Modeling spatially-correlated sensor network data

Apoorva Jindal, Konstantinos Psounis
Department of Electrical Engineering-Systems
University of Southern California
Los Angeles CA-90089
Email: apoorvaj@usc.edu, kpsounis@usc.edu

Abstract— **The physical phenomena monitored by sensor networks, e.g. forest temperature, water contamination, usually yield sensed data that are strongly correlated in space. With this in mind, researchers have designed a large number of sensor network protocols and algorithms that attempt to exploit such correlations.**

To carefully study the performance of these algorithms, there is an increasing need to synthetically generate large traces of spatially correlated data representing a wide range of conditions. Further, a mathematical model for generating synthetic traces would provide guidelines for designing more efficient algorithms. These reasons motivate us to obtain a simple and accurate model of spatially correlated sensor network data.

The model can capture correlation in data irrespective of the node density, the number of source nodes or the topology. We describe a mathematical procedure to extract the model parameters from real traces and generate synthetic traces using these parameters. Then, we validate our model by statistically comparing synthetic data and experimental data, as well as by comparing the performance of various algorithms whose performance depends on the degree of spatial correlation. Finally, we create a tool that can be easily used by researchers to synthetically generate traces of any size and degree of correlation.

I. Introduction

The wireless sensor networks of the near future are envisioned to consist of hundreds to thousands of inexpensive wireless nodes, each with some computational power and sensing capability, operating in an unattended mode. Since these sensors will be densely deployed and they detect common phenomena, it is expected that a high degree of spatial correlation will exist in the sensor network data. The presence of spatial correlation in data has been exploited by different algorithms solving different problems. Spatial correlation has been used in data aggregation and routing algorithms [1], [2], [3], [4], [5], data storage and querying [6], [7], [8], mac protocol design [9], [10], localization [11], data compression and encoding [12], and calibration [13].

The evaluation of protocols that are sensitive to the spatial features of input data requires data representing a wide range of realistic conditions. However, since very few real systems have been deployed, there is hardly any experimental data available to test the proposed algorithms. Further, no effort has been made to propose a model which captures the spatial correlation in sensor network data. As a result, sensor network researchers make different assumptions when generating data inputs to evaluate systems; some assume the data to be jointly gaussian with the correlation being a function of the distance [10], some assume that the data follows the diffusion property [8], and some assume a function for the joint entropy of the data [2]. Finally, some researchers propose the use of environmental monitoring data obtained from remote sensing [6], however the granularity of these data sets do not match the expected granularity of sensor networks' data.

The need of a proper mathematical model which can capture spatial correlation of any degree irrespective of the granularity, density, number of source nodes or topology is evident. Yan Yu et al [14] proposed a method to interpolate existing experimental data to support irregular topologies. However, they do not propose a model which can generate synthetic data traces of any granularity and density without the need of experimental data traces. On a different context, Psounis et al [15] proposed a markovian model to capture temporal correlation in web traces.

In this paper we propose a mathematical model that is similar in flavor to that in [15], in order to capture the spatial correlation in sensor network data. We present a method to generate large synthetic traces from a small experimental trace while preserving the correlation pattern, and a method to generate synthetic traces exhibiting arbitrary correlation patterns. We show that synthetic traces are very close to real traces using statistics, and by running algorithms which exploit the presence of spatial correlation, against both types of traces. We use two well

0-7803-8796-1/04/$20.00 (c) 2004 IEEE.

known algorithms, DIMENSIONS [6] and CMAC [10], for this purpose.

The remainder of the paper is organized as follows. Section II presents a statistical analysis of experimental data to motivate our model. The model is formally presented in Section III. We show through variogram[1] plots how spatial correlation properties depend on the model parameters, and then present a mathematical method to infer the model parameters from a real trace. The correctness of the model is verified by comparing the statistics of the original and the synthetic traces in Section IV-A. In Section IV-B, the accuracy of the model is validated by comparing the performance of various algorithms against real and the corresponding synthetic traces. Finally, Section V presents two tools which we have created to enable researchers generate large traces from a small input trace, or create traces having varied correlation properties by tweaking the model parameters.

II. Statistical Analysis of Experimental Data

A. Data Set Description

Since there are no sensor network data sets available to date, we base our study on environment monitoring data. This paper makes use of two data sets, the S-Pol Radar Data Set [2] and the Precipitation Data Set [16]. These data sets have been extensively used in the sensor networks literature over the last couple of years, e.g. [2], [6], [14].

1) S-Pol Radar Data Set: The resampled S-Pol radar data, provided by NCAR, records the intensity of reflectivity of atmosphere 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 500×1008 samples for each 360 degree azimuthal sweep. They were converted to the cartesian grid using the nearest neighboring resampling method [17]. In this paper, we have selected a 64×64 spatial subset of the original data and 259 time snapshots across 2 days in May 2002.

[1] a measure of correlation introduced in Section II-B

[2] S-Pol radar data were collected during the IHOP 2002 project (`http://www.atd.ucar.edu/rtf/projects/ihop_2002/spol/`). S-Pol is fielded by the Atmospheric Technology Division of the National Center for Atmospheric Research. We acknowledge NCAR and its sponsor, the National Science Foundation, for provision of the S-Pol data set.

2) Precipitation Data Set: This data set consists of the daily rainfall precipitation for the Pacific Northwest from 1949-1994. The final measurement points in the data set formed a regular grid of 50 km $\times$50 km regions over the region under study. We select a subset of data that has no missing values. Specifically, each snapshot of data is a 8×8 spatial grid data with a 50 km resolution.

B. Statistics used to Measure Correlation in Data

To study the correlation properties of data, researchers usually use the autocorrelation function. Given a two dimensional stationary process $X(x, y)$, the autocorrelation function is defined as

$$R(d_1, d_2) = E[X(x, y)X(x + d_1, y + d_2)]. \quad (1)$$

Another statistic often used to characterize spatial correlation in data is the variogram. The variogram (also called semivariance) is defined as

$$\gamma(d_1, d_2) = \frac{1}{2}E[(X(x, y) - X(x + d_1, y + d_2))^2]. \quad (2)$$

For isotropic random processes [18] the variogram depends only on the distance $d = d_1 + d_2$ between two nodes. [3] In this case, if (x_d, y_d) denotes a point which is d distance away from (x, y),

$$\gamma(d) = \frac{1}{2}E[(X(x, y) - X(x_d, y_d))^2], \quad (3)$$

where $|x - x_d| + |y - y_d| = d$.

