11). In a similar manner, one can prove the validity of the covariance function of Y in (1) using Proposition 1.ii.

IV Application: Mediterranean Surface Winds

A critical component of the interface between the atmosphere and the upper ocean occurs due to the transfer of momentum and the exchange of heat and fresh water, which is manifested through surface winds from the atmosphere. Due to a lack of direct measurements of surface wind over the ocean, wind field estimates over such regions were historically based on a blend between mechanistic models of the atmosphere and a relatively sparse global network of wind observations from buoys and ships of opportunity. The practical spatial resolution of these so-called "analysis" winds is limited to fairly large spatial and temporal scales of variability, yet they are reported on fairly high-resolution grids. The advent of space-borne scatterometer instruments in the 1990s provided the first high-volume, high-resolution in space, wind estimates over the oceans. Although these scatterometer winds have higher spatial resolution (effectively "point" scale), they are incomplete in space and time, necessitating an optimal blending approach (e.g., Wikle et al. (2001)). Milliff et al. (2011), and Wikle et al. (2013) give reviews of recent statistical approaches to generate spatially and temporally complete ocean wind fields.

As mentioned above, the weather center analysis winds do not contain spatial information commensurate with the spatial support in which they are estimated (e.g., see Milliff et al. (2011) for discussion). That is, the kinetic energy spectrum of the winds does not contain realistic variation at small spatial scales. The support given by the additional (and incomplete) scatterometer wind estimates is relatively much smaller. To date, there have been no attempts to consider an optimal spatial support for statistical wind predictions given these types of data.

In the example presented here, we consider ocean surface wind data from two sources over the Mediterranean Sea. In particular, we consider the north-south wind component for analysis winds from the European Center for Medium range Weather Forecasting (ECMWF) and satellite wind observations from the QuikSCAT scatterometer; this is a subset of the data used in the study by Milliff et al. (2011). We assume that the high resolution (25-km) scatterometer wind observations

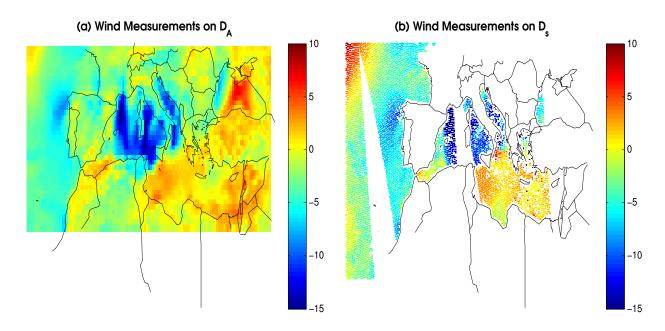


Figure 8: Wind observations from 2 February 2005 at 12:00 UTC (Universal Coordinated Time). (a) North-south (v) component of the wind from the ECMWF-analysis winds on a $0.5^{\circ} \times 0.5^{\circ}$ grid. (b) North-south wind component from the high resolution (25km), but spatially intermittent, QuickSCAT scatterometer wind retrievals.

are effectively "point" support (relative to the analysis winds). Thus, these data are recorded on both $D_s \subset \mathbb{R}^2$ and D_A . Here, D_s ranges from 30° to 48° north latitude, and -19° to 42° east longitude, and D_A consists of a $0.5^{\circ} \times 0.5^{\circ}$ resolution grid on D_s . In total, D_A consists of 4,551 areal units and D_s consists of 6,916 observations for the time of interest, resulting in a dataset of 11,467 spatial observations. Figure 8 shows these data for a 6-hour window centered on 12:00 UTC (Universal Coordinate Time) for 2 February, 2005.

In this application, we let D_B be a half-degree grid. We consider the model in Algorithm 1, where ψ is a multiresolution bisquare basis vector consisting of local bisquare functions in (6). We chose r=200 knots using a space-filling design and the plot in Figure 5 (see Section II.iii). We consider both structural hierarchical clustering and k-means to define $\mathscr C$ in (19) with $g_L=280$ and $g_U=380$; note that we these choices of $g_L=280$ and $g_U=380$ were guided by the approach discussed in Section II.iii using the k-means algorithm with initial choices of $g_L=2$ and $g_U=600$.

We also considered an equivalent analysis using the Wendland GBFs with *k*-means clustering. Here, the Wendland basis functions (Wendland, 1998) are defined as

$$\psi_j^{\text{WL}}(\mathbf{s}) = \begin{cases} (1 - d_j(\mathbf{s}))^6 (35d_j(\mathbf{s})^2 + 18d_j(\mathbf{s}) + 3)/3 & \text{if } 0 \le d_j \le 1\\ 0 & \text{otherwise; } \mathbf{s} \in D_s \end{cases}$$
(4.0)

where j = 1,...,200, $d_j(\mathbf{s}) = ||\mathbf{s} - \mathbf{c}_j^*||/w$, we choose w = 1.5 times the smallest distance between two different knots, and $\{\mathbf{c}_j\}$ consists of the same 200 knot specifications used in the bisquare basis functions. Additionally, since the latent field is interpretable on D_s , we use CAGE within the expression of D_C^{op} in (20). Following Milliff et al. (2011), the variances of $\varepsilon(\mathbf{u})$ are set equal to 1 when $\mathbf{u} \in D_s$, and set equal to 10 when $\mathbf{u} \in D_A$.

The results of the CAGE analysis of the posterior wind predictions is given in Figure 9. The top row of this figure shows that when using the standard 0.5° resolution support, there is a noticeable high CAGE "crescent" in the south central portion of the region. This would suggest that one should be concerned about assuming that statistics on the wind field over this region can be interpreted at the point level. Note that the optimal support regions with k-means and bisquare GBFs (the second row of 9) are much larger than the D_B level shown in the first row, but the predictions look qualitatively similar to the half-degree predictions, although with more smoothing and the corresponding reduction in root prediction error associated with the relatively large optimal aggregation regions. The optimal aggregation seems to pick up realistic meteorological features. For example, notice the homogeneous region centered on Corsica and Sardina, which corresponds to a region of more intense southerly winds off of the mainland (so-called "Mistral winds") that are important in forcing the ocean circulation (e.g., see Milliff et al. (2011)). Perhaps more importantly, although the higher CAGE crescent is still present, it is noticeably reduced in intensity relative to the D_B support. The Wendlend GBF predictions (third row) are similar to the bisquare predictions, but with generally larger regions and with higher CAGE values that are shifted north-

ward. Finally, the last row of Figure 9 shows the bisquare results with the structural hierarchical clustering method. These are similar to the bisquare *k*-means results, but one notices more spatial detail in the predictions.

There is a striking amount of dimension reduction that results from the CAGE analysis. That is, values of n_C^{op} are considerably smaller than the number of observations, 11,467. We have that $n_C^{op} = 323$ when using the bisquare GBFs and k-means, $n_C^{op} = 315$ when using the Wendland GBFs and k-means, and $n_C^{op} = 327$ when using the bisquare GBFs and SHC. This suggests that optimal aggregation, such as the results presented in Figure 7, may be a viable alternative approach for dimension reduction.

We note that there is quite a large amount of shrinkage in these wind predictions relative to the data, which is not surprising given the uncertainty in the winds and the fact that no temporal information is being considered here. As discussed in Wikle et al. (2013), one can gain significant prediction efficiencies if temporal dynamic information is included in the model for winds. Such an analysis is beyond the scope of this simple illustration, but the CAGE-based selection of prediction support could, in principle, be utilized in that framework.

V Technical Proofs

In Section V.i, we provide the proofs to Propositions 1–6. In addition to these proofs, we also provide results alluded to, but not explicitly stated in the main text (Section V.ii).

V.i Proof of Propositions 1−6

Proof of Proposition 1: The assumptions of Proposition 1 allow us to apply the K-L decomposition of $\{Y(\mathbf{s}) : \mathbf{s} \in D_s\}$ from Karhunen (1947). That is, from Karhunen (1947) we have

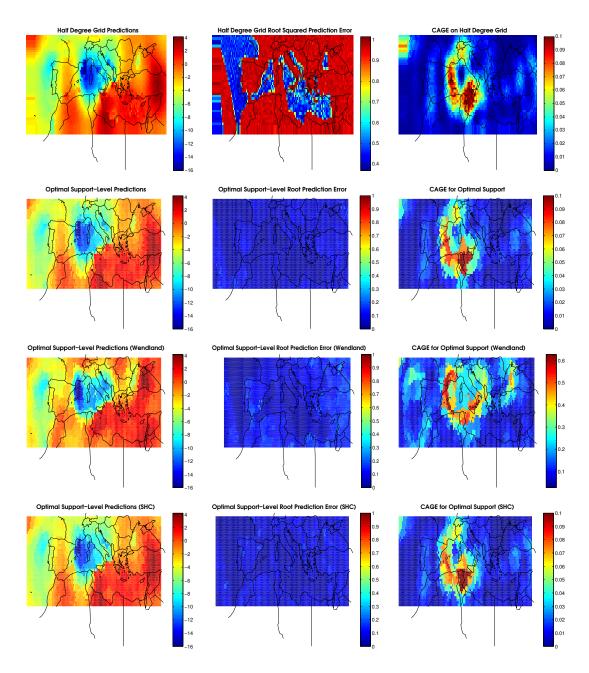


Figure 9: CAGE-based posterior summaries of the predicted north-south wind components based on the analysis and scatterometer observations from 2 February 2005 at 12:00 UTC. The first column displays the posterior mean; the second column displays the posterior standard deviations; and the third column contains the calculated CAGE. In the first row the values (i.e., posterior mean, posterior root prediction error, and CAGE) are all defined on a half degree grid. In the second row values are defined on the optimal spatial support found using *k*-means and the bisquare GBFs. In the third row values are defined on the optimal spatial support found using *k*-means and the Wendland GBFs. In the fourth row values are defined on the optimal spatial support using structural hierarchical clustering (SHC) and bisquare GBFs. Note that the colorbar for the predictions differ from the colorbar used in Figure 6.

