Title

Sensor Network Data Fault Detection Using Bayesian Maximum a Posterior Sensor Selection and Hierarchical Bayesian Space-Time Models

Authors

Ni, Kevin
Pottie, G J

Publication Date

2009-01-20

Sensor Network Data Fault Detection Using Bayesian Maximum a Posterior Sensor Selection and Hierarchical Bayesian Space-Time Models

Kevin S. Ni
University of California, Los Angeles
Electrical Engineering Department
kevinni@ucla.edu

Gregory J. Pottie
University of California, Los Angeles
Electrical Engineering Department
pottie@icsl.ucla.edu

Abstract

Data faults in sensor networks must be marked to ensure accurate inferences. We introduce a two phase semi-realtime end-to-end Bayesian fault detection system for sensor networks. The first phase selects a subset of agreeing sensors from which a model of expected behavior is derived. The second phase uses this subset to derive and tag questionable sensor data. To accurately model the data, we use a hierarchical Bayesian space-time (HBST) model, as compared to the linear autoregressive modeling used in previous works. Applying this system to simulated and real world data, results are excellent when the phenomenon is well modeled by the HBST model. It achieves high detection rates and almost zero false detection rates. Results also indicate that in cases of critically low spatial sampling density a more accurate model is required.

1 Introduction

Sensor networks have enabled new ways of observing the environment. Scientists are now able to access vast amounts of data to draw inferences about environmental behaviors [2] [3] [16] [8]. But, as sensor networks mature, the focus on data quality has also increased. Since the sensor nodes are left exposed to the sometimes harsh environment, they may fail or malfunction during a deployment, leading to faulty data and bad inferences. Thus, it is important to ensure data integrity and identify when data is faulty.

There has been significant effort in ensuring the integrity of data communications in large distributed ad-hoc networks, e.g. [5] [19] [24]. However, none of these focus on the problem of the integrity of the data itself.

Many deployment experiences show that this is a major issue that needs to be addressed. For example, with the goal of creating a simple to use sensor network application, [1] observes the difficulty of obtaining accurate sensor data. Following a test deployment, they note that failures can occur in unexpected ways and that calibration is a difficult task. Using this system, the authors of [23] deployed a sensor network with the goal of examining the micro-climate over the volume of a redwood tree. The authors discovered that there were many data anomalies that needed to be discarded post deployment. Only 49% of the collected data could be used for meaningful interpretation. Also, in a deployment at Great Duck Island, [21] 3% to 60% of data from each sensor classified as faulty. In both cases, these faults had to be manually identified with human knowledge of what was the expected beahvior of the phenomenon.

Additionally, [25] evaluates a sensor network in a volcano monitoring environment with high data quality requirements as measured by yield and data fidelity, concluding that sensor networks must still improve. Post deployment analysis of a wireless sensor network in [22] exposed many network packet losses in addition to several node level data problems.

With such high data fault rates, it is difficult to draw meaningful scientific inferences. In addition, any scientific conclusion would have an elevated uncertainty associated with it due to the high rate of sensor data faults. Therefore, to reduce this uncertainty and to aid scientists, we have developed an effective system using Bayesian methods to detect these faults.

Our approach pairs the Bayesian maximum a posterior (MAP) selection of an agreeing subset of sensors presented in [13] with the hierarchical Bayesian space-time (HBST) modeling approach first presented in [26] and adapted for sensor network fault detection in [14]. We develop a semi-realtime system to model and detect sensor network faults. With the improved accuracy that HBST modeling gives, we are able to effectively detect sensor network data faults while keeping the false detection rates low.

Also, since data faults are defined by a model of desired behavior, we will see how model accuracy and sensor deployment density factor into the performance of our fault detection system. Sensor deployment density is defined rel-

ative to the variability of the phenomenon. We will see that in cases of sufficient sampling densities, our model is accurate and our fault detection system performs well. However, when there is low sensor density, a much more accurate model must be used to capture spatial trends and patterns.

This paper is organized as follows. First, we discuss the previous work that focused on fault detection in sensor networks in section 2. We also discuss the general approaches for MAP sensor selection and HBST modeling as detailed in [13] and [14]. In section 3, we first discuss the assumptions that we make and how some may differ from previous papers. In section 4, we introduce the end-to-end Bayesian implementation of our fault detection system which is broken up into two phases. We then apply our method to the three different datasets to see the effectiveness of this end-to-end Bayesian system in section 5. The results show that this system can be very effective in some situations, but when an environment is under-sampled and the model is inaccurate, then this system has trouble.

2 Previous Work

The field of fault detection techniques in sensor networks has been growing as the use of sensor networks gains traction. To aid in fault detection, the most commonly seen sensor network data faults in environmental monitoring situations are detailed in [15] and [18]. [18] presents a few fault types and evaluates different approaches to detect these faults. [15] lays out modeling issues, features and indications of sensor data faults emphasizing that modeling expected behavior is crucial in detecting faults. Taking advantage of expected spatial and temporal correlations is a key component of effective detection systems, but this is not easily exploited. However, we will incorporate HBST modeling into our system to detect faults more effectively than existing methods.

[4] uses models of real-world processes based on sensor readings to answer queries to a sensor network for data. Using time-varying multivariate Gaussians to model data, the authors respond to a predetermined set of query types, treating the sensor network like a database. To some extent this shields the user from faulty sensors. However, the authors point out that more complex models should be used to detect faulty sensors and give reliable data in the presence of faults.

[11] discusses a cross-validation method of detecting in the presence of faults using a minimization of multiple unspecified sensor fusion functions. This requires heavy calculations with each additional datum. [10] uses a basic approach by dividing samples into temporal granules and co-located sensors into spatial granules. Each granule is then expected to be measuring similar data and anything outside of this is considered to be a fault. [12] uses linear autoregressive models to characterize data for error correction, targeting transient "soft" failures. With these linear models, the authors develop a predictive error correction technique using current data to decide past data.

