Bi-Scale Temporal Sampling Strategy for Traffic-Induced Pollution Data with Wireless Sensor Networks

Lamling Venus Shum, Stephen Hailes
University College London
United Kingdom
{v.shum, s.hailes}@ucl.ac.uk

Manik Gupta, Eliane Bodanese
Queen Mary
University of London
United Kingdom
{manik.gupta, eliane.bodanese}
@eecs.qmul.ac.uk

Pachamuthu Rajalakshmi,
Uday B. Dasai
Indian Institute of Technology
Hyderabad, India
{raji, ubdesai}@iith.ac.in

ABSTRACT

Carbon Monoxide (CO) induced by traffic pollution is highly dynamic and non-linear. In a pilot research, we collected some fine-grained 1Hz CO pollution data from a residential road and a busy motorway in Hyderabad, India, in preparation of the deployment of a larger scale, longer term wireless sensor monitoring system. Power conservation is an important issue as the sensor nodes are battery operated. We studied the characteristics of the collected data and designed an adaptive sampling algorithm, Bi-Scale temporal sampler, which adapts the sampling frequency to the statistics collected in real time. This design has incorporated practical engineering considerations including minimising electronic noise, sensor warm-up time and data characteristics. Results show that Bi-Scale sampler achieves better energy saving and statistical deviation ratio for our requirements than burst sampling and eSENSE sampling strategies, which are techniques popularly used in environmental monitoring applications.

Categories and Subject Descriptors

Algorithms, Measurement, Design

General Terms

Algorithms, Measurement, Design

Keywords

Wireless Sensor Network, Pollution Measurement, Carbon Monoxide, Adaptive Sampling, Algorithm, India

1. MOTIVATION

The impact of pollution is of significant scientific, social and economic interest to countries across the globe [1]. In developed countries, in which considerable control is exercised over emissions, air pollution from traffic constitutes one of the most significant remaining sources of exposure for individuals. In countries with rapidly developing economies, the uncontrolled rise in traffic has led to a corresponding decrease in air quality for many in urban environments. In both cases, such pollution has an effect on both the general quality of the environment and, in a direct and significant manner, on the health of the population [1].

Existing approaches to the monitoring of pollution typically use networks of few well-spaced high-quality monitoring stations. However, the dynamics of atmospheric dispersion are such that even different sides of the same road, or locations at different heights above ground, can experience very significantly different levels of pollution. The scarcity of good models means that it is difficult to estimate bounds on exposure across an area from single sample points. As a result, sensor systems that are distributed across an area are a natural way of obtaining direct measurements of the underlying spatio-temporal process and, consequently, in collecting information about the underlying physical processes, informing the construction of better models. Wireless sensor networks (WSN) make the acquisition of such data simpler and also enable the possibility of providing information (and alarms) in real time.

Whilst the system deployed was an experimental WSN system used to explore issues in data collection and analysis, the pollution data were collected for use by collaborating environmental engineers with an interest in understanding the dynamics of dispersion in urban canyons [3][4], as determined by traffic conditions, street structure and wind conditions. The wireless sensors used in this study were small bespoke battery-powered devices [7] that are simple to deploy, discreet, and relatively low cost. This approach provides an increase in the temporal and spatial resolution of the data over existing techniques at the cost of some loss of accuracy for individual readings. The sensors used were electrochemical CO sensors, chosen because CO is neither photoreactive nor naturally occurring in significant quantities.

The data obtained from per-second CO measurements are highly dynamic, showing changes that can be attributed to individual vehicles or short timescale variation in traffic flows. Given this fine grained data, we undertake a post-hoc analysis to understand the minimum (adaptive) sampling requirements needed to approximate pollution averages, and so to produce snapshots of pollution dispersion. The main motivation for this analysis is to understand how sensor battery power can most effectively be conserved in longer term deployments, since the retrieval of sensors in live environments is problematic.

This area is one of current research interest; however, the

978-1-4799-3780-6/14/$31.00 ©2014 IEEE

evaluation of many of the proposed techniques is based either on a theoretical analysis and/or on simulation that are founded on assumptions that prove not to be reasonable in this case, or on measurements from sensors that measure physical processes that are fundamentally different to that of pollution (e.g. of temperature). In both cases, the results of the relevant previous work are not generalizable to this domain.

The characteristics of the data and the application are crucial in determining the sampling strategy to be used. Whilst there is some degree of correlation between traffic dynamics and daily pollution averages, the nature of the data is sufficiently different that sampling strategies must differ. Likewise, for datasets that possess multi-scale characteristics, such as pollution data, different sampling strategies are needed for different timescales of interest. Moreover, there is a need to cater for out of the norm situations, such as traffic incidents or sports events.

This paper is, consequently, focussed on the practical development of strategies for sampling in this situation – which is one in which theoretical analysis is not readily possible (since good models of the environment or the process do not exist), nor likely to be particularly more useful than an engineered approach in view of the simplifications needed to make it tractable.

