Analysis of the Impact of Data Correlation on Adaptive Sampling in Wireless Sensor Networks

Alireza Masoum, Nirvana Meratnia, Paul J. M. Havinga
Pervasive Systems, Department of Computer Science University of Twente, The Netherlands
{a.masoum, n.meratnia, p.j.m.havinga} @utwente.nl

***Abstract*—Wireless Sensor Networks (WSNs) are often densely deployed to monitor a physical phenomenon, whose nature often exhibits temporal correlation in sequential readings. Such a dense deployment results in high correlation of sensing data in the space domain. Since WSNs suffer from sever resource constraints, temporal, spatial and spatio-temporal correlation among sensor data can be exploited to find an optimal sampling strategy, which reduces the number of sampling nodes and/or sampling rates while maintaining high data quality. In this study, we investigate the impact of the data correlation on sampling strategies, by taking both data quality and energy consumption into account.**

I. Introduction

WIRELESS sensor networks (WSNs) are new revolutionary monitoring platforms. Their high dynamicity and sever resource constrains require optimal resource management policy, as such a policy dictates the quality and quantity of sensor data and tasks executed by sensor nodes. Dealing with sampling policies is one of the possible approaches for resource management. Determining which sensor nodes and how often should collect data in such a way that application's data quality requirement is met is the challenge faced by resource management solutions. Data quality implies representativeness of the data collected at the base station compared with the real situation of the phenomena being monitored. Usually, ensuring data quality comes at the expense of energy expenditure [1]. In order to capture dynamic changes of the monitored phenomena, on the one hand, and to save energy, on the other, having fixed sampling plans should be avoided. This requires having an adaptive sampling strategy in place, which can adapt to the dynamicity observed.

Our objective is to find the best sampling nodes and the best sampling frequency in order to achieve high data equality and minimum energy consumption for a given application. To this end, we study the impact of different sampling plans on both data quality and energy consumption. We do so by changing the number of sampling nodes and sampling frequency according to correlation exhibited in sensor data and exploring impacts of utilization of temporal, spatial, and spatio-temporal data correlations on a given dataset [5].

II. Models

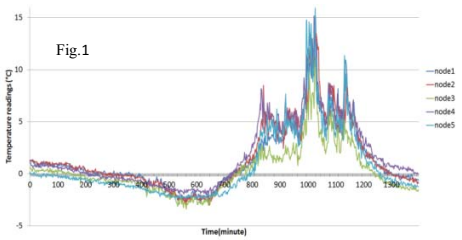
We consider a network composed of N stationary nodes and *C* cluster heads. The location of sensor nodes, cluster heads and the base station are fixed and are known a priori. All sensor nodes are homogeneous but cluster heads, which are more powerful than sensor nodes are of different type. Each sensor node belongs to only one cluster and it is assumed that each node is able to communicate only with its cluster. Each sensor node has a set of possible sampling rates denoted by $SRSet=\{sr_1, sr_2,\ldots, sr_k\}$.

To be able to leverage the benefit of data correlation, some models are necessary to have prediction about samples based on correlation that exist between sensor readings. These models can be temporal, spatial, or spatio-temporal. To model temporal correlation between sensor readings and predict future readings based on this temporal correlation, we apply Auto- Regressive Moving Average model (ARMA) [2]. To model spatial correlation among sensor data and to predict future samples based on this spatial correlation, utilize Multi-variable Normal (MVN) distribution model [3]. It is often preferable to combine the two previously mentioned models together to leverage the benefit from each. By doing so, the optimal combination of sampling nodes and their proper sampling frequency can be identified through utilization of both spatial and temporal correlations.

Energy consumption has direct relation with the number of sampling sensor nodes active in the sampling process and their sampling frequency. In order to find out this cost, we use the energy model discussed in [4]. To be able to well quantify data quality parameter, we use $Err(X)=|X-R|$, where X represents the predicted value using a prediction model and R represents the real observation. If this error is less than a defined error threshold, it means that environmental conditions cannot be covered by the current sampling rate or number of sampling nodes.

III. Impact Analysis of Data Correlation Models

To analyze the impact of utilization of temporal, spatial, and spatio-temporal correlation models, we use a real dataset [5] containing temperature readings collected for a period of two months. Fig. 1 illustrates the data for one day. It can be seen from the figure that data changes frequently between 400 and 700 minutes as well as 800 and 1200 minutes, while in some other time intervals such as time intervals 1-400, 700-800 and 1200-1400 minutes, it exhibits more stability. We focus on one cluster which consists of five

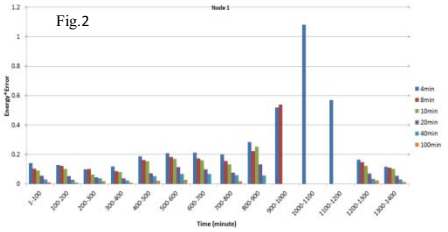


978-1-4673-1786-3/12/$31.00 ©2012 IEEE

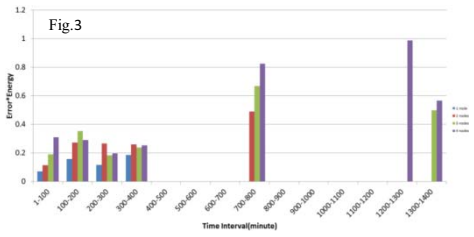
sensor nodes. The available sampling rates are *SRSet={1,2,5,10,12,15,25,50}*.

