Data Quality and Failures Characterization of Sensing Data in Environmental Applications*

Kewei Sha[1], Guoxing Zhan[2], Safwan Al-Omari[2], Tim Calappi[2], Weisong Shi[2], and Carol Miller[2]

[1] Oklahoma City University, Oklahoma City OK 73106, USA
ksha@okcu.edu
[2] Wayne State University, Detroit MI 48202, USA
{gxzhan,somari,tcalappi,weisong,cjmiller}@wayne.edu

Abstract. Environmental monitoring is one of the most important sensor network application domains. The success of those applications is determined by the quality of the collected data. Thus, it is crucial to carefully analyze the collected sensing data, which not only helps us understand the features of monitored field, but also unveil any limitations and opportunities that should be considered in future sensor system design. In this paper, we take an initial step and analyze one-month sensing data collected from a real-world water system surveillance application, focusing on the data similarity, data abnormality and failure patterns. Our major findings include: (1) Information similarity, including pattern similarity and numerical similarity, is very common, which provides a good opportunity to trade off energy efficiency and data quality; (2) Spatial and multi-modality correlation analysis provide a way to evaluate data integrity and to detect conflicting data that usually indicates appearances of sensor malfunction or interesting events; and (3) External harsh environmental conditions may be the most important factor on inflicting failures in environmental applications. Communication failures, mainly caused by lacking of synchronization, contribute the largest portion among all failure types.

1 Introduction

As new fabrication and integration technologies reduce the cost and size of wireless micro-sensors, we are witnessing another revolution that facilitates the observation and control of our physical world [3,12], just as networking technologies have changed the way individuals and organizations exchange information. Environmental monitoring, targeting at discovering and understanding the environmental laws and changes, is one of the most important sensor network application domains.

With the increasing number of deployments of sensor systems, in which the main function is to collect interesting data at the sink, it is becoming crucial to carefully analyze the large amount of collected data. However, this problem is neglected in previous research, which mainly focuses on energy efficient, reliable sensor systems design and optimization. Although data quality management attracts more and more attention in

* This work is in part supported by NSF grant CNS-0721456. The work was done when the first author was a graduate student at Wayne State University.

E. Bertino and J.B.D. Joshi (Eds.): CollaborateCom 2008, LNICST 10, pp. 679–695, 2009.

the last two years [10,15], proposing novel data quality management mechanisms is still an important and interesting research topic. We argue that sensor system optimization and data quality management are closely related to the characteristics of collected data, in other words, sensor system optimization and data quality management should take data characteristics into consideration. Thus, in this paper, we take an initial step to characterize the data quality and failures using a set of one-month data collected by a real-world water system surveillance application. The data set consists of water level, precipitation, and gauge voltage measurements from 13 gauges located around Lake Winnebago, St. Clair River and Detroit River in January 2008.

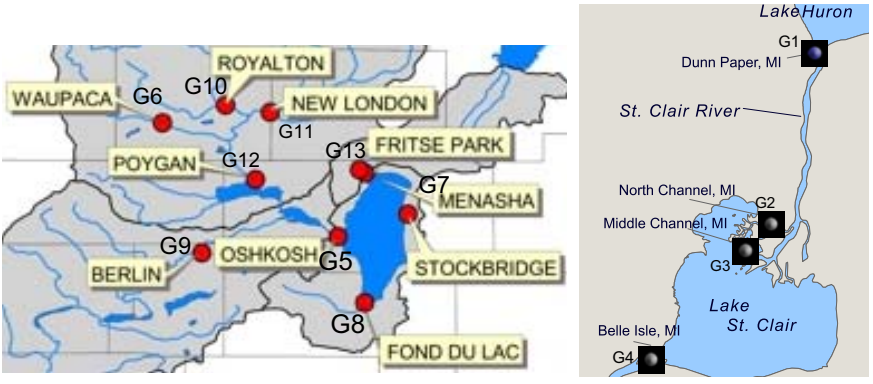
Our data analysis focuses on two aspects: *quality oriented data analysis* and *failure pattern analysis*. In quality oriented data analysis, we intend to discover two types of data, namely similarity data and abnormal data, whereas, in failure pattern analysis, we try to classify the common failure type and record failure time. The significance of our discovery is two-fold. On one hand, it helps us understand the laws and changes in the monitored field. On the other hand, it unveils the limitations in the current sensor system design, and provides us with a strong ground upon which we can base our future WSN systems design.

Our study reveals several interesting facts. First, information similarity, including pattern similarity and numerical similarity, is very common, which provides a good opportunity to trade off energy efficiency and data quality. Second, different parameters exhibit different data characteristics, which suggests that adaptive protocols using variable sampling rates can bring in significant improvements. Third, spatial correlation analysis and multi-modality correlation analysis provide a way to evaluate data integrity and to detect conflicting data that usually indicates appearances of sensor malfunction or interesting events. Fourth, abnormal data may appear all the time, and continuous appearance of abnormal data usually suggests a failure or an interesting event. Finally, external harsh environmental conditions may be the most important factor on inflicting failures in environmental applications. Communication failures, mainly caused by lack of synchronization, contribute the largest portion among all failure types.

The rest of the paper is organized as follows. A brief background of the targeting application and data is described in Section 2. We conduct quality oriented data analysis in Section 3, followed by failure pattern analysis in Section 4. Finally, related work and conclusion are listed in Section 5 and Section 6 respectively.