For a set of samples $x(x_i, y_i)$, $i = 1, 2, \ldots$, $\gamma(d)$ can be estimated as follows,

$$\hat{\gamma}(d) = \frac{1}{2m(d)} \sum_{1}^{m(d)} [x(x_i, y_i) - x(x_j, y_j)]^2, \quad (4)$$

where $m(d)$ is the number of points at a distance d within each other, i.e. the sum is over all points for which $|x_i - x_j| + |y_i - y_j| = d$.

We experimented with both metrics and found that the results were similar. Since the variogram gives a better visual representation of the variation of data with distance, in this paper we will only present variogram results.

C. Analysis of Data using Variograms

If a process is independent and identically distributed (iid), its variance will not change with distance and the variogram should be a straight line parallel to the x-axis. Figure 1 shows the variogram for an iid process with the underlying random variable being Gaussian with mean 0 and standard deviation equal to 10.

[3] We are using the L1 (Manhattan) distance here because we will later assume that points reside in a grid, but one may define distance in any meaningful manner.

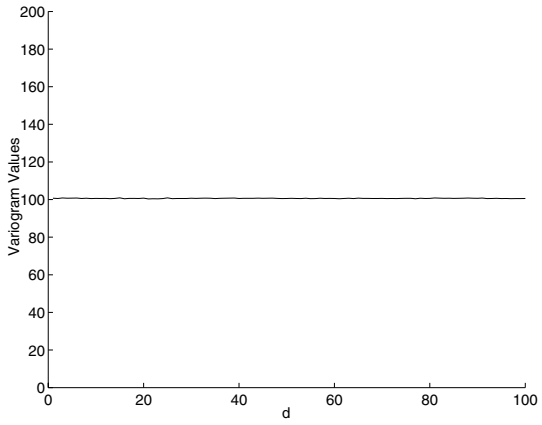

Fig. 1. Variogram for an iid process.

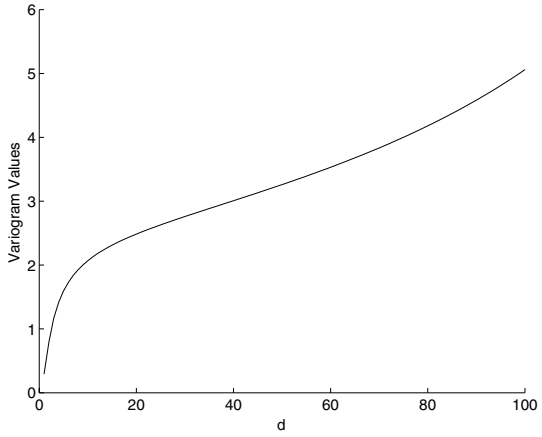

Fig. 2. Variogram for a process which follows the diffusion law ($\alpha = 1$).

When the phenomenon under observation is being emitted from a single source it usually follows the diffusion property with distance, i.e. $f(d) \propto \frac{1}{d^\alpha}$ where $f(d)$ is the magnitude of the event's effect at a distance d from the source and α is the diffusion parameter that depends on the type of the effect. Figure 2 shows the variogram for a process following the diffusion law.

But the process under observation seldom has a single source and the presence of multiple sources will require us to calculate a phasor sum of data values at a node. For atmospheric data such as temperature, precipitation and humidity, it is not even possible to define a source. The data values at nodes close to each other will be correlated, while for large d the process will start looking like an iid process. As an example, the variogram at a time snapshot of the S-Pol radar data is shown in Figure 3. As it is evident from the plot, as the distance grows from zero the spatial correlation decreases fast. Also, for distances larger than 20 correlation is quite small.

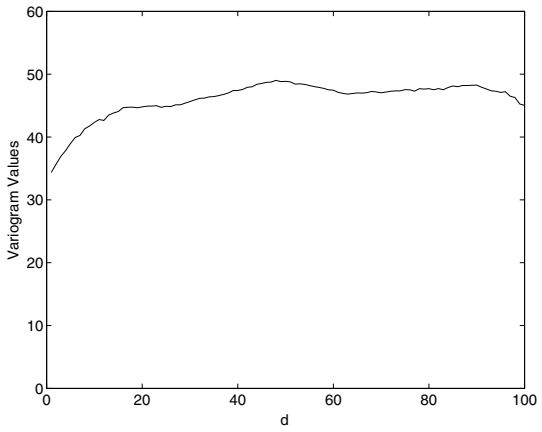

Fig. 3. Variogram of the experimental data at a time snapshot.

III. The Model

In this section we introduce our model for capturing the statistical properties of sensor networks data. Let $X(x, y)$ be the data value at a node (x, y). We assume that $X(x, y)$ is a stationary process that has a unique first order distribution whose probability density function (pdf) is denoted by $f_X(x)$. (We call this the long term distribution of the data.) For simplicity, we also assume that the nodes are located on a regular rectangular grid. (In section III-A, we comment on how to use the model with irregular topologies.)

