Separating the Wheat from the Chaff: Practical Anomaly Detection Schemes in Ecological Applications of Distributed Sensor Networks

Luís M.A. Bettencourt[1], Aric A. Hagberg[1], and Levi B. Larkey[2]

[1] Mathematical Modeling and Analysis, Theoretical Division
Los Alamos National Laboratory, Los Alamos, NM 87545
[2] Modeling, Algorithms, and Informatics, Computer and Computational Sciences Division
Los Alamos National Laboratory, Los Alamos, NM 87545

Abstract. We develop a practical, distributed algorithm to detect events, identify measurement errors, and infer missing readings in ecological applications of wireless sensor networks. To address issues of non-stationarity in environmental data streams, each sensor-processor learns statistical distributions of differences between its readings and those of its neighbors, as well as between its current and previous measurements. Scalar physical quantities such as air temperature, soil moisture, and light flux naturally display a large degree of spatiotemporal coherence, which gives a spectrum of fluctuations between adjacent or consecutive measurements with small variances. This feature permits stable estimation over a small state space. The resulting probability distributions of differences, estimated online in real time, are then used in statistical significance tests to identify rare events. Utilizing the spatio-temporal distributed nature of the measurements across the network, these events are classified as single mode failures - usually corresponding to measurement errors at a single sensor - or common mode events. The event structure also allows the network to automatically attribute potential measurement errors to specific sensors and to correct them in real time via a combination of current measurements at neighboring nodes and the statistics of differences between them. Compared to methods that use Bayesian classification of raw data streams at each sensor, this algorithm is more storage-efficient, learns faster, and is more robust in the face of non-stationary phenomena. Field results from a wireless sensor network (Sensor Web) deployed at Sevilleta National Wildlife Refuge are presented.

1 Introduction

Wireless sensor networks consist of multiple sensor-processor nodes that communicate with each other using radio frequencies. Sensor nodes, at present and in the envisioned future, are simple devices that operate within limitations in local memory storage and processing. These constraints, although by no means fundamental, are often the result of the practical considerations of producing devices that are inexpensive, small, and autonomous. In addition, sensor operations, and their communication in particular, are also limited by battery capacity or by the ability to harvest power, e.g. through solar panels.

J. Aspnes et al. (Eds.): DCOSS 2007, LNCS 4549, pp. 223–239, 2007.

Networks of distributed sensors are a promising technology because they can sense environments—natural and human made—over an unprecedented range of spatial and temporal scales [1, 2]. The large number of nodes required to cover large areas, over long times, places practical constraints on their individual cost. The drive for low-cost sensors and the need for unattended operation, frequently in harsh environments, requires simple and robust devices. Even the most robust devices, however, are subject to operational faults. Under these circumstances it is crucial that isolated errors in individual components do not jeopardize the operation of the whole network. Thus, an important issue for this emerging technology is data quality assurance and robustness of operation under point failures [3, 4, 5].

A general approach for robustness to point failures is to create partial functional redundancy among nodes in a sensor network. In some distributed sensor applications this emerges naturally because neighboring nodes measure local environments that are temporally and/or spatially correlated [6, 7]. Then, measurements at adjacent sensors, and at the same sensor over time, although potentially stochastic and non-stationary, display significant amounts of mutual information. Hence data quality can be assured through state co-inference between multiple, partially redundant and correlated readings from neighboring nodes, or from the same node at consecutive times [8, 9].

This paper presents a practical, distributed algorithm for detecting measurement anomalies - corresponding to both point failures and common mode events - and for estimating erroneous or missing data in ecological applications of wireless sensor networks. The algorithm has been designed for ecological sensing at the Sevilleta Long Term Ecological Research (LTER) site by a Sensor Web developed at NASA JPL [10, 2, 11]. Because it is designed to work under current technological constraints on memory and processing, the algorithm is intentionally simple and easy to implement. Processing can be performed locally on each node and requires only communication between proximal sensors. Such local, distributed algorithms are desirable for wireless sensor networks, where minimizing the amount of wireless communication is a necessary operational constraint [12].

The remainder of the paper is organized as follows. First, we describe related work on ecological applications of distributed sensor networks, and associated requirements for autonomous operation with emphasis on sensor measurement error detection and correction. We review related approaches in other contexts that use the distributed nature of the network for practical state co-inference, learning, and quality assurance and the performance and implementation requirements of direct Bayesian classifiers. Next, we describe the characteristics of the method, which performs automatic inference and prediction at a given sensor based on the distributions of differences of its measurements in time and in space relative to its neighbors. Finally we give several illustrations of the method's application to real data streams from a Sensor Web deployed at the Sevilleta LTER site, summarize our results, and discuss the outlook for future work.

2 Related Work

Ecological and habitat monitoring are natural applications for wireless sensor networks since the data often must be collected from remote areas that have little or no

communication infrastructure and from sensing systems that are often distributed over large geographic areas. Among other advances, wireless sensor networks permit better sensor placement, unhindered by wires, and may use on-board computational power to processing running statistics, perform hypothesis testing and even operate the experiments themselves [8, 13].

Present deployments are still far from fulfilling this promise, but have been invaluable in providing experience and highlighting the difficulties that arise from measuring data streams in the physical world [14, 15, 16]. Most of these problems arise from sensors and networks operating unattended in harsh, real-world conditions, with inadequate error identification and correction capabilities, and without sufficient algorithms to automatically quantify and actively reduce uncertainty [8, 13].

