# Bayesian Multiscale Modeling of Spatial Infrastructure Performance Predictions with an Application to Electric Power Outage Forecasting

Allison Reilly, A.M.ASCE[1]; and Seth Guikema, A.M.ASCE[2]

**Abstract:** A number of models have been developed to estimate the spatial distribution of infrastructure impact during a natural hazard event. For example, statistical approaches have been developed to estimate the percentage of customers without power attributable to a hurricane, with the estimates made at a local geography level such as census tracts. Whereas some statistical infrastructure performance models use extensive covariate data that captures a significant amount of the spatial information, others use a limited number of covariates to enhance model simplicity and to reduce the cost and time associated with obtaining these data. However, in these simpler models, the omitted covariates result in loss of spatial information in the model, leading to a situation in which predictions from adjacent regions are more dissimilar than would be expected. In this paper, a tree-based statistical mass-balance multiscale model is developed to smooth the outage predictions at granular levels by allowing spatially similar areas to inform one another with the goals of: (1) reducing spatial error in simplified prediction models, and (2) yielding estimates at other levels of aggregation in addition to the native model resolution. A generalized density-based clustering algorithm is used to extract the hierarchical spatial structure. The *noise* regions (i.e., those regions located in sparse areas) are then aggregated using a distance-based clustering approach. The authors demonstrate this approach using outage predictions from Hurricane Ivan and develop outage prediction maps at different levels of granularity. **DOI: 10.1061/(ASCE)IS.1943-555X.0000222.** 

**Author keywords:** Electric power outages; Multiscale modeling; Spatial smoothing; Clustering; Hurricanes; Risk management.



## Introduction

Many models have been developed to estimate the spatial distribution of infrastructure impact during a natural hazard event (Han et al. 2009a, b; Winkler et al. 2010; Nateghi et al. 2011). For example, statistical approaches have been developed to estimate the number of outages at a spatially detailed grid cell level (Han et al. 2009a, b; Nateghi et al. 2013) and the percentage of customers without power because of a hurricane at the census tract level (Guikema et al. 2013). Statistical infrastructure performance models must include exogenous factors to assess how widespread the damage will be and how the system will perform. There are two general approaches taken with these models. The first is to include a wide range of explanatory variables—covariates—in the model, capturing as much of the relevant spatial information as possible. Han et al. (2009a, b) is typical of this approach. The second approach is to simplify the model as much as possible by reducing the covariates used in the model (e.g., Nateghi et al. 2013, Guikema et al. 2013). This approach can maintain strong overall predictive accuracy, but at the cost of a possible reduction of spatial detail in the predictions, leading to spatially similar regions having predictions less similar than may be expected. This paper focuses primarily on the second type of models—simplified prediction models, and aims to increase the accuracy in the spatial pattern of predictions while maintaining mass balance in predicted outages.

In simplified prediction models, omitted covariates are often spatially correlated. This leads to a situation in which the spatial structure of the covariates is not fully captured, leading to possibly spatially correlated model errors. For example, some electric-power outage forecasting models exclude covariate data such as rainfall amounts, infrastructure age, and tree cover, all of which are both: (1) difficult to estimate across large regions in advance of a storm, and (2) spatially correlated. On the other hand, covariate data that are used describe the conditions at a localized level (e.g., a census tract) inaccurately describe each infinitesimally small location that collectively create that localized level, meaning single covariate observations are used rather than local distributions over the covariates. Here, information from a spatially-similar region may bolster the single observations without using fully probabilistic local input to the model.

*In this paper, the authors seek to ameliorate spatial error and imprecise localized observations in simplified infrastructure performance prediction models by identifying the region's spatial structure and smoothing (i.e., informing) predictions ex post with predictions from regions that are spatially similar.* This work contributes to both the infrastructure and mass balance modeling literature. No existing work has focused on spatially informing predictions among spatially-similar regions in a way that is mass-balanced across different levels of resolution. Although Liu et al. (2007) did include spatial correlation in the training of their hurricane power outage model, they considered only the correlation in the model prediction errors. Furthermore, the authors know of no existing mass balance literature that has combined density-based spatial clustering with mass balance modeling for hierarchical modeling on spatially similar regions.

[1]Postdoctoral Research Fellow, Dept. of Geography and Environmental Engineering, Johns Hopkins Univ., 3400 N Charles St., Ames Hall, Baltimore, MD 21218 (corresponding author). E-mail: acr@jhu.edu

[2]Assistant Professor, Dept. of Geography and Environmental Engineering, Johns Hopkins Univ., 3400 N Charles St., Ames Hall, Baltimore, MD 21218.

Note. This manuscript was submitted on October 17, 2013; approved on April 18, 2014; published online on June 23, 2014. Discussion period open until November 23, 2014; separate discussions must be submitted for individual papers. This paper is part of the ***Journal of Infrastructure Systems***, © ASCE, ISSN 1076-0342/04014036(11)/$25.00.

A second objective of this paper, in addition to ex post spatial smoothing of predictions, is the development of a method for producing predictions at spatial scales different from the model's native spatial scale. Infrastructure performance models typically produce estimates at a given spatial scale such as 12,000 foot by 8,000 foot grid cells or census tracts (Han et al. 2009a, b). However, it is desirable to be able to obtain estimates at other spatial scales based on the output at the *native* scale of the model.

Our approach first determines the spatial structure of the region using the spatial density of census tracts as a proxy for spatial similarity. Census tracts that are more densely packed (e.g., urban areas) are likely have a more similar environment than census tracts that are loosely packed (e.g., rural areas). A density-based clustering algorithm is used based on a modified version of *OPTICS* (Ankerst et al. 1999) to identify regions that contain census tracts that are approximately equally dense. A centroid-based clustering algorithm then clusters regions of equal density. This process is recursively repeated so clusters are subdivided further into more clusters. Ultimately, this creates a hierarchy, with one cluster at the top level that represents all census tracts, and numerous clusters at the bottom level that each represent a few census tracts. Bayesian mass-balanced multiscale modeling is then used to estimate the predicted number of events at each level in the spatial hierarchy in a way that allows the predictions within a cluster to inform each other. This acts to smooth the data within the clusters. Posterior predictions at a coarser level become prior estimates at more granular levels. A benefit of examining the spatial data this way is that it allows the user to aggregate census tracts and clusters for region-wide performance assessments. This is different than simply summing together damage predictions for a region; the spatial structure influences the *weights* or percentages that each cluster contributes to the whole.

This paper is divided into five sections. The next section is devoted to clustering methodology. Here, density and centroid clustering and methods for extracting clusters are discussed. Next, mass-balance multiscale modeling methodology are discussed. The fourth section discusses a case study that examines predictions of electric-power outages from Hurricane Ivan in the southern U.S. The fifth section provides conclusions. Whereas this work is described largely in the context of hurricane power outage prediction in this paper, it can apply directly to any model that produces a spatial map of predictions of count events, including, for example, traffic accident models, water distribution system pipe break models, and communication outage models.

This work builds from existing models in several areas, but it extends these and combines them in a novel manner. The clustering portion of the work builds from both *k*-means clustering and the *OPTICS* (Ankerst et al. 1999) approach but advances *OPTICS* by improving how it models certain types of spatially connected but nonconvex regions. The Bayesian multiscale model that is then applied at the spatial clusters is a standard Bayesian mass-balance Poisson multiscale model (e.g., Kolaczyk and Huang 2001). What is particularly new in this paper is the combination of the (improved) *OPTICS*/*k*-means hybrid clustering with the multiscale model to improve the spatial predictions from infrastructure performance models.

## Cluster Methodology

A cluster is simply a group of objects, be they people, ideas, locations, or events, considered similar by one or more metrics. With applications in biology (Alon et al. 1999; Fathian et al. 2007), computer science (Frey and Dueck 2007; Broder et al. 1997), and social science (Hillhouse and Adler 1997; Cook 2005), among many others, clustering analysis is an efficient data mining method for grouping objects with similar characteristics [e.g., urban crime hotspots, see Moser et al. (2007)]. As in the case here, the cluster analysis can be a preprocessing tool for use later, such as multiscale modeling.