Let $N(d)$ denote the number of nodes at a distance d from (x, y). Let X_d denote the data value at a node which is d distance away from (x, y), and X_d^k denote the data value at the k^{th} node ($1 \leq k \leq N(d)$) at a distance d from (x, y). We propose the following model for generating data values:

$$X(x,y) = \begin{cases} X_1^1 + Z & \text{with probability } \frac{\alpha_1}{N(1)} \\ \vdots & \\ X_1^{N(1)} + Z & \text{w.p. } \frac{\alpha_1}{N(1)} \\ X_2^1 + Z & \text{w.p. } \frac{\alpha_2}{N(2)} \\ \vdots & \\ X_2^{N(2)} + Z & \text{w.p. } \frac{\alpha_2}{N(2)} \\ \vdots & \\ X_h^1 + Z & \text{w.p. } \frac{\alpha_h}{N(h)} \\ \vdots & \\ X_h^{N(h)} + Z & \text{w.p. } \frac{\alpha_h}{N(h)} \\ Y & \text{w.p. } \beta \end{cases} \tag{5}$$

where Z and Y are random variables independent of each other as well as X, and their pdf's are denoted by

$f_Z(z)$ and $f_Y(y)$ respectively. Both Y and Z determine the long term distribution of X, and Z captures the small differences between neighboring data values. The above equation simply says that the probability that $X(x,y)$ is derived from the value of a node which is d distance away from (x,y) is α_d. Further, the probability that $X(x,y)$ is derived from the value of a particular such node is $\frac{\alpha_d}{N(d)}$. The parameters of the model are h, the α_i's, β, $f_Y(y)$ and $f_Z(z)$. Obviously $\beta + \sum_{i=1}^{h} \alpha_i = 1$.

We now derive the relationship between the distributions of X, Y and Z. For mathematical convenience, we define the following three random variables:

- $A_j^i = X_j^i + Z$,
- T which indicates the outcome of the toss of an $h+1$ sided, biased coin with $P(T=j) = \alpha_j$ for $1 \leq j \leq h$ and $P(T = h+1) = \beta$, and
- U_d which is a discrete uniform random variable taking values between 1 and $N(d)$.

Using this notation, we can find the probability density function $f_X(x)$ as follows:

$$P(X \leq x) = \sum_{j=1}^{h} \sum_{i=1}^{N(j)} P(A_j^i \leq x | T=j, U_j = i) P(U_j = i | T = j) P(T=j) + P(Y \leq x | T = h+1) P(T = h+1)$$

$$\Rightarrow F_X(x) = \sum_{j=1}^{h} \sum_{i=1}^{N(j)} F_{A_j^i}(x) P(T=j) P(U_j = i) + F_Y(x) P(T = h+1)$$

$$\Rightarrow f_X(x) = \sum_{j=1}^{h} \sum_{i=1}^{N(j)} f_{A_j^i}(x) P(T=j) P(U_j = i) + f_Y(x) P(T = h+1). \quad (6)$$

In stationarity, X_1^1, ..., $X_1^{N(1)}$, ..., $X_h^{N(h)}$ have the same distribution, that is $f_X = f_{X_1^1} = ... = f_{X_1^{N(1)}} = ... = f_{X_h^1} = ... = f_{X_h^{N(h)}}$. Similarly, $f_A = f_{A_j^i}$, for all i, j.

Using the above and Equation (6) the characteristic function of $f_X(x)$ can be written as

$$\Phi_X(j\omega) = (1-\beta)\Phi_A(j\omega) + \beta \Phi_Y(j\omega). \quad (7)$$

Hence, given any two distribution functions, we can find the third one using Equation (7).

Without loss of generality, from now onwards we will assume that Z is a normal random variable with mean $\mu = 0$ and standard deviation $\sigma = \sigma_z$. We use the S-Pol radar data to justify our assumption. Figure 4 shows the distribution of $X - X_1$, where X_1 represents the sample values at nodes at a unit distance away from X. As it is evident from the plot, this distribution can be very closely approximated by a gaussian distribution. Note that the distribution of Z need not necessarily be gaussian; any other distribution will not change any equations upto this point, though the succeeding analysis will be modified.

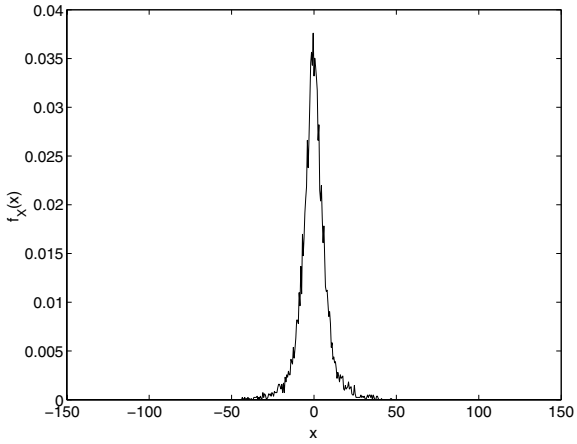

Fig. 4. Distribution of $X - X_1$ for samples from a time snapshot of the S-Pol radar data where X_1 are the sample values at nodes at a unit distance away from X.

Since X and Z are independent the characteristic function of f_A can be written as

$$\Phi_A(j\omega) = \Phi_X(j\omega)\Phi_Z(j\omega) = \Phi_X(j\omega) e^{[\frac{-{\sigma_z}^2\omega^2}{2}]}.$$

Hence, Equation (7) reduces to

$$\Phi_X(j\omega) = \frac{\beta}{1-(1-\beta)e^{[\frac{-{\sigma_z}^2\omega^2}{2}]}} \Phi_Y(j\omega). \quad (8)$$

Equation (8) describes the relationship between the characteristic functions, and hence the distribution functions, of the random variables X, Y and Z.

For mathematical convenience, we define a new random variable L having a characteristic function given by

$$\Phi_L(j\omega) = \frac{\beta}{1-(1-\beta)e^{[\frac{-{\sigma_z}^2\omega^2}{2}]}}. \quad (9)$$

(The random variable L will be used in the calculations in Section III-B.) Equation (8) can now be rewritten as

$$X = Y + L. \quad (10)$$

A. Parameters of the Model and Correlation

The presence of many parameters in the model gives us great flexibility to model processes having different correlation properties. In this section, we study how the tweaking of different parameters effect the correlation

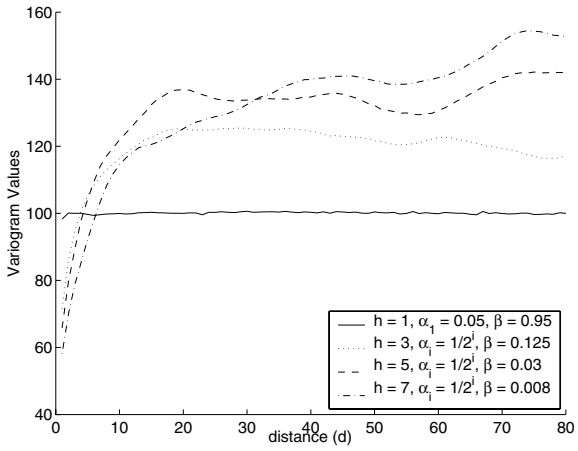

Fig. 5. Dependence of the spatial correlation properties of the data generated by the model on the model parameters: Variogram plots for different values of α_i's, β and h with $f_Y(y) \sim N(0, 10)$ and $f_Z(z) \sim N(0, 2)$.