Bayesian techniques are not new in the fault detection application. In [6], Bayesian updating for the distribution of an individual sensor's readings using prior distributions is employed. However, the prior knowledge of the phenomenon that is to provide the prior distribution is not given in detail, and the method does not explicitly take advantage of any spatial or temporal modeling. [7] uses spatio-temporal correlations to learn contextual information statistically. This contextual information is used in a Bayesian framework to detect faults. However, these Bayesian methods do not directly model the phenomenon.

In [13], a Bayesian maximum a posterior (MAP) criterion was used to select a subset of sensors whose local temporal trends agreed the best. However, this approach was limited in its success due to a lack of accurate models for the phenomenon of interest. The system relied upon linear autoregressive (AR) models which required careful selection of the window over which the data was modeled. Loose assumptions on the smoothness of the phenomenon field and the correlation among sensors also proved to be inaccurate when the system was applied to real data. We will however use a similar approach to selecting a subset of agreeing sensors for fault detection in our end-to-end Bayesian fault detection system.

In our end-to-end Bayesian sensor fault detection system, we use hierarchical Bayesian space-time modeling (HBST) to model the phenomenon of interest. The modeling method we use is based upon the framework developed in [26]. In [14], this framework was adapted for sensor network data fault detection and shown to be much more effective than linear AR modeling, which was heavily used in [13]. While HBST modeling is far more complex and computationally expensive than linear AR modeling, by auditing sensors on daily scale instead of with each new incoming data value, this computational cost becomes inconsequential.

3 System and Assumptions

In this paper, we make several assumptions on the system that are true for most current sensor networks. First we assume all of the data from S sensors is forwarded to the fusion center where data processing and fault analysis occurs. This assumption is implicitly or explicitly made in several of the works presented in section 2 including [15], [18], [10], and [11]. Corrupted or missing data communication packets are ignored and treated as unavailable.

Since we use HBST modeling to model the phenomenon, assumptions on the spatial and temporal behavior are de-

fined by the model setup as presented in section 4. The assumptions we make apply specifically to temperature data, however the HBST model can be easily changed for different types of phenomena and assumptions. Sensor measurements are assumed to include normal additive noise as well as the phenomenon process. The phenomenon process itself is composed of several components in addition to normal additive noise. The phenomenon is assumed to have a long term diurnal trend with an additional day to day linear trend. Each sensor has a site specific mean that is assumed to have a spatial linear trend. The phenomenon also has a time dynamic term which is assumed to be a diagonal vector autoregressive process.

Unlike [13], there is no assumption placed on the duration of faults. Faults may be as transient as a single outlier sample point, or they may persist throughout the deployment lifetime of the sensor network. Also, we make the restriction on the system that a minimum of $\lfloor \frac{S}{2} \rfloor + 1$ sensors must not be faulty at any given time. This is to avoid any confusion and to have a firm majority of sensors in the agreeing subset.

Also, since we use HBST modeling, we use the binning technique first introduced in [14] to have synchronous data. Real world data from all sensors are not usually synchronized so the data arriving to the fusion center cannot be easily vectorized. Most common space-time statistical tools assume that samples occur at regularly defined time intervals in a synchronized manner so that they may be easily placed in vectors at each time instant. In order to adapt real world data to such a scheme we "bin" the data by time instances for each sensor.

For one sensor, examining the regularly defined time instant at time t_i, we look at the interval surrounding this, which is of the size r, where r is the difference between time instances t_i and t_{i+1}. If within the interval $t_i - \frac{r}{2}$ and $t_i + \frac{r}{2}$ there is one sensor value for this sensor then the output at time t_i for this sensor is exactly this sensor value. If there are multiple sensor values within this interval, then the output at time t_i is the mean of all these values. However, if there is no data point, then a line between the two nearest surrounding data points is used to interpolate all values in between.

This process requires the data to be sufficiently sampled such that linear interpolation in between data points provides a good approximation. This primarily means that there are no large gaps in sensor data, otherwise interpolation will fail to effectively estimate data. We have observed this system to be effective for our data.

This process is shown to be an effective and accurate way of synchronizing data in [14] where the energy difference between the original data and synchronized data was insignificant.

4 End-to-end Bayesian implementation

We first introduce the end-to-end Bayesian implementation, and then we will discuss how this system improves upon the previous systems. We divide our end-to-end Bayesian implementation into two phases. The first phase will use a Bayesian MAP criterion to determine a set of agreeing subsets in a similar manner to [13]. The second phase will make a final determination as to whether certain data is faulty or not.

Since we use the same HBST modeling technique as [14], the primary weakness of our system is that the posterior simulation of the model parameters using MCMC techniques and Gibbs sampling is computationally expensive. Thus, we seek to minimize the frequency that we calculate parameters by specifying a semi-realtime detection system. By having this semi-realtime system, we can exploit the capabilities of HBST modeling while minimizing the impact of the high computation cost. Instead of performing calculations with each new incoming data value as is done in systems such as those in [13], [11], and [12], calculations are to be performed at regular time intervals at a time scale larger than the sensing intervals.

That is, sensor data integrity audits occur much less frequently than sensor samples are taken. For example, while the sensor data used in this paper measures the phenomenon on a scale of every 5 minutes, we will audit sensors every one day. This reflects logistical realities, in that it is unlikely for sensor replacement to be on the sensing time scale in the environmental sensing context, i.e. sensor replacement or repair does not usually occur immediately. Also it is common for a sensor to temporarily report questionable data and then return to normal [15]. Therefore, by having the audit occur at larger intervals, a sensor that returns to normal operating conditions will not be as frequently tagged.

4.1 Phase One

We begin the first phase by modeling data from all of the sensors over the course of the modeling window of size T, which is usually one day in our case. We use the HBST modeling approach of developed in [26] and adapted in [14]. Following the model development of [14], to adapt the HBST modeling to the data sets we are working with, we assume that sensors are deployed in a line. That is, we only have one spatial dimension in addition to the time dimension as we have easy access to multiple datasets with one spatial dimension. Considering two or three correlated spatial dimensions is possible, but doing so adds complexity and increases convergence time because of the multiplication of parameters.