Several algorithms have recently been proposed that utilize statistical models to selectively acquire and summarize data in distributed sensor networks [17, 18]. Because of common climatic drivers, environmental signals at neighboring sensors are usually spatially and temporally correlated. Some methods explicitly explore the correlated nature of raw signals to reconstruct missing or erroneous readings [19]. Environmental data streams pose the additional challenge that signals are non-stationary, driven by diurnal and seasonal cycles, and by climatic events that never quite repeat. These features are typical of other sensing problems measuring physical and/or social environments. Here we propose an approach based on difference techniques, similar to those found in image [20] and signal processing [21], to factor out common drivers and capture the statistics of correlations between neighboring sensors. We show that this approach, complemented with the use of statistical tests to detect anomalous measurements, naturally leads to the identification of events with different structure, that can correspond to point sensor failures, or common mode events. The common mode anomalies may be erroneous or result from real spatio-temporally coherent events. In this way, missing or erroneous measurements at a sensor can also be automatically inferred via the joint consideration of neighboring readings and learned difference probability distributions.

Because of these general properties of environmental data streams, the straightforward application of standard statistical learning methods to environmental data streams must be performed with care. For example, Bayesian classifier methods [22] are a powerful way to perform sequential estimation, and are therefore a natural formalism for devising learning algorithms in distributed sensor networks. However, the direct implementation of such methods tends to run into the *practical* limitations of these simple devices. A recent proposal for *context-aware sensors* based on Bayesian classifiers uses statistical correlations between sensor readings to detect outliers and approximate missing readings [23]. We briefly review this method in the next section in order to provide context to the conceptual differences of our approach.

3 Bayesian Classifier Method

Assume that sensor measurements take values in the interval $[l, u]$, and let $R = \{r_1, ..., r_m\}$ be a disjoint cover of this interval. Each subinterval in R is considered a discrete class, with average precision $(u-l)/m$. Each node has its own classifier, consisting of the state

of that node's previous reading, h, and of the measurements from two (indistinguishable) nearby sensors, denoted as $n \in \{(r_i, r_j) \in R \times R, i \leq j\}$.

By Bayes' theorem, the conditional probability of a reading r_i, given the previous value h at that sensor and readings n from two nearby neighbors, is

$$P(r_i|h,n) = \frac{P(h,n|r_i)P(r_i)}{P(h,n)}. \tag{1}$$

In addition, to reduce the state space for inference, it is assumed in [23] that the neighbor's spatial measurements and the temporal information contained in the previous reading are conditionally independent,[1] given the reading of the sensor at the present time, yielding the "Naive Bayes" classifier

$$P(r_i|h,n) = \frac{P(h|r_i)P(n|r_i)P(r_i)}{P(h)P(n)}. \tag{2}$$

The output of the classifier is inferred using the method of maximum *a posteriori* (MAP) estimation [24], and is given by

$$\arg\max_{r_i \in R} P(r_i|h,n) = \arg\max_{r_i \in R} \frac{P(h|r_i)P(n|r_i)P(r_i)}{P(h)P(n)} = \arg\max_{r_i \in R} P(h|r_i)P(n|r_i)P(r_i)\,, \tag{3}$$

where the denominator can be omitted from the optimization because it does not depend on r_i.

This method is exhaustive and powerful in classifying all possible states of the system and learning their likelihood, but runs into practical implementation problems. To see this, consider that each node must learn the parameters of its classifier online. To learn $P(r_i)$, a node keeps a count of the number of times r_i occurs for each of m possible values. To learn $P(h|r_i)$, a node also keeps a count of the number of times h and r_i occur together for each of m^2 possible combinations. Similarly, to estimate $P(n|r_i)$, a node must keep a tally of the number of instances n and r_i occur together, for each of $(m^3+m^2)/2$ possible states. Finally, to compute probabilities for outlier detection, a node learns $P(n)$ online by keeping a count of the number of times n occurs for each of $(m^2+m)/2$ values. $P(h)$ is given by $P(r_i)$ where $r_i = h$ and a node must also keep a count of the total number of instances observed. Thus the total number of states stored is $m^3/2+2m^2+3m/2+1$. This expression was obtained by considering the measurements of a node relative to *two* neighbors. For $k > 2$ neighbors, the corresponding expression scales with leading exponent $k+1$.

The size of the state space required for inference is important for two reasons. First, nodes typically have limited storage capacity, which in turn limits precision. Consider the example of covering a range of 100 degrees with 1 degree precision. Then a classifier would have to store 520,151 counts, or roughly 2 megabytes. Secondly, the amount of learning data required to populate the state space is prohibitive in many cases. In the same example at least 5 million learning instances would be necessary for estimation (taken here to be roughly an order of magnitude greater than the size of the state space).

[1] We note that these assumptions do not apply to ecological environmental data under most circumstances.

To put this into perspective, consider that a node taking a reading every five minutes (e.g., [2]) would require about 47 years to populate its state space.

The issue of learning is even more critical in cases involving non-stationary phenomena because the learning rate cannot be slower than the rate at which parameters evolve. For example, in the case of outdoor air temperature, conditions change throughout the day as the sun rises, moves across the sky (e.g., placing sensors in and out of shadows), and sets. In addition, conditions also change with season and from year to year, such that combinations of data that occur frequently during a hot summer appear rarely during a cold winter, and will differ to the next summer. Thus an important discriminating criterion for any data quality assurance method is that it must operate on a timescale commensurate with that of any non-stationary phenomena being measured. For ecological sensing this time scale is typically less than a few hours.

4 A Method Based on the Statistics of Differences Between Sensor Measurements

We now propose a method for performing automatic event detection and data quality assurance, in which each node learns statistical distributions of *differences* between its readings and those of its neighbor's, and also between its own measurements at different times. Such distributions, together with current measurements are then used to identify anomalous measurements and to infer missing values. The inference of statistical distributions for measurement differences helps bypass issues of non-stationarity in environmental data streams, and leads, in general, to smaller ranges of statistical variables and better sampling for smaller datasets.