that for $\mathbf{s} \in D_s$

$$Y_A(B_h) = \sum_{j=1}^{\infty} \phi_j(\mathbf{s}) \alpha_j, \tag{5.0}$$

where the eigenfunctions $\{\phi_j(\mathbf{s}): j=1,2,...\}$ have domain D_s and satisfies,

$$\int_{D_s} \phi_j(\mathbf{s}) \phi_k(\mathbf{s}) d\mathbf{s} = \delta_{jk}, \tag{5.0}$$

where δ_{jk} is the Kronecker delta function. Additionally, the random variables in the set $\{\alpha_j : j = 1, 2, ...\}$ are uncorrelated with variances $\{\lambda_j : j = 1, 2, ...\}$, and the coefficients $\{\alpha_j : j = 1, 2, ...\}$ can be found by projecting $Y_s(\cdot)$ onto the eigenfunctions. That is,

$$\alpha_j = \int_{D_s} Y_s(\mathbf{s}) \phi_j(\mathbf{s}) d\mathbf{s}, \tag{5.0}$$

for each *j*. Also, these eigenfunctions are solutions to the Fredholm integral equation (e.g., Papoulis (1965)),

$$\int_{D_s} C(\mathbf{s}, \mathbf{u}) \phi_j(\mathbf{s}) d\mathbf{s} = \lambda_j \phi_j(\mathbf{u}); \ \mathbf{u} \in D_s, j = 1, 2, ...,$$
(5.0)

where, from the statement of Proposition 1, $C(\mathbf{s}, \mathbf{u})$ is a valid covariance function for each $\mathbf{s}, \mathbf{u} \in D_s$.

The statement that

$$Y_A(A) = \sum_{i=1}^{\infty} \phi_{A,j}(A) \alpha_j, \tag{5.0}$$

in $L^2(\Omega)$ for $A \subset D_s$, is equivalent to saying that

$$\zeta_n(A) \equiv E\left\{ \left(Y_A(A) - \sum_{i=1}^n \phi_{A,j}(A) \alpha_j \right)^2 \right\}$$
 (5.0)

converges to zero as n goes to infinity. Note that in (6), the expectation is taken with respect to $(\Omega, \mathcal{F}, \mathcal{P})$. Expanding (6) we have,

$$\zeta_n(A) = E\left\{Y_A(A)^2\right\} + E\left\{\left(\sum_{i=1}^n \phi_{A,j}(A)\alpha_j\right)^2\right\} - 2E\left\{Y_A(A)\left(\sum_{i=1}^n \phi_{A,j}(A)\alpha_j\right)\right\}.$$
 (5.0)

The first term of the right-hand side of (6) can be written as

$$E\left\{Y_A(A)^2\right\} = E\left\{\frac{1}{|A|^2} \int_A \int_A Y_s(\mathbf{s}) Y_s(\mathbf{u}) d\mathbf{s} d\mathbf{u}\right\}$$
$$= \frac{1}{|A|^2} \int_A \int_A E(Y_s(\mathbf{s}) Y_s(\mathbf{u})) d\mathbf{s} d\mathbf{u}$$
$$= \frac{1}{|A|^2} \int_A \int_A C(\mathbf{s}, \mathbf{u}) d\mathbf{s} d\mathbf{u}.$$

The second term of the right-hand side of (6) can be written as

$$E\left\{\left(\sum_{i=1}^{n}\phi_{A,j}(A)\alpha_{j}\right)^{2}\right\} = E\left\{\left(\sum_{i=1}^{n}\phi_{A,i}(A)\alpha_{i}\right)\left(\sum_{j=1}^{n}\phi_{A,j}(A)\alpha_{j}\right)\right\}$$

$$= E\left\{\sum_{i=1}^{n}\sum_{j=1}^{n}\phi_{A,i}(A)\phi_{A,j}(A)\alpha_{i}\alpha_{j}\right\}$$

$$= E\left\{\frac{1}{|A|^{2}}\sum_{i=1}^{n}\sum_{j=1}^{n}\int_{A}\int_{A}\phi_{s,i}(\mathbf{s})\phi_{s,j}(\mathbf{u})\alpha_{i}\alpha_{j}d\mathbf{s}d\mathbf{u}\right\}$$

$$= \frac{1}{|A|^{2}}\sum_{i=1}^{n}\sum_{j=1}^{n}\int_{A}\int_{A}\phi_{s,i}(\mathbf{s})\phi_{s,j}(\mathbf{u})E(\alpha_{i}\alpha_{j})d\mathbf{s}d\mathbf{u}$$

$$= \frac{1}{|A|^{2}}\int_{A}\int_{A}\sum_{j=1}^{n}\phi_{s,j}(\mathbf{s})\phi_{s,j}(\mathbf{u})\lambda_{j}d\mathbf{s}d\mathbf{u},$$

since recall from the K-L decomposition that α_i and α_j are uncorrelated with variances λ_i and λ_j , respectively. Finally, the third term of the right-hand side of (6) can be written as

$$E\left\{Y_A(A)\left(\sum_{i=1}^n\phi_{A,j}(A)\alpha_j\right)\right\} = E\left\{\frac{1}{|A|^2}\int_A\int_A\sum_{i=1}^n\phi_{s,i}(\mathbf{s})Y_s(\mathbf{u})\alpha_i d\mathbf{s} d\mathbf{u}\right\},$$

Since α_i is found by projecting Y_s onto the eigenfunctions. From (6) we have that

$$\begin{split} E\left\{Y_A(A)\left(\sum_{i=1}^n\phi_{A,j}(A)\alpha_j\right)\right\} &= E\left\{\frac{1}{|A|^2}\int_A\int_A\sum_{i=1}^n\phi_{s,i}(\mathbf{s})Y_s(\mathbf{u})\int_DY_s(\mathbf{w})\phi_i(\mathbf{w})d\mathbf{w}d\mathbf{s}d\mathbf{u}\right\} \\ &= E\left\{\frac{1}{|A|^2}\int_A\int_A\sum_{i=1}^n\phi_{s,i}(\mathbf{s})\int_DY_s(\mathbf{u})Y_s(\mathbf{w})\phi_i(\mathbf{w})d\mathbf{w}d\mathbf{s}d\mathbf{u}\right\} \\ &= \frac{1}{|A|^2}\int_A\int_A\sum_{i=1}^n\phi_{s,i}(\mathbf{s})\int_DE\left\{Y_s(\mathbf{u})Y_s(\mathbf{w})\right\}\phi_i(\mathbf{w})d\mathbf{w}d\mathbf{s}d\mathbf{u} \\ &= \frac{1}{|A|^2}\int_A\int_A\sum_{i=1}^n\phi_{s,i}(\mathbf{s})\int_DC(\mathbf{u},\mathbf{w})\phi_i(\mathbf{w})d\mathbf{w}d\mathbf{s}d\mathbf{u}. \end{split}$$

From the Fredholm integral equation in (6) we have

$$E\left\{Y_A(A)\left(\sum_{i=1}^n \phi_{A,j}(A)\alpha_j\right)\right\} = \frac{1}{|A|^2} \int_A \int_A \sum_{i=1}^n \phi_{s,i}(\mathbf{s}) \int_D C(\mathbf{u}, \mathbf{w})\phi_i(\mathbf{w}) d\mathbf{w} d\mathbf{s} d\mathbf{u}$$
$$= \frac{1}{|A|^2} \int_A \int_A \sum_{i=1}^n \phi_{s,i}(\mathbf{s}) \phi_{s,i}(\mathbf{u}) \lambda_i d\mathbf{s} d\mathbf{u}.$$

Substituting (6), (6), and (6) into (6) gives

$$\zeta_n(A) = \frac{1}{|A|^2} \int_A \int_A C(\mathbf{s}, \mathbf{u}) - \sum_{i=1}^n \phi_{s,i}(\mathbf{s}) \phi_{s,i}(\mathbf{u}) \lambda_i d\mathbf{s} d\mathbf{u}.$$
 (5.-7)

Upon taking the limit as n goes to infinity on both sides of (6), it follows from Mercer's theorem (Mercer, 1909) that

$$\lim_{n \to \infty} \zeta_n(A) = 0, \tag{5.-7}$$

for each $A \subset D_s$; note that Mercer's theorem shows *uniform* convergence at the point-level, allowing one to pass the limit through the integral. This proves the result.

The proof of 1.ii follows a similar logic to (6). That is, note that

$$\sum_{i=1}^{n} \phi_{A,i}(A)\phi_{A,i}(B)\lambda_{i} - \operatorname{cov}\left\{Y_{A}(A), Y_{A}(B)\right\}$$

$$= \frac{1}{|A||B|} \int_{A} \int_{B} C(\mathbf{s}, \mathbf{u}) - \sum_{i=1}^{n} \phi_{s,i}(\mathbf{s})\phi_{s,i}(\mathbf{u})\lambda_{i}d\mathbf{s}d\mathbf{u}.$$

Upon taking the limit as n goes to infinity on both sides of (6), it follows from Mercer's theorem (Mercer, 1909) that Proposition 1.ii holds.

Proof of Proposition 2: First, we prove the following statement: If $\phi_k(\mathbf{x}_j) = \phi_{A,k}(A_j)$ for $j = 1,...,n_A$ and for any positive integer k, then $\mathbf{Y}_s^{(A)} = \mathbf{Y}_A$ almost surely. Then the continuous mapping theorem is applied to get $f(\mathbf{Y}_s^{(A)}) = f(\mathbf{Y}_A)$ almost surely.