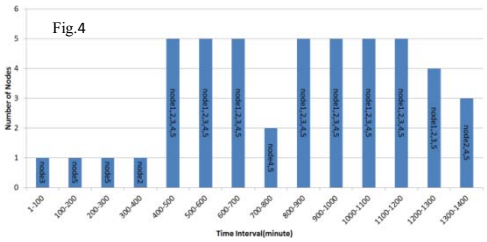
The rate at which the temperature is sampled at each sensor node influences the network performance in terms of *energy*error*. Using higher sampling rates leads to more samples, higher data resolution and better quality, which comes in the expense of higher energy consumption.

Performing a number of experiments on the dataset shows that required sampling rates are in direct relation with environment conditions. This means that the time intervals in which we face highly dynamic changes in the area conditions, demand higher sampling rates while more stable or slowly changing conditions can be covered by lower sampling rates. Fig. 2 shows more clearly how well different sampling rates impact *energy*error* value for one sensor node. Studying data readings during the times between 1 and 800 or 1200 and 1400 minutes, shown in Fig. 1, illustrates slow changes in the temperature readings. Therefore, using higher sampling rates on those periods consumes energy without improving the data quality substantially. In fact, results of Fig. 2 prove that the higher sampling rates are not appropriate. High fluctuation in temperature readings during the times between 800 and 1200 minutes leads to low temporal correlation among sensor readings. In such cases, since the temporal correlation among sensor readings is low, using lower sampling rates results in high error rates. Thus higher sampling rates need to be used.

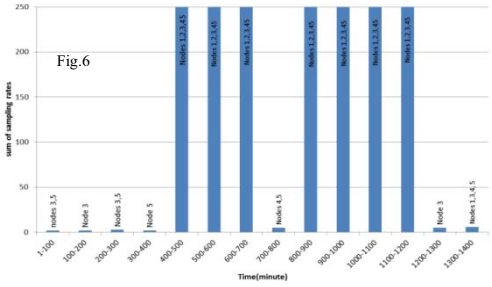


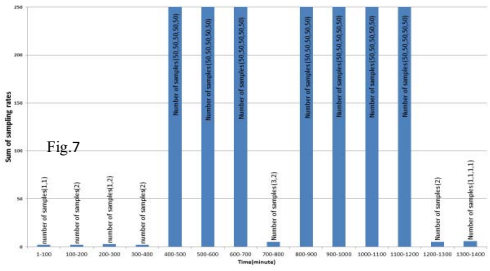
Analyzing the effects of different combination of sampling nodes on *energy*error* can helps in finding which nodes should sample for different time intervals. In case of dense sensor networks, using more sampling nodes leads to more measurements and higher energy consumption but it does not always improve the data quality. There are some time intervals in which environmental conditions change in such a way that correlation among sensor data is low. In these time intervals, it is necessary to use all sensor nodes with the maximum sampling rates. Fig. 3 shows how well different number of sampling nodes impact *energy*error* parameter for different time intervals for the given dataset. In early time intervals (1 till 400 minutes), high degree of correlation exists among sensor data. In this case, increasing number of sampling nodes increases energy consumption. However, during 700-800 minutes and 1200-1400 minutes intervals, when high fluctuations exist, it is necessary to use more sampling nodes to cover for the correlation level degradation. Fig. 3 shows that the higher number of sampling nodes for these time intervals leads to better performance. Considering readings of different sensor nodes in 400 till 700 minutes interval and 800 till 1200 minutes interval, one can see the spatial correlation level among sensor nodes is not enough to satisfy data quality requirement. Therefore, it is necessary to use all sensor nodes. The best number and combination of sampling nodes for each time interval minimizing *energy* error*, is illustrated in Fig. 4.





The impact of taking both spatial and temporal correlations into account on our dataset is depicted in Fig. 5. One can see that during the time 400-700 minutes and between 800-1200 minutes, the correlation level among sensing data in both time and space domain is low. Therefore, for these time intervals, no combination of sampling nodes and sampling frequency can best satisfy the required data quality level. Therefore, these time intervals do not have any candidate for sampling nodes. This implies that maximum sampling frequency and all sensor nodes should be utilized in these time intervals. Fig. 6 and Fig.7 summarize the results depicted in Fig. 5 and introduces the best combination of sampling nodes and their sampling rates for each time intervals. The time intervals in which environment experiences higher fluctuations must be monitor with all sensor nodes with maximum sampling rates.





REFERENCES

[1] Ch. Liu, K. Wu, J. Pei, " An energy efficient data collection framework for wireless sensor networks by exploiting spatiotemporal correlation" Parallel and distributed aystems, IEEE transactions on; July 2007 ;Volume:18 ;pp.1010 - 1023 ISSN: 1045-9219

[2] ARIMA ; http://en.wikipedia.org/wiki/Arima

[3] MVN;en.wikipedia.org/wiki/Multivariate_normal_distribution

[4] M. N. Halgamuge, M. Zukerman, K. Ramamohanarao, H. L. Vu, "An estimation of sensor energy consumption," Progress In Electromagnetics Research B, Vol. 12, 259-295, 2009.

[5] G. Aiello, G.L. Scalia, R. Micale, "Simulation analysis of cold chain performance based on time–temperature data", Production Planning & Control, DOI:10.1080/09537287.2011.564219