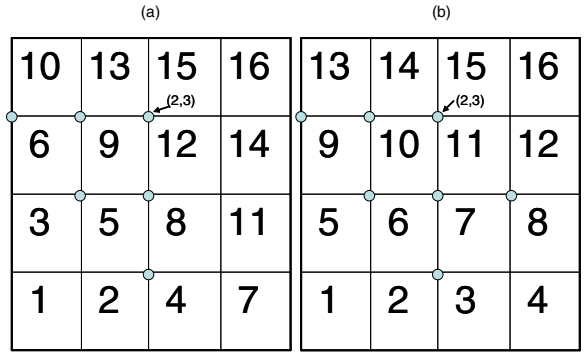

Fig. 6. Two methods to populate data. Note that the marked nodes correspond to the nodes contributing to the model in Equation (5) when node (2,3) is being visited for $h = 2$. The numbers indicate the order in which the nodes are being visited. (a) Method 1 : Data dependence in a quarter circular fashion ($N(i) = i + 1$) (b) Method 2 : Data dependence in a semi circular fashion ($N(i) = 2i$).

properties of the generated data. For example, it is easy to see that the values of β and h are going to play a major role in determining the correlation properties. Also, Equation (8) shows that $f_Y(y)$, $f_Z(z)$ and β are sufficient to determine the long term distribution of X.

We will now plot a few variograms for different values of α_i's, β and h, assuming $f_Y(y) \sim N(0, 10)$ and $f_Z(z) \sim N(0, 2)$. Figure 5 confirms that synthetic traces representing wide range of conditions can be generated from the same model by varying the model parameters. For example, for large β the process is very close to an iid process and the variogram is close to a straight line, while for smaller values of β, spatial correlation is strong for small distances.

The procedure we described can be also used to generate data on random and irregular topologies as follows. Since the granularity of the data is in our hands, we generate a data set on a regular grid but at a much higher granularity. Then, we keep random or irregularly distributed nodes as required, and throw away the rest to create a synthetic trace on an irregular topology.

B. Inferring Model Parameters

In this section we present techniques for inferring the parameters of the model from real traces.

We infer $f_X(x)$ from its empirical distribution. Inferring σ_z, α_i's and β is more involved. Suppose for now that a suitable value of h has been chosen. (We present a procedure for determining h at the end of this section.) We will compute the semivariance $\gamma(i)$, $i = 1, \ldots, h+1$, using the model, and equate it with its estimate $\hat{\gamma}(i)$ as obtained from the real trace. This will give $h + 1$ equations with unknowns σ_z, β and α_i's.

Using Equation (3),

$$\begin{aligned} \gamma(i) = & \ \tfrac{1}{2}E[(X - X_i)^2] \\ \Rightarrow \gamma(i) = & \ \tfrac{1}{2}E[E[(X - X_i^k)^2]|k] \\ \Rightarrow \gamma(i) = & \ \tfrac{1}{2N(i)}[E[(X - X_i^1)^2] + E[(X - X_i^2)^2] \\ & + ... + E[(X - X_i^{N(i)})^2]]. \end{aligned} \tag{11}$$

Let $d_{j_1 i_1}$ denote the distance between the points X_j^1 and X_i^1. Then, each of the terms in Equation (11) can be expanded as follows:

$$\begin{aligned} E[(X - X_i^1)^2] = & \frac{\alpha_1}{N(1)}E[(X_1^1 + Z - X_i^1)^2] \\ & ... + \frac{\alpha_1}{N(1)}E[(X_1^{N(1)} + Z - X_i^1)^2] + ... \\ & + \frac{\alpha_i}{N(i)}E[Z^2] + ... + \beta E[(Y - X_i^1)^2], \end{aligned}$$

where $E[(X_j^1 + Z - X_i^1)^2] = E[(X_j^1 - X_i^1)^2] + E[Z^2] = 2\gamma(d_{j_1 i_1}) + \sigma_z^2$, and $E[(Y - X_i^1)^2] = E[(Y - X)^2] = E[L^2]$. Hence,

$$\begin{aligned} E[(X - X_i^1)^2] = & \ 2(1 - \beta)\sigma_z^2 + 2\frac{\alpha_1}{N(1)}\gamma(d_{i_1 1_1}) + ... \\ & + 2\frac{\alpha_1}{N(1)}\gamma(d_{i_1 1_{N(1)}}) + ... + \beta E[L^2]. \end{aligned}$$

Using Equation (9), $E[L^2]$ is evaluated to be $\frac{(1-\beta)\sigma_z^2}{\beta}$.

To find the value of $N(i)$, the nodes contributing to the model in Equation 5 have to be specified. Also, the order in which the nodes are visited during the generation of synthetic data has to be specified.

Two methods of doing so are presented in Figure 6. As shown in the figure, the first method corresponds to a data dependence of a quarter circular fashion while the second method corresponds to a data dependence of a semi circular fashion. Which method to choose will depend on the physical phenomenon being modeled. We ran simulations for both the methods and the results were similar. So, from now on, we will use the first method to determine the order in which nodes are visited. In this case, it can be easily shown that $N(i)$ equals $i+1$.