We proceed using a proof by contradiction. Assume that $\mathbf{Y}_s^{(A)}$ is not almost surely equal to \mathbf{Y}_A . Then, for at least one \mathbf{x}_i and A_i , there exists a $\gamma > 0$ such that

$$P(|Y_s(\mathbf{x}_i) - Y_A(A_i)| \ge \gamma) > 0.$$
 (5.-8)

However, we have from Chebychev's inequality

$$P(|Y_s(\mathbf{x}_i) - Y_A(A_i)| \ge \gamma) \le \frac{E\left[\{Y_s(\mathbf{x}_i) - Y_A(A_i)\}^2\right]}{\gamma^2}.$$
 (5.-8)

Assume that $\phi_k(\mathbf{x}_j) = \phi_{A,k}(A_j)$ for $j = 1,...,n_A$ and every positive integer k. Then, upon adding and subtracting $\sum_{k=1}^{n} \phi_k(\mathbf{x}_i)$ within (6) we have:

$$P(|Y_{s}(\mathbf{x}_{i}) - Y_{A}(A_{i})| \geq \gamma) \leq \frac{1}{\gamma^{2}} E\left\{Y_{s}(\mathbf{x}_{i}) - \sum_{k=1}^{n} \phi_{k}(\mathbf{x}_{i})\alpha_{k} + \sum_{k=1}^{n} \phi_{A,k}(A_{i})\alpha_{k} - Y_{A}(A_{i})\right\}^{2}$$

$$= \frac{1}{\gamma^{2}} E\left\{Y_{s}(\mathbf{x}_{i}) - \sum_{k=1}^{n} \phi_{k}(\mathbf{x}_{i})\alpha_{k}\right\}^{2} + \frac{1}{\gamma^{2}} E\left\{\sum_{k=1}^{n} \phi_{A,k}(A_{i})\alpha_{k} - Y_{A}(A_{i})\right\}^{2}$$

$$+ \frac{2}{\gamma^{2}} E\left\{Y_{s}(\mathbf{x}_{i}) - \sum_{k=1}^{n} \phi_{k}(\mathbf{x}_{i})\alpha_{k}\right\} \left\{\sum_{k=1}^{n} \phi_{A,k}(A_{i})\alpha_{k} - Y_{A}(A_{i})\right\}\right\}.$$

It follows from Karhunen (1947) that the first term on the right-hand-side of (6) converges to zero. Likewise, from Proposition 1 the second term on the right-hand-side of (6) converges to zero as n goes to infinity. Note that since $P(|Y_s(\mathbf{x}_i) - Y_A(A_i)| \ge \gamma)$ does not depend on n we have that,

$$P(|Y_s(\mathbf{x}_i) - Y_A(A_i)| \ge \gamma)$$

$$\leq \lim_{n \to \infty} \frac{2}{\gamma^2} E\left[\left\{Y_s(\mathbf{x}_i) - \sum_{k=1}^n \phi_k(\mathbf{x}_i)\alpha_k\right\} \left\{\sum_{k=1}^n \phi_{A,k}(A_i)\alpha_k - Y_A(A_i)\right\}\right].$$

Thus, we are left to find the expression of the limit in (6). Note,

$$E\left[\left\{Y_{s}(\mathbf{x}_{i})-\sum_{k=1}^{n}\phi_{k}(\mathbf{x}_{i})\alpha_{k}\right\}\left\{\sum_{k=1}^{n}\phi_{A,k}(A_{i})\alpha_{k}-Y_{A}(A_{i})\right\}\right]$$

$$=\frac{1}{|A|}E\left\{\int_{A_{i}}\sum_{k=1}^{n}Y_{s}(\mathbf{x}_{i})\alpha_{k}\phi_{k}(\mathbf{s})d\mathbf{s}\right\}$$

$$-\frac{1}{|A|}E\left\{\int_{A_{i}}Y_{s}(\mathbf{x}_{i})Y_{s}(\mathbf{u})d\mathbf{u}\right\}$$

$$-\frac{1}{|A|}E\left\{\sum_{k=1}^{n}\sum_{j=1}^{n}\int_{A_{i}}\phi_{k}(\mathbf{x}_{i})\phi_{j}(\mathbf{s})\alpha_{j}\alpha_{k}d\mathbf{s}\right\}$$

$$+\frac{1}{|A|}E\left\{\sum_{k=1}^{n}\int_{A_{i}}\phi_{k}(\mathbf{x}_{i})\alpha_{k}Y_{s}(\mathbf{s})d\mathbf{s}\right\}.$$

For the term in (6) notice from (6) and (6) we have that

$$\frac{1}{|A|}E\left\{\int_{A_i}\sum_{k=1}^n Y_s(\mathbf{x}_i)\alpha_k\phi_k(\mathbf{s})d\mathbf{s}\right\} = \frac{1}{|A|}E\left\{\int_{A_i}\sum_{k=1}^n Y_s(\mathbf{x}_i)\int_{D_s} Y_s(\mathbf{u})\phi_k(\mathbf{u})d\mathbf{u} \ \phi_k(\mathbf{s})d\mathbf{s}\right\}
= \frac{1}{|A|}\int_{A_i}\sum_{k=1}^n \int_{D_s} E\left\{Y_s(\mathbf{x}_i)Y_s(\mathbf{u})\right\}\phi_j(\mathbf{u})d\mathbf{u} \ \phi_k(\mathbf{s})d\mathbf{s}
= \frac{1}{|A|}\int_{A_i}\sum_{k=1}^n \int_{D_s} C(\mathbf{x}_i,\mathbf{u})\phi_k(\mathbf{u})d\mathbf{u} \ \phi_k(\mathbf{s})d\mathbf{s}
= \frac{1}{|A|}\int_{A_i}\sum_{k=1}^n \phi_k(\mathbf{s})\phi_k(\mathbf{x}_i)\lambda_k d\mathbf{s}.$$

The terms in (6) and (6) can be written as

$$-\frac{1}{|A|}E\left\{\int_{A_i}Y_s(\mathbf{x}_i)Y_s(\mathbf{u})d\mathbf{u}\right\} = -\frac{1}{|A|}E\left\{\int_{A_i}C(\mathbf{x}_i,\mathbf{u})d\mathbf{u}\right\},$$

$$-\frac{1}{|A|}E\left\{\sum_{k=1}^n\sum_{j=1}^n\int_{A_i}\phi_k(\mathbf{x}_i)\phi_j(\mathbf{s})\alpha_j\alpha_k d\mathbf{s}\right\} = -\frac{1}{|A|}\int_{A_i}\sum_{k=1}^n\phi_k(\mathbf{s})\phi_k(\mathbf{x}_i)\lambda_k d\mathbf{s}.$$

For the term in (6) notice from (6) and (6) we have that

$$\frac{1}{|A|}E\left\{\sum_{k=1}^n\int_{A_i}\phi_k(\mathbf{x}_i)\alpha_kY_s(\mathbf{s})d\mathbf{s}\right\}=\frac{1}{|A|}\int_{A_i}\sum_{k=1}^n\phi_k(\mathbf{s})\phi_k(\mathbf{x}_i)\lambda_kd\mathbf{s}.$$

Thus, it follows that

$$E\left[\left\{Y_{s}(\mathbf{x}_{i}) - \sum_{k=1}^{n} \phi_{k}(\mathbf{x}_{i}) \alpha_{k}\right\} \left\{\sum_{k=1}^{n} \phi_{A,k}(A_{i}) \alpha_{k} - Y_{A}(A_{i})\right\}\right]$$

$$= \frac{2}{|A|} \int_{A_{i}} \sum_{k=1}^{n} \phi_{k}(\mathbf{x}_{i}) \phi_{k}(\mathbf{s}) \lambda_{k} - C(\mathbf{x}_{i}, \mathbf{s}) d\mathbf{s},$$

which, again by Mercer's theorem, converges to 0 as n goes to infinity. Thus, from (6) we have that

$$P(|Y_s(\mathbf{x}_i) - Y_A(A_i)| \ge \gamma) = 0,$$

which contradicts (6). One can prove forward implication of Proposition 2.ii in a similar manner.

To prove the reverse statement of Proposition 2.*i*, suppose that $f(\mathbf{Y}_s^{(A)}) = f(\mathbf{Y}_A)$ almost surely for any measurable real-valued function f. Thus, the functions $f_i(\mathbf{b}) = b_i$ for $i = 1, ..., n_A$ and $\mathbf{b} = (b_i : i = 1, ..., n_A)' \in \mathbb{R}^{n_A}$, imply that

$$Y_s(\mathbf{x}_i) = Y_A(A_i), \tag{5.-12}$$

almost surely. Multiplying both sides by α_i we have

$$Y_s(\mathbf{x}_i)\alpha_j = Y_A(A_i)\alpha_j$$

almost surely. Substituting (6) into the equation above gives,

$$Y_s(\mathbf{x}_i) \int_{D_s} Y_s(\mathbf{s}) \phi_j(\mathbf{s}) d\mathbf{s} = \frac{1}{|A_i|} \int_{A_i} \int_{D_s} Y_s(\mathbf{u}) Y_s(\mathbf{s}) \phi_j(\mathbf{s}) d\mathbf{s} d\mathbf{u}.$$

Taking the expectation on both sides we have

$$\int_{D_s} C(\mathbf{x}_i, \mathbf{s}) \phi_j(\mathbf{s}) d\mathbf{s} = \frac{1}{|A_i|} \int_{A_i} \int_{D_s} C(\mathbf{u}, \mathbf{s}) \phi_j(\mathbf{s}) d\mathbf{s} d\mathbf{u},$$

and then from (6) we have

$$\phi_j(\mathbf{x}_i)\lambda_j = \frac{1}{|A_i|} \int_{A_i} \phi_j(\mathbf{u}) d\mathbf{u} \lambda_j.$$

Dividing by λ_j

$$\phi_j(\mathbf{x}_i) = \frac{1}{|A_i|} \int_{A_i} \phi_j(\mathbf{u}) d\mathbf{u}.$$

This proves the result. One can prove the reverse statement of Proposition 2.ii in a similar manner. By the condition in Proposition 2.iii, we have that for a given ϕ_k ,

$$\phi_k(B_j) = \frac{1}{|B_j|} \int_{B_j} \phi_k(\mathbf{s}) d\mathbf{s} = \frac{1}{|B_j|} \int_{B_j} \phi_k(A_j) d\mathbf{s}$$
$$= \phi_k(A_j) \frac{1}{|B_j|} \int_{B_j} 1 d\mathbf{s} = \phi_k(A_j).$$

It follows from Proposition 2.ii that Proposition 2.iii holds.