2 Background

The United States Army Corps of Engineers (USACE) in Detroit District, has 22 data collection platforms commonly referred to as sensor nodes or gauges, deployed around the St. Clair and Detroit rivers in southeast Michigan as well as the Lake Winnebago watershed southwest of Green Bay, Wisconsin. One month data in January 2008 from 13 of the 22 gauges were made available for this study. Each sensor node collects battery voltage, water level and precipitation except the Dunn Paper gauge (G1) which collects battery voltage, air temperature and water temperature. However, precipitation data for the St. Clair/Detroit river system is not used in this work, because that "data" in the raw files is simply an artifact of the gauge programming. For convenience, we name each



(a) (b)

Fig. 1. Map of: (a) Lake Winnebago Watershed, (b) St. Clair River and Detroit River

sensor node as G1, where 'G' stands for "gauge." The gauge locations of G1, G2 and G3 are on the St. Clair River, G4 is located on the Detroit River, and G5 through G13 are spread around the Lake Winnebago watershed. Gauges G5 through G13 are shown in Figure 1(a), and the remaining gauges are shown in Figure 1(b). It is worth mentioning that G8 and G10 suffered many failures throughout the period of study. Therefore, any data analysis on G8 or G10 will mostly look like "weird" or at least different from the other gauges.

Data samples are sent from each gauge to the GOES satellite, once every hour or every four hours, depending on whether the station has a high baud rate transmitter or not. High baud rate transmitters send data every hour. Data is then sent from the satellite to a central location in Wallops Island in Virginia, where the data samples are collected and arranged in files for later download through a regular ftp service. We conducted our analysis directly using the un-decoded files. This raw data set has not been subject to any quality control procedure, and thus provide a good opportunity to study failures happening in sensor network. Water level and precipitation are sampled once every hour, whereas voltage is sampled once every hour or every two hours. Water level is measured against the IGLD Datum 1985, which works as the base to measure current water level. So negative water level means it is below the local IGLD Datum 1985, though that does not happen often. Precipitation data is supposed to be constantly increasing (except when the gauge resets as part of its normal operation). To figure out how much precipitation fell over a one-hour period, the difference between consecutive samples are calculated and reported. The measurements for voltage and precipitation are in volts and inches respectively. Water level is reported in meters, centimeters and feet. For convenience, we converted the readings for water level in meters.

3 Quality-Oriented Sensing Data Analysis

In this section, we focus on quality-oriented sensing data analysis. How to define data quality is still an open problem. Here, We define high quality data as the data that contains the most information from the monitored field.

To understand the quality of the collected data, we try to discover the spatial and temporal relationship among those data; specifically, we are mostly interested in detecting two types of data, redundant data and abnormal data. Usually, redundant data, which we name as similarity, will not affect the overall quality of the collected data when it is removed. Contrary to redundant data, abnormal data, which largely affects the data quality, should be examined more carefully, because it usually denotes sensor failures, malicious attacks or interesting events.

3.1 Time Series Analysis for Individual Parameter

For each individual parameter, we define two types of similarity, *the pattern similarity* and *the numerical similarity*. Here, we define a pattern as the continuous reappearance of the same value sensed by one sensor, and the number of continuous reappearance is called pattern length. Note that a pattern must have a minimum length of 2. Thus, each pattern is a two tuple $< key, length >$. For example, if the sensor reads a series of $4, 4, 4, 5, 5, 4, 5$, we detect the patterns as $< 4, 3 >$ and $< 5, 2 >$, and the number of appearance of each pattern is 1. We use the pattern reappearance ratio, which is the ratio of the pattern data in the whole data, to measure the pattern similarity among the data. The numerical similarity records the number of reappearance of the same numerical value. For example, if a sensor reads a series of $4, 4, 4, 5, 5, 4, 5$, we get numerical similarity as 4 times of appearance of value 4 and 3 times of appearance of value 5. Similarly, the numerical similarity ratio is used to evaluate the value similarity, which is defined as the ratio of the reappeared sensor readings in all sensor readings.

Pattern Similarity. We detect all patterns for all monitoring parameters in 13 gauges. Here, we pick up gauge G5 as a typical example to show the patterns we detected as well as reappearance times of the pattern, which is shown in Figure 2.

From the figure, we do find specific patterns in the collected data, and the number of total patterns is small for all three parameters. Water level of gauge G5 has the largest number of patterns, which is 33, whereas, precipitation has the smallest number of patterns, which is 19. Some patterns have very large pattern length. For example, the largest pattern length for water level and for precipitation are 77 and 139 respectively. This indicates that water level and precipitation values stay constant for a long period of time at the area where G5 located. The number of appearances of each pattern is mostly small, especially for patterns with large length. This is because we set up endurance interval as $[0.00, 0.00]$, thus, very small difference between two keys, such as 14.66 and 14.67, are distinguish. Here endurance interval is an interval within which the difference between two readings can be ignored, for example, if the endurance interval is $[-0.02, 0.02]$, 14.67 and 14.66 can be regarded as the same. Because of unavoidable system error in measurement and applications lowered requirements on accuracy, it is reasonable to set up an endurance interval for each monitoring parameter. Another reason is that we define different patterns even when they have the same key value but different lengths.

Figure 3(a) shows the pattern reappearance ratio, where "-" means there is no available data. In the figure, we find that voltage has the smallest pattern reappearance ratio, which suggests that the changes of the voltage are very frequent. This is also because