This paper is organised as follow: Section 2 includes a reviews on the existing sampling methodologies. Analysis on the data collected from an experiment in India is included in Section 3. In Section 4, we introduce the Bi-scale adaptive sampling strategies and the performance metrics used for the evaluation. In Section 5, a comprehensive analysis on the Bi-scale sampler, with comparison to other popular sampling techniques is presented.

2. RELATED WORKS

The ongoing research that forms the basis of this paper is a collaborative effort between teams in the UK and in India [5][6][14]. We have collected several sets of pollution data in both countries using our bespoke carbon monoxide monitors with high temporal and spatial resolution in relation to the norm for atmospheric science. Earlier work focuses on developing and calibrating sensor devices to collect fine-grained, accurate data for scientific study and on algorithm development [5]. A bias adjustment technique was developed with the aim of removing the bias caused by environmental conditions specific to the electrochemical sensors used in these experiments [7].

Regular sampling (or equi-frequency sampling) [19], Monte Carlo [9] sampling and burst sampling are simple techniques widely used in wireless sensor networks for data collection. Monte Carlo sampling has the advantage over regular sampling that regular events happening in anti-phase to a regular sample, or events happening entirely within periods, are at least theoretically detectable. Burst sampling describes a sampling technique where the sensors are turned on for a short period only to collect fine-scale data. It is commonly used where multiple time scales of the data are of interest.

Adaptive sampling algorithms designed for rare event detection have interesting objectives: the devices are scheduled to sample more often when an event is likely to occur. At times where it is perceived as normal, sampling is less frequent to conserve energy. Such algorithms include the exponentially-weighted moving average (EWMA) based adaptive sampling algorithm proposed in [18] for volcanic monitoring, and [17] for snow monitoring. These algorithms apply differences in long term and short term statistics for event prediction. The sensed parameters usually do not change much during a normal situation; a change in value, or the rate of change would indicate the possibility of an event. This kind of algorithm has some commonality with our requirements, although the sampling environment of pollution monitoring is much more dynamic.

Some algorithms take data characteristics into the adaptation of sampling rates, based typically on the prior distributions or prediction models. EDSAS [12] uses exponential double smoothing and EWMA for future values prediction. In [19] Cheng et. al. proposed an interpolation based adaptive sampling algorithm. In [16], a Kalman filter is used to determine the sampling rate (regular samples) collectively in a network.

eSENSE is a model-based stochastic sensing algorithm for wireless sensor networks [13] which uses a data-trained random walk model to construct the probability distributions of state change steps ahead to adjust sampling frequency. The random walk model makes a statistical inference on the distribution of the steps-ahead prediction and this is used to calculate the probability of sampling. eSENSE requires a well-trained model that suits the data characteristics, and the models must be maintained over time due to the non-stationary nature of the time series. In this paper, we will use eSENSE as a comparison algorithm.

Whilst we can learn from many of these techniques, practical considerations such as warm up time of the sensors and data characteristics must be considered to ensure the efficiency of the algorithm in real applications.

3. DATA CHARACTERISTICS

Although in this paper, majority of the data used were collected in the experiments conducted in India, the characteristics of traffic-induced pollution are similar in the experiments we conducted in Cyprus and the UK. Other than the environmental differences including temperature and wind conditions between the different locations, pollution level is dependent on the traffic flows and volume. The data in this paper are bias-uncorrected [7].

The basic sampling rate for all the data sets is $f_b = 1s$. Whilst it is possible to sample at a faster rate, 1Hz is adequate in capturing temporal dynamics of traffic-induced pollution and the adjacent data points are highly correlated with one another.

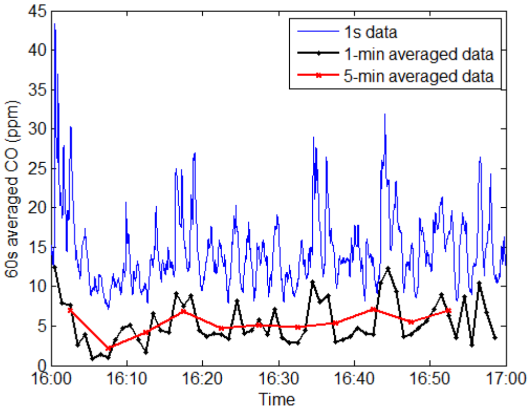

Figure 1: Time series of CO during peak-hour 4-5pm on 9/2/2012, at Kukatpally, Hyderabad, India (bias-corrected)

The CO level exhibited rapid increases when vehicles approached the sensors and decreases as the vehicles moved away. During traffic congestion, if high emission vehicles were present near the sensors, the local CO level accumulates and could be very high for a short period of time. We measured a maximum of 80ppm in our experiments. The arrival, emission levels of vehicles, and duration of the emissions can be considered as random variables, which are dependent on geographic location, traffic speed and environmental conditions.