Proof of Proposition 3: We now prove the equalities listed in Equations (8), (9), and (10) of Proposition 3. We start with Equation (8). Notice that for a given $\mathbf{s} \in D_s$, $A \in D_A$, $\{\phi_k(\cdot)\}$, and

 $\{\lambda_k\},$

$$\begin{split} &E\left[\left\{Y_{s}(\mathbf{s})-Y_{A}(A)\right\}^{2}|\{\phi_{k}\},\{\lambda_{k}\}\right] \\ &=E\left[\left\{Y_{s}(\mathbf{s})-\sum_{k=1}^{n}\phi_{k}(\mathbf{s})\alpha_{k}\right\}^{2}|\{\phi_{k}\},\{\lambda_{k}\}\right] \\ &+E\left[\left\{\sum_{k=1}^{n}\phi_{k}(\mathbf{s})\alpha_{k}-\sum_{k=1}^{n}\phi_{A,k}(A)\alpha_{k}\right\}^{2}|\{\phi_{k}\},\{\lambda_{k}\}\right] \\ &+\frac{1}{|A|}E\left[\left\{\sum_{k=1}^{n}\phi_{A,k}(A)\alpha_{k}-Y_{A}(A)\right\}^{2}|\{\phi_{k}\},\{\lambda_{k}\}\right] \\ &+2E\left[\left\{Y_{s}(\mathbf{s})-\sum_{k=1}^{n}\phi_{k}(\mathbf{s})\alpha_{k}\right\}\left\{\sum_{k=1}^{n}\phi_{k}(\mathbf{s})\alpha_{k}-\sum_{k=1}^{n}\phi_{A,k}(A)\alpha_{k}\right\}|\{\phi_{k}\},\{\lambda_{k}\}\right] \\ &+2E\left[\left\{Y_{s}(\mathbf{s})-\sum_{k=1}^{n}\phi_{k}(\mathbf{s})\alpha_{k}\right\}\left\{\sum_{k=1}^{n}\phi_{A,k}(A)\alpha_{k}-Y_{A}(A)\right\}|\{\phi_{k}\},\{\lambda_{k}\}\right] \\ &+2E\left[\left\{\sum_{k=1}^{n}\phi_{k}(\mathbf{s})\alpha_{k}-\sum_{k=1}^{n}\phi_{A,k}(A)\alpha_{k}\right\}\left\{\sum_{k=1}^{n}\phi_{A,k}(A)\alpha_{k}-Y_{A}(A)\right\}|\{\phi_{k}\},\{\lambda_{k}\}\right]. \end{split}$$

Through an application of Mercer's theorem we have that the sum of the cross-product terms in (6), (6), and (6) converge to zero as n goes to infinity. Similarly, it follows from Karhunen (1947) that (6) goes to zero as n goes to infinity, and from Proposition 1 that (6) goes to zero as n goes to infinity. Thus,

$$E\left[\left\{Y_{s}(\mathbf{s}) - Y_{A}(A)\right\}^{2} | \{\phi_{k}\}, \{\lambda_{k}\}\right] = \sum_{j=1}^{\infty} \left(\phi_{j}(\mathbf{s}) - \phi_{A,j}(A)\right)^{2} \lambda_{j}, \tag{5.-13}$$

Then, upon taking the expectation with respect to $\{\phi_k\}, \{\lambda_k\} | \mathbf{Z}$ we have the desired result.

To prove Equation (9) recall from Mercer's theorem and Proposition 1.ii that,

$$\operatorname{var}\left\{Y_{s}(\mathbf{s})\right\} = \sum_{k=1}^{\infty} \phi_{k}(\mathbf{s})^{2} \lambda_{j}$$
$$\operatorname{var}\left\{Y_{A}(A)\right\} = \sum_{k=1}^{\infty} \phi_{A,k}(A)^{2} \lambda_{j}.$$

Expanding (9) and substituting (6) we have

$$\begin{aligned} \operatorname{CAGE}(A) &= E\left[\int_{A} \frac{\sum_{j=1}^{\infty} \left\{\phi_{j}(\mathbf{s}) - \phi_{A,j}(A)\right\}^{2} \lambda_{j}}{|A|} d\mathbf{s} | \mathbf{Z}\right] \\ &= E\left\{\int_{A} \frac{\sum_{j=1}^{\infty} \phi_{j}(\mathbf{s})^{2} \lambda_{j} - 2\sum_{j=1}^{\infty} \phi_{j}(\mathbf{s}) \phi_{A,j}(A) \lambda_{j}}{|A|} d\mathbf{s} + \sum_{k=1}^{\infty} \phi_{A,k}(A)^{2} \lambda_{j} | \mathbf{Z}\right\} \\ &= E\left\{\int_{A} \frac{\sum_{j=1}^{\infty} \phi_{j}(\mathbf{s})^{2} \lambda_{j}}{|A|} d\mathbf{s} - 2\sum_{k=1}^{\infty} \phi_{A,k}(A)^{2} \lambda_{j} + \sum_{k=1}^{\infty} \phi_{A,k}(A)^{2} \lambda_{j} | \mathbf{Z}\right\} \\ &= E\left\{\int_{A} \frac{\sum_{j=1}^{\infty} \phi_{j}(\mathbf{s})^{2} \lambda_{j}}{|A|} d\mathbf{s} - \sum_{k=1}^{\infty} \phi_{A,k}(A)^{2} \lambda_{j} | \mathbf{Z}\right\} \\ &= E\left[\int_{A} \frac{\operatorname{var}\left\{Y_{s}(\mathbf{s})\right\}}{|A|} d\mathbf{s} - \operatorname{var}\left\{Y_{A}(A)\right\} | \mathbf{Z}\right]; \ A \subset D_{s}. \end{aligned}$$

This proves (9).

We now prove Equation (10). From (8) we have for any $A \subset D_s$,

$$CAGE(A) = E \left[\int_{A} \frac{\left\{ Y_{s}(\mathbf{s}) - Y_{A}(A) \right\}^{2}}{|A|} d\mathbf{s} | \mathbf{Z} \right].$$
 (5.-19)

Adding and subtracting \widehat{Y}_A ,

$$CAGE(A) = E \left[\int_{A} \frac{\left\{ Y_{s}(\mathbf{s}) - \widehat{Y}_{A}(A) + \widehat{Y}_{A}(A) - Y_{A}(A) \right\}^{2}}{|A|} d\mathbf{s} | \mathbf{Z} \right]$$

$$= E \left[\int_{A} \frac{\left\{ Y_{s}(\mathbf{s}) - \widehat{Y}_{A}(A) \right\}^{2}}{|A|} d\mathbf{s} | \mathbf{Z} \right] + E \left[\int_{A} \frac{\left\{ \widehat{Y}_{A}(A) - Y_{A}(A) \right\}^{2}}{|A|} d\mathbf{s} | \mathbf{Z} \right]$$

$$+ 2E \left[\int_{A} \frac{\left\{ Y_{s}(\mathbf{s}) - \widehat{Y}_{A}(A) \right\} \left\{ \widehat{Y}_{A}(A) - Y_{A}(A) \right\}}{|A|} d\mathbf{s} | \mathbf{Z} \right]$$

$$= E \left[\int_{A} \frac{\left\{ Y_{s}(\mathbf{s}) - \widehat{Y}_{A}(A) \right\}^{2}}{|A|} d\mathbf{s} | \mathbf{Z} \right] + E \left[\left\{ \widehat{Y}_{A}(A) - Y_{A}(A) \right\}^{2} d\mathbf{s} | \mathbf{Z} \right]$$

$$- 2E \left[\left\{ \widehat{Y}_{A}(A) - Y_{A}(A) \right\}^{2} | \mathbf{Z} \right]$$

$$= E \left[\int_{A} \frac{\left\{ Y_{s}(\mathbf{s}) - \widehat{Y}_{A}(A) \right\}^{2}}{|A|} d\mathbf{s} | \mathbf{Z} \right] - E \left[\left\{ \widehat{Y}_{A}(A) - Y_{A}(A) \right\}^{2} | \mathbf{Z} \right].$$

This proves Equation (11).

Proof of Proposition 4: The fine-scale variation term δ in (16) can be written as

$$\delta(\mathbf{u};\boldsymbol{\xi}) = \mathbf{h}(\mathbf{u})'\boldsymbol{\xi}; \ \mathbf{u} \in D_s \cup D_A,$$

where

$$\mathbf{h}(\mathbf{u}) \equiv \begin{cases} (I(\mathbf{u} \in B) : B \in D_B)' & \text{if } \mathbf{u} \in D_s \\ \left(\frac{|\mathbf{u} \cap B|}{|B|} : B \in D_B\right)' & \text{if } \mathbf{u} \in D_A, \end{cases}$$

and $I(\cdot)$ is the indicator function. Then, from Equation (15) we have that for a given ϕ_s and α ,

$$\mathbf{Y}_{s}^{(C)} = \mu \mathbf{1}_{n_{C}} + \mathbf{\Phi}_{s}^{(C)} \boldsymbol{\alpha} + \mathbf{H}_{s}^{(C)} \boldsymbol{\xi}$$
$$\mathbf{Y}_{C} = \mu \mathbf{1}_{n_{C}} + \mathbf{\Phi}_{C} \boldsymbol{\alpha} + \mathbf{H}_{C} \boldsymbol{\xi},$$

where the $n_C \times r$ matrices $\mathbf{\Phi}_s^{(C)} \equiv (\boldsymbol{\phi}_s(\mathbf{x}_j)': j=1,...,n_C)'$ and $\mathbf{\Phi}_C \equiv (\boldsymbol{\phi}(C_j; \; \boldsymbol{\phi}_s)': j=1,...,n_C)'$, and the $n_C \times n_B$ matrices $\mathbf{H}_s^{(C)} \equiv (\mathbf{h}(\mathbf{x}_j)': j=1,...,n_C)'$ and $\mathbf{H}_C \equiv (\mathbf{h}(C_j)': j=1,...,n_C)'$. Notice that for the values of $\{\mathbf{x}_j\}$ and $\{\mathbf{C}_j\}$ given in the statement of Proposition 5, we have $\mathbf{H}_s^{(C)} = \mathbf{H}_C = \mathbf{I}_{n_C}$ (the $n_C \times n_C$ identity matrix), and thus,

$$\mathbf{Y}_s^{(C)} = \mu \mathbf{1}_{n_C} + \mathbf{\Phi}_s^{(C)} \boldsymbol{\alpha} + \boldsymbol{\xi}$$

 $\mathbf{Y}_C = \mu \mathbf{1}_{n_C} + \mathbf{\Phi}_C \boldsymbol{\alpha} + \boldsymbol{\xi}$.