Many methods for determining clusters within a data set exist, each with their own benefits and downsides. Generally these clustering methods fall into one of two categories—partitioning and hierarchical algorithms. Partitioning algorithms (e.g., *k*-means, an algorithm used later in this paper) divide the dataset, whereas hierarchical algorithms decompose the dataset into a nested partition (Ankerset et al. 1999). Some algorithms are distinctively partitioning or hierarchical algorithms whereas others are hybrids, blurring the definitions of both. The clustering method chosen by the user is heavily context dependent, and the outcomes depend on the method selected. The reader is referred to Jain (2010) for a review.

Bayesian mass-balanced multiscale modeling requires both identification of spatial clusters and the intrinsic hierarchy of these clusters (e.g., clusters within clusters), so that coarser clusters can act as informative priors to more granular clusters. The ultimate objective is to use this structure to allow spatially similar objects to inform each other more intensely than spatially dissimilar objects at each level under consideration. In this section, the following methods are discussed: (1) the hybrid density- and centroid-based method by which spatial clusters are extracted, and (2) an efficient approach for identifying density-cluster hierarchy in spatial data.

### Density Clustering

A new hybrid approach for finding spatially similar regions has been developed. First, the region of interest through the use of a density-based clustering is examined; here, density serves as a proxy for spatial-likeness. Once the data's intrinsic density hierarchy is established, the clusters become nodes in a dendrogram, with subnodes, or descendant nodes, pointing to parent nodes. The leaf nodes, i.e., the nodes/clusters at the finest resolutions, are then examined through a centroid-based method of clustering to further examine the hierarchical nature of the data.

Density-based clustering methods find objects (e.g., census tracts) that are densely grouped together and segregate them from objects that are not—often considered *noise*. A physical interpretation of this is urban verses rural census tracts. Urban census tracts, typically relatively small, have centroids closer together than rural census tracts. Ideally, the condition of being a dense region versus not dense is not binary but rather a continuum that establishes sparse regions, somewhat dense regions, and so on (e.g., urban, suburban, exurban, and rural areas). A less-dense region can ring around a denser one, and so on. Fig. 1 shows an example of this concept. The object centroids are represented by the symbol ○. Cluster A is most dense and is clearly a cluster. Cluster *B* is less dense, but no less likely to be a cluster than Cluster *A*. Objects *q* and *p* are not near Clusters *A* or *B* and are only considered members of a cluster if there is one cluster that describes the whole dataset.

Density-based clustering is most appropriate for aggregation at the coarsest levels. Here, objects in the same cluster are simply more spatially similar, and no arbitrary parameter is used to predefine the number of objects allowed in the cluster, as is the case with centroid-based algorithms (Jain 2010). Predefining the size of a cluster as is done in algorithms such as *k*-means can inherently bias multiscale modeling results because it can unduly include or exclude an object. Note though, once spatially similar regions

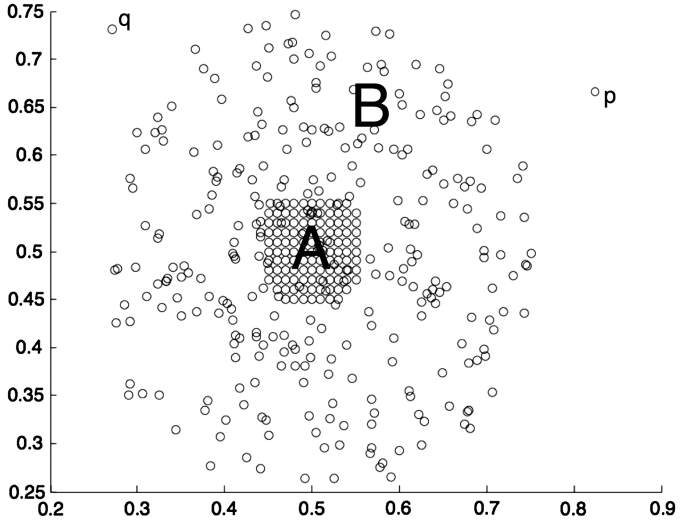


**Fig. 1.** Example of densely grouped census tracts in the center surrounded by less densely grouped census tracts; the circle symbol represents the centroid of a census tract [data from Achtert et al. (2013)]

are established, centroid-based clustering techniques are reasonable because the data's underlying spatial distribution is similar.

Before progressing, $n$ observation from the dataset used in the case study, a property of the data set that is quite general and likely will impact other similar spatial clustering problems. Imagine a cluster consisting of dense urban census tracts, and then a suburban cluster completely surrounding, or *doughnuting*, the first, as seen in Fig. 1. Assume that for postclustering, the performance predictions are smoothed within each of the two clusters. Although it is reasonable that the predictions in census tracts forming the inner cluster should inform each other, it is not the case that predictions from the western region in the outer cluster should equally inform predictions in the eastern region in the outer cluster, especially considering there is the inner cluster between them. A novel approach has been developed, described below, to aid in segregating this outer cluster into subclusters so that the density clustering is maintained and geographic separation is considered. Geographic separation is appropriate in this case because the authors seek to have geographically-like regions inform each other through multiscale modeling. A sparse region that is connected through density to another sparse region, though separated by great distance (e.g., the western and eastern regions of Cluster B in Fig. 1) may have little in common spatially. For example, it is unreasonable to assume these regions share the same topology, tree-cover, and infrastructure age and investment, especially when the distance that separates them is great. This exercise of separating clusters based on geographic separation is not appropriate when one wishes to simply find regions of like density, regardless of the spatial shape that may result.

Now the focus is turned to an overview of density-based modeling. For more in-depth reading, the reader is referred to Ankerst et al. (1999). Perhaps the most intuitive method for cluster discovering is to use a grid cell approach, whereby the spatial data are partitioned into predetermined square cells, and the number of objects (e.g., census tracts) are simply counted. Wang et al. (1997) develop a statistical algorithm, *STING*, which extends this concept and demonstrates great success in reducing computational demand from density extraction for large data sets. The user-selected parameters dictate the quality and the interpretation of the results.

A common density-based clustering algorithm for finding clusters is DBSCAN (Ester et al. 1996). *DBCAN* is the foundation of the clustering algorithm ultimately used in this paper, and, because of this, some of the technical details of *DBSCAN* and an extension of *DBSCAN* called *OPTICS* are reviewed in the following paragraphs. In this algorithm, the user defines the minimum number of points ($MP$) that must be contained in a radius, $\varepsilon$, around object $i$. Should the $MP$ criterion be met, the object $i$ is sufficiently dense. However, should the algorithm stop there, it would improperly classify objects on the border of a cluster as noise. To avoid this, *DBSCAN* distinguishes between objects in the center of a cluster and those on the exterior by introducing the concepts of directly density-reachable, density-reachable, and density-connected. The first applies to center (or core) objects, and point $p$ is directly density-reachable to point $q$ if $p$ is within radius $\varepsilon$ of $q$ and $q$ has at least $MP$ in its radius $\varepsilon$. Objects $q$ and $p$ are density-reachable if there is an ordered chain of objects, starting with $q$ and ending with $p$ such that each successive object is directly-density reachable to its former object.

If objects $p$ and $q$ are both border points, it is possible that they are neither directly density-reachable nor density-reachable, because neither term guarantees symmetric conditions for two objects. Hence, the term density-connected is established to categorize these objects. Objects $p$ and $q$ are density-connected if there exists an object $r$ that is density-reachable to both $p$ and $q$. From this, a cluster can be defined: if object $p$ is in cluster $C$ and object $q$ is density-reachable from object $p$, then object $q$ is also in cluster $C$, and all objects within a cluster are density-connected to each other. The algorithm then sorts through the spatial data to identify clusters that fit these criteria.

While computationally efficient, *DBSCAN* is highly sensitive to its input parameters $MP$ and $\varepsilon$, and cannot differentiate among clusters of varying densities (e.g., clusters within clusters are typically not identified). Ordering points to identify the clustering structure, from hereon referred to as *OPTICS*, offers a generalization of *DBSCAN*, and the appeal is that it does not explicitly find and extract clusters, but rather uses what is called reachability-distance to order the objects (Ankerst et al. 1999). It is from this ordering that clusters can be found.