Earlier work, including [10][11], showed that pollutants were *log-normally distributed* at all the time scales and an appropriate averaging time and frequency were proposed. Some of these results were used as guidelines for deriving pollution limits and standards. The variation of CO levels during the day time is much higher than at night time [20], which is also found in our collected data. Consequently, more of the sampling budgets must be allocated to day time statistics to achieve the same confidence level on the sample mean and to capture the variations in the averaged time series. While this periodicity of day and night variation is the norm, our designed algorithm must cater for the rare exceptions, such as night time events which increase the traffic levels, accidents and variations of weather conditions.

The Log-normal distribution is one of the heavy-tailed distributions. Both the frequency of high readings and their duration is small compared to low readings. With our high frequency Indian data, this lognormal distribution is observed during the day and night time with low to high traffic flow. The time series appear to have some local trends and cycles, which are spurious and disappear after some time. According to [12] this is a typical feature of long-memory process. The 1Hz time series is a non-linear 1/f process.

Samples very close to each other in time give us less information due to the underlying rate of the dispersion process. The adjacent points of 1Hz data sets are indeed strongly correlated to each other. Moreover, the dynamics of the pollution induced by passing vehicles or congestion can be clearly identified.

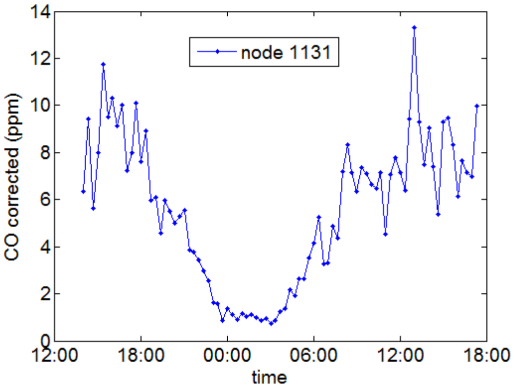

Figure 2: 20 minutes averaged series after bias removal.

The frequency analysis on the data has been reported in [20]. Periodicities observed can be related to the traffic flow caused by the sequence of traffic lights. Frequency analysis was performed on the subsampled 0.2Hz data series and $1/f$ noise is observed in addition to the major frequency component. The dominant frequency component varies with time of the day dependent on the traffic flow and is observed most clearly between the hours of 11am and 7pm. Note that this frequency characteristic was particular to the locations of the experiments carried out on that day.

More details on the experiment and analytic results can be found in [20]. The high frequency 1Hz data sets are more than adequate in capturing the traffic dynamic, the mean estimates will not be deviated much with a slightly lower frequency data set, say 0.2 Hz.

As opposed to the 1Hz data set, the 1 and 5 minutes averaged time series in Figure 2 are much smoother and more predictable than the 1 Hz data with increasing smoothness with longer averaging time. We expect to observe pollution cycles at peak-hours if longer experiments were carried out. In Figure 2, the 20-minute averages from midnight to 4am did not change much due to low traffic levels.

Auto and partial correlation results on the 20 minutes averages strongly suggests an Autoregressive integrated (AR) model of order 1-2. The AR parameters, however, varies with the time of the day.

4. BI-SCALE ADAPTIVE SAMPLER

We propose a simple probabilistic sampler based on the statistical confidences in data and the rate of change of the statistics for our application. The purpose of the sampler is to provide the best estimated statistics, while minimising the battery power used – and to do so in as simple a manner as possible to allow for the implementation on simple sensor nodes.

We utilise the characteristics of different time scales in the design of the bi-scale adaptive statistical sampling technique. At a finer scale, we evaluate the mean of the traffic pollution using a probabilistic approach, with the aim of gathering sufficient data in a statistical *block* to satisfy mean estimation to a defined confidence level. At a coarser scale, we look at the rate of change of the mean pollution levels and decide how often we need to take samples. Hence, it is possible that at the finer scale there may be a lot of traffic induced variations in the data (high variance), but if we are confident that the next evaluation blocks are statistically similar to the last block, we can skip sampling to preserve energy.

4.1 PROBLEM DEFINITION

We formulate the adaptive sampling problem as an optimisation problem to minimise the errors on the obtained expectations, while minimising the total energy expenditure.

Let $x_t = \{x_1, x_2, \dots..\}$ be the time series of data with basic sampling rate f_b , the requirement is to obtain the expectations of x_t every T statistical evaluation period, where $E(x) = \sum_T x_t p(x_t)$, $p(x_t)$ is the density function of the data x_t . As long as the sample distribution is the same as the intrinsic distribution p , the expectation can be approximated by Monte Carlo estimation, $E(x) \simeq \frac{1}{T}\sum_{l=1}^{T} x_l$. We are seeking to minimise the error on the expectation $e_x = (E(x) - \tilde{x})^2$ and the overall energy usage, where $\tilde{x}$ is the true mean of the block.

4.2 THE DESIGN

At interval , the calculated expectation is $E_m(x)$, or $\hat{X}_m$. The time series formed by the averaged mean estimates are $\hat{X}_m = \{\hat{X}_1, \hat{X}_2, \dots.\}$. The sampler needs to estimate the expected means $\hat{X}_m$ with statistical confidence and monitor the change of stationarity in the statistics.

