UCLA
Technical Reports

Title

Data Modeling and Synthetic Data Generation For Fine-Grained Networked Sensing

Permalink

https://escholarship.org/uc/item/3s75m359

Authors

Yan Yu
Deepak Ganesan
Lewis Girod
et al.

Publication Date

2003

Synthetic Data Generation to Support Irregular Sampling in Sensor Networks

Yan Yu, Deepak Ganesan, Lewis Girod, Deborah Estrin, Ramesh Govindan

November 27, 2003

Abstract

Sensor networks have drawn much attention because of their promising applications in environmental monitoring, seismology, and military surveillance. Despite increasing interest, sensor network research is still in its initial phase. Few real systems have been deployed and little data is available to test proposed protocol and data management designs. Most sensor network research to date uses randomly generated data input to simulate their systems. Some researchers have proposed using environmental monitoring data obtained from remote sensing or in-situ instrumentation. In many cases, neither of these approaches is relevant, because they are either collected from regular grid topology, or too coarse grained. This paper proposes to use synthetic data generation techniques to generate irregular data topology from data sets measured on a grid.

To tackle this problem, we investigate the use of the available sparsely sampled data sets, model the spatio-temporal correlation in these data sets, and generate irregular topology data based on empirical models of the experimental data. Our goal is to more realistically evaluate sensor network system designs before large scale field deployment. In obtaining these synthetic data sets, we draw heavily on techniques developed in geo-statistics and other spatial interpolation techniques, but appropriately modify them for the application at hand.

Our evaluation results on the radar data set of weather observations shows that the spatial correlation of the original and synthetic data are similar. Moreover, visual comparison shows that the synthetic data retains interesting properties (*e.g.*, edges) of the original data. Our case study on the DIMENSIONS system demonstrates how synthetic data helps to evaluate the system over an irregular topology, and points out the need to improve the algorithm.

1   Introduction

Despite increasing interest, sensor network research is still in its initial phase. Few real systems are deployed and little data is available to test proposed protocol designs. Most sensor network research to date uses randomly generated data input to evaluate systems. Evaluating the system with data representing real-world scenarios or representing a wide range of conditions is essential for systematic protocol design and evaluation of sensor network systems whose performance are sensitive to the spatio-temporal features of the system inputs. To our knowledge, there has been no previous work done on modeling data input in a sensor network context.

Some researchers have propose using environmental monitoring data obtained from remote sensing or in-situ instrumentation. However, these data are mostly collected from a regular grid configuration. Due to the large scale deployment, the proposed sensor networks (*e.g.*, in habitat monitoring [6]) are most likely in an irregular topology. Further, the granularity and density of those data sets does not match the expected granularity and density of future sensor network deployment. Although they cannot be directly used to evaluate the sensor network algorithms, they can, provide useful models of spatial and temporal correlations in the experimental data, which can be used to generate synthetic data sets. Because many sensor network protocols exploit spatial correlations, we are interested in synthetic data that have similar spatial correlations as that of the experimental data. In this paper we focus on modeling the experimental data to generate irregular topology data for two reasons: First, we lack ground truth data to verify that the synthetic data match some interesting statistics of the experimental data at the scale of fine granularity. Second, we cannot assume that the experimental data are generated from a band-limited spatial process.

In order to evaluate sensor network algorithms under different topologies other than the single topology associated with the available data set, we proposed to generate irregular topology data. We first apply spatial interpolation techniques, implicitly or explicitly model the spatial and temporal correlation in a data set. From this empirical model, we generate ultra fine-grained data, and then use it to generate irregular data. This technique will also allow us to study system performance under various topology, but with the same data correlation model. On the other hand, by using the same experimental data setting, and plugging in different correlation models, we are able to evaluate how the algorithms interact with various data correlation characteristics. In this paper, we use the DIMENSIONS [13] system as our case study and investigate the impact of irregular topologies on algorithm performance. DIMENSIONS provides a unified view of data handling in sensor networks, incorporating long-term storage, multi-resolution data access and spatio-temporal pattern mining. It is designed to support observation, analysis and querying of distributed sensor data at multiple resolutions, while exploiting spatio-temporal correlation. While the interplay of topology and radio connectivity has been studied in-depth in the context of sensor networks (*e.g.*, ASCENT [7], GAF/CEC [32], STEM [27] etc), there is little work on studying the interplay between in-network data processing and topology. Our models and synthetic data sets are intended to help study the coupling between the topology and data processing schemes in such networks.

In the remainder of this paper, we first review related work in section 2. In section 3, we start with how to generate fine grained spatial data maps using a model of spatial correlation as well as how to generate fine grained spatio-temporal data sets using a joint space-time model. This is an essential step in irregular data generation, which we discuss in section 4. We also present results of applying these two modeling techniques to an experimental radar data set in section 3. In section 4, we use the DIMENSIONS system as a case study to demonstrate how the synthetic data from the modeling of experimental data helps in system evaluation, and point out the need to improve the algorithm. We conclude in section 5.

2 Related work

Data modeling techniques in environmental science   To the best of our knowledge, no previous work has been done on data modeling in a sensor network context. However, in environmental science or geophysics, various data analysis techniques have been applied to extract interesting statistical features from the data. In addition, such techniques are used to estimate and interpolate the data values at un-sampled or missing data points. We have a slightly different objective: to generate fine grained synthetic data that can approximate the spatial and temporal correlation in the experimental data. Nevertheless, the data modeling techniques from the environmental sciences provide a good starting point for our data modeling purpose.

There has been extensive research on spatial and/or temporal data analysis. Various spatial interpolation techniques, such as Voronoi polygons, Triangulation, Natural neighbor interpolation, Trend Surface or Splines [30], have been proposed. Kriging, which refers to a family of generalized least-squares regression algorithms, has been used extensively in various environmental science disciplines. Kriging models the spatial correlation in the data and minimizes the estimation variance under the unbiasedness constraints of the estimator. Different Kriging estimators assume different degrees of stationarity in the data: *simple kriging* considers the mean to be known and constant throughout the study area; *ordinary kriging* relaxes the stationarity requirement by limiting the stationarity assumption only to a local neighborhood, and the local mean is deemed unknown; and *kriging with a trend model* explicitly models the trend component as a linear combination of functions of the coordinates. In this paper, we reported our experience with ordinary kriging and several non-stochastical interpolation techniques.

In addition, there is significant research devoted to time series analysis. Autoregressive Integrated Moving Average model (ARIMA) [3] explicitly considers the trend and periodic behavior in the temporal data. Because of the cycle or seasonality in the data, it is convenient to conduct the time series data analysis in the frequency domain. The wavelet model [12] has been successfully used to model the cyclic, or repeatable behavior in data. In addition, researchers have also explored neural networks [10], kernel smoothing for time series analysis.

Joint spatio-temporal models have received much attention in recent years [18, 26, 25, 20] because they inherently model the correlation between the temporal and spatial domain. The joint space-time model used in our data analysis is inspired by and simplified from the joint space-time model proposed by Kyriakidis *et al.* [19]. In [19], co-located terrain elevation values are used to enhance the spatial prediction of the coefficients in the temporal model constructed at each gauge station. However, this requires the availability of an extra environmental variable, which does not exist in our case.

Data modeling in Database and Data Mining   Theodoridis *et al.* [28] proposes to generate spatio-temporal datasets according to parametric models and user-defined parameters. However, the design space is huge, it is impossible to exhaustively visit the entire design space, *i.e.*, generate data sets for every possible set of parameter values. Without additional knowledge, we have no reason to believe that any parameter setting is more realistic or more important than others. Therefore we proposed to start with an experimental data set, and generate synthetic data that shares similar statistics with the experimental data.

Given a large data set that is beyond the computer memory constraints, data squashing [29] proposes schemes to shrink a

large data set to manageable size. Data squashing is acceptable if the queries (or the statistical modeling) have small errors on the shrinked data set compared to on the original one. Although sharing the same objective of deriving synthetic data from modeling existing data as us, they consider non-spatio-temporal data set. The spatio-temporal data can not be assumed to be drawn from the same certain probability model as assumed by [29].

TCP traffic Modeling in Internet   In a similar attempt to model the data input to the network system in an Internet context, researchers have studied TCP traffic modeling. For example, Caceres *et al.* [5] characterized and built empirical models of wide area network applications. The specific data modeling technique in their study [5] does not apply to sensor networks due to the following: (a) Sensor networks are closely coupled with the physical world, therefore the data modeling in sensor networks needs to capture the spatial and temporal correlation in a highly dynamic physical environment. (b) The characteristics of wide area TCP traffic is potentially very different from the workload or traffic in sensor networks.