V Pattern	Appearances	W Pattern	Appearances	P Pattern	Appearances
<14.09, 2>	1	<0.67, 2>	1	<1.67, 2>	1
<14.58, 2>	1	<0.68, 2>	1	<1.77, 2>	1
<14.59, 2>	1	<0.69, 2>	2	<2.28, 2>	1
<14.63, 2>	2	<0.70, 2>	1	<2.50, 2>	1
<14.64, 2>	1	<0.71, 2>	2	<2.52, 2>	1
<14.66, 2>	2	<0.72, 2>	1	<2.30, 5>	1
<14.67, 2>	2	<0.68, 3>	1	<2.47, 5>	1
<14.68, 2>	3	<0.69, 3>	2	<2.24, 7>	1
<14.69, 2>	1	<0.70, 3>	2	<2.34, 7>	1
<14.71, 2>	1	<0.71, 3>	4	<2.40, 8>	1
<14.73, 2>	1	<0.72, 3>	2	<1.69, 16>	1
<14.57, 3>	1	<0.70, 4>	2	<2.36, 22>	1
<14.60, 3>	1	<0.71, 4>	1	<2.20, 38>	1
<14.67, 3>	3	<0.72, 4>	3	<2.36, 44>	1
<14.72, 3>	1	<0.69, 5>	2	<2.11, 49>	1
<14.65, 4>	1	<0.70, 5>	4	<2.53, 67>	1
<14.67, 4>	2	<0.70, 9>	1	<2.41, 97>	1
<14.68, 4>	3	<0.72, 9>	1	<2.51, 102>	1
<14.69, 4>	2	<0.71, 11>	1	<1.29, 139>	1
<14.61, 5>	1	<0.71, 14>	1		
<14.64, 5>	1	<0.68, 16>	1		
<14.67, 5>	1	<0.71, 16>	1		
<14.68, 5>	2	<0.71, 18>	1		
<14.68, 6>	1	<0.71, 22>	1		
<14.70, 6>	1	<0.67, 28>	1		
<14.67, 7>	2	<0.68, 29>	1		
<14.69, 7>	1	<0.71, 30>	1		
<14.68, 8>	2	<0.70, 34>	1		
<14.68, 9>	2	<0.70, 38>	1		
<14.68, 10>	1	<0.69, 45>	1		
<14.68, 11>	1	<0.69, 47>	1		
<14.67, 12>	1	<0.68, 69>	1		
		<0.72, 77>	1		

Fig. 2. Detected patterns and the number of appearance in gauge G5

Gauge ID	Voltage Pattern Reappearance Ratio	Water Level Pattern Reappearance Ratio	Precipitation Pattern Reappearance Ratio
G1	0.07	-	-
G2	0.85	0.64	-
G3	0.78	0.43	-
G4	0.38	0.57	-
G5	0.50	0.88	0.92
G6	0.17	0.84	0.89
G7	0.96	0.69	0.93
G8	-	-	-
G9	0.23	0.87	0.89
G10	0.04	0.44	0.77
G11	0.10	0.83	0.90
G12	0.87	0.94	0.91
G13	0.05	0.88	0.89

(a)

Gauge ID	Voltage Pattern Reappearance Ratio [-0.1, 0.1]	Water Level Pattern Reappearance Ratio [-0.02, 0.02]	Precipitation Pattern Reappearance Ratio [-0.02, 0.02]
G1	0.59	-	-
G2	1.00	0.85	-
G3	1.00	0.74	-
G4	0.96	0.78	-
G5	0.88	0.99	0.95
G6	0.98	0.94	0.92
G7	1.00	0.99	0.95
G8	-	-	-
G9	1.00	0.94	0.92
G10	0.64	0.72	0.79
G11	0.78	0.93	0.94
G12	1.00	0.99	0.94
G13	0.80	0.95	0.94

(b)

Fig. 3. Pattern reappearance ratio with: (a) zero endurance interval, (b) increased endure intervals

we distinguished the pattern keys in extremely fine granularity; however, even in such a fine granularity, both patterns in water level and precipitation show a large ratio of pattern similarity. For example, the smallest pattern reappearance ratio is 0.43 in G3,

and the largest pattern reappearance ratio is 0.94 in G12. While precipitation shows the largest pattern similarity, which can be seen not only from the least number of patterns in Figure 2, but also from the fact that it has all pattern reappearance ratio larger than 0.77; actually, most pattern reappearance ratio of precipitation is about 0.99 for all gauges. We can expect precipitation to stay stable at most time. It may change suddenly, however, after this sudden change, it goes back to normal and stabilizes for a long period of time. Voltage has the most varying pattern reappearance ratios, which ranges from 0.04 to 0.96, showing that the performance of the power supply is really independent and highly dynamic.

The goal of the sensor network applications is to collect meaningful data, thus, most of those applications can endure a certain level of data inaccuracy, which will not affect our discovery of the rules and events in the monitoring field. We reexamine the pattern reappearance ratio after we lower the accuracy requirements on the collected data and set up different endurance intervals for three parameters. The resulted pattern reappearance ratio is depicted in Figure 3(b), where the three numbers under the title are the endurance intervals, which are mostly 10% of possible largest changes, i.e., we allow voltage to endure 0.2 volts changes, water level to endure 0.04 meter changes, and precipitation to endure 0.04 inch changes. Note that different units are used for water level and precipitation, i.e., meter for water level and inch for precipitation, which we keep the original units as in the raw data.

Comparing Figure 3(b) to Figure 3(a), we can find that almost all pattern reappearance ratios increased by increasing the endurance interval, especially for those with small reappearance ratio in Figure 3(a). After we increase the endurance interval, we can see that 50% of the voltage data pattern reappearance ratio is larger than 0.95, while water level and precipitation pattern reappearance ratio do not change too much compared to that in voltage; however, most of them are still larger than those in Figure 3(a). From both figures, we can see that there is a big pattern reappearance ratio.

In our definition, pattern length means the number of continuous appearance of the same sensor reading. Thus, we try to figure out the distribution of the pattern length in terms of variable endurance interval, as shown in Figure 4, where the x-axis is the length of the pattern and the y-axis is the CDF of the pattern length. From the figure, we find that most patterns have short patten length. For example, when the endurance interval is set to be $[0.00, 0.00]$, 90% of voltage patterns have length less than 10, and about 70%