The crucial assumption required for the method to work is that the observed phenomena are spatiotemporally coherent, so that the measurements at neighboring sensors, and at the same sensor over time, display a large amount of mutual information. This is true of ecological applications, where typical node-to-node spacings are in the range of 100-200 meters or less. Moreover, environmental variables such as air temperature, relative humidity, light flux, soil temperature, and soil moisture display a substantial amount of temporal correlation as a result of common climatic drivers. It is assumed below that measurements at different sensors are performed at time intervals which are much smaller than the temporal correlation time of acquired signals, which we measured to be of order 1 hour. This is a characteristic of Sensor Web measurements, which are synchronous across the entire network and measurements can be taken every few minutes. An additional final assumption of the method is that the probability density of the differences has a peak near the mean and tails that taper as differences deviate away from it (e.g., see Fig. 1). That is, the method assumes that the probability of observing a difference decreases with the distance between that difference and the mean of all observed differences. This is not a strong assumption and could easily be relaxed in more complex circumstances if judged necessary.

Under these circumstances spatial and temporal measurement differences display a (much more) stationary distribution when compared to individual sensor readings. This permits more stable estimation of the statistics of differences over a much smaller state space. The estimation of differences between sensors placed at different

micro-environments, or between those and experimental controls can also capture quantities of direct ecological interest [13], e.g. by comparing control plots to treatments.

To set the context and notation for the method presented below consider then a node with k neighbors. Let ϕ be the node's reading, ϕ_0 be its previous measurement, and $\phi_i, i = 1,...,k$, be the readings of its neighbors. At each new measurement the node computes the difference between its current reading and its previous measurement and between its reading and each of its neighbor's $d_i = \phi - \phi_i, i = 0,...,k$. Given knowledge of the distribution of differences each new observation can be tested for errors. The probability of observing a difference d as or more extreme than d_i is its p-value, p_i

$$p_i = \min\left[P(d \leq d_i), P(d \geq d_i)\right], \tag{4}$$

where the probability P may refer to temporal differences $i = 0$ or differences with neighbor $i > 0$. For example, consider the distribution shown in Fig. 1, in which 88 percent of differences fall between -2 and 3, with 7 percent of differences less than or equal to -2, and 5 percent greater than or equal to 3. If $d_i = -2$, then $P(d \leq -2) = 0.07$ and $P(d \geq -2) = 0.93$. Thus $p_i = \min[0.07, 0.93] = 0.07$. Similarly, if $d_i = 3$, then $P(d \leq 3) = 0.95$ and $P(d \geq 3) = 0.05$. Thus $p_i = \min[0.95, 0.05] = 0.05$. The value of p_i in each instance is compared to a chosen significance level α. The measurement is flagged as anomalous if $p_i < \alpha$. We discuss how the combination of such p-tests between a sensor and all its neighbors identifies types of events below.

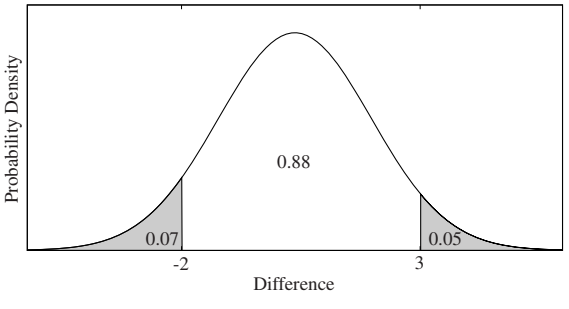

Fig. 1. An example of a probability density distribution illustrating the likelihood of observing an extreme difference. In this example, 88% of differences are between -2 and 3, with 7% of differences less than or equal to -2, and 5% greater than or equal to 3.

4.1 Statistical Inference

Each probability distribution $P(d)$ is learned from observed differences. There are several standard ways to implement this estimation, depending on the degree of prior knowledge. If the distributions are known to be well described by particular class of functions, then learning consists of estimating corresponding parameters. Filters, which specify sequential rules for parameter estimation, can then usually be constructed and optimized in order to minimize memory storage. If no parametric representation is adequate standard methods to construct non parametric distributions, in terms of frequency histograms, are employed.

Parametric estimation. If the distributions of differences are well fit by known distributions, estimation can be cast in terms of computation of distribution parameters from data. From the point of view of minimizing storage, estimation should be performed sequentially, so that only distribution parameters and current measurements are kept in memory at each single time. This can be achieved via the construction of filters to update estimators for distribution parameters [25].

Because distributions of differences of environmental variables are usually characterized by a small variance it is suggestive that, for sufficient number of observations, their shape may be well described by Gaussians. For a normal distribution $P(d)$ is defined by its mean and variance, which may be computed via standard maximal likelihood (unbiased) standard estimators, from t measurements as

$$\mu_{i,t} = \frac{1}{t}\sum_{k=1}^{t} d_{i,k}, \qquad \sigma_t^2 = \frac{1}{t-1}\sum_{k=1}^{t}(d_{i,k}-\mu_{i,t})^2, \tag{5}$$

which can be written using sequential updates as

$$\mu_{i,t} = \frac{(t-1)\mu_{i,t-1}+d_{i,t}}{t} \equiv \mu_{i,t-1} + K_t\,(d_{i,t}-\mu_{i,t-1}), \tag{6}$$

$$\begin{aligned}\sigma_{i,t}^2 &= \frac{1}{t-1}\left[(t-2)\sigma_{i,t-1}^2 + \frac{t}{t-1}(d_{i,t}-\mu_{i,t})^2\right] \\ &\equiv \frac{1}{1-K_t}\left[(1-2K_t)\sigma_{i,t-1}^2 + \frac{K_t}{1-K_t}(d_{i,t}-\mu_{i,t})^2\right],\end{aligned} \tag{7}$$

where t indexes times when differences are observed (for simplicity, assumed here to be synchronous across the network), and $\mu_{i,0} = \sigma_{i,0}^2 = \sigma_{i,1}^2 = 0$. K_t is usually referred to as the gain factor in the context of filters. In the familiar case of t observations which are equally weighted the maximum likelihood estimator implies that $K_t = 1/t$.