System components modeling in wireless ad-hoc networks and sensor networks   Previous research has been carried out on modeling system components in ad-hoc networks and sensor networks, however, to our knowledge, none of this research has focused on modeling the data input to the system.

Among the work on modeling system components in the context of ad-hoc networks, [4, 8] use regular or uniform topology setups, and "random waypoint" models in their protocol evaluations, and [24, 15] discuss multiple topology setups and mobility patterns for more realistic scenarios. In modeling wireless channels, [23] discuss propagation models for different indoor and outdoor environments, and various fading channel models. Konrad *et al.* [16] study non-stationary behavior of packet loss in the wireless channel and modeled the GSM (Global System for Mobile) traces with a MTA (Markov based Trace Analysis) algorithm. The packet loss rate from low power radios that is used in sensor networks could be much more bursty than what is observed in a GSM trace. Therefore, it is not clear whether the MTA algorithm applies to the sensor network context.

Ns-2[2] and GloMoSim [33] provide flexibility in simulating various layers of wired networks or wireless ad-hoc networks. However, they do not capture many important aspects of sensor networks, such as sensor models, or channel models. In contrast, Sensorsim [21, 22] directly targets sensor networks. It provides basic models for various components of sensor network simulations. In addition to a few topology and traffic scenarios, they introduce the notion of a sensor stack and sensing channel. The sensor stack is used to model the signal source, and the sensing channel is used to model the medium which the signal travels through. Sensorsim also provides parametric models for the power consumption of various system components, battery capacity and battery drain rate. Our work could be used as a new model in Sensorsim.

3   Synthetic data generation based on empirical models of experimental data

Before delving into irregular topology data generation, we start with the problem of generating fine-grained synthetic data, which is an essential step in our irregular topology data generation. Our proposed synthetic data generation includes both

spatial and spatio-temporal data types. To generate spatial data, we start with an experimental data set which is a collection of data measurements from a study area. Assuming the data is a realization of an ergodic and local stationary random process, we use spatial interpolation techniques to generate synthetic data at unmonitored locations. This allows us to generate synthetic data with arbitrary granularity and arbitrary topologies, where the spatial correlation of the synthetic data approximates that of the experimental data.

Similarly, to generate synthetic spatio-temporal data, we again start with an experimental space-time data set, which includes multiple snapshots of data measurements from a study area at various times. If we were only interested in data at recording time, we could apply our proposed spatial interpolation techniques to each snapshot of data separately, then generate a collection of spatial data sets at each recording time. However, this does not allow us to generate synthetic data at times other than the recording times. In addition, the joint space-time correlation is not fully modeled and exploited if we model each snapshot of spatial data separately. Therefore, we propose to model the joint space-time dependency and variation in the data. Inspired by a joint space-time model in [19], we model the data as a joint realization of a collection of space indexed time series, one for each spatial location. Time series model coefficients are space-dependent, and so we further spatially model them to capture the space-time interactions. Synthetic data are then generated at unmonitored locations and time from the joint space-time model. This allows us to generate synthetic data at arbitrary spatial and temporal configurations.

In the remainder of this section, we first discuss spatial interpolation techniques and present the results of apply them to a radar data set. Then we discuss a joint spatio-temporal model and the result of applying it to the same radar data set.

3.1 Generating Synthetic Spatial Data Sets

We start with an experimental data set, which is typically sparsely sampled. To generate a large set of samples at much finer granularity, a spatial interpolation algorithm is used to predict at unsampled locations. The spatial interpolation problem has been extensively studied. Both stochastic and non-stochastic spatial interpolation techniques exist, depending on whether we assume the observations are generated from a stochastic random process. In general, the spatial interpolation problem can be formulated as: Given a set of observations $\{z(k_1), z(k_2), ..., z(k_n)\}$ at known locations k_i, $i = 1, ..., n$, spatial interpolation is used to generate prediction at an unknown location u. However, if we take a stochastic approach, the above spatial interpolation problem can be formulated as the following estimation problem. A random process, Z, is defined as a set of dependent (here spatially dependent) random variables $Z(u)$, one for each location u in the study area A, denoted as $\{Z(u), \forall u \in A\}$. Assuming Z is an ergodic process, the problem is defined to estimate some statistics (*e.g.*, mean) of $Z(u)$ ($u \in A$) given a realization of $\{Z(u)\}$ at locations u_i, $i = 1, ..., n$, $u_i \in A$. A lies in one dimensional or high dimensional space.

Kriging [14] is a widely used geostatistics technique to address the above estimation problem. Kriging, which is named after D. G. Krige [17], refers to a range of least-squares based estimation techniques. It has both linear and non-linear forms. In this paper, ordinary kriging, which is a linear estimator, is used in our spatial interpolation and joint spatio-temporal modeling example.

In ordinary kriging, at an unmonitored location, the data is estimated as a weighted average of the neighboring samples.

There are different ways to determine the weights, *e.g.*, assign all of the weight to the nearest data, as used in the nearest neighbor interpolation approach; assign the weights inversely proportional to the distance from the location being estimated. Assuming the underlying random process is locally stationary, Kriging uses a variogram[1] to model the spatial correlation in the data. The weights are determined by minimizing the estimation variance, which is written as a function of the variogram (or covariance). The primary factors that affect the Kriging weights are: closeness of the sample locations to the location being estimated, redundancy among the data, and the variogram model. In addition to providing least squares based estimate, Kriging also provides estimation variance, which is one of the important reasons that Kriging has been popular in geostatistics. However, as we will explain shortly, estimation is not our ultimate goal; our goal is to generate fine grained sensing data which can be used to effectively evaluate sensor network protocols. Therefore we also study other non-stochastic spatial interpolation algorithms: Nearest neighbor interpolation, Delaunay triangulation interpolation, Inverse-distance-squared weighted average interpolation, BiLinear interpolation, BiCubic interpolation, Spline interpolation, and Edge directed interpolation [11].

Most interpolation algorithms estimate the value at an un-monitored location as a weighted average of the neighboring observations. They differ in the number of neighbors involved in the interpolation of a point and the manner used to determine the weight of each neighbor. Nearest neighbor assigns all the weight to the nearest neighbor. In Delaunay triangulation, the entire network is triangulated based on the locations of all observations and, on any triangle, the predictor is a planar interpolant of the surrounding three data values. In Inverse-distance-squared weighted average, neighboring data values are weighted according to how far they are from the location being estimated. Bilinear interpolates at the unmonitored location as a linear average of its 4 closest neighbors, first interpolate along each row, then interpolate along each column. In contrast, Bicubic interpolation is a weighted average of its 16 nearest neighbors, weights are a cubic function of the relative location from the unknown point to each neighbor. Edge-directed interpolation [11] was originally designed to increase an image's resolution. Its basic case is to increase the resolution of an image in each spatial dimension by a factor of 2, and it can be applied recursively. It uses 4th-order linear interpolation, *i.e.*, the weighted sum of its 4 diagonal neighbors. Under the assumption of local stationary Gaussian process, according to the classical Weiner filtering theory, the MSE linear interpolation coefficients are given by $\alpha = inv(R) * r$. where α is a vector of weights, R (4x4 matrix) and r (1x4 vector) are local covariance of high resolution image. Again, under the assumption of local stationary Gaussian process, high resolution covariance is a function of its corresponding low resolution covariance, which in turn can be computed from the observation.

Spline is a piecewise polynomial function that can have a locally very simple form, yet at the same time be globally smooth. Cubic Spline, which is used in our study, is a piecewise cubic function, and the first and second derivative of the piecewise cubic functions exist at internal sample points (*i.e.*, the piecewise function is continuous and smooth).

---

[1] Please refer to Appendix A for a brief introduction to variogram, and appendix B for an introduction of ordinary kriging.

3.1.1 Evaluation of synthetic data generation

Data set description   To apply the spatial interpolation techniques described above, we consider the resampled S-Pol radar data provided by NCAR[2], which records the intensity of reflectivity in dBZ, where Z is proportional to the returned power for a particular radar and a particular range. The original data were recorded in the polar coordinate system. Samples were taken at every 0.7 degrees in azimuth and 1008 sample locations (approximately 150 meters between neighboring samples) in range, resulting in a total of 500 x 1008 samples for each 360 degree azimuthal sweep. They were converted to the Cartesian grid using the nearest neighbor resampling method. A grid point is only assigned a value from a neighbor when the neighbor is within 1km and 10 degree range. If none of its neighbors are within this range, the grid point is labeled as missing value, *e.g.*, the NaN value is assigned. Resampling, instead of averaging, was used to retain the critical unambiguous and definitive differences in the data. In this paper, we select a subset of the data that has no missing values to perform our data analysis. Specifically, each snapshot of data in our study is a 60 x 60 spatial grid data with 1 km spacing.