The measurement process Z_t is composed of the phenomenon process with additive normal noise, giving Z_t a

distribution:

$$Z_t|\{Y_t, \sigma_Z^2\} \sim \mathcal{N}(Y_t, \sigma_Z^2\mathbf{I})$$

The phenomenon process, Y_t, consists of a site specific mean, $\mu(s)$, which has a first order linear trend in space, a daily harmonic with an additional linear day-to-day linear trend in time, $M_t(s)$, a time dynamic "diagonal" vector autoregressive process X, and some additive normal noise with variance σ_Y^2. This gives

$$Y_t|\{\mu, M_t, X_t, \sigma_Y^2\} \sim \mathcal{N}(\mu + M_t + X_t, \sigma_Y^2\mathbf{I})$$

We define the site specific mean to be a simple first order spatial regression. At a given physical location, l_s for sensor s we define:

$$\mu(s) = \mu_1 + \mu_2 l_s$$

Here, μ_1 is the overall mean of the phenomenon and μ_2 represents small corrections according to spatial trends. These two parameters of μ are modeled as independent normal random variables with fixed and specified priors.

$$\mu_1 \sim \mathcal{N}(\bar{\mu_1}, \sigma_{\mu_1}^2)$$
$$\mu_2 \sim \mathcal{N}(\bar{\mu_2}, \sigma_{\mu_2}^2)$$

The daily harmonic with spatially varying amplitudes and phases with an additional linear trend is defined as:

$$M_t(s) = (f_1 + f_2 l_s)\cos(\omega t) + (g_1 + g_2 l_s)\sin(\omega t) + h_1 t$$

where $\omega = 2\pi$ for a daily harmonic (when t is defined in units of days). f_1,f_2,g_1,g_2 define how the harmonic varies spatially. We add the h_1 term to account for the day to day weather trend over the modeling window.

We assume all of the parameters in M_t to be independent normal random variables with fixed and specified priors.

$$f_1 \sim \mathcal{N}(\bar{f_1}, \sigma_{f_1}^2)$$
$$f_2 \sim \mathcal{N}(\bar{f_2}, \sigma_{f_2}^2)$$
$$g_1 \sim \mathcal{N}(\bar{g_1}, \sigma_{g_1}^2)$$
$$g_2 \sim \mathcal{N}(\bar{g_2}, \sigma_{g_2}^2)$$
$$h_1 \sim \mathcal{N}(\bar{h_1}, \sigma_{h_1}^2)$$

We model the time dynamic term as a "diagonal" vector autoregressive process:

$$X_t = \mathbf{H}X_{t-1} + \epsilon_X$$

where

$$\mathbf{H} = a\mathbf{I}$$

giving

$$X_t|\{X_{t-1}, \mathbf{H}, \sigma_X\} \sim \mathcal{N}(\mathbf{H}X_{t-1}, \sigma_X^2\mathbf{I})$$

We assume that a is the same for all locations and it is normally distributed:

$$a \sim \mathcal{N}(\bar{a}, \sigma_a^2)$$

We specify the variances of the X, Y and Z to have an inverse gamma distribution, which is the conjugate prior to the normal distribution.

$$\sigma_Z^2 \sim \Gamma^{-1}(\alpha_Z, \beta_Z)$$
$$\sigma_Y^2 \sim \Gamma^{-1}(\alpha_Y, \beta_Y)$$
$$\sigma_X^2 \sim \Gamma^{-1}(\alpha_X, \beta_X)$$

The prior values for the parameters of these distributions are fixed and specified. Note that the prior distribution choices were made for ease in analytically deriving the conditional distributions. Alternative prior distributions can be used but may affect the complexity in the derivations of the conditional distributions for use in the Gibbs sampler as discussed in the following section.

With this model, we can generate samples from the distributions for each of the parameters using a Gibbs sampler. Gibbs sampling is a form of Markov chain Monte Carlo simulation and is a computationally efficient way of drawing samples from a target joint posterior distribution. Further information on the Gibbs sampler can be found in [9]. Details on the derivations of the full conditional distribution required by the Gibbs sampler are can be found in [14].

Once we have the parameters of this model, we can then begin the selection of the subset of agreeing sensors. For this step, we use the time dynamic term X_t as the basis of our decision. The time dynamic term removes all modeled trends, and only leaves unmodeled time dynamics to compare. This assumes that we have captured all of the space-time processes in our HBST model, resulting in the assumption that all time dynamic variations should be the same for each location.

We let the symbol $\hat{}$ represent the sample mean across all samples for the simulated posterior parameters. The first task is to estimate the covariance of this $\hat{X}$ term. Note that in our model, we modeled the vector autoregressive noise term to be independent spatial white noise for simplicity and simulation convergence reasons. However, here we no longer assume that X_t has no instantaneous spatial interaction, correcting for this deficiency. So, we first must determine the expected mean before we can estimate the covariance for $\hat{X}$.

We define the mean for sensor s to be determined by all other sensors excluding s. This allows for each sensor to be judged relative to the collective group without the possibility of self influence. That is, for sensor $s = 1, \ldots, S$ at time $t = 1, \ldots, T$ we define the mean to be:

$$\bar{X}_t(s) = \frac{1}{S-1}\Big(\sum_{r=1}^{s-1}\hat{X}_t(r) + \sum_{r=s+1}^{S}\hat{X}_t(r)\Big)$$

With this mean, we estimate the components of the covariance, Λ for $\hat{X}$ as:

$$\Lambda_{nm} = \frac{1}{T}\sum_{t=1}^{T}(\hat{X}_t(n) - \bar{X}_t(n))(\hat{X}_t(m) - \bar{X}_t(m))$$