Substituting all of the above in Equation (11) leads to the following:

$$\gamma(i) = (1-\beta)\sigma_z^2 + \frac{1}{2(i+1)} \sum_{j=1}^{i+1} \sum_{k=1}^{h} \sum_{l=1}^{k+1} \frac{\alpha_k}{k+1} \gamma(d_{i_j k_l}). \quad (12)$$

Equating $\hat{\gamma}(i) = \gamma(i)$ for $1 \leq i \leq h+1$ gives $h+1$ equations. These equations along with the equation $\beta + \sum_{i=1}^{h} \alpha_i = 1$ form a system of $h+2$ non linear equations with $h+2$ unknowns, the α_i's, β and σ_z^2. After solving the above system, $f_Y(y)$ can be obtained through Equation (8).

What remains is a procedure for determining h. In theory, overestimating h, which results in a larger matrix, would still find the correct parameters. However, in practice, larger h values leads to more rounding and statistical errors, hence to small negative α_i's in the solution of the non linear system. A solution to this is to start from an overestimated h, and lower its value until all the α_i's are positive.

IV. Model Verification and Validation

In this section, the model parameters are inferred from the traces described in Section II-A and then used to generate synthetic traces. First, we verify our model by comparing the statistics of the original experimental data trace and the corresponding synthetic trace. Then, we validate our model by comparing the performance of algorithms which exploit spatial correlation, against real traces and their synthetic counterparts.

A. Model Verification

1) S-Pol Radar data set: We choose a snapshot in time of the S-Pol Radar data as the experimental data trace. The S-Pol radar data set is a 64×64 spatial subset of the original data. The parameters of the model for the trace are inferred using the method described in Section III-B. Figure 7 presents the values of α_i's, β and σ_z inferred from the trace. Using these parameters,

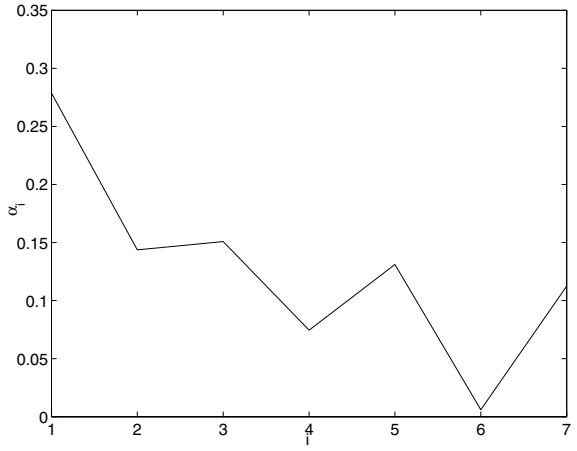

Fig. 7. α_i's for the S-Pol Radar data trace ($h = 7$, $\beta = 0.1$ and $\sigma_z = 1.25$).

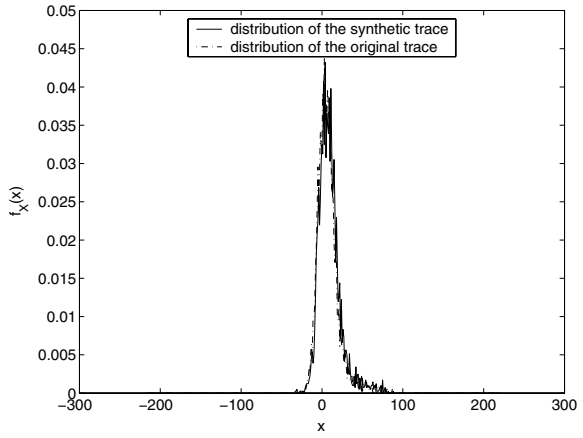

Fig. 8. S-Pol Radar data : Comparison of the distribution of the original and synthetic traces.

we generate a synthetic counterpart of the original trace. Then, we compare the statistics of both the traces. Figure 8 shows the long term distribution of the two traces and Figure 9 shows their respective variograms. Note the presence of edge effects due to insufficient number of samples for $d > 80$. As expected, the long term distribution of the two traces match closely. A look at the variograms tell us that the model is slightly underestimating the degree of spatial correlation in the data, though the exhibited correlation pattern is similar.

2) Precipitation data set: We choose a snapshot in time of the precipitation data as the experimental data trace. The precipitation data set is a 8×8 spatial subset of the original data. The parameters inferred for the trace are $h = 1$, $\alpha_1 = 0.72$, $\beta = 0.28$ and $\sigma_z = 2.61$. Using these parameters we generate a synthetic counterpart of the original trace. Then we compare the statistics of both

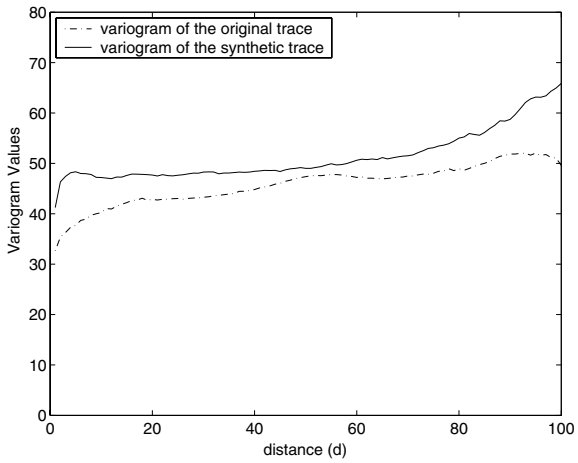

Fig. 9. S-Pol Radar data : Comparison of the variogram of the original and synthetic traces.

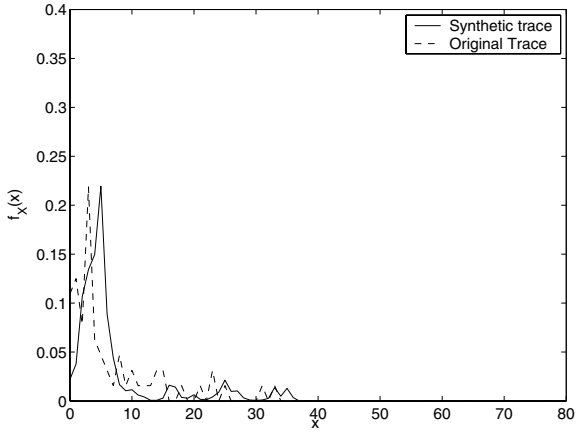

Fig. 10. Precipitation data : Comparison of the distribution of the original and synthetic traces.

the traces. Figure 10 shows the long term distribution of the two traces and Figure 11 shows their respective variograms. Both the distribution and variograms match closely.