Spatial interpolation algorithms implementation   We apply the above-mentioned eight interpolation algorithms to the selected spatial radar data sets. We use the *spatial* package in R [1] to achieve Kriging. Nearest neighbor, Bilinear, Bicubic, Spline interpolation results were obtained from the interp2() function in Matlab. Since Bilinear and Bicubic interpolations provide no prediction for edge points, we use results from Nearest Neighbor interpolation for edge points in bilinear or bicubic interpolation. Edge directed interpolation is based on [11]. Inverse-distance-squared weighted average interpolation, and Delaunay triangulation interpolation were implemented in Matlab following the interface of interp2(). The *s*patial package in R and the interp2() function in Matlab generate output for a grid region. This motivates us to use the resampled grid data, instead of the raw data from the polar coordinate system.

Evaluation metrics   Having obtained fine-grained synthetic data from different interpolation algorithms, we need a metric to help us evaluate whether the synthetic data have desirable features to test sensor network algorithms. For our synthetic data generation, we are interested in how close the synthetic data can approximate the interesting statistical features of the original data. The set of statistical features selected as evaluation metrics should be of interest to the algorithm and applications for which the synthetic data are intended to be used. It is hard to define a statistical feature set that is generally applicable to most algorithms and data sets, nevertheless, quite a few existing sensor network protocols (including DIMENSIONS, which is used as our case study) exploit spatial correlations in the data. In general, since sensor networks are envisioned to be deployed in the physical environment and deal with data from the geometric world, we believe that many sensor network protocols will exploit spatial correlation in the data. Therefore, besides visual comparison, we use spatial correlation (which is measured by its variogram values) of the synthetic data versus original data to assess the applicability of this synthetic data generation technique to the sensor network algorithm being evaluated. Suppose two data sets A and B, and their variogram values are

---

$\{\hat{\gamma}_1(h_i)\}$ and $\{\hat{\gamma}_2(h_i)\}$ respectively, where h_i are sample separation distances between two observations; i =1, ..., m. The Mean Square Difference of variogram values of two data sets is defined as: $\sum_{i=1}^{m}(\hat{\gamma}_1(h_i) - \hat{\gamma}_2(h_i))^2$. Next we use DIMENSIONS as an example to support our earlier claim that the spatial correlation of a data set affects the performance of the protocol being evaluated. DIMENSIONS is based on wavelet compression. We apply wavelet compression (using package provided by [9]) to the synthetic data sets, which had been generated from the above interpolation algorithms based on 100 snapshots of radar data. Each point in Figure 1 represents a synthetic data set generated by applying an interpolation algorithm to a spatial data set. The x-axis is the Mean Square Difference in variogram values between the synthetic data and the original data and the y-axis is the absolute difference between the MSE of applying wavelet compression on this synthetic data and on the original data. The majority of the data in Figure 1 indicates a trend of that the synthetic data set with closer (*i.e.*, smaller Mean Square Difference in) variogram values relative to the original data set often suggest that its Mean Square Error from applying wavelet compression is also closer to that of the original data.

Interpolation resolution We studied two extremes of interpolation resolutions: (1) Coarse grained interpolation, in which case, we start from the down-sampled data (which reduces the data size in half in each dimension), increase the interpolation resolution by 4, compare the variogram value of the interpolated data with that of the original data. Note that the original data can be considered as ground truth in this case. The coarse grained interpolation is used to evaluate how the synthetic data generated by different interpolation algorithms approximate the spatial correlation of the experimental data. (2) Fine grained interpolation. Starting with a radar data set with 1km spacing, we increase the resolution by 10 times in each dimension, resulting in a 590x590 grid with 100m spacing. Fine grained interpolation is an essential step in generating irregular topology data. Based on a snapshot of S-Pol radar data (a 60 x 60 grid with 1 km spacing) shown in Figure 2, Figure 3 and 4 are presented here as two examples of fine-grained interpolation using the ordinary kriging and the nearest neighbor interpolation respectively. The interpolation result from applying the nearest neighbor is visually very similar to the original one; whereas the Kriging result is a smoothed version of the original data, however, it maintains interesting features (such as the edges) of the original data.

Evaluation Results First we visually present how the spatial correlation (*i.e.*, variogram values) of the synthetic data approximates that of the original data in the case of coarse-grained interpolation. Figure 5 shows the variogram plot of several synthetic data sets *vs.* the original data for one single snapshot. It demonstrates that the variogram curves of most synthetic data (except the one from Inverse-distance-squared weighted average interpolation) closely approximate that of the original one. At the long lag distances, the synthetic data may appear slightly under-estimating the long-range dependency in the original data. The source of this under-estimate may be due to the smoothing effect of the interpolation algorithms.

Further, we use the Mean Square Difference between the variogram values of the original data and the synthetic data as a quantitative measure of how closely the synthetic data approximates the original data in terms of variogram values. Table 1 lists the Mean Square Difference results averaged over 100 snapshots of radar data in increasing order. For this radar data set, in both coarse grained interpolation and fine-grained interpolation, the Nearest neighbor interpolation best matches with the

MSD for coarse-grained interpolationi	
Name of method	MSD for coarse-grained interp.
Nearest neighbor	8.354218e+01 (1.836358e+01)
Edge directed	1.970850e+02 (2.129320e+01)
Cubic	2.000790e+02 (1.694163e+01)
Delaunay triangulation	3.406270e+02 (4.795614e+01)
Linear	3.941510e+02 (2.876476e+01)
Spline	7.148526e+02 (5.5949e+01)
Kriging	1.469954e+03 (1.913371e+04)
Inv.-dist.-sq.-weighted avg	1.682726e+03 (3.617214e+02)

MSD for fine-grained interpolation	
Name of method	MSD for fine-grained interp.
Nearest neighbor	1.213458e+00 (6.4e-02)
Spline	5.447618e+01 (1.8)
Cubic	1.363040e+02 (4.3)
Edge directed	2.522578e+02 (1.05e+01)
Delaunay triangulation	3.275830e+02 (1.4e+01)
Linear	3.993235e+02 (1.2e+01)
Inv-dist-sq-weighted avg	1.429660e+03 (4.9e+02)

Table 1: The table lists the Mean Square Difference of variogram values for different interpolation algorithms in the increasing order of MSD. Here we use median from 100 snapshots instead of mean to get rid of outliers, and list 95% confidence interval in the brackets.

original variogram, the Inverse-distance-squared weighted averaging appeared the worst in preserving the original variogram, while the order of other interpolation algorithms changes between two different interpolation resolutions. We observe the same inconsistency with another precipitation data set [31]. Note that in the case of fine-grained interpolation, we verify whether the spatial correlation in the synthetic data matches that of the experimental data at the extent of coarse granularity, not at the scale of fine granulairty, where we do not have ground truth data.

Based on these results we do not recommend one single interpolation algorithm over others, but propose to use spatial correlation as the evaluation metric for our synthetic data generation purpose and a suite of interpolation algorithms. Given a new synthetic data generation task, we would test with different interpolation algorithms, select one that can best suit the current application and experimental data set at hand. Note that although the Nearest neighbor interpolation appears best matching with the original variogram model, it is not appropriate in the case of ultra-fine grained interpolation, since it assigns all nodes in a local neighborhood the same value from the nearby sample. However, most physical phenomena have some degree of variation even in a small local neighborhood, thereby, we would not expect all sensors deployed in a local neighborhood report the same sensor readings as in the case of the nearest neighbor interpolation.

As shown in Table 1 and our study of a rainfall data set [31], there is no significant increase in the MSD of variogram values from coarse grained interpolation (*i.e.*, 4x increase in resolution) to fine grained interpolation (*i.e.*, 100x increase in resolution). This may suggest that the interpolation algorithms can be applied to generate ultra-fine grained synthetic data set.