Because our observations are correlated, we use the functional freedom introduced by K_t to optimize inference of missing or erroneous values. (Similar procedures can be applied to parameters of other distributions.) By varying the gain factor K_t we obtain the best estimator for the distribution parameters under the joint constraints of a limited number of samples and non-stationary data. The limit as $K_t \to 0$ corresponds to no update of the distribution resulting from the current reading. Even if perfect *a priori* knowledge of the parameters is given at some time, this eventually fails because of the non-stationarity of environmental data streams. As a consequence, the error between actual and predicted data must increase, eventually, as $K_t \to 0$. On the other extreme, when $K \to 1$, only the current measurement is used in predicting the distribution. This fails because of the standard estimation problem that a small sample of realizations generates imprecise parameter determinations. This reasoning indicates that there is an intermediate value for K_t that minimizes the error between actual and inferred measurements. We illustrate these features in the next section with data from the Sensor Web deployed at the Sevilleta LTER site.

Estimation of non-parametric distributions. When the distributions are not known to belong to a particular class, non-parametric estimation is still straightforward, although resulting in larger memory requirements [26].

Here we perform the estimation of the probability density for differences as a simple frequency histogram by dividing the interval of possible differences, $[l_d, u_d]$, into m subintervals, dictated in most cases by the corresponding sensor resolution. In this sense discretization of measurements is unavoidable in practice and the non-parametric estimation introduces no further approximation. We should nevertheless keep in mind that binning of data to construct frequency histograms is usually acceptable only when the underlying distribution $P(d)$ is approximately constant over the bin size [26]. As discussed below (see Fig. 3) the sensor precision may suffice to satisfy this criterion Figs. 3 (b)-(d), or have single bins with considerable excess of observations [Fig. 3 (a)].

The average precision, $(u_d - l_d)/m$ achieved in the estimation of differences, is generally much higher than that of the Bayesian classifier, $(u-l)/m$, because $u_d - l_d$ is typically much less than $u - l$. For example, while temperature readings may range from 0 to 100 degrees, differences between temperature readings at neighboring sensors may only vary between -5 and 5 degrees. Thus using 100 subintervals yields an average precision of 0.1 degrees for this method versus 1 degree for the Bayesian classifier.

Sample size and memory requirements. The advantage of using a parametric estimation, whenever it is applicable, is that a node is not required to store previously observed differences; only the current estimates for the distribution parameters and the number of utilized instances are required. For a normal distribution this is μ_i and σ_i^2 for differences in time and differences in space relative to each neighbor, and also t. Thus the total storage required in this case is $2(k+1)$ floating point numbers and an integer, roughly 24 bytes for a node with two neighbors. In addition, the mean and variance can be approximated from as little as 10 observed differences. Other distributions which may be relevant in sensing problems such as Laplace, Poisson, or negative binomial, require similar or smaller estimation effort and memory storage.

To approximate $P(d)$ without parametric assumptions, as a frequency histogram, a node keeps a count of the number of times observed differences fall in each subinterval. The probability $P(d \leq d_i)$ is the sum of counts for subintervals overlapping $(-\infty, d_i]$, normalized by the sum of all counts. Therefore, in the non-parametric case, a node needs to store $m(k+1)$ integers or roughly $4m(k+1)$ bytes. For example, to cover a range of differences spanning 10 degrees with one degree precision, a node with 2 neighbors would have to store 30 states or roughly 120 bytes, whereas the Bayesian classifier would have to store roughly 2 megabytes. In addition, the amount of learning data required to populate the counts is much smaller than for the Bayesian classifier. For example, to cover a range of differences spanning 10 degrees with 1 degree precision would require about 100 observations (roughly an order of magnitude greater than the size of the state space), versus about 5 million learning instances for the Bayesian classifier. In terms of learning time for a node taking a reading every five minutes, this method would require about 9 hours, versus 47 years for the Bayesian classifier. In some cases, a number of measurements commensurate with the size of the state space may suffice, resulting in learning times an order of magnitude below these numbers; however, the ratio between the learning times for each method would be the same.

4.2 Statistical Anomalies: Error and Event Detection

The estimated distributions of differences enable the acceptance or rejection of new measurements based on their likelihood. We adopt a simple p-value test, as described above, to determine if a new measurement difference is significant. If the new difference fails the significance test it is flagged as anomalous. Table 4.2 illustrates how different event types are encoded in the structure of these tests between a reference node l and the ensemble of its neighbors. We consider three characteristic situations.

First, for a standard measurement all observed differences at all nodes are significant. We refer to this situation as a global significance consensus because all tests agree and are significant. In this situation readings should be accepted and used to update statistics. Next, if there is a single point failure at sensor l then it will observe a global failure consensus, indicating an anomaly in time, relative to its earlier reading, and to each of its neighbors. In this situation sensor l identifies its measurement as anomalous, and may discard it. Furthermore, and assuming no other point failures for simplicity, each of the neighbors of l observes that each of its observed differences is significant, except for that to sensor l. This allows them to identify an error at l and produce their estimate of l's correct reading. We return to this point below. Finally, if there is a common mode event across the network, an anomaly may be detected for temporal differences but a spatial significance consensus will still be observed. Each sensor observers this same structure of p-value tests. This type of event may indicate a common mode failure or a real event, such as rain. Such discrimination may be identifiable through the consideration of correlations across different types of sensors (air temperature, relative humidity, soil moisture) but lies beyond the scope of this work. Ambiguous events may also take

Table 1. Determination of event types from combined p-value tests

Event type	Pod l	Neighboring Pods
Standard measurement	$p_0 > \alpha$, $p_{i\neq 0} > \alpha$	$p_0 > 0$, $p_{j\neq 0} > \alpha$
Point failure	$p_0 < \alpha$, $p_{i\neq 0} < \alpha$	$p_{j=l} < \alpha$, $p_{j\neq l} > \alpha$
Common event	$p_0 < \alpha$, $p_{i\neq 0} > \alpha$	$p_0 < \alpha$, $p_{j\neq 0} > \alpha$

place, where a fraction of all differences may fail significance tests, but not be easily classifiable as a single point failure or common mode event.