B. Model Validation

There are several algorithms which try to exploit the presence of spatial correlation in sensor network data. A few of these algorithms were mentioned in the introduction. We selected two amongst them (DIMENSIONS [6] and CMAC [10]) to evaluate our model by comparing their performance for the original and the corresponding synthetic traces. We did not choose any of the algorithms which use entropies because we did not have enough time snapshots of sensor data traces to calculate the joint entropies. Though we have enough time snapshots of the precipitation data, the number of samples in space are

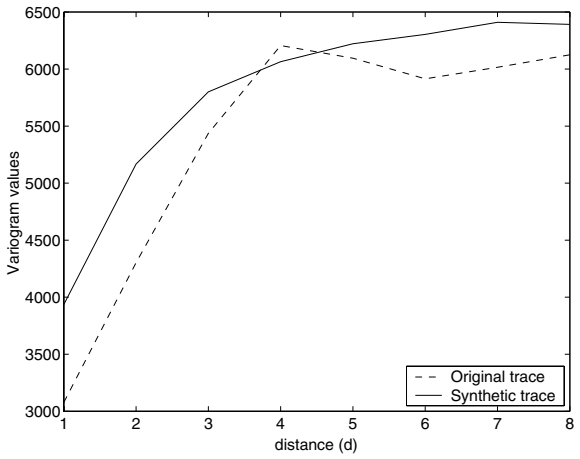

Fig. 11. Precipitation data : Comparison of the variogram of the original and synthetic traces.

not sufficient to be able to evaluate any algorithm. We went for algorithms which clearly defined how to process the sensor data and provided a meaningful metric for comparison. The algorithms along with the comparisons are discussed below.

1) DIMENSIONS: This is a data storage and querying algorithm. DIMENSIONS proposes wavelet based multi-resolution summarization and drill down querying. Summaries are generated in a multi-resolution manner, corresponding to different spatial scales. Queries on such data are posed in a drill down manner, that is, they are first processed on coarse, highly compressed summaries corresponding to larger spatial volumes, and the approximate results obtained are used to focus on regions in the network that are most likely to contain relevant information. A variety of queries can be posed on the data set; we present the performance results for the query $average(X)$. The evaluation metric used is the query error which is defined as $QueryError = (QueryResponseOverDimensions - ActualValue)/ActualValue$. In the DIMENSIONS hierarchy, each lower level stores twice the amount of data as the higher level. Therefore, as the query processing proceeds down the hierarchy gaining access to more detailed information, the query error should drop down gradually.

We only use the S-Pol radar data trace for evaluating algorithms because the precipitation data does not have sufficient spatial samples. We first choose a snapshot in time of the S-Pol radar data as the experimental data trace. After inferring the parameters of the model for the original trace, we generate a synthetic counterpart of the original trace. Figure 12 shows the result of running

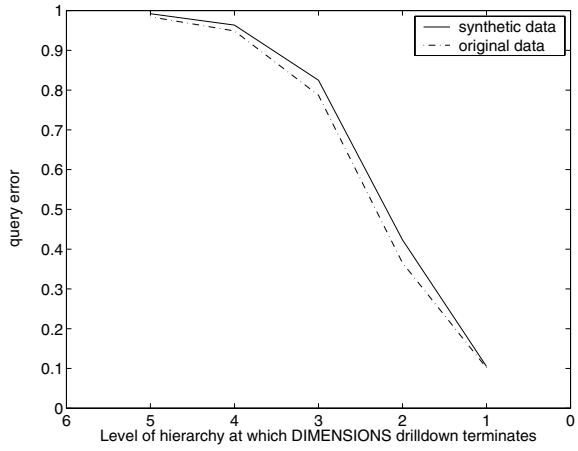

Fig. 12. Error vs Query Termination Level : Comparison of the performance of DIMENSIONS on original and synthetic trace.

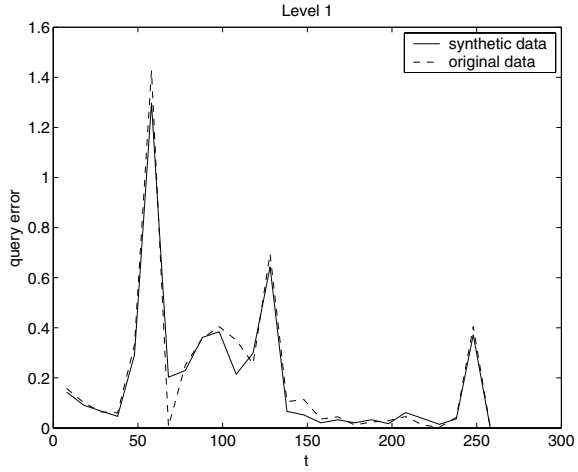

Fig. 13. Error at query termination level 1 at different snapshots : Comparison of the performance of DIMENSIONS on original and synthetic trace.

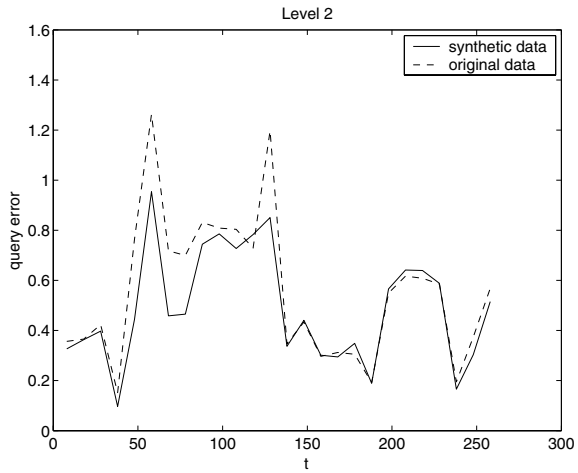

Fig. 14. Error at query termination level 2 at different snapshots : Comparison of the performance of DIMENSIONS on original and synthetic trace.