Summary: As shown above, most interpolation algorithms can approximate the original variogram models. However, it can only be used to interpolate at unsampled locations, not unsampled time. Furthermore, spatial interpolation algorithms, including Kriging, is not able to characterize the correlation between the spatial domain and temporal domain of the data, such as how the time trend varies at each location and how the spatial correlation changes as time progresses. Next, we wish to use the joint space-time model to address the limitations of the spatial interpolation techniques alone.

3.2 Joint space-time model

When considering the time and space domains together, a spatial-temporal random process was often decomposed into a mean component modeling the trend, and a random residue component modeling the fluctuations around the trend in both the time and space domains. Formally,

$$Z(u_\alpha, t_i) = M(u_\alpha, t_i) + R(u_\alpha, t_i) \tag{1}$$

where $Z(u_\alpha, t_i)$ is the attribute value under study, u_α is the location, t_i is the time, $M(u_\alpha, t_i)$ is the trend, and $R(u_\alpha, t_i)$ is the stationary residual component.

For the trend component, we borrowed the model from [19] where Kyriakidis *et al.* built a space-time model for daily precipitation data in northern California coastal region. $M(u_\alpha, t_i)$, in Equation 1 is further modeled as the sum of $(K+1)$ basis functions of time, $f_k(t_i)$: $M(u_\alpha, t_i) = \sum_{k=0}^{K} b_k(u_\alpha) f_k(t_i))$ where $f_k(t_i)$ is a function solely dependent on time t_i, with $f_0(t_i) = 1$ by convention. $b_k(u_\alpha)$ is the coefficient associated with the k-th function, $f_k(t_i)$, which is solely dependent on location u_α. $B(u_\alpha)$ and $F(t_i)$ can be computed as follows.

We first describe the guidelines to compute $f_k(t_i)$. [19] suggested that any temporal periodicities in the data should be incorporated in $f_k(t_i)$. Alternatively, $f_k(t_i)$ could also be identified as a set of orthogonal factors from Empirical Orthogonal Function (EOF) analysis of the data, or the spatial average of data at a time snapshot. In this paper, we use two basis functions: $f_0(t_i) = 1$ by convention; for $f_1(t_i)$, we take the spatial average of each time snapshot of the data. Formally, $F(t_i)$ (for illustration convenience, we write $f_k(t_i)$ and $b_k(u_\alpha)$ in matrix formats) can be written as:

$$\begin{pmatrix} 1 & \frac{1}{n} \sum_u z(u, t_1) \\ 1 & \frac{1}{n} \sum_u z(u, t_2) \\ \\ 1 & \end{pmatrix} \tag{2}$$

Next let us see $B(u_\alpha)$. If we ignore the residue component in Equation 1 for now, $z(u_\alpha, t_i)$ can be written as

$$Z(u_\alpha, t_i) = B(u_\alpha) \cdot F(t_i) \tag{3}$$

The vector of coefficients $B(u_\alpha)$ can be written as a weighted linear combination of the data vector $Z(u_\alpha, t_i)$: $B(u_\alpha) = H(u_\alpha) \cdot Z(u_\alpha)$, where $H(u_\alpha)$ is a matrix of weights assigned to each data component of $Z(u_\alpha)$ and $Z(u_\alpha)$ is a vector consisting of a time series data at location u_α. If the matrix F is of full rank, we have $H = (F' \cdot F)^{-1} \cdot F'$ from the ordinary least squares analysis (OLS).

The joint spatio-temporal trend model is constructed at each monitored location. The resulting trend parameters, $\{b_k(u_\alpha)\}$, are spatially correlated since they are derived from the same realization of the underlying spatio-temporal random process. Therefore, we spatially model and interpolate the trend parameters, $\{b_k(u_\alpha)\}$, using Kriging (Note that other spatial interpolation techniques could also be used) to obtain the value of $\{b_k(u_\alpha)\}$ at unsampled location u_α. Similarly, $\{F_{t_i}\}$ can be modeled

10

and interpolated to obtain the value of F_t at unsampled time point t.

In [19], co-regionalizing (so called co-kriging) is suggested to account for any cross-correlation between the b_k coefficients. To simplify our analysis, we choose to spatially model each coefficient separately. Kyriakidis *et al.* [19] also suggested to enhance the spatial prediction of these coefficients $\{b_k(u_\alpha)\}$ by considering relevant ancillary information. Co-located terrain elevation values are then used to enhance the prediction of daily precipitation level. However, this requires the availability of an extra environmental variable, which does not exist in our case.

As mentioned above, a spatio-temporal random process was decomposed into a trend and a random residue component. The decomposition is subject to the particular model we select. Therefore, the residue component is not guaranteed to be stationary and free of any spatio-temporal trend. It is possible to further construct a spatio-temporal model on the residue. In the following example, we treat the residue component as stationary noise, without additional modeling.

3.2.1 Evaluation of joint space-time modeling

To apply the joint space-time model described above, we considered a subset of the S-Pol radar data provided by NCAR. We selected a 70 x 70 spatial subset of the original data with 1km spacing, and 259 time snapshots across 2 days in May 2002. As mentioned above, the synthetic data is desired to have similar spatial correlation as the original data. Here we use one snapshot to shed some light on how the synthetic data generated from the joint space-time model captures the spatial correlation in the original data. Figure 7 shows the variogram plot of the synthetic data (which is generated from the joint space-time model) *vs.* original data, where the variogram value is normalized by the variance of the data, and the lag distance[3] is normalized by the maximum distance in horizontal or vertical directions. For comparison, we also show the results from spatial Kriging for the same snapshot in Figure 6. In Figure 7, the variogram of the synthetic data approximates that of the original one in the range [0, 0.6], which is less than half of the maximum distance between any two nodes; while in the spatial interpolation case (Figure 6), similar trends between two variogram curves (of synthetic vs. original data) are observed except in lags with range [1.1, 1.3], where the last few points in the variogram may be less accurate due to the fact that it is based on less data compared to other portions of the variogram. The mean square difference between the variograms of the synthetic *vs.* original data in the case of spatial modeling is 0.0633, while the mean square difference between two variograms in the case of joint space-time model is 0.2424. The larger discrepancy between the synthetic and original data in Figure 7 suggests that the joint space-time model does not capture the spatial correlation in the original data as precisely as the purely spatial interpolation does.

Spatial modeling vs. Joint space-time model   A joint spatio-temporal model can capture the correlation between the temporal trend and spatial variation and generate synthetic data at unmonitored times and locations. Further, compared with spatial interpolation at each time snapshot, joint space-time modeling is faster when we generate large amounts of snapshots of spatial data. For example, the precipitation data set [31] to be introduced next in section 4, includes daily precipitation data from the Pacific Northwest for approximately *45* years, or *16801* days. If we interpolate each daily snapshot separately, we

---

[3]lag distance is defined as the distance between two points

would have to apply spatial interpolation technique 16801 times. However, using the joint space-time model described above, we need only to interpolate the coefficients $\{B_k(u_\alpha)\}$ ($k = 1, 2$) in the spatial domain. This reduces the application of spatial interpolation to only two times and interpolating $F(t)$ in the temporal domain to only once.

The joint space-time model come at a cost. It combines spatial and temporal domains that have completely different characteristics. Therefore, a joint space-time model is usually more complicated and typically results in greater error than a space model. When we compare the prediction accuracy of spatial interpolation to the joint space-time model at a fixed point in time, a spatial interpolation technique usually can more closely capture the original data than a joint space-time model. In addition, spatial interpolation is often used as a component in the joint spatio-temporal models. Therefore, even though a joint space-time model naturally better captures the temporal and spatial characteristics in the data, the spatial modeling and interpolation is still important and popular in practice.

4   Case study: Using synthetic data to better evaluate a sensor network protocol

In this section, we use DIMENSIONS [13] as an example of how synthetic data generated from empirical models aids in evaluating the performance of an algorithm. In particular, when experimental data sets are available only from regular grid topology, we show how synthetic data generation allows us to evaluate DIMENSIONS over irregular topologies. Since we use DIMENSIONS only as a case study, our proposed approach to irregular topology data generation is by no means tied to the DIMENSIONS system. The irregular topology data generated from the procedure described next can be used to evaluate other sensor network algorithms.