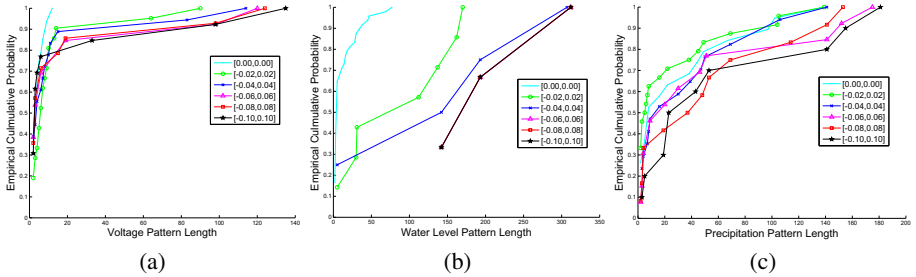


Fig. 4. CDF of pattern length: (a) Voltage, (b) water level, (c) precipitation

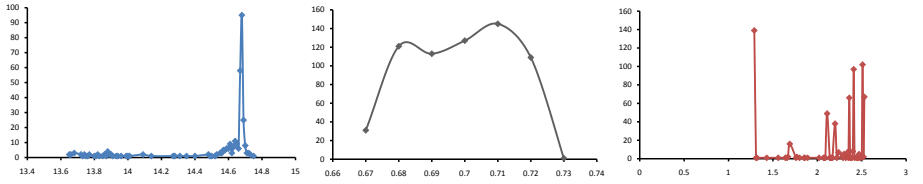


Fig. 5. The number of appearances for each numerical values: (a) Voltage, (b) water level, (c) precipitation

of water level patterns and about 60% of precipitation patterns have length less than 10. However, different parameters have different pattern lengths. In the figure, we can see that voltage, which has almost all pattern length less than 20, has more short length patterns than water level and precipitation, while precipitation has the longest length among the three parameters, where about 30% of the precipitation pattern has length longer than 20. This observation shows that precipitation is stable at most of the time, but the reading of the voltage has high dynamics. By increasing the endurance interval, more patterns have longer length appear. For example, when water level endurance interval is increased to $[-0.04, 0.04]$, more than 30% of the patterns have length between 140 to 180.

Numerical Similarity. Having studied pattern similarity, we move on to check the numerical similarity. Numerical similarity focuses on the numerical value reappearance of the sensing data, which differs from pattern similarity in that numerical similarity does not intend to detect any pattern. For the numerical similarity, we identify the number of appearance for each individual value. Figure 5 shows the numerical distribution of the collected data, where the x-axis is the numerical value of the sensing reading and the y-axis denotes the number of appearance of the corresponding numerical value. Note that we pick up the data collected by gauge G5 as an example.

In the figure, we find that those three parameters exhibit totally different distributions. The reading of the voltage and water level are very close to normal distribution with $\mu = 14.22, \sigma = 0.38$ and $\mu = 0.7, sigma = 0.02$ respectively. The voltage readings are more centralized to value 14.7, while water level readings are more broadly distributed from 0.68 to 0.72 and centralized at 0.71. The reading of precipitation shows no obvious distribution. It spreads from about 1.25 to 2.55. Some precipitation values such as 1.29 and 2.51, appear much more times than others, which means no rain or snow falls for a long time after the precipitation value is read, while other precipitation readings only appear several times, which mainly depicts some transitional states during a continuous rain or snow falling. Like the pattern similarity ratio, numerical reappearance ratio is used to evaluate the numerical similarity.

Figure 6(a) presents the numerical reappearance ratio of the sensing data at all gauges. We can see that all parameters exhibit very high reappearance ratio. Compared to pattern reappearance ratio, numerical reappearance ratio is much larger for voltage, fairly larger for water level, and comparable for precipitation. For example, the voltage pattern reappearance ratio in G1, G10 and G11 is less than 10%, while the voltage numerical pattern reappearance is close to 80%. The large difference implies that

Gauge ID	Voltage Numerical Reappearance Ratio	Water Level Numerical Reappearance Ratio	Precipitation Numerical Reappearance Ratio
G1	0.84	-	-
G2	0.99	0.64	-
G3	0.97	0.43	-
G4	0.91	0.57	-
G5	0.82	0.88	0.92
G6	0.94	0.84	0.89
G7	0.99	0.69	0.93
G8	-	-	-
G9	0.97	0.87	0.89
G10	0.77	0.44	0.77
G11	0.82	0.83	0.90
G12	0.99	0.94	0.91
G13	0.89	0.88	0.89

(a)

Gauge ID	Voltage Numerical Reappearance Ratio [-0.1, 0.1]	Water Level Numerical Reappearance Ratio [-0.02, 0.02]	Precipitation Numerical Reappearance Ratio [-0.02, 0.02]
G1	0.95	-	-
G2	0.99	0.97	-
G3	0.99	0.94	-
G4	0.97	0.95	-
G5	0.94	0.99	0.95
G6	0.98	0.98	0.93
G7	1.00	1.00	0.95
G8	-	-	-
G9	0.99	0.96	0.93
G10	0.89	0.93	0.90
G11	0.92	0.96	0.94
G12	1.00	0.99	0.95
G13	0.95	0.97	0.94

(b)

Fig. 6. Numerical reappearance ratio with: (a) endurance interval [0.00, 0.00], (b) increased endurance intervals

although the numerical readings of the voltage have a large similarity, they fluctuate very frequently and there are no obvious patterns in voltage readings. The two reappearance ratios of precipitation do not differ too much, which suggests that reappearance patterns play an important role in precipitation.

Similar to what we have done in the pattern similarity analysis, we increase the endurance interval to a certain level. Here, we set the endurance interval to the same value as we did in the pattern similarity analysis. As a result, most numerical appearance ratios are increased by increasing the endurance interval; however, the increasing rate is not as big as the one in pattern similarity analysis. From Figure 6(b), we really find that the numerical redundancy is very high in all three types of sensing data. For instance, after we increase the endurance interval, the numerical reappearance ratio is mostly over 90%. We also try to mine the pattern of the data change in terms of the time series. We calculate the coefficiency in the time series with different time periods such as 24 hours, 48 hours and so on, however, we find that all the coefficiencies are very low, thus, we believe that there is no strong clues showing the periodically reappearance pattern in data changes.

Abnormal Data Detection. Abnormal data may result from sensor malfunction, data loss during the communication, faked data inserted by malicious nodes, or the appearance of an interesting event. We try to detect abnormal data based on the presented numerical value of the data. Basically, two types of abnormal data can be detected. One is the out-of-range data, and the other is dramatic changing data.