The reachability-distance, $RD_{MP,\varepsilon}(i,j)$, is a function of the location of two objects, $i$ and $j$, and the parameters $MP$ and $\varepsilon$. $RD_{MP,\varepsilon}(i,j)$ is undefined when the number objects (or neighbors) located in radius $\varepsilon$ around object $i$ is less than $MP$ and is the maximum of the distance between objects $i$ and $j$ and the distance between object $i$ and object $k$, the $MP$th closest neighbor of object $i$, otherwise. Fig. 2 demonstrates this definition. When $RD_{MP,\varepsilon}(i,j)$ is undefined, object $j$ is noise; specifically it is located in a sparse area relative to object $i$. When $RD_{MP,\varepsilon}(i,j)$ is the distance between object $i$ and object $k$, the $MP$th closest neighbor of object $i$, object $j$ is a core object. Otherwise, object $j$, simply the distance between objects $i$ and $j$, may be a border object (Ankerst et al. 1999).

The *OPTICS* algorithm iterates through all objects in the dataset, computes their reachability-distance, and then orders the objects based on the order in which they are processed. Using the reachability-distance and the object ordering, one can extract (1) clusters, and (2) any hierarchy within the clusters. Because of how reachability-distances are defined and how the clusters are exacted, the method is less sensitive to the parameters $MP$ and $\varepsilon$ than *DBSCAN*. Fig. 3 shows the reachability-distance plot for the same set of objects shown in Fig. 1. The valleys in the reachability plot correspond to clusters, and the deeper the valley, the denser the cluster. Like all clustering algorithms, *OPTICS* is a heuristic intended for computational feasibility for a problem that is otherwise intractable. The best ordering of the objects is not guaranteed in the reachability-distance plot, however, from the authors' perspective and experience, the ordering does provide a good representation of the spatial density. The outcome of the *OPTICS*

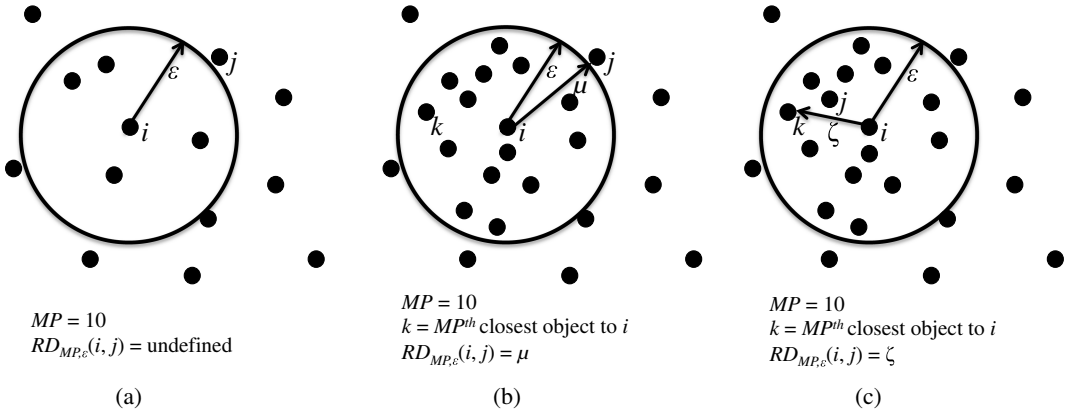


**Fig. 2.** Demonstration of reachability-distance, $RD_{MP,\varepsilon}(i,j)$: (a) $RD_{MP,\varepsilon}(i,j)$ is undefined when $MP$ objects do not exist within radius $\varepsilon$ of object I; (b) when $MP$ objects exist within radius $\varepsilon$ of object, then $RD_{MP,\varepsilon}(i,j)$ exists; Object $k$ is the $MP$th closed object to $i$; because distance$(i,j) >$ distance$(i,k)$, $RD_{MP,\varepsilon}(i,j) =$ distance$(i,j) = \mu$; (c) because distance$(i,j) <$ distance$(i,k)$, $RD_{MP,\varepsilon}(i,j) =$ distance$(i,k) = \zeta$

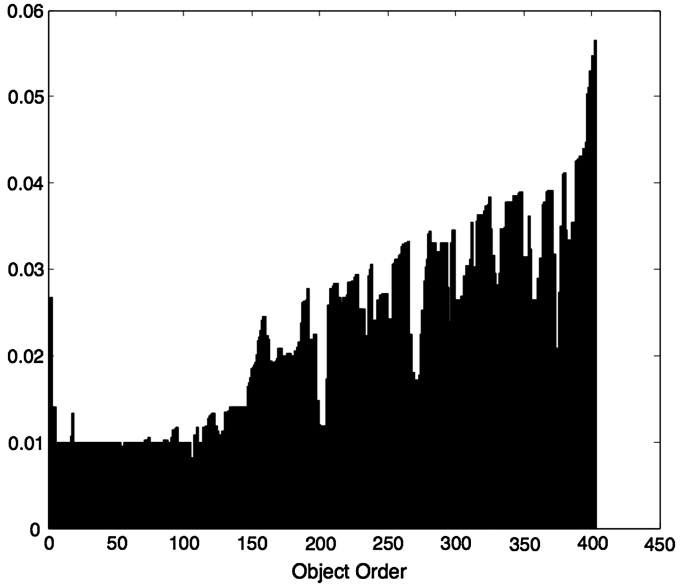


**Fig. 3.** Sample reachability-distance graph of the dataset in Fig. 1

algorithm is repeatable only when starting with the same object; initialization using a different object may lead to different object ordering and reachability-distances. Testing should be conducted ex post to ensure that clustering is consistent. The reader is referred to Ankerst et al. (1999) for details of the *OPTICS* algorithm.

Before concluding, an additional output from the algorithm that is recorded for later use merits mentioning. The object $p$ that precedes object $q$ in the reachability-distance plot need not be density-reachable to object $q$ and can simply be density-connected. As such, one should record also the object $o$ that both is density-reachable to object $q$ and produces the reachability-distance described in the output for object $q$. This will help to segregate *doughnut* clusters in the new extension to *OPTICS* described below.

### Extracting Clusters and Their Intrinsic Hierarchy

An algorithm that extracts the intrinsic hierarchy of the data structure and is insensitive to input parameters is necessary. Intrinsic hierarchy indicates something quite specific—the algorithm must be able to identify regions within regions that have distinct densities (e.g., find cities within suburban sprawl). Once this is done, a dendrogram of the hierarchy can be created, with the top node referring to the entire data set and branches below being clusters within the broader cluster.

The authors adapt the algorithm developed in Sander et al. (2003) to extract the density hierarchy. It uses the reachability-distance plot—specifically the order by which objects are plotted and the objects' reachability-distances—to determine the clusters. Significant valleys within the plot become clusters, valleys within valleys become subclusters that point to the first cluster, and so on.

The reader is referred to Fig. 4 for the flow chart of the algorithm. The modified nodes within the flow chart, i.e., the new contribution to the clustering approach, are shaded gray, and the rest of the algorithm is attributed to Sander et al. (2003). A brief description follows: First a list is created of all the objects that have a greater reachability-distance relative to the objects to its right and left in the ordered objects array. This list of local maxima is sorted in ascending order. The first object from the list is selected and this object serves as a divider in the reachability-plot. If the reachability-distance of the local maximum is less than parameter $\alpha$ times the average reachability-distance in either the right or the left sections, then the algorithm is recursively called using the next smallest local maximum. The sections are too sparse. Otherwise, two new clusters are added to the dendrogram, and the objects in the local maxima list are divided whether object $n$ of $N$ is to the left or to the right of local maximum that divides the plot. This local maximum is not added to either list.

The algorithm also determines to which parent node the new clusters ought to point in the dendrogram—either to the cluster from which the new clusters are formed, or to the parent node of that cluster. Essentially, this step determines whether the local maximum that separates the parent node is *similar* to the local maximum that separates the objects in the clusters of interest. If so, the new clusters bypass the cluster that created them and point to the creating cluster's parent cluster.