Procedure to generate an irregular topology data set

1. Generating Ultra Fine-grained Data: In Step 1, we create a grid topology at a much finer granularity than our target topology. We model the correlation in the experimental data using the joint spatio-temporal model described in Section 3.2, and further generate a much finer grained data set based on this empirical model. In this case study, we consider a rainfall data set that provides 50km resolution daily precipitation data for the Pacific NorthWest from 1949-1994 [31]. The spatial setup comprises a 15x12 grid of nodes, each node recording daily precipitation values. Since the data set covers 45 years of daily precipitation data, it is rich enough in the temporal dimension. Thus, we increase the data granularity only in the spatial dimension, leaving the granularity of the temporal dimension unchanged. This fine-grained model was used to interpolate 9 points between every pair of points in both horizontal and vertical dimension, thus resulting in a 140 x 110 grid data.

2. Creating Irregular Topology Data: In Step 2, our objective is to down-sample the fine-grained data set from Step 1, and generate a data set for an arbitrary topology. We overlay the target topology on this ultra fine-grained grid data. Each node in the target topology is assigned a value from the nearest grid data. Note that our joint space-time model could be directly used to generate synthetic data at an arbitrary location and time. However, providing an ultra fine-grained data

set allows the protocol designers to derive arbitrary topology data as they wish from the fine-grained data, simplifying the generation of synthetic data in their chosen configuration.

To create a random topology with a predefined number of nodes, we select grid points at random from the fine-grained grid data. In our case study, 2% of the nodes in the 140 x 110 fine-grained grid were chosen at random (i.e., each node was chosen with probability 2%). This results in approximately 14x11x2 nodes being chosen in the network.

Next we demonstrate how the synthetic data generated from the above procedure help to expose problems, and gain more insights into the current DIMENSIONS system design.

DIMENSIONS   We now return to our case study to illustrate how data processing algorithms can be sensitive to topology features. DIMENSIONS [13] proposes wavelet-based multi-resolution summarization and drill-down querying. Summaries are generated in a multi-resolution manner, corresponding to different spatial and temporal scales. Queries on such data are posed in a drill-down manner, *i.e.*, they are first processed on coarse, highly compressed summaries corresponding to larger spatio-temporal volumes, and the approximate results obtained are used to focus on regions in the network that are most likely to contain result set data.

The standard wavelet compression algorithm used in DIMENSIONS assumes a grid topology in the data, and cannot directly handle an irregular placement of nodes. However, sensor network topologies are more likely to be irregular. To make the standard wavelet compression algorithm work with irregular topologies (without changing the wavelet algorithm itself), DIMENSIONS first derives a regular grid data set from the original irregular data set before applying the standard wavelet compression algorithm. However, this data regularizing process may skew the original data and introduce error in the query processing. We use a simple example to illustrate why DIMENSIONS can be sensitive to topology.

Figure 8 shows an irregular topology. The solid black circles represent real nodes, each labeled with a sensor reading. To convert the topology into a grid, we overlay a grid on top of the original irregular topology. In general, when we overlay a grid on an irregular topology, one of the following scenarios will be observed: (1) multiple data will appear in an overlaid grid cell when the region is divided coarsely; (2) some grid cells will be empty when the region is divided in a very fine-grained manner; or (3) most likely, a combination of the above. In standard grid-based wavelet processing, at the base level of the data hierarchy, one data point is required from each grid. Thus we need to address the cases in which multiple data points are packed into one cell, or a cell is empty. Possible solutions are:

1. When there are multiple data points in a grid cell, we need a single value representing all data in the same cell. Here, we average all data values in the same cell to obtain a single representative figure. However, this will smooth the original data, and likely lead to an inaccurate reply for queries such as *maximum value* in the sensor field.

2. In the case of an empty cell, we need to fill the cell with some artificial data. Here, we use the value from the nearest non-empty neighbor. If multiple neighbors are equidistant, we randomly select one.

Using the above approach, we establish a grid over the irregular topology in Figure 8, where the dotted circle represents

the virtual node with values computed from the averaging or nearest neighbor interpolation procedure described above. In this data configuration, for a query about *the maximum value in the sensor field*, the wavelet operation will generate a reply *28* (assuming no compression error), but the real maximum sensor reading is *38*. The discrepancy is due to smoothing in the interpolation. Taking another example: for a query about *average sensor reading*, the wavelet operation on this will reply *14* (assuming no compression error), but the real average is *15*. The discrepancy is again due to interpolation and averaging when converting the original data into a grid.

In summary, the procedure to convert an irregular topology to grid data will introduce errors. Different irregular topologies will likely introduce different amounts of errors. Further, for a certain topology, different approaches used to construct a grid over the original data are also expected to deliver different results.

Next, we compare the algorithm performance under regular and irregular topologies. For regular topologies, we use an experimental grid data set; for irregular topologies, we use synthetic irregular data sets generated from the joint spatio-temporal model explained in section 3.2.

4.1 Evaluate DIMENSIONS using the grid data set

DIMENSIONS [13] has been evaluated using a rainfall data set that provides 50km resolution daily precipitation data for the Pacific NorthWest from 1949-1994 [31]. The spatial setup comprises a 15x12 grid of nodes with 50-km spacing, each node recording daily precipitation values. While presumably this data set is at a significantly larger scale than what we would envision in a densely deployed sensor network, the data exhibits spatio-temporal correlations, providing a useful performance case study. A variety of queries can be posted on such a data set, such as range-sum queries, *e.g.*, total precipitation over a specified period from a single node or a region, or drill-down extreme value queries. In our comparison, we use maximum value queries as an example to demonstrate the algorithm performance. We present performance results for two queries: (a) GlobalDailyMaxQuery: which node receives max precipitation for a day in year X? (b) GlobalYearlyMax Query: which node receives the max precipitation for year X? The evaluation metric we used in this paper is mean square error. The error is defined as the difference between the measured query answer by dimensions, and the real answer as calculated by an optimal global algorithm with the same data input, normalized by the real answer. If we consider 15x12 grid data input as a coarse sample of the original phenomena, this mean square error can be deemed as the error in the system output compared with an approximate ground truth. Figure 9 and 10 show the mean square error vs. the level at which the drill-down query processing terminates.

In the DIMENSIONS hierarchy, each lower level stores twice the amount of data as the higher level. Therefore, as query processing proceeds down the DIMENSIONS hierarchy, consequently gaining access to more detailed information, as expected, the mean square error drops down gradually (shown in Figure 9 and 10). If the query accuracy followed the storage ratios (*i.e.*, 1:2:4:8:16), the accuracy improvement should be linear, however, it is usually super-linear, although the marginal improvement decreases with increasing levels. For instance, as shown in Figure 10, the marginal improvement in accuracy with just having the topmost level is approximately 66%, having the next level is 15%, followed by 6% and 2% respectively.

Figure 9 exhibits similar results.

4.2 Evaluate DIMENSIONS using the irregular data set

In this section, we present the algorithm performance over an irregular topology. The precipitation data set we used to evaluate DIMENSIONS in section 4.1 is from a grid topology, thus it cannot be directly used here. On the other hand, the performance of the wavelet compression algorithm is sensitive to the correlation pattern in the data and the space of data correlation is too big to simulate exhaustively. Therefore, a randomly generated data set is considered unrealistic and of little value in the DIMENSIONS evaluation. Instead, we use a synthetic irregular topology derived from models of the precipitation data set we used in section 4.1. Next, we describe the procedure to generate this irregular data set.

4.2.1 Generate irregular topology based on models of rainfall data set of grid topology

Regularizing the irregular topology Following the procedure described in the beginning of section 4 , we generate a random topology consisting of 14x11x2 nodes from the precipitation data. The current implementation of the wavelet compression algorithm in the DIMENSIONS system works only with a regular grid topology. Thus, a pre-processing step is added to convert the irregular data set into regular grid data. The regularizing procedure works as follows:

- Choose a coarse grid: choose a grid size that is coarser than the fine-grained data. In this case, 14x11 grid regions were chosen overlaying on the 140x110 grid acquired in step 1, and an average of 2 nodes per grid are expected.

- Averaging: For each such grid, average data from all nodes in the grid. Averaging data from multiple nodes will smooth the data and also reduce the noise in data. If there are N sensors in a certain square, the original noise per sensor is σ^2, and assuming the noise distribution at each sensor node is *i.i.d.*, then averaging the data obtains an aggregated measurement with variance σ^2/N (from the CLT theorem).

- Interpolation for empty grid cells: If there are no nodes in a certain cell, we use simple nearest neighbor interpolation to fill in these grids. Nearest neighbor interpolation fills the cell with data from the nearest non-empty cell. This can be easily implemented in a distributed setting: a cluster head that needs to receive data from 4 lower level nodes: and receives, say, only 3, fills in the empty grid with interpolated data. Finally, the resulting regular data is passed to the standard grid-based wavelet compression.