Figure 7 shows the appearance of the out-of-range data, which is the data out of the possible valid range defined by the domain scientists. Based on the figure, we figure out that most sensing data are within the normal range. We find out-of-range data only at two gauges, G6 and G10, and G6 only has one invalid reading. Considering the failure patterns to be discussed in Section 4, we find that G10 has a maximum number of failures as well. So, we believe there are some relations between the probability of abnormal readings and the probability of failures.

Gauge ID	Parameter	Position	Value
G6	Water Level	Reading # 45	62.79
G10	Voltage	Reading #47	1.00
G10	Voltage	Reading #119	1.00
G10	Voltage	Reading #126	1.00
G10	Voltage	Reading #127	1.00
G10	Voltage	Reading #323	1.00

Fig. 7. Detected out-of-range readings

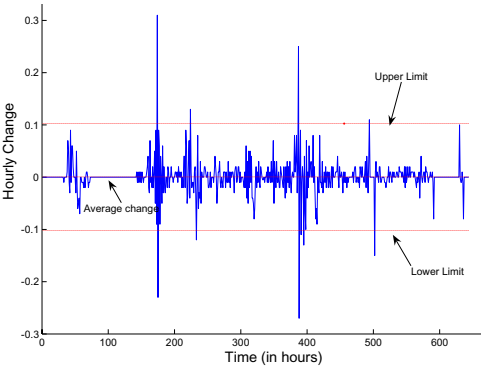


Fig. 8. Limit control of G3's hourly water level changes

Figure 8 explains hourly water level changes in gauge G3, where we find that their distributions are close to normal distribution based on normal probability plot, which is a graphical technique for assessing whether or not a data set is approximately normally distributed. For such data, 3-sigma limits is a common practice to base the control limit, i.e., whenever a data point falls out of 3 times the standard deviation from its average value, it is assumed that the process is probably out of control. In the figure, two horizontal lines depict the upper and lower 3-sigma limits. We find that only several points are out of the two limits, which means domain scientists do not need to check the cause of water level changes at most time. The similar patterns are detected in all gauges as shown in Figure 9, where most gauges have water level changes within 3-sigma limits. Investigation is deserved when out-of-limit changes are detected to find the cause of the abnormality.

Implications. Learned from above similarity and abnormal analysis, we argue that we need to revisit system protocol design by integrating the intrinsic features of the monitoring parameters.

First, we can take advantage of the large amount of data similarity. Because data similarity is common, it is not necessary to transfer all the collected data to the gateway.

G2	G3	G4	G5	G6	G7	G9	G10	G11	G12	G13
2.17%	1.55%	2.95%	0.00%	0.31%	0.16%	0.00%	1.40%	0.16%	6.06%	1.86%

Fig. 9. Out-of-Limit ratio for hourly water level changes

Quality-assured local data processing, aggregation and compression algorithms are necessary to remove redundant data and reduce overall data volume but keep the quality of the collected data at a satisfactory level. By enduring a certain level of data inaccuracy, we can reduce the total amount of collected data up to 90% according to the pattern and numerical reappearance ratios. In addition, strong patterns are helpful to estimate the future data and detect abnormal data.

Second, we can use different data sampling rates for different monitoring parameters. For example, we discover that the changes in voltage is much more frequent than those in precipitation. Thus, we need to increase the sampling rate to sense voltage data more frequently, whereas, decrease the sampling rate for precipitation. Furthermore, in the sensor readings for precipitation, some of them reappear a large amount of times, while others only appear once. Usually, the readings that only appear once or twice imply a high dynamic environment. Therefore, it is better to increase sampling rate so that we can detect the details in changes.

Third, there may be a lot of abnormal data existing in the sensor reading. Basically, they can be classified to two categories. One type is transitional, which disappears very quickly. We can mostly ignore this type of data without affecting overall data quality by replacing it with a reasonable value. The other type is continuous, which typically lasts a longer period of time. This type of abnormal data usually implies malicious data or interesting events. When continuous abnormal data is detected, more attention should be paid to them at the early stage. For example, more data should be sampled and reported to the gateway as fast as possible.

Finally, various data sampling rates may result in different amount of data traffic. Samplings for different parameters and detected abnormal data may have different priorities in their delivery to the gateway. A well designed data collection protocol is necessary to achieve this goal.

3.2 Multi-Modality and Spatial Sensing Data Analysis

In this subsection, we analyze the relationship between two types of sensing data, water level and precipitation. Moreover, we try to explore the spatial relationship at different locations.

Although water level can be affected by many factors, including moisture when it starts raining, rainfall intensity, and even temperature and slope of the land, we believe

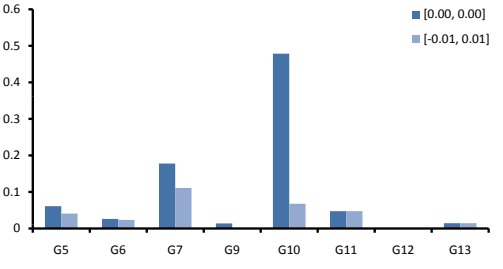


Fig. 10. Conflict ratio of water level and precipitation

that there is a relationship between water level and precipitation. Mostly when precipitation increases, water level should also increase. We count the ratio of the conflict, which is defined as the appearance when precipitation increases but water level decreases, to verify this relationship. Figure 10 records the conflict ratio between water level and precipitation. In the figure, the x-axis depicts the gauge ID, and the y-axis shows the conflict ratio. The dark blue bar and the gray bar denote the conflict ratio with endurance interval $[0.00, 0.00]$ and $[-0.01, 0.01]$ independently. From the figure we observe that in most cases the conflict ratio is less than 6%, which verifies that water level is closely related to precipitation; however, there are two gauges with conflicts larger than 10%, i.e., G7 has conflict ratio of 18% and G10 has conflict ratio of 48%. After a carefully examination, we figure out that G10's high conflict ratio is related with lots of failures it has. While G7's high conflict ratio may be caused by other reasons, because when precipitation increases only a little, other factors, such as moisture and temperature, may play major roles to determine water level. This is verified by the fact that when we increase the endurance interval a little, the conflict ratio decreases very fast, and it eventually disappears when we set endurance interval to $[-0.01, 0.01]$ for water level and $[-0.02, 0.02]$ for precipitation.