At this point in the algorithm, one of two modifications is proposed. This algorithm does not guarantee that all of the objects within the new clusters are density-reachable, and only guarantees that clusters are density-connected. This can create spatial clusters that are seemingly amiss. For example, a cluster could appear to be divided in half by another cluster of a different density, or a cluster can doughnut around another cluster of a different density, as is the case in Fig. 1. The seemingly awry clusters are accurate because

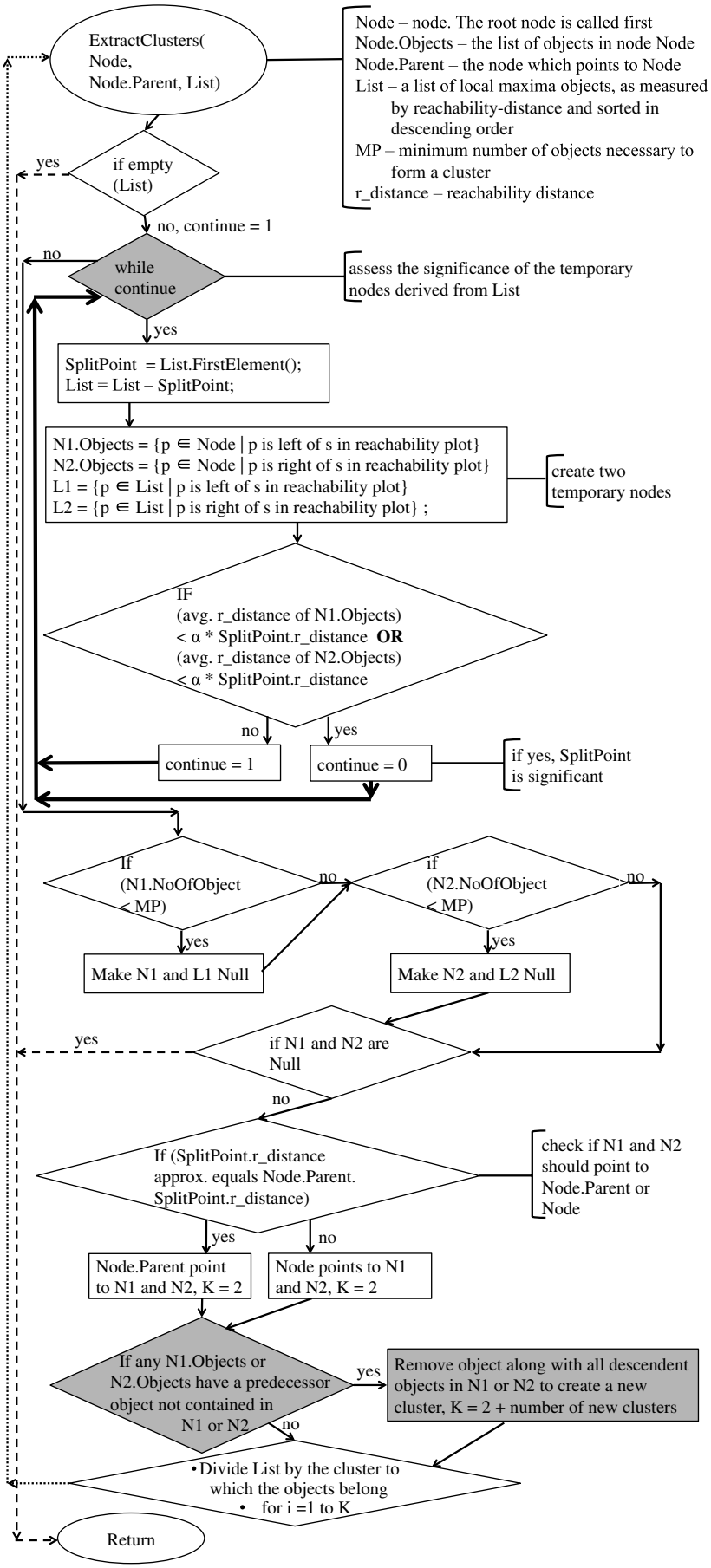


**Fig. 4.** Cluster extraction algorithm [adapted from Sander et al. (2003)]; the shaded boxed indicate modifications by the authors

they are similar in density, and the *OPTICS* algorithm guarantees that they be density-connected. However, it may be unreasonable to consider them as one spatial cluster for the multiscale modeling. As mentioned earlier, predictions from an eastern region should not inform predictions of an equally dense western region if there is a cluster dividing the regions.

To reduce the likelihood of divided clusters being in the output, we incorporate each object's predecessor from the *OPTICS* algorithm. An object's predecessor is the object that defines its outputted reachability-distance, and it may or may not precede the object in object ordering in the reachability-distance plot. Once a cluster is added to the dendrogram, all of the ordered objects are successively checked, except the first, that lay within the cluster if their predecessor object is also within this the cluster. The predecessor of the first ordered object in a cluster will not be in the cluster. If an object's predecessor is not within the cluster, the object is removed from the cluster along with all objects that consider the object to be their predecessor, and so on. These objects form their own clusters, each with equivalent densities, and the list of local maxima is divided appropriately among the clusters. Once this is complete, the program is recursively called for both of the new clusters using the next smallest local maximum in each new cluster. This algorithm is equivalent to a recursive program that seeks to minimize the number of unique descendent clusters within a parent cluster such that all the predecessors of the ordered objects in a descendent cluster (but the first ordered object) are contained within that descendent cluster.

As mentioned in Ankerst et al. (1999) and Ester et al. (1996), clustering algorithms do not scale well with size, particularly attributable to the recursive nature of these algorithms. This is of particular concern in large data sets. For example, the one used in the case study later examines the spatial structure of 904 census tracts and, depending on the chosen parameters, can employ hundreds of recursions. When predicting infrastructure performance in large geographic areas, the number of census tracts can easily be in the tens of thousands. Hence, the algorithm is modified slightly to reduce to the number of recursions necessary. Rather than recursively call the program when a local maximum is deemed insufficient, this is replaced with a while loop that iteratively examines the list of local maxima and stops once a local maximum is found that adequately divides two valleys of the reachability-distance plot.

### *k*-Means and Hierarchy

Once the intrinsic hierarchy of the density clustering is established, a centroid-based partitioning algorithm, *k*-means, is used to further group the objects, that is, to develop the smaller subclusters within each of the density-based clusters. *k*-means groups objects into *k*-clusters and each object is assigned to the cluster with the closest cluster center. The cluster centers are optimized to minimize the overall distance between the centers and objects.

There is some criticism of *k*-means clustering; primarily it can lead to clusters of approximately equal size (even if this is visually unwarranted), and it can neglect the natural boundaries, or lack thereof, of clusters (Jain 2010). The result is that some visually obvious clusters are divided. Despite this, wthe *k*-means is deemed appropriate here because it is established a priori that the clusters from which the *k*-means algorithm starts are spatially similar based on density. Here, the goal is to add smaller clusters to the dendrogram. That is, the process starts from the intrinsic, density-based clusters identified with the approach above and creates further subdivision to allow performance of multiscale modeling at various granular levels.

Then *k*-means are recursively called for all clusters that are leaves in the dendrogram, recalling that new clusters form new leaves. Termination occurs once the number of objects in the cluster is less than $k$. In this case, the choice for parameter $k$ must balance two factors. A larger $k$ is more likely to create accurate clusters. In fact, $k$ equal to the number of objects always guarantees optimal clustering, because each object is in its own cluster. However, this does not allow for any aggregation through multiscale modeling and defeats the purpose of clustering. Conversely, a smaller $k$ allows for more partitioning—meaning more granularity and levels for the multiscale model.

## Multiscale Modeling

Multiscale modeling has a wide variety of application areas—from epidemiology (Louie and Kolaczyk 2006) to collecting oceanographic data (Menemenlis et al. 1997). It is useful in any situation where different instruments collect observations at different scales (e.g., satellite data for an entire region and sensor data at a specific coordinate) or where spatial data are to be aggregated in a manner that accounts for spatial inhomogeneity, as is the situation here. Ultimately, observations are combined to reduce observational uncertainty and to reveal the latent mean process.

A variety of methods exist for multiscale modeling—from explicit probabilistic multiscale models like Gaussian and Poisson mass-balance multiscale models to multiscale decomposition models like convolutions and wavelets. Each method has an appropriate usage and is context dependent. This paper focuses on Bayesian mass-balance multiscale modeling.

Mass-balance multiscale modeling was originally developed in Kolaczyk and Huang (2001) as a generalization of a Gaussian multiscale model to include both a variable number of descendants in the dendrogram and Poisson count observations. The objective is to use the observations to estimate the latent mean process within each spatial cluster. In this case, the data consist of strictly nonnegative count events, making a Poisson distribution a reasonable distribution for the latent process. Using a Poisson distribution also helps to prevent the oversmoothing that can occur if a Gaussian distribution is assumed instead (Ferreira and Lee 2007).