4.2.2 Results and comparison

Due to the time consuming nature of wavelet operations, we illustrate our results with only a single topology. The results presented in this section are averaged for one irregular topology and multiple (*i.e.,* 44) queries.

Figure 11 and 12 show the results for the same queries as presented in section 4.1, GlobalDailyMax and GlobalYearlyMax, but for the irregular topology generated from the procedure described above in section 4.2.1. As in the regular grid case, we present the results in the form of the mean square error vs. the level at which the drill-down query processing terminates. The

error here is the difference between the output of DIMENSIONS system with 14x11x2 irregularly placed nodes as input and the approximate ground truth. Here, the ground truth is approximated in the following way: Since the 14x11x2 irregularly placed nodes were selected from the 140x110 fine grained data, we first overlay a 14x11 grid over the 140x110 fine grained data, and for each grid, average 100 fine grained data points in the grid and obtain a value for each of 14x11 grid, and input this 14x11 data to the global algorithm. The ground truth was approximated by the output of the global algorithm.

As shown in Figure 11 and 12, DIMENSIONS behaves differently in an irregular setting from a regular one. In regular case (as shown in Figure 9 and 10) , we observe gradual error improvement with more levels of drill-down. However, in the irregular case (Figure 11 and 12), there is no consistent error improvement with more levels of drill down. For the topology that we studied, in some cases (Figure 11 and 12), the query error actually increases as more levels are drilled down. Such a behavior would not be expected in a regular grid topology. This may be due to effects caused by holes in topology and our corresponding interpolation procedure. An irregular topology might result in large regions where there is no data, and as described above, interpolation is used to fill in the empty cells in the regularization process. However, interpolation may smooth and therefore skew the data, which will introduce error in the query processing. Note that if the graphs are averaged over multiple irregular topologies, some of the kinks in Figure 11 and 12 might be straightened out. However, we do not expect the same consistent improvement or the same order of improvement as with the regular grid data case.

Discussions   The error that the user perceived and approximated by our computed mean square error is subject to multiple factors. We just name a few: the physical node topology, *i.e.,* the actual position of nodes; the data correlation statistics, *e.g.,* the form of the spatio-temporal model, and its interaction with wavelet compression; and the mechanism used to interpolate (in our case, nearest neighbor) in the regularizing procedure. If our objective is to study how irregular topology affects system performance, ideally we want to vary only the topology, and keep other parameters fixed. One way to achieve this is to model the data correlation in the experimental data set, and then use the same empirical model to generate regular and irregular topologies, which we will use to evaluate the algorithm's performance over regular and irregular data respectively. More specifically, we could use the output of the data modeling process as described in section 3 (*i.e.*, fine grained data set) as the empirical model. We sub-sample the same fine grained data set to obtain the regular and irregular sampled data as input to the DIMENSIONS system. The common model here is fine grained data. Since it is overly interpolated, it is deemed as an approximate model of the phenomena and used as the approximate ground truth to compute the mean square error. The error is thereby computed as: $(QueryResponseOverDimesions - QueryResponseOverModel)/QueryResponseOverModel$. We plan to explore this in our future work.

In summary, we used the synthetic data sets generated from modeling the spatio-temporal correlation in the experimental data set to evaluate DIMENSIONS over an irregular data set. DIMENSIONS in an irregular setting exhibited different behavior from the regular setting. This exposes the problem of DIMENSIONS's current regularization scheme. Thus our proposed synthetic data set helped to systematically evaluate the algorithm, and point out needed improvements.

In our case study, the precipitation data is from a regular grid topology. However, even if we have experimental data sets from irregular topologies, our proposed synthetic data generation approach will enable evaluating algorithms over different

irregular topologies other than the particular setting used in the experimental data set. More importantly, it also allows us to evaluate algorithms over the same underlying data correlation model but different topology settings, so that we can study how the algorithm performance interacts with different parameters independently.

5   Conclusions

In this paper, we proposed to generate synthetic data from empirical models of experimental data, and use it to evaluate sensor network algorithms. In modeling the spatio-temporal correlation in the data, we draw heavily on spatial interpolation and geo-statistical techniques to implicitly or explicitly model the spatial correlation in the data, and generate fine-grained data that approximates the spatial correlation in the experimental data. To capture the correlation between the spatial and temporal domain, we use a joint space-time model inspired by [19]. We apply the proposed modeling techniques to a S-Pol radar data set. We propose spatial correlation (*i.e.*,variogram values) as a quantitative metric to evaluate synthetic data sets, which will be used to test sensor network protocols. Evaluation results show that most spatial interpolation techniques can closely approximate the spatial correlation and interesting properties (*e.g.*, edges) of the original data. The joint space-time modeling provides advantages over spatial modeling alone, such as enabling quick data generation at any location and any time. However, evaluation results show that the joint space-time model cannot capture the spatial correlation in the original data as closely as the spatial modeling alone. We also use the DIMENSIONS system as a case study to show how synthetic data can help to evaluate the system over irregular topologies.

6   Acknowledgements

References

[1] The r project for statistical computing. In *http://www.R-project.org/*.

[2] Sandeep Bajaj, Lee Breslau, Deborah Estrin, Kevin Fall, Sally Floyd, Padma Haldar, Mark Handley, Ahmed Helmy, John Heidemann, Polly Huang, Satish Kumar, Steven McCanne, Reza Rejaie, Puneet Sharma, Kannan Varadhan, Ya Xu, Haobo Yu, and Daniel Zappala. Improving simulation for network research. Technical Report 99-702b, University of Southern California, March 1999. revised September 1999, to appear in IEEE Computer.

[3] G. Box and G. Jenkins. Time series analysis: forecasting and control. Holden-Day, 1976.

[4] J. Broch, D. A. Maltz, D. B. Johnson, Y.-C. Hu, and J. Jetcheva. A Performance Comparison of Multi-Hop Wireless Ad-Hoc Network Routing Protocols. In *Proceedings of the Fourth Annual ACM/IEEE International Conference on Mobile Computing and Networking (Mobicom'98)*, Dallas, TX, 1998.

[5] Ramon Caceres, Peter B. Danzig, Sugih Jamin, and Danny J. Mitzel. Characteristics of wide-area tcp/ip conversations. In *SIGCOMM*. ACM, 1991.

[6] Alberto Cerpa, Jeremy Elson, Deborah Estrin, Lewis Girod, Michael Hamilton, and Jerry Zhao. Habitat monitoring: Application driver for wireless communications technology. In *2001 ACM SIGCOMM Workshop on Data Communications in Latin America and the Caribbean*, April 2001.

[7] Alberto Cerpa and Deborah Estrin. Ascent: Adaptive self-configuring sensor networks topologies. In *Infocom '02*, New York, June 2002. IEEE.

[8] Samir R. Das, Charles E. Perkins, and Elizabeth M. Royer. Performance comparison of two on-demand routing protocols for ad hoc networks. In *INFOCOM*, Israel, March 2000.

[9] Geoff Davis. Baseline wavelet transform coder construction kit. Technical report, 1/29/97.

[10] Georg Dorffner. Neural networks for time series processing. *Neural Network World*, 6(4):447–468, 1996.

[11] Xin Li et al. New edge-directed interpolation. In *IEEE Trans. on Image Processing*, Oct. 2001.

[12] P. Fryzlewicz, S. Van Bellegem, and R. von Sachs. A wavelet-based model for forecasting non-stationary processes. In *Submitted for publication.*, 2002.

[13] Deepak Ganesan, Deborah Estrin, and John Heidemann. Dimensions: Why do we need a new data handling architecture for sensor networks? In *Proceedings of the First Workshop on Hot Topics In Networks (HotNets-I)*, Oct 2002.

[14] Pierre Goovaerts. Geostatistics for natural resources evaluation. Oxford University Press, Inc., 1997.

[15] Per Johansson, Tony Larsson, Nicklas Hedman, Bartosz Mielczarek, and Mikael Degermark. Scenario-based performance analysis of routing protocols for mobile ad-hoc networks. In *Proc. ACM Mobicom*, Seattle, Washington,, 1999.

[16] Almudena Konrad, Ben Y. Zhao, Anthony D. Joseph, and Reiner Ludwig. A markov-based channel model algorithm for wireless networks. In *Proceedings of Fourth ACM International Workshop on Modeling, Analysis and Simulation of Wireless and Mobile Systems, ACM MSWiM*, July 2001.