We analyze spatial correlation for all of the three parameters. Because there are no direct communications among sensors at different locations in this application, we do not expect spatial correlation among voltage readings at the different gauges, which is validated by the collected data. The calculated co-efficiency value between any two gauges is less than 0.54 and 99% of them is less than 0.32. However, we do find some spatial correlation for both water level and precipitation based on the data sensed from various gauges. The results are depicted in Figure 11(a) and 11(b) respectively.

In the figure for precipitation, we only have data for listed gauges. We can see that all gauges with precipitation data have very large co-efficiency value because they are all located at Lake Winnebago, which means that the weather in that area is pretty uniform. When there is a rain fall at the location of one of the gauges, it is most probably raining at the locations of the other gauges as well. Water level also exhibits the similar pattern. In Figure 11(a), gauges located closely usually have high co-efficiency values, which results in similarity in water level changes, while gauges located far away usually have no obvious similarity in terms of water level changes. For instance, gauges can be grouped into several small groups with similar water level changes based on the calculated large co-efficiency values. Thus, gauge G2 and G3 are within one group with co-efficiency value larger than 92%. We can see that both of them are located in St. Clair River. G4 is the only gauge in Detroit River, so it has no high co-efficiency

Water Level	G2	G3	G4	G5	G6	G7	G9	G11	G12	G13
G2	1	0.924	0.174	0.344	0.083	0.314	0.64	0.471	0.513	0.569
G3	0.924	1	0.306	0.428	0.254	0.377	0.601	0.505	0.525	0.591
G4	0.174	0.306	1	0.486	0.413	0.35	0.35	0.54	0.436	0.528
G5	0.344	0.428	0.486	1	0.159	0.916	0.699	0.888	0.875	0.555
G6	0.083	0.254	0.413	0.159	1	0.113	-0.1	0.002	-0.01	0.32
G7	0.314	0.377	0.35	0.916	0.113	1	0.659	0.832	0.848	0.474
G9	0.64	0.601	0.35	0.699	-0.1	0.659	1	0.825	0.835	0.7
G11	0.471	0.505	0.54	0.888	0.002	0.832	0.825	1	0.909	0.627
G12	0.513	0.525	0.436	0.875	-0.01	0.848	0.835	0.909	1	0.652
G13	0.569	0.591	0.528	0.555	0.32	0.474	0.7	0.627	0.652	1

(a)

Precipitation	G5	G6	G7	G9	G11	G12	G13
G5	1	0.977	0.996	0.995	0.985	0.996	0.996
G6	0.977	1	0.959	0.99	0.996	0.967	0.968
G7	0.996	0.959	1	0.983	0.969	0.999	0.998
G9	0.995	0.99	0.983	1	0.992	0.988	0.988
G11	0.985	0.996	0.969	0.992	1	0.975	0.975
G12	0.996	0.967	0.999	0.988	0.975	1	0.999
G13	0.996	0.968	0.998	0.988	0.975	0.999	1

(b)

Fig. 11. Spatial correlation of: (a) water level, (b) precipitation

with any other gauges. Moreover, gauge G5, G7, G11, and G12 show high similarity because they are located closely. Thus, we believe that geographical similarity exists in the sensed data for water level and precipitation.

Implication: Multi-modality and spatial sensing data analysis helps us to find the correlation between different parameters and geological correlation of the same parameter. Therefore, data collected by the correlated sensors can be used as a reference to calibrate the sensing data. For example, an increase in precipitation mostly results in an increase in water level. When there are some conflicts between them, we need to take a close look and figure out the reason of the conflict. Furthermore, we can take advantage of similarity in different parameters or sensors located in different locations. Quality-assured aggregation can be applied in this scenario to reduce the volume of sensing data. Thus, multi-modality models and spatial models are very useful in quality-assured data collection protocol design.

4 Failure Analysis

In this section we study the failure patterns of the sensor system including communication related failures and sensing hardware related failures. First, we present a few important definitions.

TTF denotes Time To Failure and represents the time between two consecutive failures. Mean TTF (MTTF) is a measure of the system reliability.
TTR denotes Time to Repair that is the time it takes the system to recover from a failure. A system that exhibits a small Mean TTR (MTTR) typically maintains high availability.
Total time: represents the total system lifetime including functioning as well as failing periods.
Uptime: the total time a system is in the functioning mode, in contrast, **Downtime** is the total time, in which the system is un-available. The following two equations illustrate the relationship between these values:

$$MTTF = \frac{Total\ time}{number\ of\ failures} \tag{1}$$

$$Downtime = MTTR \cdot number\ of\ failures \tag{2}$$

4.1 Methodology

For each sensing parameter (i.e., Water level or Precipitation), we organize the readings as a discrete time series and locate missing or corrupted readings. Each missing or corrupted reading is considered a failure. For each time series, we record the **Number of Failures**, **TTFs**, and **TTRs** for each individual failure type independently.