The condition for the forward implication of Proposition 4.i is that $\Phi_s^{(C)} = \Phi_C$; thus, from (6) we have that

$$\mathbf{Y}_{s}^{(C)} = \mu \mathbf{1}_{n_{C}} + \mathbf{\Phi}_{s}^{(C)} \boldsymbol{\alpha} + \boldsymbol{\xi} = \mathbf{Y}_{C}.$$
 (5.-27)

When applying any real-valued measurable f to both sides of (6), we obtain that $f(\mathbf{Y}_s^{(C)}) = f(\mathbf{Y}_C)$ almost surely. One can prove forward implication of Proposition 4.ii in a similar manner.

To prove the reverse statement of Proposition 4.*i*, suppose that $f(\mathbf{Y}_s^{(C)}) = f(\mathbf{Y}_C)$ almost surely for any real-valued function f. Thus, the functions $f_i(\mathbf{b}) = b_i$ for $i = 1, ..., n_A$ and $\mathbf{b} = (b_j : j = 1, ..., n_A)' \in \mathbb{R}^{n_A}$, imply that

$$\mathbf{Y}_{s}^{(C)} = \mathbf{Y}_{C},\tag{5.-27}$$

almost surely. From (6) and (6) we see that

$$\mathbf{\Phi}_s^{(C)} \boldsymbol{\alpha} = \mathbf{\Phi}_C \boldsymbol{\alpha}, \tag{5.-27}$$

almost surely. Multiply both sides of (6) by α' , and take the expectation with respect to $Y|\phi_s, \Lambda$ to

obtain

$$\mathbf{\Phi}_s^{(C)} \mathbf{\Lambda} = \mathbf{\Phi}_C \mathbf{\Lambda}. \tag{5.-27}$$

Provided that $\lambda_j > 0$ for all j, we can take the inverse of Λ on both sides of (6) so that $\Phi_s^{(C)} = \Phi_C$, which is the desired result. One can prove the reverse statement of Proposition 4.ii in a similar manner.

By the condition in Proposition 4.iii, we have that for a given ϕ_s and α ,

$$\boldsymbol{\phi}_{s}(\mathbf{x}_{j})'\boldsymbol{\alpha} = \boldsymbol{\phi}(C_{j}; \ \boldsymbol{\phi}_{s})'\boldsymbol{\alpha}; \ j = 1,...,n_{C}.$$
 (5.-27)

Integrating (6) with respect to \mathbf{x}_i we have

$$\phi(B_j; \phi_s)'\alpha = \phi(C_j; \phi_s)'\alpha; j = 1,...,n_C.$$

Since $\lambda_j > 0$ for all j, this leads to the condition for the forward implication of Proposition 4.ii, and thus, it follows that Proposition 4.iii holds.

Proof of Proposition 5: From Equation (1) we see that for $Y(\cdot; \phi_s^{OC})$ to be a multiscale truncated K-L expansion, we only need to show that $Y_s(\cdot; \phi_s^{OC})$ is a truncated K-L expansion. Many of the following equations can be found in Obled and Creutin (1986).

To show that $Y_s(\cdot; \phi_s^{OC})$ is a truncated K-L expansion, we need to establish three items: the eigenvalues must be nonnegative with at least one eigenvalue strictly positive; the Fredholm integral equations must hold; and the eigenvectors must be orthonormal. Notice that

$$\operatorname{cov}\left[Y_{s}\left\{\mathbf{s};\boldsymbol{\phi}_{s}^{\operatorname{OC}}(\cdot;\ \mathbf{F})\right\},Y_{s}\left\{\mathbf{u};\boldsymbol{\phi}_{s}^{\operatorname{OC}}(\cdot;\ \mathbf{F})\right\}\right]$$

$$=E\left[\left\{\sum_{k=1}^{r}\sum_{i=1}^{r}\psi_{i}(\mathbf{s})F_{ik}\alpha_{k}\right\}\left\{\sum_{q=1}^{r}\sum_{p=1}^{r}\psi_{q}(\mathbf{u})F_{qp}\alpha_{p}\right\}\right]$$

$$=\sum_{k=1}^{r}\lambda_{k}\left\{\sum_{i=1}^{r}\psi_{i}(\mathbf{s})F_{ik}\right\}\left\{\sum_{q=1}^{r}\psi_{q}(\mathbf{u})F_{qk}\right\}.$$

Substituting (6) into the Fredholm integral equation we have, for k = 1, ..., r,

$$\int_{D_s} \left\{ \sum_{i=1}^r \sum_{k=1}^r \sum_{q=1}^r F_{qk} \lambda_k F_{ik} \psi_i(\mathbf{s}) \psi_q(\mathbf{u}) \right\} \left\{ \sum_{m=1}^r \psi_m(\mathbf{s}) F_{mp} \right\} d\mathbf{s} = \omega_p \left\{ \sum_{q=1}^r \psi_q(\mathbf{u}) F_{qp} \right\}, \quad (5.-29)$$

where $\{\omega_k\}$ represents the eigenvalues of $Y_s(\cdot; \phi_s^{OC})$. Distributing the sums and integral through (6), we obtain

$$\sum_{q=1}^{r} \psi_q(\mathbf{u}) \left\{ \sum_{i=1}^{r} \sum_{k=1}^{r} \sum_{m=1}^{r} F_{qk} \lambda_k F_{ik} \right\} \int_{D_s} \psi_i(\mathbf{s}) \psi_m(\mathbf{s}) F_{mp} d\mathbf{s} = \omega_p \left\{ \sum_{q=1}^{r} \psi_q(\mathbf{u}) F_{qp} \right\}. \tag{5.-29}$$

Matching terms in (6), we have

$$\sum_{i=1}^{r} \sum_{k=1}^{r} \sum_{m=1}^{r} F_{qk} \lambda_k F_{ik} W_{im} F_{mp} = \omega_p F_{qp}; q = 1, ..., r.$$
 (5.-29)

In matrix form, (6) becomes,

$$\mathbf{F}\mathbf{\Lambda}\mathbf{F}'\mathbf{W}\mathbf{F} = \mathbf{F}\mathbf{\Omega},\tag{5.-29}$$

where $\Lambda \equiv \text{diag}(\lambda_k)$ and $\Omega \equiv \text{diag}(\omega_k)$. The assumption that $\mathbf{F}'\mathbf{WF} = \mathbf{I}$ and (6) implies that the Fredholm-integral equation holds provided that

$$\mathbf{F}\mathbf{\Lambda} = \mathbf{F}\mathbf{\Omega}.\tag{5.-29}$$

Since, **F** is invertible we have that (6) verifies that the eigenvalues of $Y_s(\cdot; \phi_s^{OC})$ are nonnegative with $\mathbf{\Lambda} = \mathbf{\Omega}$ (and at least one eigenvalue is strictly positive), and that the Fredholm integral equations for $Y_s(\cdot; \phi_s^{OC})$ hold. The orthogonality of ϕ_s^{OC} holds by assumption since

$$\int \phi_i^{\text{OC}}(\mathbf{s}; \ \mathbf{F}) \phi_j^{\text{OC}}(\mathbf{s}; \ \mathbf{F}) d\mathbf{s} = \sum_{k=1}^r \sum_{p=1}^r F_{ki} F_{pj} \int \psi_k(\mathbf{s}) \psi_p(\mathbf{s}) d\mathbf{s}$$
$$= \sum_{k=1}^r \sum_{p=1}^r F_{ki} W_{kp} F_{pj} = I(i=j),$$

which results in the relation,

$$\mathbf{F}'\mathbf{WF} = \mathbf{I}.$$

This completes the proof.

Proof of Proposition 6: Let $\mathbf{W} = \mathbf{P}_W \mathbf{\Lambda}_W \mathbf{P}_W'$ be the spectral decomposition of \mathbf{W} . It follows that the Cholesky square root of \mathbf{W} and \mathbf{W}^{-1} is given by $\mathbf{P}_W \mathbf{\Lambda}_W^{1/2}$ and $\mathbf{P}_W \mathbf{\Lambda}_W^{-1/2}$, respectively. It follows immediately that $\mathbf{G}'(\mathbf{P}_W \mathbf{\Lambda}_W^{-1/2})' \mathbf{W} \mathbf{P}_W \mathbf{\Lambda}_W^{-1/2} \mathbf{G} = \mathbf{I}$.

V.ii Additional Results

In the main-text, three results were discussed, but not formally stated. Thus, in this section we state and prove these results. In particular, at the end of Remark 6, we mentioned that the CAGE identities in Proposition 3 also hold for DCAGE; this extension of Proposition 3 is referred to as *Result 1*. Also, at the end of Section 3.2 we mention that a version of Proposition 3 exists for CAGE in (17) and DCAGE in (18); these two extensions are referred to as *Result 2* and *Result 3*, respectively.