We evaluate the likelihood for all size $\lfloor \frac{S}{2} \rfloor + 1$ subsets, ϕ. Each subset is represented by a vector ϕ whose components are $\{0, 1\}$ which represent exclusion and inclusion in the subset. We define the covariance for a particular subset ϕ, Λ_ϕ to be Λ with the appropriate rows and columns removed as indicated by ϕ. Similarly, the mean $\bar{X}_{t,\phi}$ and data $\hat{X}_{t,\phi}$, has the appropriate values of the vector removed indicated by ϕ. For each subset, we calculate the likelihood $f(\hat{X}_t(s)|\phi)$ to be:

$$\begin{aligned} f(\hat{X}_t(s)|\phi) &= \frac{1}{(2\pi)^{\frac{K}{2}}|\Lambda_\phi|^{\frac{1}{2}}}\exp\Big(-\frac{1}{2}\times \\ &\quad (\hat{X}_{t,\phi} - \bar{X}_{t,\phi})^T\Lambda_\phi^{-1}(\hat{X}_{t,\phi} - \bar{X}_{t,\phi})\Big) \end{aligned}$$

K is the number sensors in the subset.

In order to calculate the posterior probability, $P(\phi|\hat{X}_t)$, at time instant t to use in the maximum a posteriori criterion we now only need the prior probabilities $P(\phi)$. Initially, this can be set to be a uniform distribution indicating that there is no prior knowledge as to whether one sensor or subset is better than the others. This distribution is then updated each successive day according to the results of the fault detection in the second phase of the end-to-end system. Given the prior and the likelihood, the MAP criterion for a time instant t can be computed as:

$$P(\phi|\hat{X}_t) = \frac{f(\hat{X}_t(s)|\phi)P(\phi)}{\sum_{all\ \phi} f(\hat{X}_t(s)|\phi)P(\phi)} \quad (1)$$

Finally, to select the overall best agreeing subset, $\tilde{\phi}$, used to develop a model of expected behavior, we average over all $t = 1, \ldots, T$ this posterior value and select the maximum:

$$\tilde{\phi} = \arg\max_{all\ \phi} \mathbf{E}_t[P(\phi|\hat{X}_t)]$$

4.2 Phase Two

We can now use the agreeing subset of sensors, $\tilde{\phi}$, to develop the model of expected behavior for all sensors. Using just the sensors included in $\tilde{\phi}$, we reapply our HBST modeling technique to obtain new samples from the distributions for the model parameters. The new parameters based off of this subset, are averaged across all of the Gibbs sampling draws to produce the mean of all distributions, $\tilde{X}_t$, $\tilde{\sigma}_X^2$, $\tilde{\sigma}_Y^2$, $\tilde{\sigma}_Z^2$, $\tilde{\mu}_1$, $\tilde{\mu}_2$, $\tilde{f}_1$, $\tilde{f}_2$, $\tilde{g}_1$, $\tilde{g}_2$, and $\tilde{h}_1$.

For the determination of a data fault, we modify the fault detection method presented in [14]. We determine the upper and lower bounds of the process $Z_l(s,t)$ and $Z_u(s,t)$ and compare the actual data to these bounds.

We calculate 95% confidence intervals around the time dynamic term for each sensor in $\tilde{\phi}$ because each sensor was originally assumed to have the same dynamics. However, we recognize that our model is imperfect and there are still unmodeled dynamics. So we take the minimum and maximum limits of these 95% confidence intervals to determine an overall lower and upper bound on X_t. Thus, if a sensor's data ultimately lies within the bounds $Z_l(s,t)$ and $Z_u(s,t)$, then it behaves similarly to at least one of the sensors in the agreeing subset that defines the expected behavior.

The lower and upper bounds of the time dynamic term at a time t using the estimates of $\tilde{X}_t$ for the sensors in the subset $\tilde{\phi}$ are:

$$X_l(s,t) = \min_{all\ n\in\tilde{\phi}}(\tilde{X}_t(n) - 2\tilde{\sigma}_X)$$

$$X_u(s,t) = \max_{all\ n\in\tilde{\phi}}(\tilde{X}_t(n) + 2\tilde{\sigma}_X)$$

Since we are comparing real data Z_t with the bounds, we must also correct for spatial and temporal trends. So, for each sensor s at location l_s, we estimate the following:

$$\widetilde{\mu}(s) = \tilde{\mu}_1 + \tilde{\mu}_2 l_s$$

$$\widetilde{M}_t(s) = (\tilde{f}_1 + \tilde{f}_2 l_s)cos(\omega t) + (\tilde{g}_1 + \tilde{g}_2 l_s)sin(\omega t) + \tilde{h}_1 t$$

where $\omega = 2\pi$. Finally, the lower and upper bounds are:

$$Z_l(s,t) = \widetilde{\mu}(s) + \widetilde{M}_t(s) + X_l(s,t) - 2(\tilde{\sigma}_Y + \tilde{\sigma}_Z)$$

$$Z_u(s,t) = \widetilde{\mu}(s) + \widetilde{M}_t(s) + X_u(s,t) + 2(\tilde{\sigma}_Y + \tilde{\sigma}_Z)$$

With the bounds calculated, we then compare the data $Z_t(s)$ to see if it is within the bounds. If data is not within the bounds, it is tagged as being faulty.

Once all of the data is tagged, we can then update the prior distribution of the subsets for the next day's posterior calculation using the proportion of data for each sensor not tagged as being faulty. Using the actual results of how frequently a sensor is not tagged is a good indicator of a set's probability of agreeing.

The prior distribution update occurs as follows. Given the vector of data with the rates that each individual sensor has been tagged, τ, it is simple to calculate the fraction of samples that are correct for each sensor: $\mathbf{1} - \tau$. We can normalize this into the probability that each sensor should be included in the agreeing subset:

$$\eta_i = \frac{1 - \tau_i}{\sum_{all\ j}(1 - \tau_j)}$$

To determine the next day's subset prior we simply apply these probabilities to each set and normalize them:

$$P(\phi) = \frac{\phi^T \eta}{\sum_{all\ \phi} \phi^T \eta}$$

This distribution can then be used for the following day's phase one calculation of equation 1.