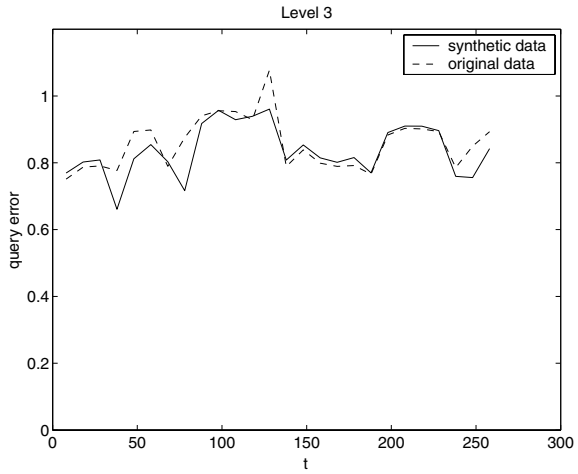

Fig. 15. Error at query termination level 3 at different snapshots : Comparison of the performance of DIMENSIONS on original and synthetic trace.

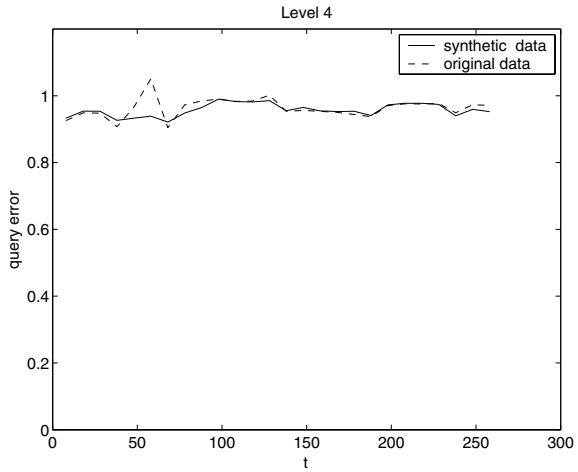

Fig. 16. Error at query termination level 4 at different snapshots : Comparison of the performance of DIMENSIONS on original and synthetic trace.

DIMENSIONS on both the traces. It is evident that the two plots match very well, thus we can infer that the synthetic data is able to capture the spatial correlation in the original data. To confirm the observation, we then infer the model parameters for different snapshots in time and run DIMENSIONS on both the original and synthetic traces. Figures 13-16 show the comparison for different query termination levels. It is obvious that the performance of the algorithm for both the traces is similar.

2) Spatial Correlation based Collaborative Medium Access Control (CMAC): Vuran et al [10] have argued that due to the presence of spatial correlation between sensor observations, it is not necessary for every node

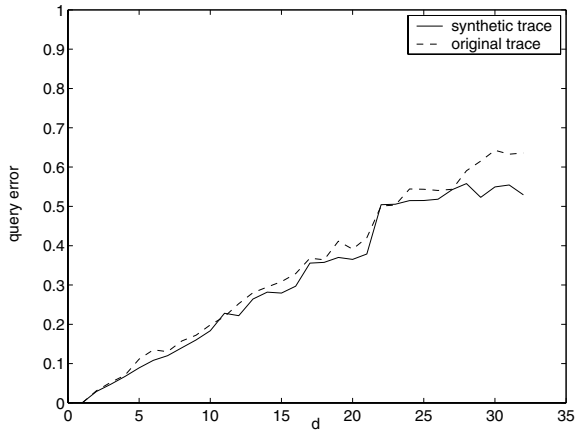

Fig. 17. Variation of error with d : Comparison of the performance of CMAC on original and synthetic trace.

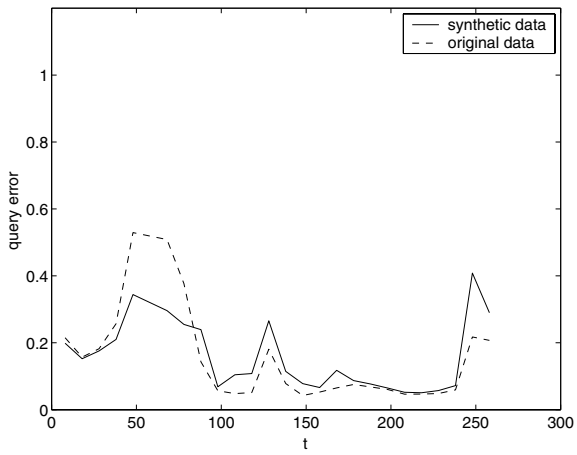

Fig. 18. Error for $d = 8$: Comparison of the performance of CMAC on original and synthetic trace.

to transmit its data. Amongst a cluster of sensor nodes, one of them can act as a representative of all other nodes. We refer to the node that sends information to the sink as the *representative node* of the cluster. Thus, a smaller number of sensor measurements are adequate to communicate the event features to the sink within a certain acceptable error.

Our simulation scenario has a 64×64 rectangular grid of sensor nodes. We present the performance results for the query *average(X)*. Since only a few of these sensor nodes will be transmitting data, the query result at the sink will not be accurate. So, as before the evaluation metric we use is the query error.

The cluster structure is assumed to be a square having a side d. Amongst all the nodes within this square, the representative node is selected randomly. Only one node in the cluster (the representative node) will transmit its data to the sink. The larger the value of d, the smaller is the number of nodes transmitting data to the sink, and hence larger is the error. For a given snapshot in time, Figure 17 plots the variation of error as a function of d for both traces. Then, we fix the value of d to 8 and run CMAC for traces at different snapshots in time and plot the error in Figure 18. It is obvious that the performance of the algorithm for both the traces is similar as the plots match closely. From the above plots, we claim that the model captures spatial correlation in the data.

V. Tools to Generate Large Synthetic Traces

In this section we describe two tools which we have created to help researchers generate synthetic traces of any size and degree of correlation. The tools are freely available at `http://www-scf.usc.edu/~apoorvaj`.