Result 1: Assume that the conditions of Proposition 1 hold. Assume that the stochastic process $Z: D_s \times \Omega \to \mathbb{R}$ is generated based on any generic probability space $(\Omega, \mathcal{F}, \mathcal{P})$ such that the conditional probability density function of $Y(\mathbf{u})|\mathbf{Z}$ exists for each $\mathbf{u} \in D_s \cup D_A$, where Z is defined in Remark 2. Then, DCAGE in (7) has the following alternative expressions:

$$DCAGE(C) = E\left[\sum_{h \in H} \frac{\left\{Y_A(B_h) - Y_A(C)\right\}^2}{|C|} | \mathbf{Z}\right]$$

$$DCAGE(C) = E\left[\sum_{h \in H} \frac{\operatorname{var}\left\{Y_A(B_h)\right\}}{|C|} - \operatorname{var}\left\{Y_A(C)\right\} | \mathbf{Z}\right]$$

$$DCAGE(C) = E\left[\sum_{h \in H} \frac{\left\{Y_A(B_h) - \widehat{Y}_A(C)\right\}^2}{|C|} | \mathbf{Z}\right] - E\left[\left\{\widehat{Y}_A(C) - Y_A(C)\right\}^2 | \mathbf{Z}\right],$$

where $C = \bigcup_{h \in H} B_h$, $H \subset \{1,...,n_B\}$, and $B_h \in D_B$ for each $h \in H$.

Proof of Result 1: We now prove the equalities listed in Equations (6), (6), and (6) of Proposition 3. We start with Equation (6). Notice that for a given $B_h \in D_B$, $C = \bigcup_{h \in H} B_h$, $H \subset \{1, ..., n_B\}$, $\{\phi_k(\cdot)\}$, and $\{\lambda_k\}$,

$$\begin{split} &E\left[\left\{Y_{A}(B_{h})-Y_{A}(C)\right\}^{2}|\{\phi_{k}\},\{\lambda_{k}\}\right]\\ &=E\left[\left\{Y_{A}(B_{h})-\sum_{k=1}^{n}\phi_{A,k}(B_{h})\alpha_{k}\right\}^{2}|\{\phi_{k}\},\{\lambda_{k}\}\right]\\ &+E\left[\left\{\sum_{k=1}^{n}\phi_{A,k}(B_{h})\alpha_{k}-\sum_{k=1}^{n}\phi_{A,k}(C)\alpha_{k}\right\}^{2}|\{\phi_{k}\},\{\lambda_{k}\}\right]\\ &+\frac{1}{|C|}E\left[\left\{\sum_{k=1}^{n}\phi_{A,k}(C)\alpha_{k}-Y_{A}(C)\right\}^{2}|\{\phi_{k}\},\{\lambda_{k}\}\right]\\ &+2E\left[\left\{Y_{A}(B_{h})-\sum_{k=1}^{n}\phi_{A,k}(B_{h})\alpha_{k}\right\}\left\{\sum_{k=1}^{n}\phi_{A,k}(B_{h})\alpha_{k}-\sum_{k=1}^{n}\phi_{A,k}(C)\alpha_{k}\right\}|\{\phi_{k}\},\{\lambda_{k}\}\right]\\ &+2E\left[\left\{Y_{A}(B_{h})-\sum_{k=1}^{n}\phi_{A,k}(B_{h})\alpha_{k}\right\}\left\{\sum_{k=1}^{n}\phi_{A,k}(C)\alpha_{k}-Y_{A}(C)\right\}|\{\phi_{k}\},\{\lambda_{k}\}\right]\\ &+2E\left[\left\{\sum_{k=1}^{n}\phi_{A,k}(B_{h})\alpha_{k}-\sum_{k=1}^{n}\phi_{A,k}(C)\alpha_{k}\right\}\left\{\sum_{k=1}^{n}\phi_{A,k}(C)\alpha_{k}-Y_{A}(C)\right\}|\{\phi_{k}\},\{\lambda_{k}\}\right]. \end{split}$$

Through an application of Mercer's theorem we have that the sum of the cross-product terms in (6), (6), and (6) converge to zero as n goes to infinity. Similarly, it follows from Proposition 1 that (6) and (6) go to zero as n goes to infinity. Thus,

$$E\left[\left\{Y_{A}(B_{h})-Y_{A}(C)\right\}^{2}|\left\{\phi_{k}\right\},\left\{\lambda_{k}\right\}\right]=\sum_{j=1}^{\infty}\left(\phi_{A,j}(B_{h})-\phi_{A,j}(C)\right)^{2}\lambda_{j},\tag{5.-33}$$

Then, upon taking the expectation with respect to $\{\phi_k\}, \{\lambda_k\} | \mathbf{Z}$ we have the desired result.

To prove Equation (6) recall from Proposition 1.ii that,

$$\operatorname{var}\left\{Y_{A}(B_{h})\right\} = \sum_{k=1}^{\infty} \phi_{A,k}(B_{h})^{2} \lambda_{j}$$
$$\operatorname{var}\left\{Y_{A}(C)\right\} = \sum_{k=1}^{\infty} \phi_{A,k}(C)^{2} \lambda_{j}.$$

Expanding (6) and substituting (6) we have

$$CAGE(C) = E\left[\sum_{h \in H} \frac{\sum_{j=1}^{\infty} \left\{\phi_{A,j}(B_{h}) - \phi_{A,j}(C)\right\}^{2} \lambda_{j}}{|C|} | \mathbf{Z}\right]$$

$$= E\left\{\sum_{h \in H} \frac{\sum_{j=1}^{\infty} \phi_{A,j}(B_{h})^{2} \lambda_{j} - 2\sum_{j=1}^{\infty} \phi_{A,j}(B_{h}) \phi_{A,j}(C) \lambda_{j}}{|C|} + \sum_{k=1}^{\infty} \phi_{A,k}(\mathbf{s})^{2} \lambda_{j} | \mathbf{Z}\right\}$$

$$= E\left\{\sum_{h \in H} \frac{\sum_{j=1}^{\infty} \phi_{A,j}(B_{h})^{2} \lambda_{j}}{|C|} - 2\sum_{k=1}^{\infty} \phi_{A,k}(C)^{2} \lambda_{j} + \sum_{k=1}^{\infty} \phi_{A,k}(C)^{2} \lambda_{j} | \mathbf{Z}\right\}$$

$$= E\left\{\sum_{h \in H} \frac{\sum_{j=1}^{\infty} \phi_{A,j}(B_{h})^{2} \lambda_{j}}{|C|} - \sum_{k=1}^{\infty} \phi_{A,k}(C)^{2} \lambda_{j} | \mathbf{Z}\right\}$$

$$= E\left\{\sum_{h \in H} \frac{\operatorname{var}\{Y_{A}(B_{h})\}}{|C|} - \operatorname{var}\{Y_{A}(C)\} | \mathbf{Z}\right\}; A \subset D_{s}.$$

This proves (6).

We now prove Equation (6). From (6) we have,

CAGE(C) =
$$E\left[\sum_{h \in H} \frac{\{Y_A(B_h) - Y_A(C)\}^2}{|C|} | \mathbf{Z}\right]$$
. (5.-39)

Adding and subtracting \widehat{Y}_A ,

$$\begin{split} \operatorname{CAGE}(C) &= E\left[\sum_{h \in H} \frac{\left\{Y_A(B_h) - \widehat{Y}_A(C) + \widehat{Y}_A(C) - Y_A(C)\right\}^2}{|C|} | \mathbf{Z}\right] \\ &= E\left[\sum_{h \in H} \frac{\left\{Y_A(B_h) - \widehat{Y}_A(C)\right\}^2}{|C|} | \mathbf{Z}\right] + E\left[\sum_{h \in H} \frac{\left\{\widehat{Y}_A(C) - Y_A(C)\right\}^2}{|C|} | \mathbf{Z}\right] \\ &+ 2E\left[\sum_{h \in H} \frac{\left\{Y_A(B_h) - \widehat{Y}_A(C)\right\} \left\{\widehat{Y}_A(C) - Y_A(C)\right\}}{|C|} | \mathbf{Z}\right] \\ &= E\left[\sum_{h \in H} \frac{\left\{Y_A(B_h) - \widehat{Y}_A(C)\right\}^2}{|C|} | \mathbf{Z}\right] + E\left[\left\{\widehat{Y}_A(C) - Y_A(C)\right\}^2 | \mathbf{Z}\right] \\ &- 2E\left[\left\{\widehat{Y}_A(C) - Y_A(C)\right\}^2 | \mathbf{Z}\right] \\ &= E\left[\sum_{h \in H} \frac{\left\{Y_A(B_h) - \widehat{Y}_A(C)\right\}^2}{|C|} | \mathbf{Z}\right] - E\left[\left\{\widehat{Y}_A(C) - Y_A(C)\right\}^2 | \mathbf{Z}\right]. \end{split}$$

This proves Equation (6).

Result 2: For Z defined in (14) and $Y(\cdot; \phi_s)$ defined in (13), we have that CAGE in (17) has the following alternative expressions:

$$\begin{split} CAGE(A) &= E\left[\int_{A} \frac{\left\{Y_{s}(s; \ \boldsymbol{\phi}_{s}) - Y_{A}(A; \ \boldsymbol{\phi}_{s})\right\}^{2}}{|A|} ds |\mathbf{Z}\right] \\ CAGE(A) &= E\left[\int_{A} \frac{\operatorname{var}\left\{Y_{s}(s; \ \boldsymbol{\phi}_{s})\right\}}{|A|} ds - \operatorname{var}\left\{Y_{A}(A; \ \boldsymbol{\phi}_{s})\right\} |\mathbf{Z}\right] \\ CAGE(A) &= E\left[\int_{A} \frac{\left\{Y_{s}(s; \ \boldsymbol{\phi}_{s}) - \widehat{Y}_{A}(A)\right\}^{2}}{|A|} ds |\mathbf{Z}\right] - E\left[\left\{\widehat{Y}_{A}(A) - Y_{A}(A; \ \boldsymbol{\phi}_{s})\right\}^{2} |\mathbf{Z}\right], \end{split}$$

where A is a generic areal unit (i.e., $A \subset D_s$), and $\widehat{Y}_A(A) \equiv E\{Y_A(A)|\mathbf{Z}\}$.