4.3 Comparison system

The end-to-end Bayesian system presented here is based upon elements of [13] and [14]. For the testing of the end-to-end Bayesian system we will compare our system to the HBST modeling system as presented in [14]. We compare against this system because it was already shown to be effective in detecting faults while reducing the false detection rates in comparison to linear autoregressive modeling used in [13].

The HBST modeling system introduces a rudimentary tagging technique upon which we have the second phase of the end-to-end Bayesian system. The HBST modeling system determines the lower and upper bounds of $Z_t(s)$ using just the two spatially adjacent sensors to determine minimum and maximum $X(s,t)$ values.

This simplistic system has two drawbacks which are eliminated by the use of a MAP agreeing sensor subset selection phase. The first is that the HBST modeling system may include the use of faulty sensors in the judgment of another sensor. This is non-ideal, and as we will see in the results, this can have a detrimental effect on the accuracy of fault detection. The MAP agreeing sensor subset does not contain faulty sensors.

The second issue is that edge sensors in the HBST modeling system see significantly higher false detection rates. This is due to the fact that only one sensor defines the bounds for an edge sensor instead of two. However, now that we have a whole subset of sensors that are trusted, there are multiple sensors defining the bounds, reducing the false detection significantly. Note though, for cases where an edge sensor is not in the agreeing subset, this sensor still may have slightly elevated false detection rates than the other sensors just because it is on the edge.

If one were to consider two and three dimensional sensor deployments, the fraction of nodes which are on the edge increases. Thus, use of a system as proposed here that employs a larger subset of the sensors to judge the edge sensors becomes even more important.

5 Results

To show the gains that we get from having an end-to-end Bayesian system, we apply our method to three datasets and compare the results to the HBST modeling system.

One data set is artificially generated and used as a toy example to illustrate under ideal conditions the performance of our system. The second data set is the cold air drainage data set from sensors that have been deployed at James Reserve in California. The last set of data is from a series of buoys deployed at Lake Fulmor, also at James Reserve. For this last set of data, we use the temperature measurements that are at the surface of the water.

Since we use the same underlying modeling as [14], all gains in performance are a direct result of the selection of an agreeing subset of sensors using MAP selection.

5.1 Simulated Data

We first use simulated data to test the system under ideal and well defined conditions. Spatial structure is well defined and matches very well to the assumptions made in our fault tagging scheme. We show results from data with no faults as well as injected faults to show the best performance of each system. Data from six sensors over three days exhibit a site specific mean, diurnal harmonics, a long term trend, and an additional unmodeled harmonic. A sample from three sensors of this data is presented in figure 1.

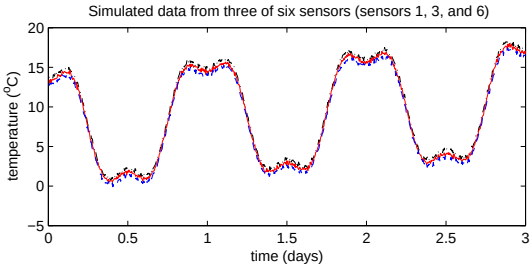


Figure 1. Simulated data. A sample of three days from three sensors.

We apply our end-to-end Bayesian fault detection system and the HBST Modeling system to the data presented. Prior distributions are calculated and updated for the next day. Results for false detection rates are presented in table 1.

Table 1. False detection rates for simulated data with no faults

	End-to-end	HBST
Overall	0	0.2079
Excluding Edge Nodes	0	0.0014
Just Edge Nodes	0	0.6210

The end-to-end Bayesian system makes no errors in false

detections. In each of the days, the agreeing subset consisted of the four middle nodes while neither edge sensor ever was in the agreeing subset. This is expected behavior because edge nodes agree the least with rest of the sensors. Even though these nodes were not in the agreeing set, there is still a lack of any false detection with these nodes indicating that the added sensors in the decision helped reduce the false detection rates. Also note the elevated false detection rates due to the edge node cases for the HBST modeling system. The end-to-end Bayesian system does not suffer from this deficiency.

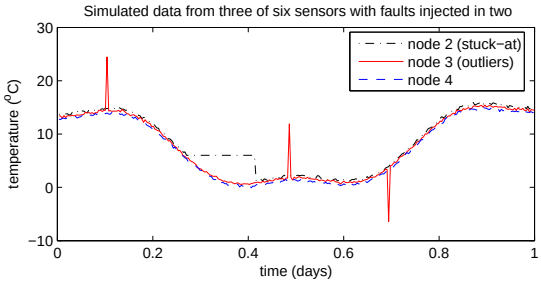


Figure 2. Simulated data with injected faults.

We next compare the the performance of our method when there are faults included in the data. One day was selected to have faults injected, and we tested the detection of each fault independently. We inject two types of common faults as defined in [18] and [15] at arbitrary locations. In figure 2 we show three sensors, two with faults, and one with no faults. One sensor has a "stuck-at" fault injected, and the other has outliers. After applying the end-to-end Bayesian system to this data, we obtain results that are summarized in figure 3.

Figure 3. Fault detection rates for simulated data with injected faults

The detection success rate for both fault cases is the same for both fault detection systems indicating that we have not lost any detection capabilities using the end-to-end Bayesian system. At the same time, the false detection rate has significantly dropped. There were no false detections in the case of the outlier, and the stuck-at fault case saw a 96.3% reduction in the false detection rate. Also, we note that when there were faults, the selection of the best agreeing subset always excluded the sensors containing the fault. These results show that the end-to-end Bayesian system is capable of detecting faults with a very low false detection rate.

5.2 Cold Air Drainage Data

Now, we examine the application of the end-to-end Bayesian system to real data. First we examine the case where data does not exhibit any apparent errors. We examine the false detection rate of six sensors deployed in James Reserve over the course of five days. Figure 4 shows temperature data from the first three sensors starting on September 17, 2005 over the course of five days.