In our investigation of the raw data traces, we discovered several failure types. Some of these failures are related to communication failures, while others are pertinent to the sensing hardware itself. Figure 4.1 lists all the failure types we encountered in the raw data along with a simple description for each one of them. The first four failures in the figure (i.e., Comm-T1 to Comm-T4) are communication failures between the

Failure Type	Message in the raw traces	Description
Comm-T1	"ADDRESS ERROR CORRECTED"	Unknown reasons. We could not reach technical people who could provide explanation.
Comm-T2	"MISSING SCHEDULED DCP MESSAGE"	Communication failure due to lack of time synchronization between the gauge station and the satellite unit.
Comm-T3	"MESSAGE RECEIVED ON WRONG CHANNEL"	The message was not received on the channel that has been assigned to that particular gauge station.
Comm-T4	"MESSAGE OVERLAPPING ASSIGNED TIME WINDOW"	Communication failure due to lack of time synchronization between the gauge station and the satellite unit.
H/W-T1	Blank	Sensor failure, no reading was reported by the sensor on time. This represents a fail-stop sensor failure
H/W-T2	Corrupted reading	Sensor failure, the data format is corrupted, unreadable reading

Fig. 12. Failure types and their description

gauge station and the satellite unit. The last two failures (i.e., H/W-T1 and H/W-T2) are sensing hardware malfunctioning. H/W-T1 represents a fail-stop failure, where the sensor simply fails to report a reading on time, whereas, in H/W-T2, the sensor reports a corrupted reading (i.e., unreadable values).

4.2 Failure Analysis by Type

To understand the relative importance of these failure types, we draw their relative occurrence in the raw data traces for all locations combined in Figure 13(a) and the total downtime due to the particular failure type in Figure 13(b). Figure 13(a) gives an idea of how frequent a particular failure type is, whereas, Figure 13(b) clarifies how severe that failure is, in other words, how long it takes to recover from the failure.

In Figure 13(a), we observe that 56% of the total number of failures are of type Comm-T2 communication failure, all other communication related failures (i.e., Comm-T1, Comm-T3, and Comm-T4) collectively account for only 6% of the total number of failures. Comm-T2 as well as Comm-T4 are directly related to the lack of time synchronization between the gauge station and the satellite unit, thus, the lack of time synchronization constitutes 58% of the total number of failures. Failure to report measurements by the sensor hardware on time (i.e., H/W-T1 failure) accounts for 34% of the total number of failures, in contrast, reporting corrupted data (i.e., H/W-T2 failure) accounts for only 4%. This observation suggests that fail-stop failures are more common in the real world environmental applications.

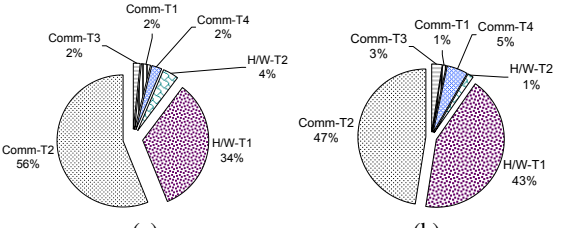


Fig. 13. Understanding relative importance of different failure types. (a) shows their relative frequency (b) shows their contribution to system total downtime.

Figure 13(b) allows us to observe the importance of the different failure types from a different perspective, in particular, how long it takes to recover from a particular failure type. For example, Although Comm-T4 and Comm-T1 failure types account for the same percentage of the total number of failures (i.e., 2% as shown in Figure 13(a)), it seems that recovering from a Comm-T4 failure takes more time compared to recovering from a Comm-T1 failure, which allows us to conclude that Comm-T4 failures are more important than Comm-T1 failures, in other words, Comm-T4 failures contribute 5% to the system total downtime, whereas, Comm-T1 contributes only 1% as shown in Figure 13(b). We also, observe that sensing hardware failures (i.e., H/W-T1 and H/W-T2 failures) account for 44% of the total system downtime.

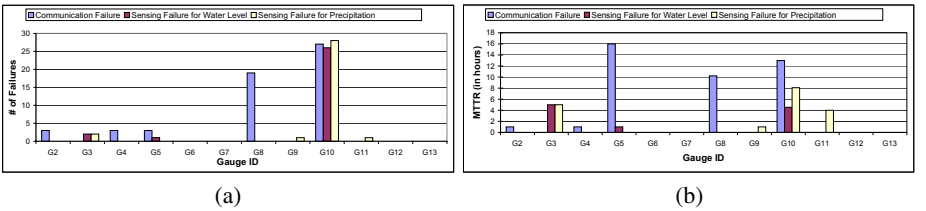
We observe in Figure 13 that we have two equally important major failure types, communication related failures and sensing hardware failures. Based on these findings and after consultation with field experts, we decide to merge these failure categories and abstract them into two failure classes: communication failures and sensing hardware failures in the rest of this section.

4.3 Failure Analysis by Location

In this subsection, we study failure characteristics at different locations, which allows us to understand the effect of the environment on inflicting failures on the system. At each location, we record the Number of failures and MTTR for each failure type and draw them in Figure 14(a) and Figure 14(b) respectively. Note that MTTF is directly proportional to the Number of failures (refer to Equation 1), therefore, including MTTF offers no insight in our analysis.

Figure 14 allows us to observe that gauge G10 experience much more failures compared to the other gauges, because of the hostile environment surrounded G10. This suggests that the external environment plays a significant role in the failure frequency and pattern of the system. We also observe in Figure 14(a) that the environmental impact is uniform in inflicting different failure types. For example, Figure 14(a) shows that gauge G10 suffered around 26 communication failures, 26 water level sensor failure, and 28 precipitation sensor failures, whereas, Gauge G3 experienced 1 water level sensor failure, 1 precipitation sensor failure, and 0 communication failures.

In Figure 14(b), we observe that different failure types need different recovery time. For example, at gauge G10 in Figure 14(b), precipitation sensor failure takes more time



(a) (b)

Fig. 14. Understanding effect of external environment on inflicting failures. (a) effect of environment on failure frequency (b) effect of environment on MTTR.

on average to recover from a failure compared to water level sensor failure. Surprisingly, communication failures exhibit much longer repair time.

4.4 Summary and Implications

Based on our findings, we believe that the lack of time synchronization is a major source of communication failures, this makes time synchronization algorithms of particular importance in remotely deployed environmental monitoring sensor applications. Sensor hardware failures are also a major source of failures in these applications, we found that fail-stop failure is the most common failure pattern in this category. We further observed that different sensor hardware exhibit different failure characteristics, in particular, different repair times. Finally, we found that external environmental factors, perhaps, are the most important factor on inflicting failures in environmental applications. This makes deployment-based failures particularly important, in which failures are independent of aging.