Proof of Result 2: We now prove the equalities listed in Equations (6), (6), and (6) of Proposition 5. We start with Equation (6). Notice that for a given $\mathbf{s} \in D_s$, $A \in D_A$, α , ϕ_s , and Λ ,

$$\frac{1}{|A|}\left\{Y_s(\mathbf{s}; \ \boldsymbol{\phi}_s) - Y_A(A; \ \boldsymbol{\phi}_s)\right\}^2 = \frac{1}{|A|}\left\{\boldsymbol{\phi}_s(\mathbf{s}) - \boldsymbol{\phi}(A; \ \boldsymbol{\phi}_s)\right\}' \boldsymbol{\alpha} \boldsymbol{\alpha}' \left\{\boldsymbol{\phi}_s(\mathbf{s}) - \boldsymbol{\phi}(A; \ \boldsymbol{\phi}_s)\right\}.$$

Taking the expectation with respect to $\alpha | \phi_s$, Λ we have

$$\frac{1}{|A|}E\left[\left\{Y_s(\mathbf{s};\ \boldsymbol{\phi}_s) - Y_A(A;\ \boldsymbol{\phi}_s)\right\}^2 |\boldsymbol{\phi}_s, \boldsymbol{\Lambda}\right] = \frac{1}{|A|}\left\{\boldsymbol{\phi}_s(\mathbf{s}) - \boldsymbol{\phi}(A;\ \boldsymbol{\phi}_s)\right\}' \boldsymbol{\Lambda}\left\{\boldsymbol{\phi}_s(\mathbf{s}) - \boldsymbol{\phi}(A;\ \boldsymbol{\phi}_s)\right\}. \tag{5.-45}$$

Then, upon taking the expectation of (6) with respect to ϕ_s , $\Lambda | \mathbf{Z}$ and integrating \mathbf{s} over A, we obtain Equation (6).

To prove Equation (6) notice that

$$\operatorname{var}\{Y_s(\mathbf{s}; \ \boldsymbol{\phi}_s)\} = \boldsymbol{\phi}_s(\mathbf{s})' \Lambda \boldsymbol{\phi}_s(\mathbf{s})$$
$$\operatorname{var}\{Y_A(A; \ \boldsymbol{\phi}_s)\} = \boldsymbol{\phi}(A; \ \boldsymbol{\phi}_s)' \Lambda \boldsymbol{\phi}(A; \ \boldsymbol{\phi}_s).$$

Expanding (6) and substituting (6) we have

$$CAGE(A) = E\left[\int_{A} \frac{\{\phi_{s}(\mathbf{s}) - \phi(A; \phi_{s})\}' \Lambda \{\phi_{s}(\mathbf{s}) - \phi(A; \phi_{s})\}}{|A|} d\mathbf{s} | \mathbf{Z}\right]$$

$$= E\left\{\int_{A} \frac{\phi_{s}(\mathbf{s})' \Lambda \phi_{s}(\mathbf{s}) - 2\phi_{s}(\mathbf{s})' \Lambda \phi(A; \phi_{s})}{|A|} d\mathbf{s} + \phi(A; \phi_{s})' \Lambda \phi(A; \phi_{s}) | \mathbf{Z}\right\}$$

$$= E\left\{\int_{A} \frac{\phi_{s}(\mathbf{s})' \Lambda \phi_{s}(\mathbf{s})}{|A|} d\mathbf{s} - 2\phi(A; \phi_{s})' \Lambda \phi(A; \phi_{s}) + \phi(A; \phi_{s})' \Lambda \phi(A; \phi_{s}) | \mathbf{Z}\right\}$$

$$= E\left\{\int_{A} \frac{\phi_{s}(\mathbf{s})' \Lambda \phi_{s}(\mathbf{s})}{|A|} d\mathbf{s} - \phi(A; \phi_{s})' \Lambda \phi(A; \phi_{s}) | \mathbf{Z}\right\}$$

$$= E\left[\int_{A} \frac{\operatorname{var}\{Y_{s}(\mathbf{s}; \phi_{s})\}}{|A|} d\mathbf{s} - \operatorname{var}\{Y_{A}(A; \phi_{s})\} | \mathbf{Z}\right]; A \subset D_{s}.$$

This proves (6).

We now prove Equation (6). From (6) we have for any $A \subset D_s$,

CAGE(A) =
$$E\left[\int_{A} \frac{\left\{Y_{s}(\mathbf{s}; \boldsymbol{\phi}_{s}) - Y_{A}(A; \boldsymbol{\phi}_{s})\right\}^{2}}{|A|} d\mathbf{s} |\mathbf{Z}\right].$$
 (5.-51)

Adding and subtracting \widehat{Y}_A ,

$$\begin{aligned} \text{CAGE}(A) &= E\left[\int_{A} \frac{\left\{Y_{s}(\mathbf{s};\ \boldsymbol{\phi}_{s}) - \widehat{Y}_{A}(A) + \widehat{Y}_{A}(A) - Y_{A}(A;\ \boldsymbol{\phi}_{s})\right\}^{2} d\mathbf{s}|\mathbf{Z}\right] \\ &= E\left[\int_{A} \frac{\left\{Y_{s}(\mathbf{s};\ \boldsymbol{\phi}_{s}) - \widehat{Y}_{A}(A)\right\}^{2} d\mathbf{s}|\mathbf{Z}\right] + E\left[\int_{A} \frac{\left\{\widehat{Y}_{A}(A) - Y_{A}(A;\ \boldsymbol{\phi}_{s})\right\}^{2} d\mathbf{s}|\mathbf{Z}\right] \\ &+ 2E\left[\int_{A} \frac{\left\{Y_{s}(\mathbf{s};\ \boldsymbol{\phi}_{s}) - \widehat{Y}_{A}(A)\right\} \left\{\widehat{Y}_{A}(A) - Y_{A}(A;\ \boldsymbol{\phi}_{s})\right\} d\mathbf{s}|\mathbf{Z}\right] \\ &= E\left[\int_{A} \frac{\left\{Y_{s}(\mathbf{s};\ \boldsymbol{\phi}_{s}) - \widehat{Y}_{A}(A)\right\}^{2} d\mathbf{s}|\mathbf{Z}\right] + E\left[\left\{\widehat{Y}_{A}(A) - Y_{A}(A;\ \boldsymbol{\phi}_{s})\right\}^{2} d\mathbf{s}|\mathbf{Z}\right] \\ &- 2E\left[\left\{\widehat{Y}_{A}(A) - Y_{A}(A;\ \boldsymbol{\phi}_{s})\right\}^{2}|\mathbf{Z}\right] \\ &= E\left[\int_{A} \frac{\left\{Y_{s}(\mathbf{s};\ \boldsymbol{\phi}_{s}) - \widehat{Y}_{A}(A)\right\}^{2} d\mathbf{s}|\mathbf{Z}\right] - E\left[\left\{\widehat{Y}_{A}(A) - Y_{A}(A;\ \boldsymbol{\phi}_{s})\right\}^{2}|\mathbf{Z}\right]. \end{aligned}$$

This proves Equation (6).

Result 3: For Z defined in (14) and $Y(\cdot; \phi_s)$ defined in (13), we have that DCAGE in (18) has the

following alternative expressions:

$$DCAGE(C) = E\left\{\sum_{h \in H} \frac{(Y_A(B_h; \boldsymbol{\phi}_s) - Y_A(C; \boldsymbol{\phi}_s))^2}{|C|} | \mathbf{Z}\right\}$$

$$DCAGE(C) = E\left(\sum_{h \in H} \frac{\operatorname{var}(Y_A(B_h; \boldsymbol{\phi}_s))}{|C|} - \operatorname{var}(Y_A(C; \boldsymbol{\phi}_s)) | \mathbf{Z}\right)$$

$$DCAGE(C) = E\left\{\sum_{h \in H} \frac{(Y_A(B_h; \boldsymbol{\phi}_s) - \widehat{Y}_A(C))^2}{|C|} | \mathbf{Z}\right\} - E\left\{(\widehat{Y}_A(C) - Y_A(C; \boldsymbol{\phi}_s))^2 | \mathbf{Z}\right\},$$

where $C = \bigcup_{h \in H} B_h$, $H \subset \{1,...,n_B\}$, and $B_h \in D_B$ for each $h \in H$.

Proof of Result 3: In the proof of Result 2, replace the integral with sums, and replace $\phi_s(\mathbf{s})$ and $Y_s(\mathbf{s}; \phi_s)$ with $\phi_A(B_h; \phi_s)$ and $Y_A(B_h; \phi_s)$, respectively.

References

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2015). *Hierarchical Modeling and Analysis for Spatial Data*, 2nd edn. Boca Raton, FL: Taylor and Francis Group.

Berliner, L. M. (1996). *Hierarchical Bayesian time-series models*. Kluwer Academic Publishers, Dordrecht, NL.

Blank, R. M., Groves, R. M., Mesenbourg, T. L., Jackson, A. A., Hogan, H. R., Matos, M. A., and Weinberg, D. H. (2011). "2010 Census redistricting data (public law 94-171) summary file." Tech. rep., US Census Bureau.

Bradley, J., Cressie, N., and Shi, T. (2014a). "A comparison of spatial predictors when datasets could be very large." *arXiv preprint:* 1410.7748.