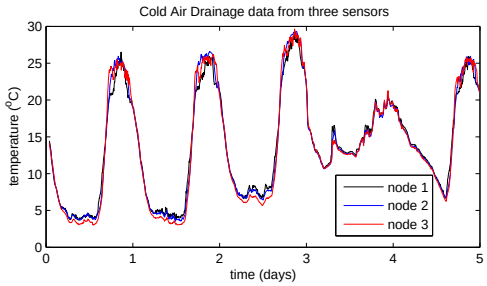


Figure 4. Data from three deployed sensors

The results after application of the end-to-end Bayesian system to this data are summarized in figure 5 We see that

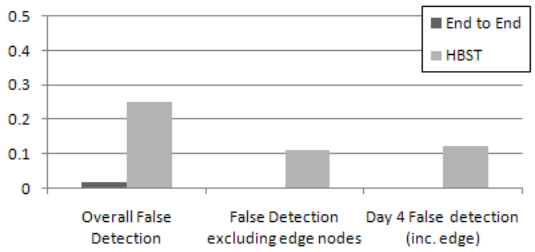


Figure 5. Fault detection rates for cold air drainage data with no faults

there is a significant difference in the overall false detection rate over the course of the five days. The end-to-end Bayesian system reduces the false detection rate from 25.2% for the HBST modeling system down to 1.8%.

Looking further, dropping the edge nodes from consideration to get the HBST modeling system's best overall performance, the HBST modeling system still has a 11.0%

overall false detection rate. Alternatively, when we disregard the edge nodes the end-to-end Bayesian system has no false detection over the course of five days. Also, in days 4 and 5 an edge node was included in the agreeing subset. This resulted in no false detection for the included edge node and no increase in false detection for any middle node.

Examining the temperature at the peak of each day, we see that the temperature is highly variable and dynamic. One possible cause of this is the passage of sunflecks where the sensor may be exposed to sun and shade alternatively due to forest coverage, wind, clouds, and passage of time during the day. This causes temperature readings to rise and fall in unexpected ways [20] [17]. The HBST system of [14] is unable to cope with these dynamics and sees higher false detection rates during these periods. However, the end-to-end Bayesian system is able to handle these dynamics.

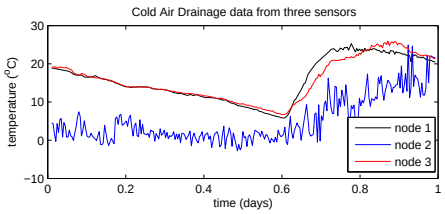
Day 4 does not exhibit these highly variable peak temperatures likely due to a either overcast weather or rain. The HBST modeling system performs its best during this day. For this day, we see that the end-to-end Bayesian system performance gives a 99.0% decrease in false detection rate. The end-to-end Bayesian system drops the false detection from 11.9% down to 0.1%.

The results for this non-faulty cold air drainage data show that the end-to-end Bayesian system is very capable of reducing false detection, even when there are many small unmodeled dynamics. Through the use of more sensors in developing a model of expected behavior, these unmodeled dynamics end up being averaged out leading to lower false detection rates. If the goal is to detect such phenomena, then increasing the spatial sampling density is required to detect these small scale dynamics.

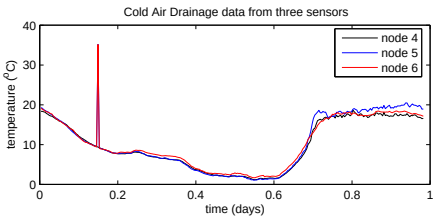
To ensure that this false detection reduction does not come at the expense of a decrease in detection rate, we apply the end-to-end Bayesian system to the two faults depicted in figure 6. Figure 6(a) shows data from three sensors for one day, Sept. 25, 2005, with one sensor giving likely faulty data, with high noise and readings distant from other sensors. The other two sensors that are physically located around this sensor are also shown. Figure 6(b) shows data from three neighboring sensors on Sept. 16, 2005 where two independent neighboring sensors exhibit outliers at the same instant. There is no conclusive reason for why this happened, but it is important to tag such an anomaly.

The results of applying our end-to-end Bayesian method to this data are summarized in figure 7. We see that both systems are very capable of detecting the faults. In both cases, the outliers from sensors 5 and 6 were detected perfectly even though they occurred at the same time, but the end-to-end Bayesian system had no false detections while the HBST modeling system had a false detection rate of 1.5%.

When examining the data with the noisy sensor, we see that both the end-to-end Bayesian system and the HBST



(a) Data from a faulty sensor



(b) Data with an outlier

Figure 6. Two examples of faults in real data

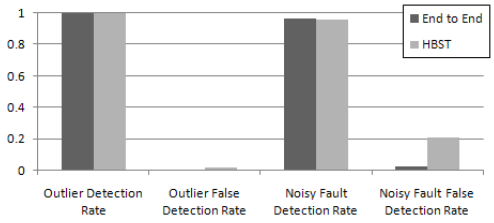


Figure 7. Fault detection rates for cold air drainage data with faults

modeling system were able to detect the faulty sensor very well. The end-to-end Bayesian system was slightly better at detection with 96.5% detection versus 95.8% detection with the HBST modeling system.

However, the biggest improvement was in the reduction of false detections in the presence of the noisy data of figure 6(a). The end-to-end Bayesian system improved on the HBST modeling system performance by 88.7%, reducing the false detection rate down to 2.3%. Note that all of the false detection for the end-to-end system was in sensor 1, with an individual rate of 11.6% false detection. This is expected because it is an edge sensor and since sensor 2 is faulty, the closest sensor in the agreeing subset is sensor 3. This is much improved in comparison to the HBST modeling system result. For the HBST modeling system, most (but not all) of the false detection was also in sensor 1, but the rate of false detection in this individual sensor was 97.9%. This is due to the fact that the only sensor involved in judging sensor 1 is the faulty sensor 2 in the HBST modeling system.