[17] D. G. Krige. Two-dimensional weighted moving average trend surfaces for ore-valuation. In *Journal of the South Africa Institute of Mining and Metallurgy*, volume 66, 1966.

[18] P.C. Kyriakidis and A.G. Journel. Geostatistical space-time models. In *Mathematical Geology*, volume 31, 1999.

[19] P.C. Kyriakidis, N. L. Miller, and J. Kim. A spatial time series framework for modeling daily precipitation at regional scales. In *82nd Annual Meeting of the American Meteorological Society*, January 2002.

[20] Yosihiko Ogata. Space-time point-process models for earthquake occurences. In *Ann. Inst. Statist. Math.*, volume 50, 1998.

[21] S. Park, A. Savvides, and M. B. Srivastava. Sensorsim: A simulation framework for sensor networks. In *MSWiM*. ACM, August 2000.

[22] S. Park, A. Savvides, and M. B. Srivastava. Simulating networks of wireless sensors. In *the 2001 Winter Simulation Conference*. ACM, 2001.

[23] Theodore Rappaport. Wireless communications: Principles and practice. ACM, 1999.

[24] N. Abu-Ghazaleh S. Tilak and W. Heinzelman. Infrastructure tradeoffs for sensor networks. In *ACM 1st International Workshop on Sensor Networks and Applications (WSNA '02)*. ACM, Sep. 2002.

[25] F. P. Schoenberg. Consistent parametric estimation of the intensity of a spatial-temporal point process. In *Ann. Inst. Stat. Math. (in review)*. Wiley, NY.

[26] F. P. Schoenberg, D. R. Brillinger, and P. M. Guttorp. Point processes, spatial-temporal. In *Encyclopedia of Environmetrics*, volume 3. Wiley, NY.

[27] Curt Schurgers, Vlasios Tsiatsis, and Mani Srivastava. Stem: Topology management for energy efficient sensor networks. In *IEEE Aerospace Conference*, Big Sky, MT, March 2002. IEEE.

[28] Y. Theodoridis and Mario Nascimento. Generating spatiotemporal data sets in the www. In *SIGMOD record 29 (3)*, 2000.

[29] DuMouchel W, Volinsky C, Johnson T, Cortes C, and Pregibon D. Squashing flat files flatter. In *Proc. KDD*, 1999.

[30] Richard Webster and Margaret A. Oliver. Geostatistics for environmental scientists. John Wiley & Sons, Inc., 2001.

[31] M. Widmann and C.Bretherton. 50 km resolution daily preciptation for the Pacific Northwest, 1949-94, http://tao.atmos.washington.edu/data_sets/widmann/.

[32] Ya Xu, John Heidemann, and Deborah Estrin. Geography informed energy conmservation for ad hoc routing. In *MOBICOM'01*, Rome, Italy, July 2001. ACM.

[33] Xiang Zeng, Rajive Bagrodia, and Mario Gerla. Glomosim: a library for parallel simulation of large-scale wireless networks. In *Proceedings of the 12th Workshop on Parallel and Distributed Simulations – PADS '98*, May 1998.

A Appendix 1: Brief introduction of variogram

A variogram is used to characterize the spatial correlation in the data. The variogram (also called semivariance) of a pair of points x_i and x_j is defined as $\gamma(x_i, x_j) = \frac{1}{2}\{Z(x_i) - Z(x_j)\}^2$. We can also define semivariance as a function of lag (*i.e.*, the separation between two points, is distance in one dimension, or a vector with both distance and direction in two and three dimensions), h:

$$\gamma(h) = \frac{1}{2}E[\{Z(x) - Z(x + h)\}^2] \tag{4}$$

For a set of samples, $z(x_i)$, i=1, 2, ..., $\gamma(h)$ can be estimated by

$$\hat{\gamma}(h) = \frac{1}{2m(h)} \sum_{i=1}^{m(h)} \{z(x_i) - z(x_i + h)\}^2 \tag{5}$$

where $m(h)$ is the number of samples separated by the lag distance h.

Data in high dimensions might add complexity in modeling variograms. If data lie in a high dimensional space, variograms are first computed in different directions separately. If variagrams in different directions turn out to be more or less the same, the data are isotropic, then sample variograms can be averaged together. Otherwise, data in different directions need to be modeled separately.

B Appendix 2: Brief introduction of ordinary kriging

Kriging refers to a range of least squares based estimation techniques, which are used to address the above estimation problem. It has both linear and non-linear form. we use its linear form, *ordinary kriging*, to illustrate the principle of Kriging. The ordinary kriging is a weighted linear estimator.

For convenience of reference, we restate a formulation of the estimation problem in spatial modeling defined in Section 3.1. A random process, Z, is defined as a set of dependent (here spatially dependent) random variables $Z(u)$, one for each location u in the study area A, denoted as $\{Z(u), \forall u \in A\}$. Assuming Z is an ergodic process, the problem is defined to estimate some statistics (*e.g.*, mean) of $Z(u)$ ($u \in A$) given a realization of $\{Z(u)\}$ at locations u_i, $i = 1, ..., n$, $u_i \in A$.

In the context of the above estimation problem, given a realization of random process Z, *i.e.*, a collection of samples $z(x_i)$ ($i = 1$ to N) at locations x_i, $i = 1, ..., n$, we estimate Z at an unknown point x_0. We use $\hat{Z}(x_0)$ to denote the prediction at point x_0, and $Z(x_0)$ to denote the true value at point x_0. $\hat{Z}(x_0)$ is written as a weighted average of the neighboring samples: $\hat{Z}(x_0) = \sum_{i=1}^{N} \lambda_i z(x_i)$. There are different ways to determine the weights: assign all of the weight to the nearest data, as used in the nearest neighbor resampling approach; assign the weights inversely proportional to the distance from the location being estimated; or use variogram and minimize the estimation variance, which is written as a function of variogram. The last approach is used by Kriging. In Kriging, the weights are determined by minimizing the estimation variance and an additional constraint, $\sum_{i=1}^{N} \lambda_i = 1$, to ensure that the estimate is unbiased.

As mentioned, the estimation variance can be written as a function of variogram (Please refer to Appendix A for a brief introduction of variogram) matrix.

$$var[\hat{Z}(x_0)] = E[\{\hat{Z}(x_0) - Z(x_0)\}^2] = 2\sum_{i=1}^{N} \lambda_i \gamma(x_i, x_0) - \sum_{i=1}^{N}\sum_{j=1}^{N} \lambda_i \lambda_j \gamma(x_i, x_j) \tag{6}$$

Since the optimization criteria is minimizing the estimation variance, the weights λ_i are derived by setting the partial derivative of Equation 6 with respect to each weight to be 0.

In Equation 6, $\gamma(x_i, x_j)$ can be computed based on samples, whereas there is no way to directly compute $\gamma(x_i, x_0)$. $\gamma(x_i, x_0)$ can be derived based on modeling the sample variogram $\gamma(x_i, x_j)$ (where $i, j = 1, ..., N$) using the least squares method [14], for example.