Electrochemical sensors require a start-up/warm up time τ_w to reach 90% of the full reading, as discussed in [5]. Data taken before the warm up time are unreliable. The value of τ_w also depends on the amount of time the sensor has been powered off; thus, we found that if the sensors in our experiments have been switched off for $\tau_{off} \geq 1s$, then $\tau_w \geq 30s$. Consequently, simply reducing sampling frequency may not lead to power savings. Power consumed by the sensor module makes up a significant proportion of the whole unit (~0.09W). Intuitively, it is beneficial to further partition the evaluation periods into blocks and take multiple samples during the on-times since the marginal cost of doing so is low. Whilst the samples are not independent; the additional data can still improve the confidence in the estimate. Moreover, power can only be conserved if the sensor nodes are put to sleep for more than 30s at a time.

In the light of this, the statistical evaluation period is divided into fixed length sampling blocks with length B, where $T = \eta B$ and η is a positive integer, such that we can calculate the expected mean easily. The actual length of sampling ω in each block can be less than B if the confidence requirement $|UCL - LCL| \leq \beta$ is satisfied, and $B - w > \tau_w$. The sensor module will then be turned off for the rest of $t_{rest} = B - \omega - \tau_w$ seconds. The lower bound of ω to be a third of the sensor warm up time τ_w, such that $\frac{\tau_w}{3} \leq \omega \leq B$.

4.3 STATISTICS WITH CONFIDENCE

The statistics (expectations and their confidence levels) are evaluated per sampling block. For a lognormal distribution p , the mean of x is defined by, $mean\ (x) = \exp\ (\mu + \frac{\sigma^2}{2})$. Hence, the mean of the time series depends not only on the log-normal location parameter μ, but also the variance parameter of the σ^2.

In [15], Parkin et al. investigated five methods of calculating confidence intervals (CI) for the mean of a log-normally distributed variable and concluded that the method developed by Land [16] was the best at estimating the lower (*LCL*) and upper confidence limits (*UCL*), which are given by

$$LCL = \exp\ (\bar{\mu} + \frac{\widehat{\sigma^2}}{2} + \hat{\sigma} C_L \sqrt{(n-1)}) \qquad (1)$$

$$UCL = \exp\ (\bar{\mu} + \frac{\widehat{\sigma^2}}{2} + \hat{\sigma} C_U \sqrt{(n-1)}) \qquad (2)$$

where C_L and C_U are factors calculated from a function that depends on the number of observations n in T, the standard deviation of the log-transformed values $\hat{\sigma}$ and the confidence α-level selected ($\alpha = 0.1$ is used in the paper). The values of C_L and C_U used in this paper are based on 90 percentile values based on the methods and tables in [15][16]. In Bi-scale sampling algorithm, Equation (1) and (2) are used in the calculation of CI of the data where $\hat{\mu}$ and $\hat{\sigma}$ are estimated from the log of experimental data.

4.4 SAMPLING PROBABILITY

Once the statistic is gathered with confidence, we then estimate the probability of a state change in the statistics in the next few block intervals.

The rate of change in the averaged time series δ_t is tracked with Exponentially Weighted Moving Average (EWMA), which is very simple and requires little memory to implement in a microcontroller,

$$D_t = \alpha * \delta_t + (1 - \alpha) * D_{t-1} \qquad \textbf{(3)}$$

where α is the moving average parameter, and $\delta_t = |X_t - X_{t-1}|$ is the first absolute difference of the data. α is an

EWMA parameters and determines how much weight we give to the latest reading compared to the historical average. D_t captures the average changes and represent how confidence we are in the next blocks without sampling. D_t is then used to determine the probability of sampling.

The probability of sampling in the k blocks ahead is calculated by the equation:-

$$p_{t+k} = D_t/(D_t + q), 0 \le p_{t+k} \le 1 \qquad \textbf{(4)}$$

where q is the state change threshold we wish to detect, or the resolution of the trend of the pollution data. The probability of sampling p_{t+k} increases with the moving average difference D_t.

The adaptation of block length B is based on the parameter D_t and is formulated as follow.

$$\begin{cases} D_t < \frac{q}{2}, & B = B * 2 \\ D_t \ge \frac{q}{2}, & B = B/2 \end{cases} \qquad \textbf{(5)}$$

The block size is bounded with $B_{min} \le B \le B_{max}$. When the averages of adjacent blocks are similar, the block length can be lengthened to further conserve energy, allowing a longer sleep period .

The missing blocks are interpolated by *sampled and hold* technique because it is simple to be performed in real time on processing-limited sensor nodes. Linear interpolation and AR interpolation has also been experimented but do not give significantly better results. It is noted that more advance technique for interpolation may be used offline to further improve the estimates.

4.5 EVALUATION CRITERIA

The performances of the algorithms are evaluated based on the statistical deviations of the 20-minute sampled averages and the power saving achieved.