The results of the application of the end-to-end Bayesian

system to the cold air drainage, both non-faulty and faulty, indicate that this approach can be an effective way of determining sensor network data faults. This system is capable of effectively detecting data faults while maintaining low false detection rates.

5.3 Lake Fulmor Data

We now apply the end-to-end Bayesian system to the data from five nodes deployed across Lake Fulmor in James Reserve. We use temperature data collected at the surface from sensors deployed on buoys between August 28th and September 1st, 2006. Figure 8 shows data from three of the five sensors used in this test. Node 3 shows aberrant behavior starting at approximately day 2.65. This fault at the end of the data set is due to the battery failing on this particular node.

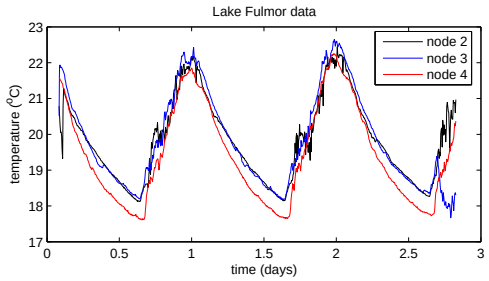


Figure 8. Data from three buoys at lake Fulmor

The results of the application of the two fault detection systems to this data are presented in figure 9.

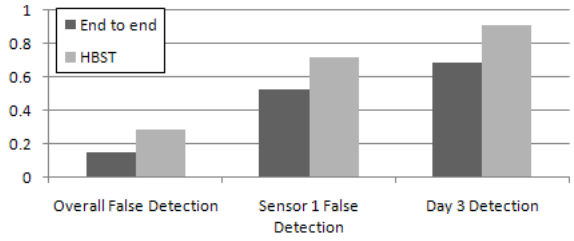


Figure 9. Detection rates for Lake Fulmor Data

The end-to-end Bayesian system has decreased the overall false detection in comparison to the HBST modeling system. The higher overall false detection in HBST modeling system is also linked to the higher detection rate, partially explaining the lower detection rate of the end-to-end Bayesian system. Also, we note that the false detection rate for sensor 1 is very high for both end-to-end Bayesian and HBST modeling systems.

Closer examination of the data reveals that the temperatures do not behave in a consistent clear spatially linear trend. To illustrate this point, figure 10 shows sensors 1 and 4 from the original data dataset. Sensor 1 clearly goes between being approximately $\frac{2}{3}^{\circ}$C warmer than sensor 4 to the same or even $\frac{1}{2}^{\circ}$C cooler than sensor 4.

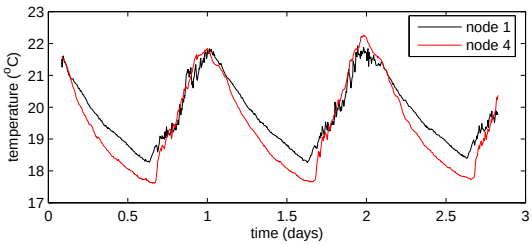


Figure 10. Data from two nodes in the Lake Fulmor data showing a inconsistent spatial trend

This behavior suggests that the phenomenon is incorrectly modeled. The linear spatial trend in the data is inconsistent, so it may be more useful to consider a higher order model for this data. However, the type and form of this model is unclear and not easily determined from the given data because of the low deployment density. This suggests the need for a second experiment with higher deployment densities to better derive a spatial model of the phenomenon. The main effect of this inaccurate spatial model is an increase in false detection coupled with a decrease in detection.

Sensor 1 is the most egregious of sensors causing the data to exhibit this non-linear behavior. If we reapply the end-to-end Bayesian system to the data disregarding sensor 1 we get significantly different results. With just sensors 2 through 4, the detection rate of the fault rises from 68.6% to 71.4%. The false detections for all sensors have all dropped to zero. The increase in detection and decrease in false detection indicates that the four sensors behave in a much more spatially linear manner.

The results of the application of the end-to-end Bayesian system to the Lake Fulmor data give mixed results. While overall fault detection rates dropped even in comparison to the HBST modeling system, this also resulted in a drop in detection rate.

These results also emphasize that a higher deployment density is required to accurately model a phenomenon that has high variability. The relatively low deployment density was sufficient to model the temperature field for the cold air drainage experiment in section 5.2 due to the low spatial variability. Meanwhile, the deployment density for the

Lake Fulmor experiment was not high enough to capture the higher spatial variability of the phenomenon. Without a higher deployment density it is difficult to determine what kind of spatial model should be used to sufficiently capture the phenomenon.

5.4 Effect of Prior Distribution

To show the effect of having the prior distribution updated by the final tagged proportion, we also tested the Lake Fulmor data without the prior distribution updating that was described in section 4.2. Table 2 show the sensors included in the agreeing subset when we exclude and include the use of the calculated prior distribution updates.

Table 2. Sensors included in the agreeing subset with and without prior distributions being used.

Day	No Priors	Using Priors
1	2,4,5	2,4,5
2	2,3,4	2,3,4
3	1,4,5	2,4,5

Examination of the results show that on the third day, sensor 1 is included in the agreeing subset. This is the sensor that does not follow the linear spatial trend. This sensor also exhibited a 75.7% tag rate on day 2 which is very high and suggests that this sensor should not be in an agreeing subset. Additionally, when we excluded the use of prior distributions, the detection rate of the fault in day three drops to 22.9%.

These results show that the inclusion and updating of prior distributions in the manner presented in section 4.2 is effective and crucial to the success of our algorithm. It validates the usage of a Bayesian MAP selection approach. It also ensures the exclusion of faulty or ill-modeled sensors from the agreeing subset, and because of this it improves the accuracy of the fault detection system.

6 Conclusion

We have shown the effectiveness of an end-to-end Bayesian modeling system using a MAP sensor subset selection combined with HBST modeling. By combining and improving elements of the approaches [13] and [14], several issues hurting detection performance are resolved.