As shown in Equation 6, there are two factors affecting the Kriging estimation variance: (1) Variogram (or covariance) model: the estimation variance depends on the complexity of the spatial correlation of the data as modeled by its variogram; (2) Data topology: it accounts for the relative geometry of data locations and their distances to the location being estimated.
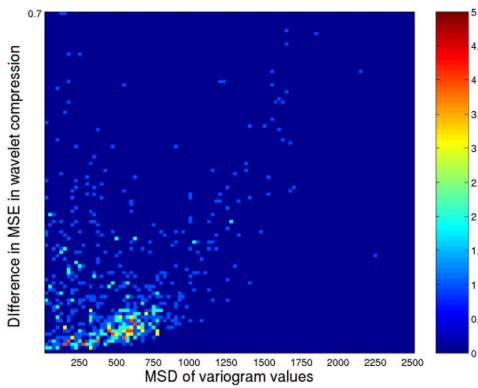
Figure 1: MSD in spatial correlation vs. difference in MSE in wavelet compression between synthetic data and the original data
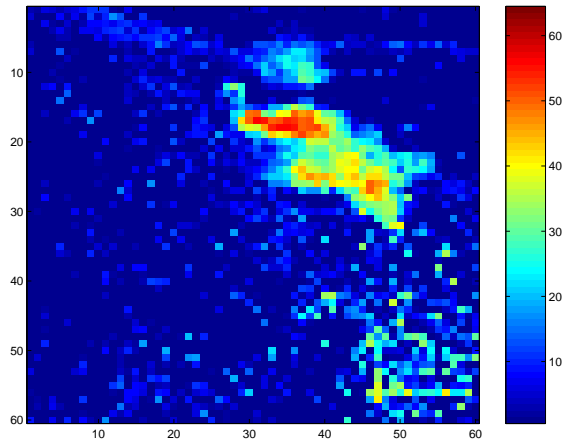
Figure 2: Spatial modeling example: original data map (60x60)
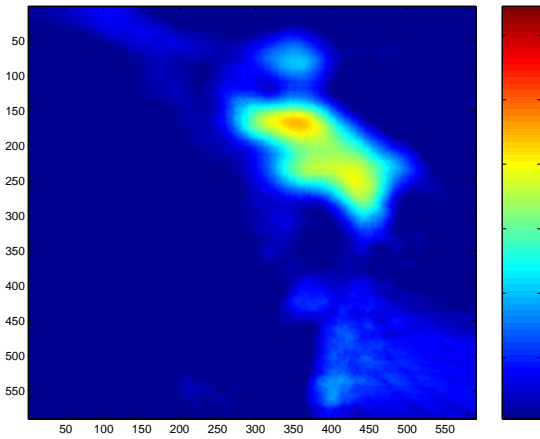
Figure 3: Spatial modeling example: Fine grained interpolation data map (590x590) from Kriging
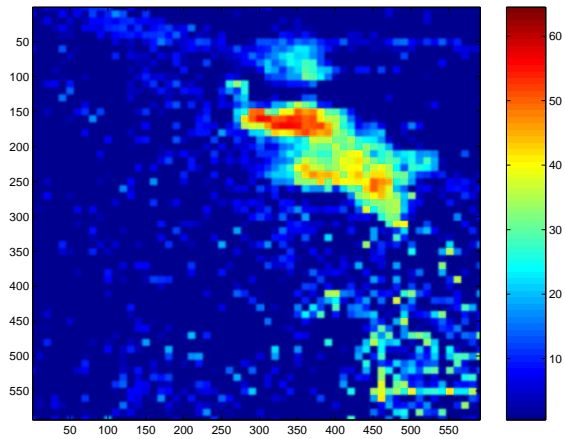
Figure 4: Spatial modeling example: Fine grained interpolation data map (590x590) from the nearest neighbor interpolation
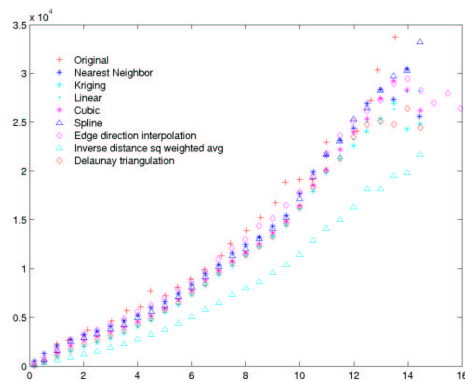
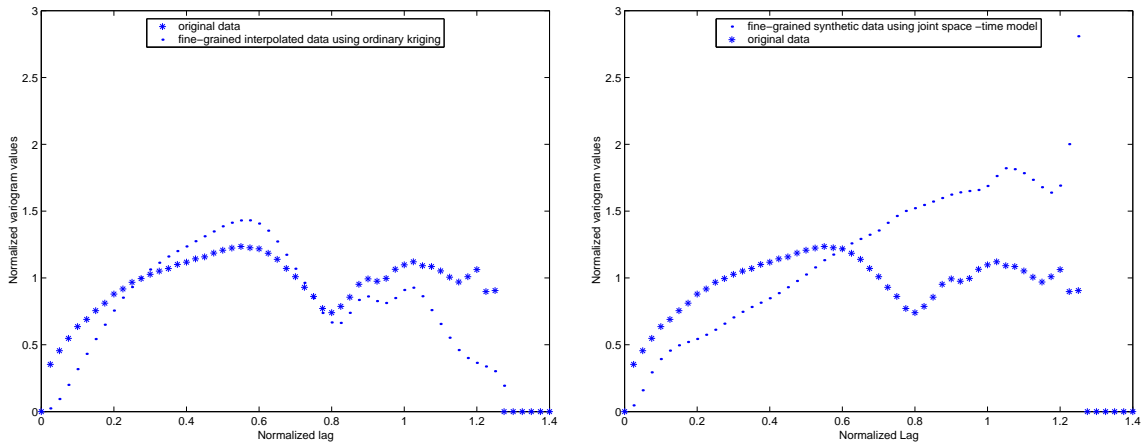Figure 5: MSD of variogram values: Coarse grained interpolation results on a snapshot of radar data

Figure 6: Spatial modeling example: Variogram of the fine-grained synthetic data and the original data

Figure 7: Joint modeling example: Variogram of the fine-grained synthetic data and the original data
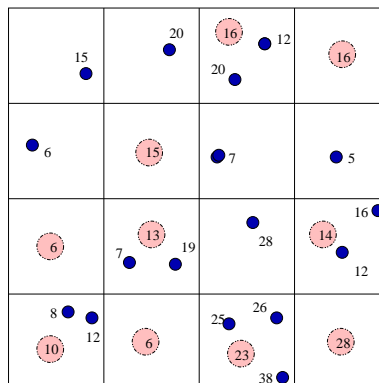
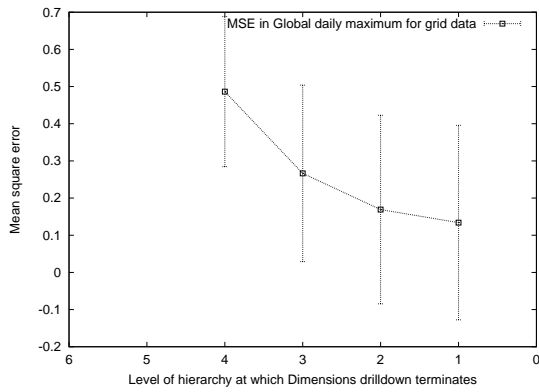Figure 8: Irregular topology: converting an irregular topology to a grid will skew the data

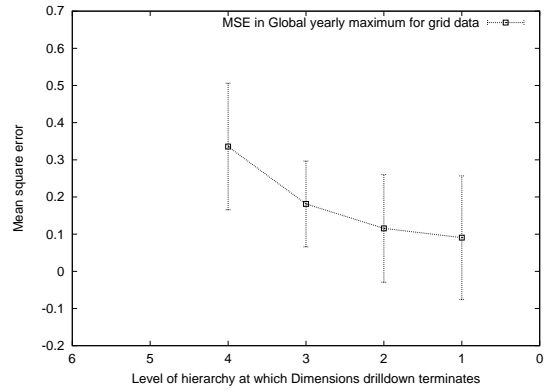Figure 9: Error vs. Query termination level: Global daily maximum over original rainfall data

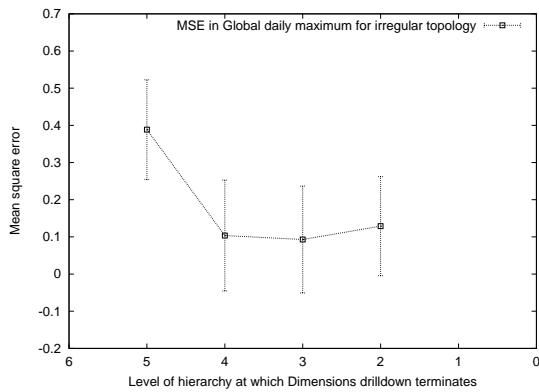Figure 10: Error vs. Query termination level: Global yearly maximum over original rainfall data

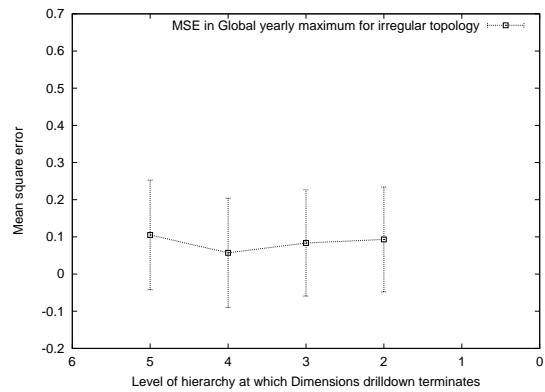Figure 11: Error vs. Query termination level: Global daily maximum over irregular topology

Figure 12: Error vs. Query termination level: Global yearly maximum over irregular topology

23