The energy consumption of the sensor board for a time unit τ_{on} is e_s, which is proportional to the current drawn to power the signal-conditioning circuit and the sensor. The signal conditioning circuit can be powered on and off separately from the main controller unit and the energy requirement can be evaluated independently. The energy consumption P_q for taking q consecutive 1Hz values at period i is,

$$P_q^i = (\tau_w + q * \tau_{on}) * e_s = T_q^i * e_s \qquad \textbf{(6)}$$

We aim to minimise $\sum Q_q$ over the monitoring period and satisfy the confidence requirement β. The energy unit e_s is similar for the sensor nodes, and hence, the on-time T_q^i (in seconds) can be used in the evaluation instead.

The algorithms are evaluated based on these three metrics:

(i) Fraction of sensor on time (f_{on}) – the fraction of time that the sensor module is switched on over the total deployment time,

$$f_{on} = (\sum_{i=1}^{N_{on}} T_q^i)/N \qquad \textbf{(7)}$$

where N_{on} denotes the number of statistical blocks and N is the total time in second of the experiment.

(ii) Root Mean Square (RMS) deviation (ϵ_{RMS}) – At interval m, the deviation of the sampled block means ($\tilde{X}_i^m$) from the basic rate sampler ($\hat{X}_i^m$) is defined as ϵ_m, the number of blocks at the interval is $N_m = \frac{T}{B_m}$, and $\epsilon_m = \sqrt{\sum_{i=1}^{N_m} \frac{(\tilde{X}_i - \hat{X}_i)^2}{N_m}}$. The overall RMS performance ϵ_{RMS} is evaluated over the mean of ϵ_m and

$$\epsilon_{RMS} = mean(\epsilon_m) \qquad \textbf{(8)}$$

(iii) Rated-RMS (r-RMS) – In order to compare the different algorithms, we define *rated-RMS* (or *r-RMS*) to take into account both f_{on} and ϵ_{RMS}. *r-RMS* can be viewed as the cost of sampling and is defined as:

$$r\text{-}RMS = \frac{\epsilon_{RMS}}{1 - f_{on}} \ (ppm) \qquad \textbf{(9)}$$

r-RMS indicates the error incurred per sleep unit and provides a common ground for the performance evaluation of difference algorithms.

5. RESULTS

5.1 INDIVIDUAL RESULTS

The performance of Bi-scale sampler is compared to *Minimum Sampling*, which is defined as burst sampling with fixed burst length T_b and burst interval M_b, and an adapted version of eSENSE. In order to achieve energy saving with the sensor warm up time requirement, eSENSE adaptation is applied to the time series made up of 30-second averages, such that the sleeping intervals are at least 30 seconds long. In the evaluation, eSENSE parameter *tolerance level* F_N [1]is altered to obtained the performance metrics. In Bi-scale sampler, β is set to 0.5ppm and q is altered to obtain the difference performance metrics. All results in this section are obtained from the mean of 20 iterations.

The results of comparison on one of the data sets are shown in Figure 3 and Figure 4. In Figure 3, ϵ_{RMS} naturally decreases with less sleep time and more samples. The shape of the curves are quite similar among the three algorithms, that ϵ_{RMS} drops rapidly at low f_{on} with just a small increase of f_{on}; the decrease of ϵ_{RMS} slows with further increase of f_{on}. In Figure 4 the region of f_{on} with the lowest *r-RMS*

[1] F_N is defined such that the miss ratio $\gamma \le F_N$. Please refer to [13] for detail definition.

cost is $0.2 \leq f_{on} \leq 0.8$. Below 0.2 on time, the loss of accuracy in the statistics is too high to justify the additional sleep time and beyond 0.8, the additional energy requirement outweighs the benefit of the improvement on statistic quality.

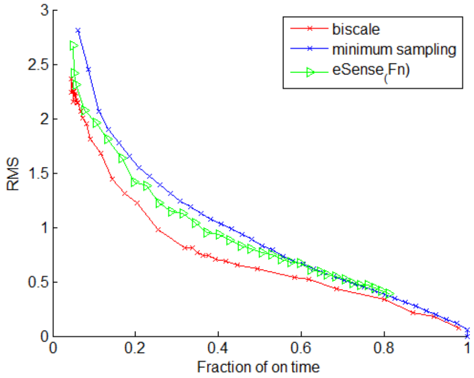

Figure 3: RMS vs f_{on}, data set 11 on 9/2/2012

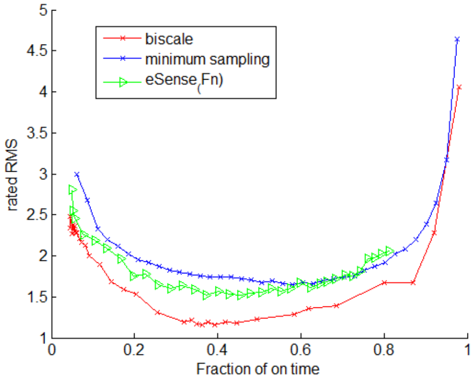

Figure **4: rated RMS, data set 11 on 9/2/2012**

5.2 COMBINED RESULTS

In this section, we repeat the tests in section 5.1 and evaluate the three algorithms against the data sets collected in two separate testbeds in India. Details of the experiments in Sections 5.3.1 and 5.3.2, including descriptions of the area, traffic conditions and locations of the sensors can be found in [7] and [20].