- *generateLargeTraceFromSmall* will create large synthetic traces having the same correlation structure as the input real data trace. It takes the real data trace and the dimensions of the output synthetic trace as its input. It also requires the user to specify the data dependence pattern. The user can choose either of the methods described in Section III-B.
- *generateSyntheticTraces* will create large synthetic traces representing a wide range of conditions by tweaking the model parameters. It takes the model parameters, h, α_i's, β, σ_z and $f_X(x)$ as its input in addition to the dimensions of the synthetic trace and the data dependence pattern.

Data collected from a testbed having a few sensor nodes is not sufficient to evaluate protocols. The first tool can generate a large trace having similar correlation properties as the real trace, and hence, help researchers to evaluate protocols with real data. The second tool will enable researchers to evaluate their protocols with data having varied correlation structures. Hence, these two tools will help researchers to evaluate their protocols with data representing wide range of realistic conditions without the need of actual dense deployment of sensor nodes.

VI. Conclusion and Future Work

In this paper, we have proposed a model to capture the spatial correlation in sensor network data. This model can generate synthetic traces representing a wide range of conditions and exhibiting any degree of correlation. We also described a mathematical procedure to extract the parameters of the model from a real data set. These model parameters are then used to generate synthetic

traces having similar correlation properties as the real data trace.

We verified our model by showing that the statistics of the synthetic trace is similar to the real data trace. We validated our model by showing that the performance of sensor network algorithms exploiting spatial correlation is similar for both the traces. For this purpose, we used the sensor network data storage and querying algorithm DIMENSIONS and the Spatial Correlation based Collaborative Medium Access Control algorithm CMAC. Finally, we have created two freely available tools to enable researchers to generate data representing real world scenarios and wide range of conditions.

This work assumes that the sensor nodes reside on a grid. This is a somewhat unrealistic assumption used to simplify the analysis. We are currently working on extending the model to cases where the nodes' placement follows a more natural pattern, e.g. a two dimensional poisson process. Finally we plan to evaluate our methodology with real sensor networks data as soon as such data becomes available.

ACKNOWLEDGMENT

We are grateful to NCAR for preparing the S-Pol data set. We also wish to thank Deborah Estrin and Yan Yu for providing us with the S-Pol radar data.

REFERENCES

[1] A. Goel and D. Estrin, "Simultaneous optimization for concave costs: single sink aggregation or single source buy-at-bulk," in *SODA*, 2003, pp. 499–505.

[2] S. Pattem, B. Krishnmachari, and R. Govindan, "The impact of spatial correlation on routing with compression in wireless sensor networks," in *Symposium on Information Processing in Sensor Networks (IPSN)*, Apr. 2004.

[3] C. Intanagonwiwat, D. Estrin, R. Govindan, and J. Heidemann, "Impact of network density on data aggregation in wireless sensor networks," in *ICDCS*, 2002.

[4] B. Krishnamachari, D. Estrin, and S. B. Wicker, "The impact of data aggregation in wireless sensor networks," in *ICDCS Workshop on Distributed Event-based Systems (DEBS)*, 2002.

[5] D. Petrovic, R. Shah, K. Ramchandran, and J. Rabaey, "Data funneling: routing with aggregation and compression for wireless sensor networks," in *Proceedings of the First IEEE Sensor Network Protocols and Applications (SNPA)*, May 2003, pp. 156–162.

[6] D. Ganesan, D. Estrin, and J. Heidemann, "Dimensions: Why do we need a new data handling architecture for sensor networks?" in *First Workshop on Hot Topics in Networks (Hotnets-I)*, Oct. 2002.

[7] D. Ganesan, B. Greenstein, D. Perelyubskiy, D. Estrin, and J. Heidemann, "An evaluation of multi-resolution storage for sensor networks," in *Proceedings of the First ACM Conference on Embedded Networked Sensor Systems (SenSys)*, Nov. 2003.

[8] J. Faruque and A. Helmy, "Rugged: Routing on fingerprint gradients in sensor networks," in *IEEE International Conference on Pervasive Services (ICPS'2004)*, July 2004.

[9] I. F. Akyildiz, M. C. Vuran, and O. B. Akan, "On exploiting spatial and temporal correlation in wireless sensor networks," in *Proceedings of WiOpt'04: Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks*, Mar. 2004, pp. 71–80.

[10] M. C. Vuran and I. F. Akyildiz, "Spatial correlation-based collaborative medium access control in wireless sensor networks," *submitted for publication, 2004*.

[11] N. Patwari and A. O. Hero, "Manifold learning algorithms for localization in wireless sensor networks," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2004.

[12] J. Chou, D. Petrovic, and K. Ramchandran, "Tracking and exploiting correlations in dense sensor networks," in *Conference Record of the Thirty-Sixth Asilomar Conference on Signals, Systems and Computers, 2002*, Nov. 2002.

[13] K. Whitehouse and D. Culler, "Calibration as parameter estimation in sensor networks," in *Proceedings of The First ACM International Workshop on Wireless Sensor Networks and Applications (WSNA'02)*, Sept. 2002.

[14] Y. Yu, D. Ganesan, L. Girod, D. Estrin, and R. Govindan, "Synthetic data generation to support irregular sampling in sensor networks," in *Geo Sensor Networks 2003*, Oct. 2003.

[15] K. Psounis, A. Zhu, B. Prabhakar, and R. Motwani, "Modeling correlations in web-traces and implications for designing replacement policies," *Computer Networks Journal, Elsevier*, vol. 45, no. 4, pp. 379–398, 2004.

[16] M. Widmann and C. Bretherton. 50 km resolution daily precipitation for the pacific northwest, 1949-1994. [Online]. Available: http://tao.atmos.washington.edu/data_sets/widmann

[17] W. Venables and B. Ripley, *Modern applied statistics with S*, 4th ed. Springer, 2002.

[18] R. A. Olea, *Geostatistics for engineers and earth scientists*. Kluwer Academic Publishers, 1999.