It may be desirable to combine various combinations of p-value tests in time and in space to each sensor's neighbor into a single significance test, that e.g. identifies consensus. The combination of multiple p-value tests into a single significance test has a long history in statistics going back to the work of Tippett and Fisher in the early 1930s [27]. Fisher's method is still probably the most widely used procedure. It assumes that the p_i are independent and uniformly distributed and so the combination

$$-2\sum_{i=1}^{k} \ln(p_i)\,, \tag{8}$$

is distributed as a χ^2_{2k} distribution with $2k$ degrees of freedom. The significance of the joint p-value tests is then computed as the probability of obtaining a value as or more

extreme than that of expression (8) for a χ^2_{2k} distribution. Because this method of combining likelihood tests involves the geometric average of the p_i it is biased towards lower values of p_i and is not a good identifier of global or spatial consensus which, as indicated in Table 4.2, are the salient features of our expected events [27].

Several combinations of the set p_i which avoid these biases and are good identifiers of consensus have been proposed to address this issue. Among these, the z-transform test and the sum of p-values are the most widely used [28]. The z-transform test averages normal variables z each corresponding to a p_i and evaluates the significance level of this combination for a Gaussian distribution. Although the z-transform method is feasible, a much simpler method is the consideration of the sum

$$\overline{p} = \frac{1}{k}\sum_{i=0}^{k} p_i, \tag{9}$$

which can be compared to a desired significance level, typically of order α. This is the procedure we adopt below, guided essentially by simplicity. We emphasize, however, that many subtleties arise when taking into account the possible dependence of the several tests, which conditions the distribution of the variable combining the p_i, and consequently the nature of its significance test and choice of significance level as a function of those for individual tests. We intend to study these issues in future work with expanded datasets.

As a final remark, we note that if it is practical to perform the temporal and spatial (relative to neighboring sensors) significance tests independently, then a simple hierarchical structure for event classification becomes apparent. A temporal anomaly $p_0 < \alpha$ indicates an event. The event can be a point failure at the present sensor if there is also a spatial failure consensus, or a common mode if there is a spatial significance consensus. If no spatial consensus of either type is present the event is ambiguous and may be flagged for further study and possible creation of a new event class.

4.3 Inference of Missing Readings

As mentioned above the structure of temporal and spatial anomalies in the statistics of differences between a node and its neighbors allow a sensor to identify an error in its own measurement (global failure consensus) and its neighbors to identify the offending sensor and supply it with their estimation of its probable correct reading.

The most natural estimator of a sensor's missing or incorrect reading by neighbor i is simply

$$\hat{\phi}(i) = \phi_i + d_i, \tag{10}$$

where d_i is drawn from the distribution of differences between the two nodes. Averaging over d_i and over all neighbors leads to

$$\hat{\phi}_{\text{av}} = \frac{1}{k}\sum_{i=0}^{k}(\phi_i + \mu_i), \tag{11}$$

where $\hat{\phi}$ is the reading estimate and μ_i is the mean difference relative to the ith neighbor, or if $i = 0$, ϕ_0 is the previous reading and μ_0 is the mean difference between the

current and previous measurements. A weighted average based on a measure of mutual information (e.g. smaller variance) between the nodes could also be adopted, but we use the simplest scheme here. In the case where the distribution class is known, μ_i is a stored value. If instead the distribution class is not known, the mean difference can be approximated by the usual maximum likelihood estimator

$$\mu_i = \frac{1}{m} \sum_{j=1}^{m} c_j m_j, \tag{12}$$

where c_j is the count for the jth subinterval and m_j is the midpoint of the jth subinterval.

5 Application to Ecological Data from Sevilleta LTER Site

In this section, we test the method using ecological data collected by a Sensor Web, developed at NASA/JPL [11,2], deployed at the Sevilleta LTER site. A Sensor Web is a spatially distributed macro instrument, where every component sensor node (or "pod") shares its readings, at each measurement cycle, with all other pods in the system. The Sensor Web is designed to maintain synchronicity among all component pods which makes it ideal for the type of correlated statistical analysis proposed in the previous section.

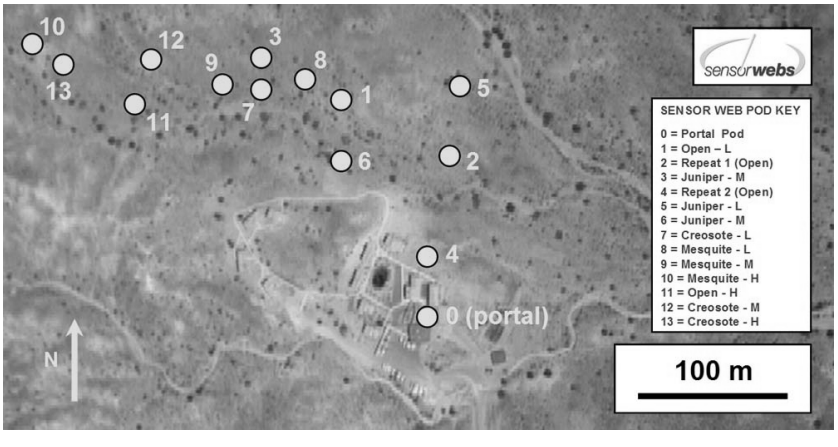

Fig. 2. Aerial photograph showing the Sensor Web layout at the Sevilleta LTER site. Fourteen sensor pods are distributed over a range of a few hundred meters to measure microclimate effects of the surrounding arid land plants. At regular time intervals the pods transmit data wirelessly to nearby pods. Sensor measurements eventually reach pod 0 where they are recorded.