The reader is referred to Ferreira and Lee (2007) for an in depth review of mass-balance multiscale modeling. Here, a general overview is provided of the terminology and assumptions from Kolaczyk and Huang (2001) and in Fig. 5 provide a flowchart of their algorithm. A multiscale model starts from an existing dendrogram with $L$ levels, here developed through the clustering algorithms, with level 1 being the coarsest level and level $L$ the finest (i.e., census tract level). Each level $l$ has $n_l$ clusters (or nodes), and all clusters at levels higher than $L$ have $m_{lj}$ descendants, $D$, where $m_{lj}$ is greater than or equal to one. If a cluster has one descendant, i.e., $m_{lj} = 1$, the descendant cluster is identical to the parent cluster.

Observations are made only at the finest resolution, $L$, though every cluster $j$ at level $L$, has an observation, $y_{L,j}$. Because the predictions are counts of events, the distribution of these observations, $y_{L,j}$, is assumed to be Poisson, e.g., $y_{L,j} \sim \text{poisson}(\mu_{L,j})$, for all $L$ and $j$. This then becomes the likelihood function for the explicit Bayesian model. Explicit, though noninformative, priors are used at each level, and the full probability model used is specified in Fig. 4. The objective is to determine the mean of the latent Poisson process $\mu(s):s \in S$, where $S \subset \mathbb{R}^k$ is the domain, for each cluster in the dendrogram and $\mu_{L,j}$, the mean of the latent process at the finest level, in particular. In this case of spatial predictions, $k$ is 2 and $\mu_{L,j} = E(y_{L,j}), j = 1, \ldots, n_L$. The parameter $\gamma_{lj}$ is used to control the degree to which clusters that are descendants from the same

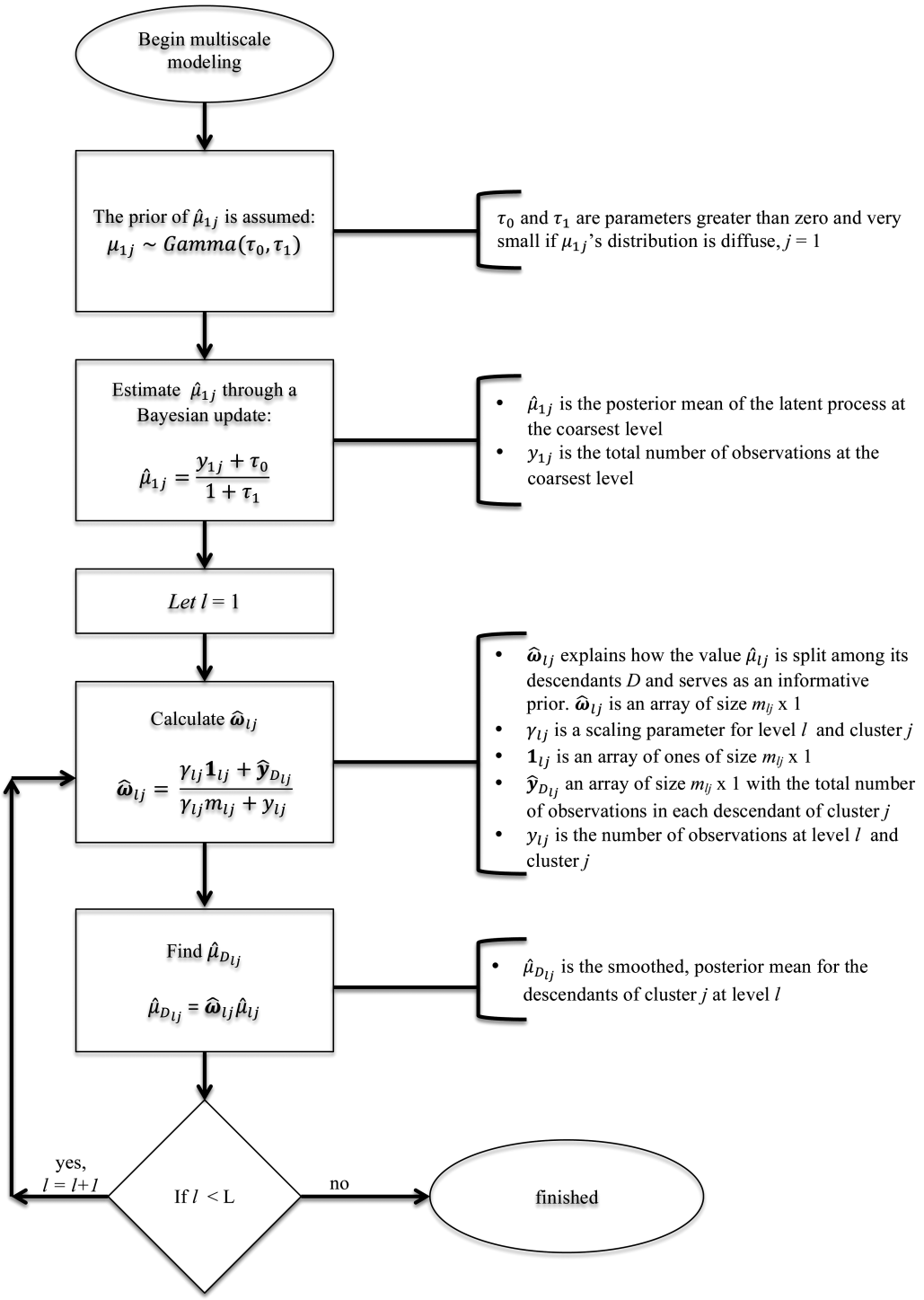


**Fig. 5.** Mass balanced multiscale model algorithm based on Ferreira and Lee (2007)

node are smoothed at each level $l$. A larger $\gamma_{lj}$ produces more homogenous results. The algorithm for finding $\hat{\mu}_{l,j}$, as described in Ferreira and Lee (2007) is presented in Fig. 5.

## Case Study

Guikema et al. (2013) developed a statistical learning theory model to estimate power outages from hurricanes at the census tract level for storms making landfall anywhere along the U.S. coastline. This model captured spatial information through the covariates used as the basis of the prediction. However, the model is simplified relative to previous work (e.g., Han et al. 2009a, b; Nateghi et al. 2013) in that a simpler set of variables is used to restrict the model to only publicly available information. There is spatial information not captured by the model, leading to some adjacent regions having seemingly discontinuous predictions. This study shows here how the intrinsic spatial structure of the region can be used to smooth the output and to aggregate the output for regional assessments.

The model's output for Hurricane Ivan is examined. Hurricane Ivan was a Category 5 hurricane that struck the Gulf Coast and produced more than $18 billion USD (in 2004 dollars) in damage

and losses. Predictions at the census tract level for the number of customers expected to be without power because of Hurricane Ivan, were developed based on the Guikema et al. (2013) model. These census tracts, 904 in total, are entirely located in one state along the Gulf Coast and comprise a large fraction of the total number of census tracts in the state, but not all of them. Along with the outage prediction at the census tract level, with the coordinates, in latitude and longitude, of the centers of the census tracts. These center coordinates that are used in the clustering. Fig. 6 shows a subset of these census tracts. They are shaded based on the percentage of customers predicted to be without power in each census tract prior to any spatial smoothing. No outage predictions exist for approximately 20% of the census tracts attributable to a lack of input data for the predictive model. These tracts are filled with diagonal lines.

The percentage of the population in each census tract without power is reported as the measure rather than the number of people without power owing to the differing populations in census tracts. This avoids highlighting population density rather than the severity of the outages. However, this study uses a mass-balance multiscale model based on the predicted *number* of people without power, and then is converted back to fraction without power. This will preserve the observation counts and guarantee the underlying data are being used consistently.

Fig. 6 clearly shows that outages are expected to be extensive. There appears to be little tapering of outages traveling north from the coast. Some outage rates seem inconsistent given surrounding census tract predictions. For example, the census tract encircled with the dotted line in Fig. 6 is predicted to have approximately 35% of its customers without power, whereas two of its neighboring census tracts are predicted to have more than 70% of their customers without power—a seeming inconsistency. Or take for instance, the set of census tracts encircled by the white solid line in Fig. 6. Within this circle lies a wide distribution: one census tract with less than 10% of customers without power, one with 15%, two with 25%, three with 40%, one with 65%, two with 75%, and finally one with more than 90% of customers without power. Although outages may be attributable to features intrinsic to these regions, even this level of variability is unexpected. Spatial smoothing allows for predictions that provide more regional consistency.