The parameters chosen for Bi-scale sampler are $q = \beta = 0.5ppm$, $60s \leq B \leq 120s$; for eSENSE, the parameters are $F_N = 0.5$, $\delta = 0.5ppm$; and for Minimum Sampling, $T_b = 300s$. These parameters are chosen to give comparable values of ϵ_{RMS} and f_{on} within the *basin* region observed in Figure 4.

5.2.1 CASE STUDY I: HYDERABAD KUKATPALLY RESIDENTIAL STREET

Data sets I was taken from a narrow residential street in the Kukatpally area of Hyderabad. The street was busy with buses and vehicles during peak hours, but the traffic was considerably less than that at the Hyderabad Centre in case study II. All the sensor nodes were mounted at 1.5m high.

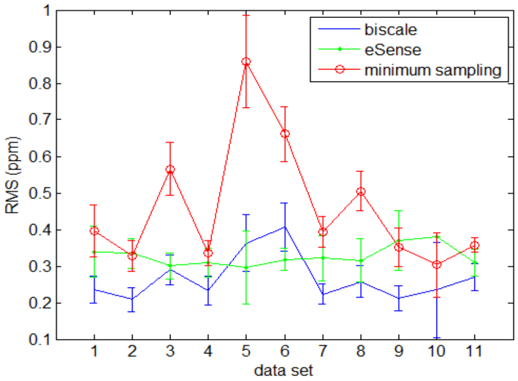

Figure 5: RMS, Hyderabad Centre data

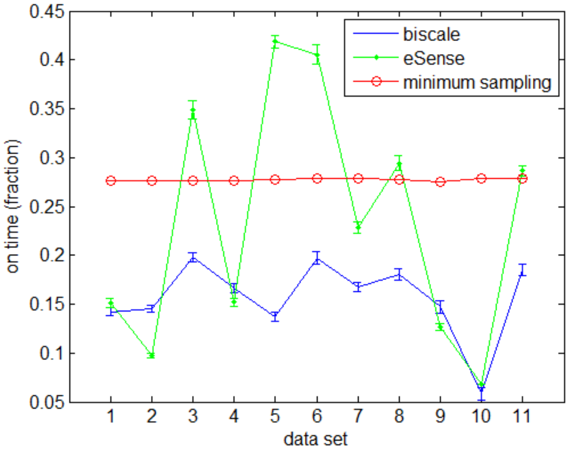

Figure 6: Fraction of on-time, Hyderabad Centre data

In Figure 5, it is observed that ϵ_{RMS} of data set 5 has the most state changes, that are least captured by Minimum Sampling. In Figure 6 we found that eSENSE has the best response to rapid state changes and ϵ_{RMS} remains at a similar level (0.3-0.4ppm) for all the data sets as shown in Figure 5.

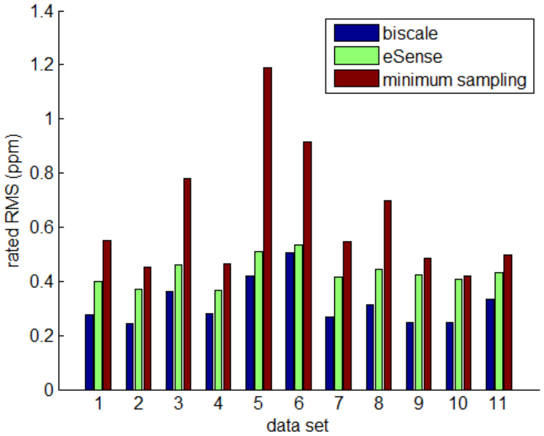

Figure 7: rated RMS, Hyderabad Centre data

The advantage of using Minimum Sampling is that f_{on} stays nearly the same for all the data sets disregarding the characteristics of data for energy budgeting. But overall when taken into account both energy saving and error performance in the *r-RMS* results in Figure 7, Bi-scale sampler out-performs the other two algorithms and is the most efficient in allocating energy budget and gives the smallest error per sleep time ratio among the datasets. The energy saving using Bi-scale sampler for all the data sets is 84-99.2% on the 1Hz data.

5.2.2 CASE STUDY II: HYDERABAD MOTORWAY

The data sets evaluated in this section were taken at the wide and busy motorway outside Hyderabad Centre with 4-6 traffic lanes in both directions. The pollution level measured was the highest in this location amongst all our experiments carried in the UK, Cyprus and India. The sensors were mounted at 1.5m and 2.5m high.