The Sensor Web was initially deployed at the Sevilleta LTER site in 2003 as part of an ongoing effort to measure canopy microclimate effects of three arid land plant species: *Juniperus monosperma* (one-seeded juniper), *Larrea tridentata* (creosote bush), and *Prosopis glandulosa var. torreyana* (honey mesquite) [13]. The deployed Sensor Web

consists of 14 sensor pods (see Fig. 2) which measure temperature, humidity, light flux, soil temperature, and soil moisture and transmit the data wirelessly to nearby pods.

The method for inferring missing readings, presented in the previous section, was tested by comparing inferred values to actual measurements. In this example, see Figs. 2 and 3, we selected an environmental variable (air temperature), a pod (pod 5), a set of neighbors (pods 8, 9, 11, 12, and 13), and a period of time (the first 2 days of July, 2005). We used the parametric version of the method [Equations (6) and (7)] because the distributions of differences are approximately normal (e.g., see Fig. 3). Figure 4(a) shows the inferred and actual readings for pod 5. The average error over the time period was 0.717 degrees Celsius.

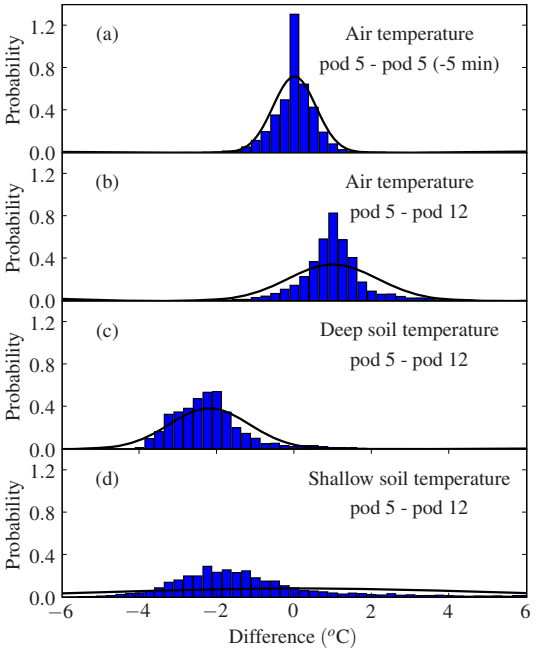

Fig. 3. A histogram of measurement differences recorded at the Sevilleta LTER site during July of 2004. (a) air temperature differences between pod 5 its previous reading (5 minutes earlier), (b) synchronous air temperature differences between pod 5 and pod 12, and (c) deep and (d) shallow soil temperature between the same two pods. The solid line shows a normal distribution with the same mean and variance as the data.

Because nodes have different placements, corresponding to distinct micro-climates, the distributions of differences are still weakly non-stationary. During warmer parts of the day, the more exposed nodes are warmer, but during cooler parts of the day (e.g. at night) the the more exposed nodes are cooler. Under these non-stationary conditions the average measurement error can be reduced by using Eqs. (6) and (7) with the appropriate value of K_t that optimizes the learning rate. Figure 4(c) shows the average error

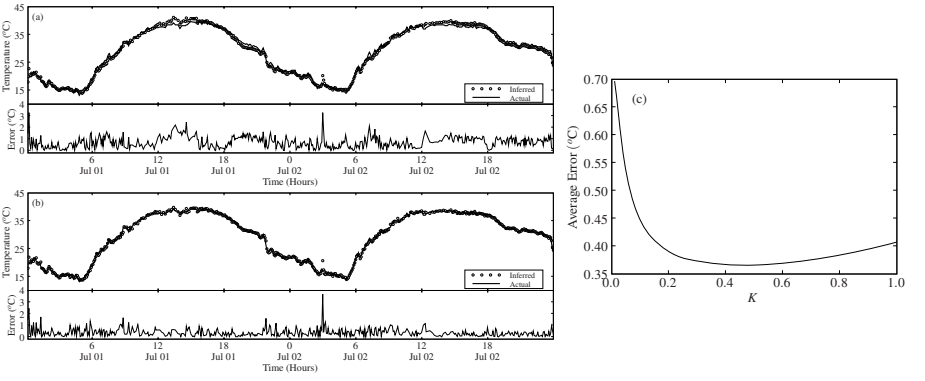

Fig. 4. Actual versus inferred air temperatures at sensor pod 5 for measurements taken in July 2005. The inferred measurements were computed using Eq. (11), with the average estimated via Eq. (6) with (a) $K_t = 1/t$, (b) $K_t = K = 0.46$. (c) The average error between the actual and inferred air temperature data as a function of the learning rate, K. The average error is computed using the entire two-day period of measurements. The minimum average error of 0.366 degrees Celsius is obtained for $K = 0.46$.

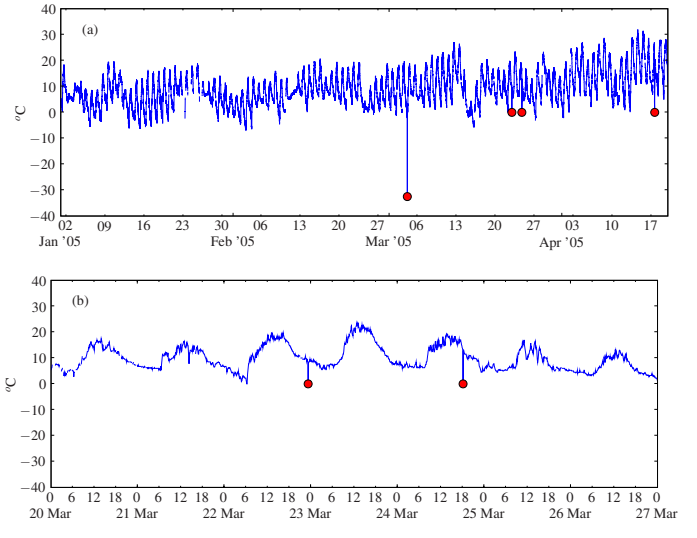

Fig. 5. Detected anomalies (marked by circles) in the pod 13 air temperature measurements for a period at the beginning of 2005. To detect the anomalies difference distributions for pod 13 (time difference) and pods 5, 11, 12, and 13 (space differences) were recorded for all of 2004 and the significance was computed using the combined p-value test of Eq. (9) with $\alpha = 0.005$. The method clearly captures the anomalies, as seen in (a), including some that are within the range of valid measurements. The two anomalies on March 23 and 24, shown in more detail in (b) are near zero degrees which is a common nighttime low temperature during that time of year.