*OPTICS* groups census tracts based on geographic size—urban census tracts tend to be smaller than rural census tracts leading to urban census tracts to be more densely grouped. The optimal neighborhood radius, $\varepsilon$, is the smallest distance that allows most objects to have a defined reachability-distance. All objects that have a defined reachability-distance, regardless of how large, are members of the same cluster at the top level. Those that are undefined are considered noise and are placed as a cluster of one census tract. For this data set, setting $\varepsilon$ to 8 miles allowed 99.9% of objects to be clustered together at the top level.

When extracting the hierarchy within the density clusters two parameters were chosen. The first compares the average reachability-distance of a potential cluster to the reachability-distance of the object that divides this cluster from another. Sander et al. (2003) recommends a ratio less than 0.75, and this seemed reasonable for this data set. When this ratio is increased, more clusters are permissible. The second parameter determines to which cluster the two new clusters will point—either to the cluster that formed the new clusters or to that cluster's parent cluster. To do this, the standard score of the reachability-distances of objects contained within the new clusters is identified. Then let $\beta$ be 1, and if the score is within $\beta$, the new clusters bypass the cluster that created them, and they point to the creating cluster's parent cluster. If $\beta$ is reduced, new clusters are less likely to bypass the cluster that creates them, and the dendrogram will have more levels.

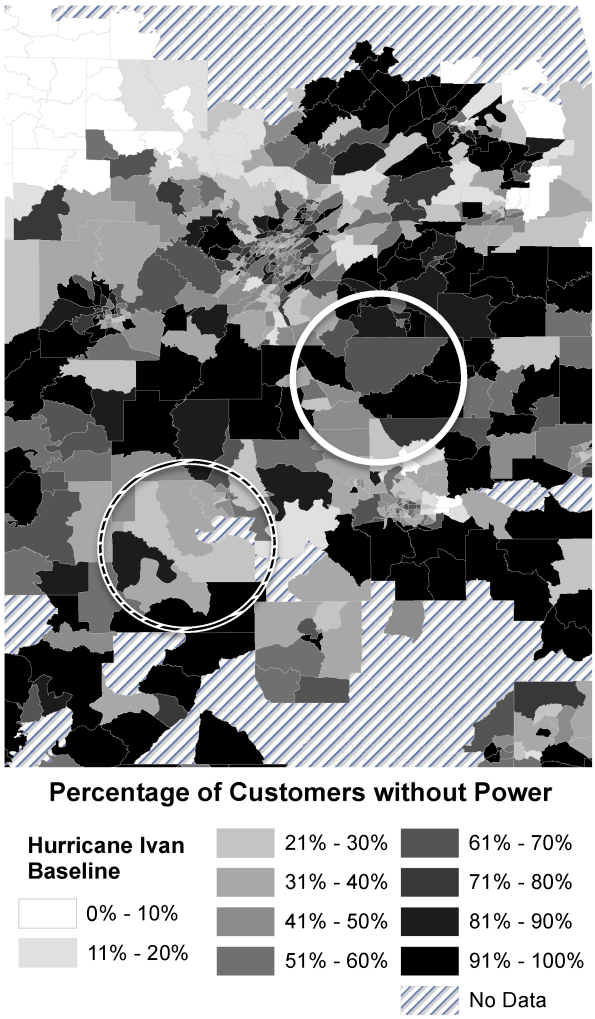


**Fig. 6.** Baseline estimates for percentage of customers without power for Hurricane Ivan; these estimates are made prior to multiscale modeling

The $k$-means clustering requires parameter $k$—the number of clusters created during each iteration. $k$ is a somewhat arbitrary parameter, which iteratively divides the objects in the clusters into $k$ groups based on distance. Setting $k$ to be equal to six balances computational requirements, cluster accuracy, and granularity of the results. This forms a six-branch tree below the initial tree formed from the density-clustering.

The final parameter that is selected is $\gamma_{lj}$, what is called the scaling parameter for the multiscale model. This parameter controls the level of smoothing that the multiscale model performs at level $l$ for cluster $j$. A value close to zero performs almost no smoothing, whereas a large value creates dramatic smoothing. Also, $\gamma_{lj}$ has a greater impact on census tracts with relatively fewer observations than census tracts with more observations. The authors let $\gamma_{lj} = 10$ to be equal for all levels and clusters, though there could be level-specific values used. Because absolute values were used when performing the multiscale modeling and later report the relative values, if the $\gamma_{lj}$ that is selected for a particular census tract is too large relative to the numbers of customers without power in that census tract, it may result in the reporting of a relative number that is unusually high.

The dendrogram for this region created a tree with six levels and a total of 432 clusters. The largest cluster is at the top of the

dendrogram, with all 903 census tracts contained within, whereas the smallest cluster, a leaf node, has only one census tract.

As mentioned earlier, Fig. 6 shows the baseline percentage of customers predicted to be without power. Fig. 7 shows the percentage of customers without power at Level 6 of the dendrogram, the level with the finest granularity. Fig. 7 appears similar to Fig. 6—the outages appear to be widespread, and there is relatively little decrease in outage rate with increasing distance from the coastline. However, there is less localized variability. Take for instance the solid white circle in Fig. 7, which is the same region circled in Fig. 6. In Fig. 7, there are fewer differences in the color gradation among adjacent census tracts. Whereas some of these census tracts now have less than 20% of customer without power, they are adjacent to other census tracts with similar prediction. Likewise, census tracts with relatively many customers without power are adjacent to like census tracts.

The goal of multiscale modeling is not to make localized regions have the same prediction; it is to allow spatially similar census tracts to inform one another. It permits census tracts with many predicted outages to exploit input from census tracts with few predicted outage, and vice versa, so that they both may be more accurate. Fig. 8 illustrates how most census tracts were not dramatically affected. Fig. 8 shows the absolute difference between baseline predictions and those at Level 6. Generally, the predicted outage adjustments were minor; only 23% of all census tracts had adjustments greater than 20%, and only 3% of census tracts were adjusted by 50% or more. Of these census tracts, 40% have fewer than 50 customers. It is possible that $\gamma_{lj} = 10$ is too large for these census tracts, and provides too large an adjustment.

This perhaps is seen most dramatically in the sole census tract in the northwestern portion of the state predicted to have more than 90% of customers without power at Level 6. It is predicted to have fewer than 10% of its customer without power from the baseline assessment. (In actuality, 30% of the customer-base lost power.) This census tract has only 11 customers, is clustered only with itself at Level 6. $\gamma_{lj} = 10$ provides too large of an adjustment for this particular tract.

Another feature of using multiscale modeling and clustering analysis is the ability to view aggregated data. Fig. 9 shows projected outputs at Level 3 of the dendrogram. At this level, many urban areas are considered spatially similar based on spatial density, and hence are located in the same cluster and have the same predictions. For example, the urban area of approximate population 200,000 and circled in Fig. 9 is aggregated into one cluster. Consequently, one can derive, for this area, a metropolitan-wide prediction for outages—approximately 60%. Notewworthy also is that at Level 3, a few census tracts form their own individual cluster. These census tracts tend to be large and in spatially sparse areas.