5 Related Work

The work presented in this paper is inspired by many previous work in WSNs, although, to our knowledge, it is the first work on detailed data quality and failures characterization of sensing data in WSNs. Next, we will list the most relevant previous efforts.

SenseWeb [14] has provided a venue for people to publish their data, but we have not seen any analysis yet. Our next step will use more data set from SenseWeb. Data aggregation is an important way to reduce the volume of the collected data. A few data aggregation approaches have been proposed. These approaches make use of cluster based structures [6] or tree based structures [2,5,7,11,18]. Tang and Xu propose to differentiate the precisions of data collected from different sensor nodes to balance their energy consumption [16].

Adaptive sampling has been proposed to match sampling rate to the properties of environment, sensor networks and data stream. Jain and Chang propose an adaptive sampling approach called backcasting [8]. Gedik and Liu proposed a similar way of data collection, selective sampling [4]. Although many approaches have been proposed to reduce energy while maintaining data quality, there exists rare study on the pattern of raw data collected by sensor nodes in the real world. In addition, most studies adopt the way of simulation, whereas, our quality-oriented sensing data analysis gives a chance to take a fresh look at how the data behaves.

Failure pattern is another important issue for WSNs. There have been a few efforts that focus on understanding failure patterns of computer hardware components, in particular hard disks [9,13]. Our work employs similar techniques and investigates a different type of system that is deployed in an open environment, therefore, we believe that its failure behavior is totaly different from classical computer systems. To the best of our knowledge, there is no prior work that specifically studies and analyzes real sensor failure traces so that our work is a leading exploration step in this direction. Most existing work on WSN reliability assumes exponential lifetime distribution of sensor nodes [1,17], and here we take a second look of such assumption by investigating real

sensor system failure traces. Although the current data set spans a short period of time and does not permit us to draw strong conclusions regarding lifetime distributions, we believe that our work brings new insights in understanding sensor device failure patterns.

6 Conclusion

In this work, we use real sensor data sets collected by 13 sensor nodes to study and analyze data quality properties as well as failure patterns.We found that data redundancy is very high in the water level and precipitation data sets. This provides us with an opportunity to design more aggressive energy-efficient data collection protocols. We also found that the lack of time synchronization is a major source of communication failures in the system, which suggests that we should pay more attention to time synchronization protocols in remotely deployed WSN environmental applications. In our future work, we will focus on designing energy-efficient sensor networks with high-quality data taking advantage of what we have learned here about information redundancy, similarity between data series, and abnormal data detection.

References

1. Al-Omari, S., Shi, W.: Availability modeling and analysis of autonomous in-door wsns. In: Proceedings of IEEE MASS 2007 (September 2007)
2. Ding, M., Cheng, X., Xue, G.: Aggregation tree construction in sensor networks. In: Proceedings of the 58th IEEE Vehicular Technology Conference 2003 (October 2003)
3. Estrin, D., Culler, D., Pister, K., Sukhatme, G.: Connecting the physical world with pervasive networks. IEEE Pervasive Computing 1(1), 59–69 (2002)
4. Gedik, B., Liu, L.: Energy-aware data collection in sensor networks: A localized selective sampling approach. Technical report, Georgia Institute of Technology (2005)
5. Goel, A., Estrin, D.: Simultaneous optimization for concave costs: Single sink aggregation or single source buy-at-bulk. In: Proceedings of the 14th Annual ACM-SIAM Symposium on Discrete Algorithms 2003 (2003)
6. Heinzelman, W., Chandrakasan, A., Balakrishnan, H.: An application-specific protocol architecture for wireless microsensor networks. IEEE Transactions on Wireless Communications 1(4), 660–670 (2002)
7. Intanagonwiwat, C., Estrin, D., Govindan, R., Heidemann, J.: Impact of network density on data aggregation in wireless sensor networks. In: Proceedings of IEEE ICDCS 2002 (July 2002)
8. Jain, A., Chang, E.: Adaptive sampling for sensor networks. In: Proceedings of the First Workshop on Data Management for Sensor Networks(DMSN 2004) (August 2004)
9. Jiang, W., Hu, C., Zhou, Y., Kanevsky, A.: Are disks the dominant contributor for storage failures? a comprehensive study of storage subsystem failure characteristics. In: Proceedings of USENIX FAST 2008 (January 2008)
10. Li, M., Ganesan, D., Shenoy, P.: PRESTO: Feedback-driven data management in sensor networks. In: Proc. of the NSDI 2006 (May 2006)
11. Luo, H., Luo, J., Liu, Y.: Energy efficient routing with adaptive data fusion in sensor networks. In: Proceedings of the Third ACM/SIGMOBILEe Workshop on Foundations of Mobile Computing 2005 (August 2005)

12. Pottie, G., Kaiser, W.: Wireless integrated network sensors. Communications of the ACM 43(5), 51–58 (2000)
13. Schroeder, B., Gibson, G.: Disk failures in the real world: What does an MTTF of 1,000,000 hours mean to you? In: Proceedings of USENIX FAST 2007 (February 2007)
14. Microsoft research
15. Sha, K., Shi, W.: Consistency-driven data quality management in wireless sensor networks. Technical Report MIST-TR-2007-001, Wayne State University (January 2007)
16. Tang, X., Xu, J.: Extending network lifetime for precision-constrained data aggregation in wireless sensor networks. In: Proc. of IEEE International Conference on Computer Communications (INFOCOM 2006) (April 2006)
17. Yu, S., Yang, A., Zhang, Y.: Dada: A 2-dimensional adaptive node schedule to provide smooth sensor network services against random failures. In: Workshop on Information Fusion and Dissemination in Wireless Sensor Networks (2005)
18. Zhang, W., Cao, G.: Optimizing tree reconfiguration for mobile target tracking in sensor networks. In: Proceedings of INFOCOM 2004 (March 2004)