However, we also see that accurate modeling is still very important as evidenced in the Lake Fulmor data of section 5.3. Also the relationship between accurate models and sampling density plays an important role in fault detection. High density deployments relative to the spatial variability of the phenomenon generally do not require complex accurate models. However the cost and challenges in deploying many sensors have limited this type of deployment in practice. Thus with lower density deployments we require better models. HBST modeling works best for deployments such as the ones presented here when sampling density is relatively low and simple modeling methods are no longer effective.

Fortunately, Bayesian techniques are particularly amenable to model updates as new data is collected. In the future, we seek to include an update to the model definition in an overall design approach and incorporate this into an iterative design for a sensor network deployment. This design update may include increasing deployment density to allow for better modeling of phenomena with high variability. In future work, we can further include Bayesian elements in this system by using Bayesian decisions in the selection of the models used in specific situations. Also, there may be other advanced statistical modeling techniques that may prove useful for other types of phenomena.

References

[1] P. Buonadonna, D. Gay, J. M. Hellerstein, W. Hong, and S. Madden. Task: Sensor network in a box. Technical Report IRB-TR-04-021, Intel Research Berkeley, Jan. 2005.

[2] G. Chen, N. Yau, M. Hansen, and D. Estrin. Sharing sensor network data. Technical Report 71, CENS, Mar. 2007.

[3] D. Culler, D. Estrin, and M. Srivastava. Guest editors' introduction: Overview of sensor networks. *Computer*, 37(8):41–49, Aug. 2004.

[4] A. Deshpande, C. Guestrin, S. R. Madden, J. M. Hellerstein, and W. Hong. Model-driven data acquisition in sensor networks. In *Proc. of Very Large Databases*, 2004.

[5] A. Durresi, V. Paruchuri, R. Kannan, and S. Iyengar. A lightweight protocol for data integrity in sensor networks. In *Intelligent Sensors, Sensor Networks and Information Processing Conference, 2004.*, pages 73–77, Dec. 2004.

[6] E. Elnahrawy and B. Nath. Cleaning and querying noisy sensors. In *Proc. of International Workshop on Wireless Sensor Networks and Applications (WSNA)*, 2003.

[7] E. Elnahrawy and B. Nath. Context aware sensors. In *Proc. of the First European Workshop on Wireless Sensor Networks (EWSN 2004)*, Jan. 2004.

[8] D. Estrin, L. Girod, G. Pottie, and M. Srivastava. Instrumenting the world with wireless sensor networks. In *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2001)*, June 2001.

[9] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC, second edition, 2004.

[10] S. R. Jeffery, G. Alonso, M. J. Franklin, W. Hong, and J. Widom. Declarative support for sensor data cleaning. In *4th International Conference on Pervasive Computing*, 2006.

[11] F. Koushanfar, M. Potkonjak, and A. Sangiovanni-Vincentelli. On-line fault detection of sensor measurements. In *Proc. of IEEE Sensors*, 2003.

[12] S. Mukhopadhyay, D. Panigrahi, and S. Dey. Model based error correction for wireless sensor networks. In *Proc. Sensor and Ad Hoc Communications and Networks SECON 2004.*, pages 575–584, Oct 2004.

[13] K. Ni and G. Pottie. Bayesian selection of non-faulty sensors. In *IEEE International Symposium on Information Theory*, June 2007.

[14] K. Ni and G. Pottie. Sensor network data fault detection using hierarchical bayesian space-time modeling. *Submitted to ACM Transactions on Sensor Networks*, 2008.

[15] K. Ni, N. Ramanathan, M. N. H. Chehade, L. Balzano, S. Nair, S. Zahedi, G. Pottie, M. Hansen, and M. Srivastava. Sensor network data fault types. *ACM Transactions on Sensor Networks*, 2008. accepted for publication.

[16] G. J. Pottie and W. J. Kaiser. Wireless integrated network sensors. *Communications of the ACM*, 43(5):51–58, May 2000.

[17] J. Ross, M. Sulev, and P. Saarelaid. Statistical treatment of the par variability and its application to willow coppice. *Agricultural and Forest Meteorology*, 91(1-2):1–21, 1998.

[18] A. Sharma, L. Golubchik, and R. Govindan. On the prevalence of sensor faults in real-world deployments. In *IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks (SECON)*, June 2007.

[19] E. Shi and A. Perrig. Designing secure sensor networks. *IEEE Wireless Communications Magazine*, 11(6), Dec. 2004.

[20] W. Smith, A. K. Knapp, and W. A. Reiners. Penumbral effects on sunlight penetration in plant communities. *Ecology*, 70(6):1603–1609, 1989.

[21] R. Szewczyk, A. Mainwaring, J. Polastre, J. Anderson, and D. Culler. An analysis of a large scale habitat monitoring application. In *2nd International Conference on Embedded Networked Sensor Systems (SenSys)*, Nov. 2004.

[22] R. Szewczyk, J. Polastre, A. Mainwaring, and D. Culler. Lessons from a sensor network expedition. In *Proc. of the 1st European Workshop on Sensor Networks (EWSN)*, Jan. 2004.

[23] G. Tolle, J. Polastre, R. Szewczyk, D. Culler, N. Turner, K. Tu, S. Burgess, T. Dawson, P. Buonadonna, D. Gay, and W. Hong. A macroscope in the redwoods. In *Proc. 3rd international conference on Embedded networked sensor systems (SenSys '05)*, 2005.

[24] H. Vogt. Integrity preservation for communication in sensor networks. Technical Report 434, ETH Zrich, Institute for Pervasive Computing, Feb. 2004.

[25] G. Werner-Allen, K. Lorincz, J. Johnson, J. Lees, and M. Welsh. Fidelity and yield in a volcano monitoring sensor network. In *7th USENIX Symposium on Operating System Design and Implementation*, Nov. 2006.

[26] C. K. Wikle, L. M. Berliner, and N. Cressie. Hierarchical bayesian space-time models. *Environmental and Ecological Statistics*, 5(2):117–154, Feb. 1998.