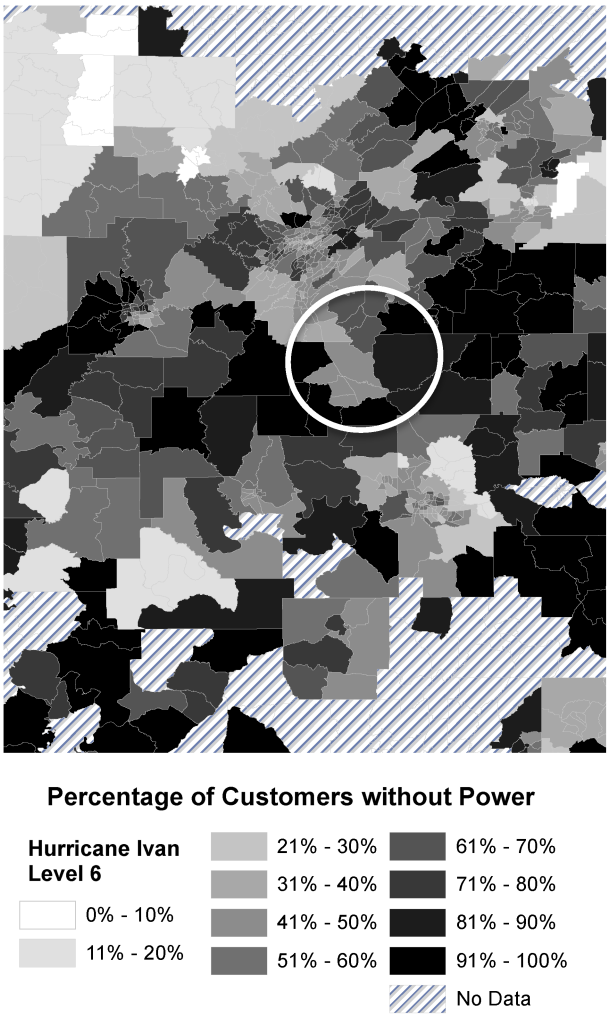


**Fig. 7.** Percentage of customers without power at Level 6 of clustering and after multiscale modeling; Level 6 is the finest level of clustering

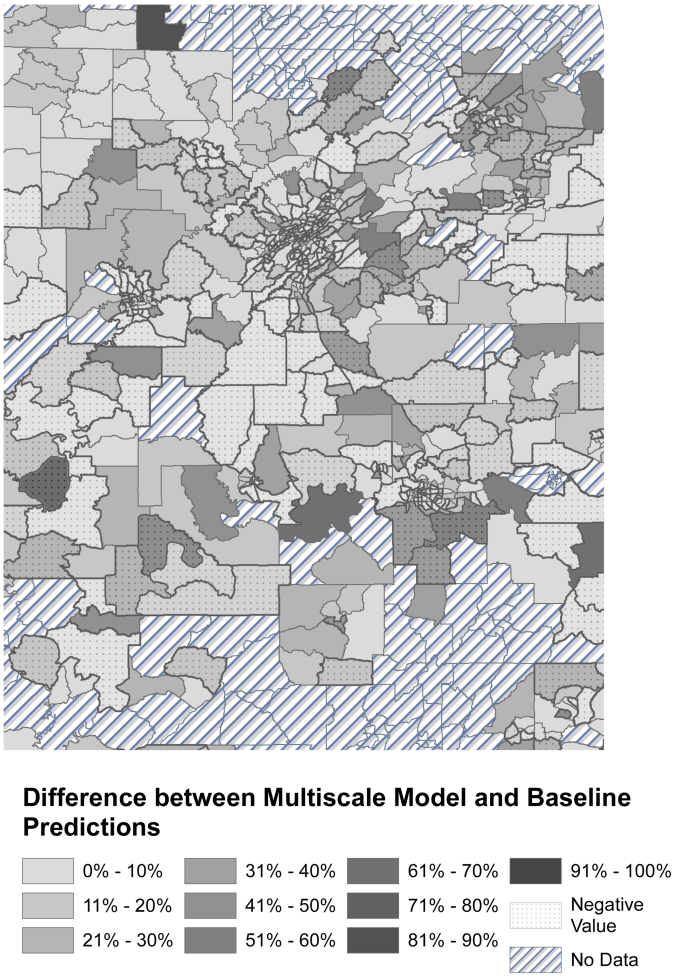


**Fig. 8.** Absolute difference between baseline predictions and those from Level 6

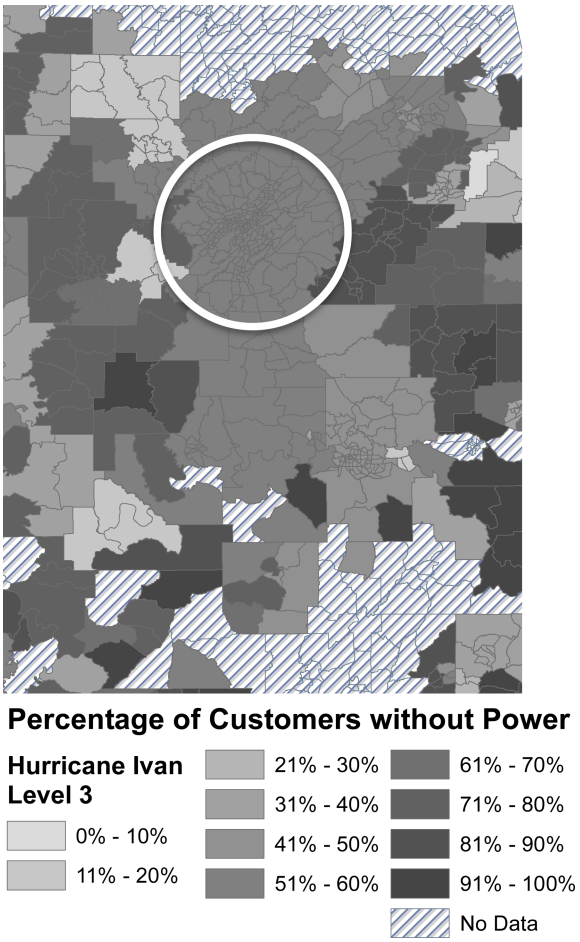


**Fig. 9.** Percentage of customers without power at Level 3 of clustering and after multiscale modeling; Level 3 among the coarser aggregation levels; the circle highlights an urban area which is clustered into one cluster

### *Validation*

To measure improvement in predictive accuracy, if any, produced by using multiscale modeling, the investigators compare the baseline predictions and the predictions derived from multiscale modeling at the most granular level to the observed outages for Hurricane Ivan and for four other storms—Hurricanes Katrina, Dennis, Georges, and Danny—that impacted the area. Two metrics were used for comparison: (1) the root mean square value, (i.e., the sum of the squared differences between the actual observations and the predicted values), and (2) the empirical cumulative distribution function. Generally the performance improves, and the improvement is typically dependent on $\gamma_{lj}$. See first the root mean squared value in Table 1.

Table 1 shows that the accuracy of the predictions are dependent on the parameter $\gamma_{lj}$. This is relevant, especially because it is impossible to know which $\gamma_{lj}$ is best for a particular set of conditions without comparing the predictions with the actual event. Generally, though, it was found that $\gamma_{lj}$ is best when it provides some smoothing, but not so large that the predictions are essentially uniform. Table 1 also shows that the accuracy of the predictions made by multiscale modeling are most dependent on the accuracy of the baseline predictions. If the baseline predictions possess inaccuracies, so will the multiscale modeling predictions. The goal is to be less inaccurate.

**Table 1.** Comparison of Multiscale Modeling Predictions at the Most Granular Level, Level 6, to the Baseline Predictions for Various Values of Smoothing Parameter $\gamma_{lj}$

| Metrics for comparison | Hurricane | | | | |
|---|---|---|---|---|---|
| | Danny | Dennis | Georges | Ivan | Katrina |
| Root mean squared error, level 6 | | | | | |
| Baseline | 0.135 | 0.354 | 0.234 | 0.455 | 0.405 |
| $\gamma_{lj} = 5$ | 0.133 | 0.347 | 0.264 | 0.423 | 0.380 |
| $\gamma_{lj} = 10$ | 0.144 | 0.345 | 0.206 | 0.423 | 0.379 |
| $\gamma_{lj} = 50$ | 0.201 | 0.348 | 0.237 | 0.422 | 0.401 |
| % reduction more than baseline | | | | | |
| $\gamma_{lj} = 5$ | 1.6 | 1.8 | −12.7 | 7.2 | 6.3 |
| $\gamma_{lj} = 10$ | −6.7 | 2.3 | 12.1 | 7.1 | 6.4 |
| $\gamma_{lj} = 50$ | −49.2 | 1.5 | −1.4 | 7.2 | 1.0 |

Note: $l = 1, \ldots, L$, $j = 1, \ldots, n_L$. The root mean squared value is shown in rows 3–6 and is the sum of the squared differences between the prediction and the actual. Rows 7–9 show the reduction in root mean squared error over the baseline prediction. A positive number indicates a decrease in error.