as a function of $K_t = K$, assumed constant. The minimum average error of 0.366 degrees Celsius is achieved for $K = 0.46$. Figure 4(b) shows the inferred and actual readings for pod 5, using $K = 0.46$.

More generally we tested the anomaly detection and event type identification scheme on air and soil temperature measurements recorded during the first part of 2005. The difference distributions were computed from measurements of pod temperatures recorded during 2004. With the significance level for the combined p-value tests in Eq. (9) set to $\alpha = 0.005$, the method detects measurements that appear likely to be anomalous, as shown in Fig. 5, with no obvious false positives. Increasing α leads to the detection of more events, which may in some cases be due to instrument noise instead of outright failure. These effects can in principle be assessed if a model of instrument noise, and how it couples to true physical measurements, is provided. In this case, the distributions of differences between any two sensors may be understood in terms of the composition of failures, instrument noise, and physical measurements. Prior knowledge, or estimation, of the former may allow their subtraction from truly physical data streams. Here we have shown that, even in the absence of this knowledge, failures and instrument noise can be excluded from recordings and automatically corrected for at a chosen level of significance.

To understand these effects more clearly we have also applied our procedure to synthetically generated data containing diurnal and seasonal cycles, and with added small amplitude random white noise (to simulate instrument measurement imprecision) and larger amplitude infrequent fluctuations (to account for true sensor errors). The algorithm, with suitably adjusted significance, performed flawlessly at identifying sensor errors, over a variety of noise and failure amplitude and frequencies, provided the amplitude of errors is larger than the instrument noise.

6 Discussion and Outlook

We presented a practical, distributed algorithm for detecting statistical anomalies in ecological applications of distributed sensor networks. Both point failures and common mode events of sensors are identified and distinguished as statistical anomalies in the spatio-temporal structure of measurements between a sensor and its neighbors. Specifically, to avoid issues of non-stationarity, each sensor-processor learns the statistical distribution of differences between its measurements and each of its neighbors, as well as between its own measurements at consecutive times. Anomalies are detected, and their structure identified, in terms of statistical p-value significant tests for new measured differences relative to the expectations from these distributions.

The method is intentionally simple to cope with the limited memory and processing capabilities that characterize current sensor network technology. For this reason there are several directions for improvement. First, the operation of differencing, aimed here at factoring out the effects of common diurnal and seasonal drivers and reducing the size of the estimation space, can be achieved in principle by more sophisticated and accurate methods that are inspired by similar problems in image processing [20], signal processing [21] or component decomposition [29]. Methods for meta-analysis [30] to

combine a variety of statistical tests can also be constructed to take into account external information about sensor or environmental specificities.

While these are interesting directions for future research we also emphasize that, for the empirical environmental data streams discussed above, the methods developed here suffice. They have the added bonus of being simple and implementable in sensors with very small amounts of memory and processing. The consideration of further constraints such as hard energy limitations, specific network and routing geometries, etc., is not necessary for most practical ecological distributed sensing problems. Instead the real challenge typical of ecological applications (and shared by others that measure physical and/or social environments) is the unpredictable, non-stationary nature of data streams and the fact that measurements tend to relate only indirectly to the hypotheses of interest. These issues place the emphasis on methods that use the rich spatiotemporal structure collected by networks of sensors to provide reliable statistical inference and to identify multi-variable event structures that may allow the testing of high level hypotheses. We believe that differencing, broadly understood, combined with sequential real-time estimation and meta-analysis of simultaneous statistical tests are important ingredients of any method concerned with automatic event detection and error correction in distributed sensor networks.

From the practical point of view, we have also shown that the combination of these ingredients, when compared to an alternative method based on Bayesian classifiers, leads to algorithms that are more storage-efficient, learn faster, and are more robust to non-stationary phenomena. In addition, the storage, processing, and communication requirements are such that it can be implemented in a distributed fashion, on each of the nodes in the network, thus reducing remote communication. Because of these qualities, this class of algorithms can provide data quality assurance for current generation of wireless sensor networks, such as the Sensor Web deployed at the Sevilleta LTER site. In the process of learning distributions of differences for data quality assurance, the algorithm also produces statistics that compare different microclimate environments [13], to each other and to control experiments, which are of immediate scientific ecological interest.

Acknowledgments

This work was supported in part by a DCI Postdoctoral Fellowship to L. B. Larkey. We thank S. Collins and R. Brown at Sevilleta LTER, and K. Delin at NASA JPL, for enabling and encouraging this work, and R. Nemzek at LANL for discussions. The image in Fig. 2 was taken by the USDA-ARS Remote Sensing Research Unit, Subtropical Agricultural Research Laboratory, Weslaco, TX.