— (2015a). "Comparing and Selecting Spatial Predictors Using Local Criteria (with discussion)." *TEST*, 24, 1–28.

- Bradley, J., Holan, S., and Wikle, C. (2014b). "Mixed effects modeling for areal data that exhibit multivariate-spatio-temporal dependencies." *arXiv preprint:* 1407.7479.
- Bradley, J., Wikle, C. K., and Holan, S. H. (2015b). "Bayesian spatial change of support for count-valued survey data." *Journal of the American Statistical Association*, forthcoming.
- Bradley, J. R., Cressie, N., and Shi, T. (2011). "Selection of rank and basis functions in the Spatial Random Effects model." In *Proceedings of the 2011 Joint Statistical Meetings*, 3393–3406. Alexandria, VA: American Statistical Association.
- Bradley, J. R., Holan, S. H., and Wikle, C. K. (2015c). "Multivariate Spatio- Temporal Models for High-Dimensional Areal Data with Application to Longitudinal Employer-Household Dynamics." *The Annals of Applied Statistics*, forthcoming.
- Cressie, N. (1993). Statistics for Spatial Data, rev. edn. New York, NY: 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.
- Cressie, N. and Wikle, C. K. (2011). Statistics for Spatio-Temporal Data. Hoboken, NJ: Wiley.
- Darby, S., Deo, H., and Doll, R. (2001). "A parallel analysis of individual and ecological data on residential radon and lung cancer in south-west England." *Journal of the Royal Statistical Society, Series A*, 164, 193–203.
- Duque, J., Anselin, L., and Rey, S. (2012). "The max-p-regions problem." *Journal of Regional Science*, 52, 397–419.
- Ferreira, J. C. and Menegatto, V. A. (2009). "Eigenvalues of integral operators defined by smooth positive definite kernels." *Integral Equations and Operatory Theory*, 61–81.
- Ferreira, M., Holan, S., and Bertolde, A. (2011). "Dynamic multiscale spatio-temporal models for Gaussian areal data." *Journal of the Royal Statistical Society, Series B*, 73, 663–688.

- Ferreira, M. and Lee, K. (2007). *Multiscale Modeling: A Bayesian Perspective*. New York: Springer.
- Folch, D. and Spielman, S. (2014). "Identifying regions based on flexible user defined constraints." *International Journal of Geographic Information Science*, DOI:10.1080/13658816.2013.848986.
- Gehike, C. and Biehl, K. (1934). "Certain effects of grouping upon the size of the correlation coefficient in census tract material." *Environmental and Ecological Statistics*, 11, 31–54.
- Guo, D. (2008). "Regionalization with dynamically constrained agglomerative clustering and partitioning (REDCAP)." *International Journal of Geographical Information Science*, 22, 801–823.
- Hartigan, J. and Wong, M. (1979). "A k-means clustering algorithm." *Applied Statistics*, 28, 100–108.
- Higham, N. (1988). "Computing a nearest symmetric positive semidefinite matrix." *Linear Algebra* and its Applications, 105, 103–118.
- Hodges, J. (2013). *Richly Parameterized Linear Models: Additive, Time Series, and Spatial Models Using Random Effects*. Boca Raton, FL: Chapman & Hall/CRC.
- Karhunen, K. (1947). "Uber lineare Methoden in der Wahrscheinlichkeitsrechnung." *Ann. Acad. Sci. Fennicae. Ser. A. I. Math.-Phys*, 37, 1–49.
- King, G. (1997). A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data. Princeton, NJ: Princeton University Press.
- Kolaczyk, E. and Huang, H. (2001). "Multiscale statistical models for hierarchical spatial aggregation." *Geographical Analysis*, 33, 95–118.
- Kolaczyk, E., Ju, J., and Gopal, S. (2005). "Multiscale, multigranular statistical image segmentation." *Journal of the American Statistical Association*, 100, 1358–1369.

- Kolaczyk, E. and Nowak, R. (2004). "Multiscale likelihood analysis and complexity penalized estimation." *The Annals of Statistics*, 32, 500–527.
- Loève, M. (1978). *Probability Theory Vol. II, 4-th ed.*. Princton, NJ: Graduate Texts in Mathematics 46 Springer-Verlag.
- Logan, J. (2011). "Identifying and bounding ethnic neighborhoods." *Urban Geography*, 32, 334–359.
- Marsland, S. (2009). *Machine Learning: An Algorithmic Perpsective*. Boca Raton, FL: Chapman & Hall/CRC.
- Martin, D. (2002). "Geography for the 2001 census in England and Wales." *Population Trends*, 108, 7–15.
- Mearns, L., Bukovsky, M., Pryor, S., and Magana, V. (2014). "Downscaling of climate information." In *Climate Change in North America, Regional Climate Studies.*, ed. G. Ohring, 201–250. Springer International Publishing: Cham.
- Mercer, J. (1909). "Functions of positive and negative type and their connection with the theory of integral equations." *Philosophical Transactions of the Royal Society A*, 209, 415–458.
- Milliff, R., Bonazzi, A., Wikle, C., Pinardi, N., and Berliner, L. (2011). "Ocean ensemble fore-casting. Part I: Ensemble Mediterranean winds from a Bayesian hierarchical model." *Quarterly Journal of the Royal Meteorological Society*, 137, 858–878.
- Mugglin, A., Carlin, B., Zhu, L., and Conlon, E. (1998). "Bayesian areal interpolation, estimation, and smoothing: An inferential approach for Geographic Information Systems." *Environment and Planning A*, 31, 1337–1352.
- Murtagh, F. (1992). "Contiguity-constrained clustering for image analysis." *Pattern Recognition Letters*, 13, 677–683.

- Nychka, D., Haaland, P., OConnell, M., and Ellner, S. (1998). "FUNFITS, Data Analysis and Statistical Tools for Estimating Functions." In *Case Studies in Environmental Statistics, Lecture Notes in Statistics*, eds. D. Nychka, W. Piegorsch, and L. Cox, 159–179. Springer-Verlag.
- Nychka, D. and Saltzman, N. (1998). "Design of Air Quality Monitoring Networks." In *Case Studies in Environmental Statistics*, *Lecture Notes in Statistics*, eds. D. Nychka, W. Piegorsch, and L. Cox, 51–76. Springer-Verlag.
- Obled, C. and Creutin, J. (1986). "Some developments in the use of empirical orthogonal functions for mapping meteorological fields." *Journal of Applied Meteorology*, 25, 1189–1204.
- Oehlert, G. (1992). "A note on the delta method." *The American Statistician*, 46, 27–29.
- Openshaw, S. (1977). "A geographical solution to scale and aggregation problems in region-building, partitioning and spatial modelling." *Transactions of the Institute of British Geographers*, 2, 459–472.
- Openshaw, S. and Taylor, P. (1979). "A million or so correlation coefficients: Three experiments on the modifiable areal unit problem." In *Statistical Applications in the Spatial Sciences*, ed. N. Wrigley, 48–78. London: Pion.
- Papoulis, A. (1965). *Probability, Random Variables, and Stochastic Processes*. New York, NY: McGraw-Hill.
- Robinson, S. (1950). "Ecological correlations and the behavior of individuals." *American Sociological Review*, 15, 351–357.
- Ruppert, D., Wand, M., and Carroll, R. (2003). *Semiparametric Regression*. Cambridge: Cambridge University Press.
- Sorbye, S. and Rue, H. (2014). "Scaling intrinsic Gaussian Markov random field priors in spatial modelling." *Spatial Statistics*, 8, 39–51.

- Speilman, S., Folch, D., and Nagle, N. (2013). "Patterns and causes of uncertainty in the American Community Survey." *Applied Geography*, 46, 147–157.
- Spielman, S. and Logan, J. (2013). "Using high-resolution population data to identify neighborhoods and establish their boundaries." *Annals of the Association of American Geographers*, 103, 67–84.
- (2015). "Reducing Uncertainty in the American Community Survey through Data-Driven Regionalization." *PLOSOne*, 10, 0115626. doi:10.1371/journal.pone.0115626.
- Stein, M. (2013). "Limitations on low rank approximations for covariance matrices of spatial data." *Spatial Statistics*, 8, 1–19.
- Trevisani, M. and Gelfand, A. (2013). "Sampling designs and prediction methods for Gaussian spatial processes." In *Advances in Theoretical and Applied Statistics*, eds. N. Torelli, F. Pesarin, and A. Bar-Hen, 269–279. Springer-Verlag Berlin Heidelberg.
- Wakefield, J. (2004). "A critique of statistical aspects of ecological studies in spatial epidemiology." *Environmental and Ecological Statistics*, 11, 3154.
- Waller, L. and Gotway, C. (2004). *Applied Spatial Statistics for Public Health Data*. New York: Wiley.
- Wendland, H. (1998). "Error estimates for interpolation by compactly supported radial basis functions of minimal degree." *Journal of Approximation Theory*, 93, 258–272.
- Wikle, C. and Berliner, M. (2005). "Combining information across spatial scales." *Technometrics*, 47, 80–91.
- Wikle, C., Milliff, R., Nychka, D., and Berliner, L. (2001). "Spatiotemporal hierarchical Bayesian modeling tropical ocean surface winds." *Journal of the American Statistical Association*, 96, 454, 382–397.

- Wikle, C. K. (2010). "Low-rank representations for spatial processes." In *Handbook of Spatial Statistics*, eds. A. E. Gelfand, P. J. Diggle, M. Fuentes, and P. Guttorp, 107–118. Boca Raton, FL: Chapman & Hall/CRC Press.
- Wikle, C. K., Milliff, R. F., Herbei, R., and Leeds, W. B. (2013). "Modern statistical methods in oceanography: A hierarchical perspective." *Statist. Sci.*, 28, 4, 466–486.
- Yang, R. and Berger, J. (1994). "Estimation of a covariance matrix using the reference prior." *Annals of Statistics*, 22, 1195–1211.