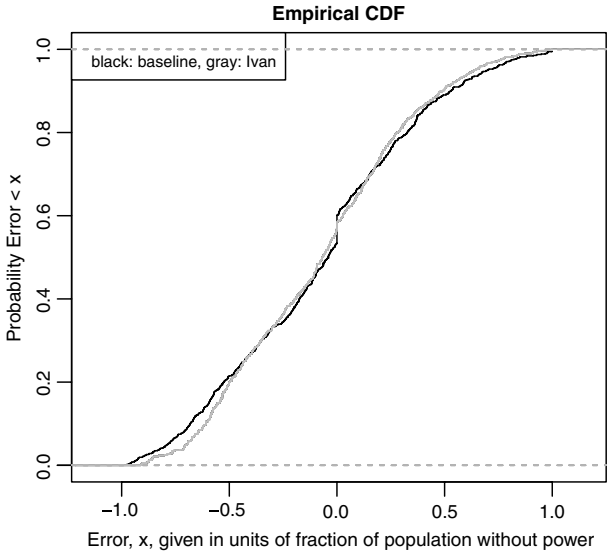


**Fig. 10.** The empirical CDF comparing the baseline predictions with the most granular level of predictions offered by multiscale modeling (Level 6) for Hurricane Ivan

Another method by which to examine possible improvement derived from multiscale modeling is the empirical cumulative distribution function (CFD) plots of the residuals, i.e., the prediction errors. The empirical CDF for Hurricane Ivan is shown in Fig. 10. As Fig. 10 shows, the improvement is slightly concentrated in the error distribution's tails. That is, the multiscale modeling approach helps improve the predictions most for those census tracts with particularly high or particularly low predicted outage rates.

## Conclusion

This paper has developed a novel approach for modeling predictions of count events such as power outages at multiple spatial scales and for spatially smoothing the predictions at any given scale. This involves combining an improved version of *OPTICS*-based clustering with *k*-means clustering and Bayesian multiscale modeling. This is novel primarily in the combination of the

techniques but also in the advance made within the *OPTICS* method. This paper shows how this new approach can improve infrastructure performance assessments. In the case of electric-power outages, improved predictions inform utility operators how to preposition assets and to assess what resources are necessary to provide power quickly again. Informing consumers of the likelihood of outages in their area may allow for better planning prior to the storm. The same model can be used to examine other spatial infrastructure performance estimates, including predictions of pipe breaks in urban areas, communication delays in communications systems, and road closures attributable to weather events. Like the electric-power case, the prediction improvements in these cases allow for users, owners, and operators to better plan and prepare for service disruptions.

Whereas this paper shows enhancements over previous methods, some improvements are possible. First, clustering algorithms and multiscale models rely on recursive techniques. Although the authors did not experience computational limits, this may occur for larger geographic areas given the memory intensiveness of recursions. A nonrecursive heuristic to approximate clusters may suffice, especially in this type of application. Second, incorporating regional and temporal data (e.g., predicted peak wind speeds, tree conditions) into a higher dimensional clustering and smoothing may improve predictions further. The approach developed here has potential use across a number of different types of infrastructures, particularly by those using performance models to help in managing infrastructure systems.

## Acknowledgments

This work is funded by grants from the National Science Foundation (NSF), grants 0968711 and 1149460. The support of the sponsor is gratefully acknowledged. Any opinions, findings, conclusions or recommendations presented in this paper are those of the authors and do not necessarily reflect the view of the National Science Foundation.

## References

Achtert, E., Kriegel, H. P., and Zimek, A. (2013). "Example data sets for ELKI." *ELKI: Environment for developing KDD applications supported by index structures*, ⟨http://elki.dbs.ifi.lmu.de/wiki/DataSets⟩ (Apr. 28, 2014).

Alon, U., et al. (1999). "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays." *Proc. National Acad. Sci.*, 96(12), 6745–6750.

Ankerst, M., Breunig, M. M., Kriegel, H.-P., and Sander, J. (1999). "OPTICS: Ordering points to identify the clustering structure." *ACM SIGMOD Rec.*, 28(2), 49–60.

Broder, A. Z., Glassman, S. C., Manasse, M. S., and Zweig, G. (1997). "Syntactic clustering of the web." *Comput. Networks ISDN Syst.*, 29(8), 1157–1166.

Cook, T. D. (2005). "Emergent principles for the design, implementation, and analysis of cluster-based experiments in social science." *ANNALS Am. Acad. Political Soc. Sci.*, 599(1), 176–198.

Ester, M., Kriegel, H.-P., and Xu, X. (1996). "A density-based algorithm for discovering clusters in large spatial databases with noise." *Proc., Second Int. Conf. on Knowledge Discovery and Data Mining*, AAAI, Palo Alto, CA, 226–231.

Fathian, M., Amiri, B., and Maroosi, A. (2007). "Application of honey-bee mating optimization algorithm on clustering." *Appl. Math. Comput.*, 190(2), 1502–1513.

Ferreira, M. A., and Lee, H. K. (2007). *Multiscale modeling: A Bayesian perspective*, Springer, New York, NY.

Frey, B. J., and Dueck, D. (2007). "Clustering by passing messages between data points." *Science*, 315(5814), 972–976.

Guikema, S. D., Nateghi, R., and Quiring, S. (2013). "Predicting infrastructure loss of service from natural hazards with statistical models: Experiences and advances with hurricane power outage prediction." *European Safety and Reliability Conf. (ESREL)*, ESRA, Amsterdam, Netherlands.

Han, S.-R., Guikema, S. D., and Quiring, S. M. (2009a). "Improving the predictive accuracy of hurricane power outage forecasts using generalized additive models." *Risk Anal.*, 29(10), 1443–1453.

Han, S.-R., Guikema, S. D., Quiring, S. M., Lee, K.-H., Rosowsky, D., and Davidson, R. A. (2009b). "Estimating the spatial distribution of power outages during hurricanes in the Gulf coast region." *Reliab. Eng. Syst. Safety*, 94(2), 199–210.

Hillhouse, J. J., and Adler, C. M. (1997). "Investigating stress effect patterns in hospital staff nurses: Results of a cluster analysis." *Soc. Sci. Med.*, 45(12), 1781–1788.

Jain, A. K. (2010). "Data clustering: 50 years beyond *K*-means." *Pattern Recognit. Lett.*, 31(8), 651–666.

Kolaczyk, E. D., and Huang, H. (2001). "Multiscale statistical models for hierarchical spatial aggregation." *Geograph. Anal.*, 33(2), 95–118.

Liu, H., Davidson, R. A., and Apanasovich, T. V. (2007). "Statistical forecasting of electric power restoration times in hurricanes and ice storms." *IEEE Trans. Power Syst.*, 22(4), 2270–2279.

Louie, M. M., and Kolaczyk, E. D. (2006). "A multiscale method for disease mapping in spatial epidemiology." *Stat. Med.*, 25(8), 1287–1306.

Menemenlis, D., Fieguth, P., Wunsch, C., and Willsky, A. (1997). "Adaptation of a fast optimal interpolation algorithm to the mapping of oceanographic data." *J. Geophys. Res.*, 102(C5), 10573–10584.

Moser, F., Ge, R., and Ester, M. (2007). "Joint cluster analysis of attribute and relationship data without a-priori specification of the number of clusters." *Proc., 13th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, AMC, New York, NY, 510–519.

Nateghi, R., Guikema, S. D., and Quiring, S. M. (2011). "Comparison and validation of statistical methods for predicting power outage durations in the event of hurricanes: Comparison and validation of statistical methods." *Risk Anal.*, 31(12), 1897–1906.

Nateghi, R., Guikema, S. D., and Quiring, S. M. (2013). "Power outage estimation for tropical cyclones: Improved accuracy with simpler models." *Risk Analysis*, 10.1111/risa.12131.

Sander, J., Qin, X., Lu, Z., Niu, N., and Kovarsky, A. (2003). "Automatic extraction of clusters from hierarchical clustering representations." *Advances in knowledge discovery and data mining*, Springer, Berlin, Heidelberg, 75–87.

Wang, W., Yang, J., and Muntz, R. (1997). "STING: A statistical information grid approach to spatial data mining." *Proc., 23th Int. Conf. on Very Large Data Bases*, Morgan Kaufmann Publishers, San Francisco, CA, 186–195.

Winkler, J., Dueñas-Osorio, L., Stein, R., and Subramanian, D. (2010). "Performance assessment of topologically diverse power systems subjected to hurricane events." *Reliab. Eng. Syst. Safety*, 95(4), 323–336.