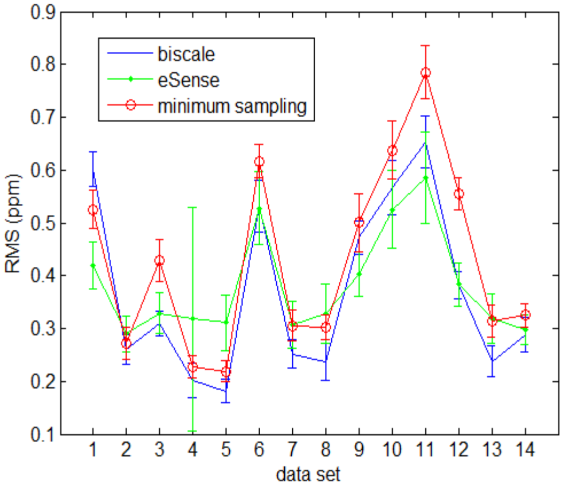

Figure 8: RMS, Kukatpally data

The results in Figure 8, Figure 9 and Figure 10 are comparable to Case Study I, however, the overall f_{on} and ϵ_{RMS} are considerably higher for these data sets due to the high variations in the CO data. The result f_{on} with Minimum Sampling remains similar among all the sets.

In terms of the energy-error efficiency measured by *r-RMS*, eSENSE has the poorest performance in the highly varying data sets 6, 10 and 11, and interestingly, even Minimum sampling out-performs eSENSE with these data sets. This is a situation when adaptation can make the results worse if it is not applied correctly, without a good understanding of the applications and challenges faced. Bi-scale sampler has the best *r-RMS* among the three algorithms for all data sets and he energy saving using Bi-scale sampler is between 30 – 96%.

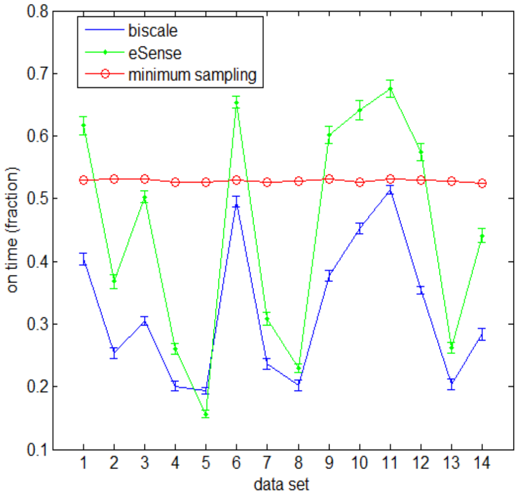

Figure 9: Fraction of on-time, Kukatpally data

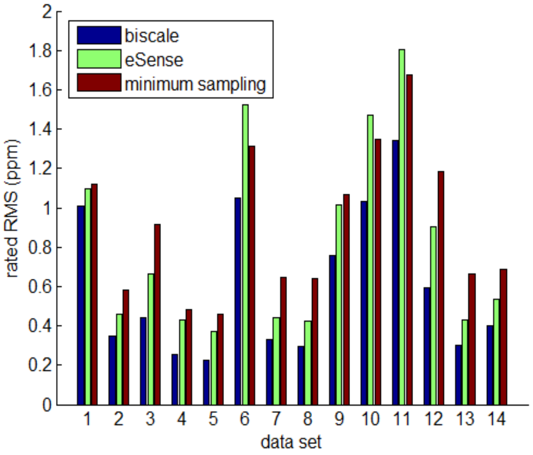

Figure 10: rated-RMS, Kukatpally data

6. CONCLUSION

Bi-scale adaptive sampling algorithm is an effective energy saving strategy for gathering good quality statistics in traffic pollution monitoring, which has data that exhibit multi-scale characteristics. It is designed based on the shorter-term probabilistic nature of traffic-induced pollution, and longer-term dynamics of the time series. The performance of Bi-scale sampler was evaluated against burst sampling and the eSENSE algorithms, with their abilities to conserve energy and preserve meaningful statistics. A new parameter *rated-RMS,* defined as the root mean square error per sleep unit was introduced for cross comparison among the algorithms.

We evaluated the algorithms with 2 sets of real data collected in India. For both sets of data, burst sampling is consistent with the energy usage as it does not take into account data characteristics and has no mean of adaptation, which is an advantage for energy budgeting in the system; eSense performs better in some of the data sets than burst sampling technique, and maintains a consistent error level over most data sets. Curiously in some highly varying data sets, burst sampling actually out-performs eSENSE in terms of rated-RMS; indicating that improper adaptation, where data properties are not fully considered, can make the results poorer in some situations. In all the data sets, Bi-scale sampler has the best *rated-RMS* score, hence the best energy to error efficiency. This means that Bi-scale sampler is the best algorithm among the three compared algorithms at allocating the sampling instances to where they are needed the most.

Although Bi-scale sampler has been designed based on the characteristics of pollution data and is particular to our application, it can be adapted to be used for other data sets that exhibit multi-scale characteristics. It is a very simple algorithm that requires no prior model training and the processing and memory requirements in sensor nodes are very low. Finally, we truly believe that understanding the data and the requirements of an application are the keys to the success of any real deployments.