References

1. Szewczyk, R., Osterweil, E., Polastre, J., Hamilton, M., Mainwaring, A., Estrin, D.: Habitat monitoring with sensor networks. Communications of the ACM 47(6), 34–40 (2004)
2. Delin, K.A.: Sensor Webs in the wild. In: Bulusu, N., Jha, S. (eds.) Wireless Sensor Networks: A Systems Perspective. Artech House (2005)

3. Marzullo, K.: Tolerating failures of continuous-valued sensors. ACM Trans. Comput. Syst. 8(4), 284–304 (1990)
4. Elnahrawy, E., Nath, B.: Cleaning and querying noisy sensors. In: Proceedings of the Second ACM International Workshop on Wireless Sensor Networks and Applications (September 2003)
5. Bychkovskiy, V., Megerian, S., Estrin, D., Potkonjak, M.: A collaborative approach to in-place sensor calibration. In: Zhao, F., Guibas, L.J. (eds.) IPSN 2003. LNCS, vol. 2634, pp. 301–316. Springer, Heidelberg (2003)
6. Sharma, A., Leana Golubchik, R.G.: On the prevalence of sensor faults in real world deployments. In: Proceedings of the IEEE Conference on Sensor, Mesh and Ad Hoc Communications and Networks (SECON) (June 2007)
7. Jeffery, S.R., Alonso, G., Franklin, M.J., Hong, W., Widom, J.: Declarative support for sensor data cleaning. In: Fishkin, K.P., Schiele, B., Nixon, P., Quigley, A. (eds.) PERVASIVE 2006. LNCS, vol. 3968, pp. 83–100. Springer, Heidelberg (2006)
8. Estrin, D., Culler, D., Pister, K., Sukhatme, G.: Connecting the physical world with pervasive networks. IEEE Pervasive Computing 1(1), 59–69 (2002)
9. Tulone, D., Madden, S.: An energy-efficient querying framework in sensor networks for detecting node similarities. In: MSWiM'06 (2006)
10. Delin, K.A.: The Sensor Web: A macro-instrument for coordinated sensing. Sensors 2, 270–285 (2002)
11. Delin, K.A., Jackson, S.P., Johnson, D.W., Burleigh, S.C., Woodrow, R.R., McAuley, J.M., Dohm, J.M., Ip, F., Ferre, T.P.A., Rucker, D.F., Baker, V.R.: Environmental studies with the Sensor Web: Principles and practice. Sensors 5, 103–117 (2005)
12. Meguerdichian, S., Slijepcevic, S., Karayan, V., Potkonjak, M.: Localized algorithms in wireless ad-hoc networks: Location discovery and sensor exposure. In: Proceedings of MobiHOC 2001, Long Beach, CA, pp. 106–116 (2001)
13. Collins, S.L., Bettencourt, L.M.A., Hagberg, A., Brown, R.F., Moore, D.I., Delin, K.A.: New opportunities in ecological sensing using wireless sensor networks. Frontiers in Ecology 4(8), 402–407 (2006)
14. Szewczyk, R., Polastre, J., Mainwaring, A., Culler, D.: Lessons from a sensor network expedition. In: Karl, H., Wolisz, A., Willig, A. (eds.) Wireless Sensor Networks. LNCS, vol. 2920, Springer, Heidelberg (2004)
15. Ramanathan, N., Balzano, L., Burt, M., Estrin, D., Harmon, T., Harvey, C., Jay, J., Kohler, E., Rothenberg, S., Srivastava, M.: Rapid deployment with confidence:calibration and fault detection in environmental sensor networks. Technical Report 62, CENS, UCLA (2006)
16. Werner-Allen, G., Lorincz, K., Johnson, J., Lees, J., Welsh, M.: Fidelity and yield in a volcano monitoring sensor network. In: Proceedings of the 7th USENIX Symposium on Operating Symposium (OSDI 2006) (2006)
17. Deshpande, A., Guestrin, C., Madden, S.R., Hellerstein, J.M., Hong, W.: Model-driven data acquisition in sensor networks. In: 30th International Conference on Very Large Data Bases, pp. 588–599 (2004)
18. Liu, K., Sayeed, A.: Asymptotically optimal decentralized type-based detection in wireless sensor networks. In: Acoustics, Speech, and Signal Processing, IEEE International Conference (ICASSP '04), vol. 3, pp. 873–876 (2004)
19. Gupta, H., Navda, V., Das, S.R., Chowdhary, V.: Efficient gathering of correlated data in sensor networks. In: MobiHoc '05: Proceedings of the 6th ACM international symposium on Mobile ad hoc networking and computing, pp. 402–413. ACM Press, New York (2005)
20. Radke, R.J., Andra, S., Al-Kofahi, O., Roysam, B.: Image change detection algorithms: A systematic survey. IEEE Trans. on Image Proc. vol. 14(3) (2005)
21. Markou, M., Singh, S.: Novelty detection: A review - part 1: Statistical approaches. Signal Process. 83(12), 2481–2497 (2003)

22. Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B.: Bayesian Data Analysis, 2nd edn. CRC Press, Boca Raton (2003)
23. Elnahrawy, E., Nath, B.: Context-aware sensors. In: Karl, H., Wolisz, A., Willig, A. (eds.) Wireless Sensor Networks. LNCS, vol. 2920, pp. 77–93. Springer, Heidelberg (2004)
24. DeGroot, M.H.: Optimal Statistical Decisions. Wiley, Chichester (2004)
25. Maybeck, P.S.: Stochastic Models, Estimation, and Control. In: Mathematics in Science and Engineering, vol. 141, Academic Press, San Diego (1979)
26. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification, 2nd edn. Wiley, Chichester (2000)
27. Rice, W.R.: A consensus combined p-value test and the family-wide significance of component tests. Biometrics 46(2), 303–308 (1990)
28. Folks, L.J.: Combination of independent tests. In: Krishnaiah, P.R., Sen, P.K. (eds.) Handbook of Statistics 4. Nonparametric Methods, North Holland, New York (1984)
29. Lakhina, A., Crovella, M., Diot, C.: Diagnosing network-wide traffic anomalies. SIGCOMM Comput. Commun. Rev. 34(4), 219–230 (2004)
30. Hedges, L.V., Olkin, I.: Statistical Method for Meta-Analysis. Academic Press, San Diego (1985)