7. ACKNOW LEDGMENTS

The authors would like to thank our partners at Civil, Environmental and Geomatic Engineering Department at University College London. This project is funded by EPSRC (Grant no. EP/J016721/1) research grant.

8. REFERENCES

[1] World's Worst Pollution Problem Reports 2008 – The Top Ten of the Toxic Twenty, Blacksmith Institute, http://www.worstpolluted.org.

[2] Styliani Karra, Liora Malki-Epshtein, Marina Neophytou, "The Dispersion of Traffic Related Pollutants Across a Non-Homogeneous Street Canyon", Urban Environmental Pollution 2010, Procedia Environmental Sciences 4 (2011) 25–34. Travel, P. 2007. *Modelling and Simulation Design*. AK Peters Ltd., Natick, MA.

[3] Styliani Karra, Liora Malki-Epshtein, "Influence of local parameters on the dispersion of traffic-related pollutants within street canyons", American Physical Society, 64th Annual Meeting of the APS Division of Fluid Dynamics, November 20-22, 2011.

[4] L. V. Shum, S. Hailes, G. Mcphillips, S. Karra, and L. Malki-Epshtein , "Experience in Carbon Monoxide Measurements with Wireless Sensor Network Technology", 92nd American Meteorological Society Annual Meeting, Jan 2012.

[5] Lamling Venus Shum, Pachamuthu Rajalakshmi, Ayo Afonja, Graeme McPhillips, Russell Binion, Lawrence Cheng, Stephen Hailes "On the Development of a Sensor Module for real-time Pollution Monitoring", International Conference on Information Science and Applications (ICISA) 2011, South Korea.

[6] Lamling Venus Shum, Stephen Hailes, Graeme Mcphillips, Lawrence Cheng, "Making Sense of Sensor Data in Practical Wireless Sensor Network Designs", ICNC workshop 2012, Hawaii.

[7] Lamling Venus Shum, Manik Gupta, Eliane Bodanese, Styliani Karra, Nina Glover, Liora Malki-Epshtein, Stephen Hailes, "Bias Adjustment of Spatially-distributed Wireless Pollution Sensors for Environmental studies in India", IEEE SECON, June 2013, USA.

[8] LV Shum, M Gupta, P Rajalakshmi, "Data Analysis on the High-Frequency Pollution Data Collected in India", arXiv preprint arXiv:1301.7231.

[9] Chen Liyan, "Monte Carlo Multi-object Tracking in Wireless Sensor Networks," cesce, vol. 2, pp.162-165, 2010 International Conference on Challenges in Environmental Science and Computer Engineering, 2010.

[10] Allan H. Marcus, "Air Pollutant Averaging time: Notes on a Statistical Model", Atmospheric Environment Pergamon Press, Vol..7, pp. 265-270, 1973.

[11] R. I. Larsen, "A New Mathematical Model of Air Pollutant Concentration Averaging Time and Frequency", Journal of the Air Pollution Control Association, vol. 19, no. 1, January, 1969.

[12] Jan Beran, "Statistics for Long Memory Process", Chapman & Hall, US, 1994.

[13] H. Liu, A. Chandra, J. Srivastava, "eSENSE: energy efficient stochastic sensing framework for wireless sensor platforms", The Fifth International Conference on Information Processing in Sensor Networks, IPSN 2006. , vol., no., pp.235-242.

[14] Gupta, M.; Shum, L.V.; Bodanese, E.; Hailes, S.; , "Design and evaluation of an adaptive sampling strategy for a wireless air pollution sensor network," Local Computer Networks (LCN), 2011 IEEE 36th Conference on , vol., no., pp.1003-1010, 4-7 Oct. 2011

[15] T.B. Parkin, S.T. Chester and J.A. Robinson, "Calculating Confidence Intervals for the Mean of a Lognormlaly Distributed Variable," vol. 54, no. 321–326, 1990.

[16] C. E. Land, "Confidence Intervals for Linear Functions of the Normal Mean and Variance," Annals of mathematical Statistics, vol. 42, no. 4, pp. 1187-1205, 1971.

[17] C. Alippi et al., "Adaptive Sampling for Energy Conservation in Wireless Sensor Networks for Snow Monitoring Applications," in Mobile Adhoc and Sensor Systems, 2007. 1-6.

[18] M. Alan, C. David, P. Joseph, S. Robert, and A. John. "Wireless sensor networks for habitat monitoring". In Proceedings of the 1st ACM international workshop on Wireless sensor networks and applications, pages pp. 88–97, Atlanta, Georgia, USA, 2002.

[19] Siyao Cheng, Jianzhong Li, Zhipeng Cai, "O(∈)-Approximation to Physical World by Sensor Networks", Infocom 2013.

[20] Christopher M. Bishop, "Pattern Recognition and Machine Learning", Springer, 2009.

[21] L.V. Shum, M. Gupta, P. Rajalakshmi, "Data Analysis on the High-Frequency Pollution Data Collected in India", arXiv:1301.7231, Jan 2013.