THE UNIVERSITY OF QUEENSLAND

AUSTRALIA

Environmental Sensing and Modelling using Wireless Sensor Networks

Siddhartha Raj Bhandari

Master of Science

Master of Engineering

Bachelor of Computer Engineering

A thesis submitted for the degree of Master of Philosophy at

The University of Queensland in 2019

School of Information Technology and Electrical Engineering

Abstract

Wireless sensor networks have gained significant traction in environmental signal monitoring and analysis. For a battery powered system, the lifetime of the system typically depends on the frequency at which environmental phenomena are monitored. If energy harvesting is added to provide indefinite lifetime, then the size and cost of the energy harvesting hardware is similarly affected by the sampling frequency. Typically, each data sample requires the node to wake up from a low-energy sleep mode. If sampling rates are reduced, then the node duty cycle can be reduced, and energy can be saved. This is particularly true when the measured quantity has slow dynamics, such as temperature. Using empirical datasets collected from environmental monitoring sensor networks, this work performs time series analyses of measured temperature time series. Unlike previous works which have concentrated on suppressing the transmission of some data samples by time-series analysis but still maintaining high sampling rates, this work investigates reducing the sampling rate (and sensor wake up rate) and looks at the effects on accuracy. Results show that the sampling period of the sensor can be increased up to one hour while still allowing intermediate and future states to be estimated with interpolation RMSE less than 0.2°C and forecasting RMSE less than 1°C.

Depending on the desired spatio-temporal resolution, the number of sensor nodes to be deployed will vary. Selecting an optimal number, position and sampling rate for an array of sensor nodes in environmental monitoring is a challenging question. Most of the current solutions are either theoretical or simulation-based where the problems are tackled using random field theory, computational geometry or computer simulation, limiting their specificity to a given sensor deployment. Using an empirical dataset from a mine rehabilitation monitoring sensor network, this work proposes a data-driven approach where co-integrated time series analysis is used to select the number of sensors from a short-term deployment of a larger set of potential node positions. Analyses conducted on temperature time series show 75% of sensors are co-integrated. Using only 25% of the original nodes can generate a complete dataset within a 0.5°C average error bound for the estimated temperature from neighbours' measurements compared to the measured temperature at each position. Our data-driven approach to sensor position selection is applicable for spatiotemporal monitoring of spatially correlated environmental parameters to minimize deployment cost without compromising data resolution.

.

Declaration by author

This thesis is composed of my original work, and contains no material previously published or written by another person except where due reference has been made in the text. I have clearly stated the contribution by others to jointly-authored works that I have included in my thesis.

I have clearly stated the contribution of others to my thesis as a whole, including statistical assistance, survey design, data analysis, significant technical procedures, professional editorial advice, financial support and any other original research work used or reported in my thesis. The content of my thesis is the result of work I have carried out since the commencement of my higher degree by research candidature and does not include a substantial part of work that has been submitted to qualify for the award of any other degree or diploma in any university or other tertiary institution. I have clearly stated which parts of my thesis, if any, have been submitted to qualify for another award.

I acknowledge that an electronic copy of my thesis must be lodged with the University Library and, subject to the policy and procedures of The University of Queensland, the thesis be made available for research and study in accordance with the Copyright Act 1968 unless a period of embargo has been approved by the Dean of the Graduate School.

I acknowledge that copyright of all material contained in my thesis resides with the copyright holder(s) of that material. Where appropriate I have obtained copyright permission from the copyright holder to reproduce material in this thesis and have sought permission from co-authors for any jointly authored works included in the thesis.

Publications included in this thesis

Peer Reviewed Journal Papers

S. Bhandari, N. Bergmann, R. Jurdak, and B. Kusy, "Time Series Data Analysis of Wireless Sensor Network Measurements of Temperature," *Sensors,* vol. 17, no. 6, pp. 21, 2017.

incorporated as Chapter 3, with conclusions section as part of Chapter 5. See contribution statement on page 18.

S. Bhandari, N. Bergmann, R. Jurdak, and B. Kusy, "Time Series Analysis for Spatial Node Selection in Environment Monitoring Sensor Networks," *Sensors,* vol. 18, no. 1, pp. 11, 2017.

incorporated as Chapter 4, with conclusions section as part of Chapter 5. See contribution statement on page 40.

Submitted manuscripts included in this thesis

No manuscripts submitted for publication.

Other publications during candidature

Refereed Conference Papers

B. Kusy, C. Richter, S. Bhandari, R. Jurdak, V.J. Neldner, M.R. Ngugi, "Evidence-based landscape rehabilitation through microclimate sensing," In *Proceedings of the 12th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON 2015)*, Seattle, WA, USA, 2015; pp 372-380

S. Bhandari, N. Bergmann, "An Internet-of-Things system architecture based on services and events," In *Proceedings of IEEE Eighth International Conference on Intelligent Sensors, Sensor Networks and Information Processing*, Melbourne, Australia, 2013; pp 339-344.

Contributions by others to the thesis

No contributions by others.

Statement of parts of the thesis submitted to qualify for the award of another degree

No works submitted towards another degree have been included in this thesis

Research Involving Human or Animal Subjects

No animal or human subjects were involved in this research.

Acknowledgements

One of the best things happen to my life was to meet my supervisor Neil Bergmann. I worked with many scholars before however no one has influenced me as he did. He cared for me personally when I had a health problem, rewrote my reports and graphs, but more than that I learned what research is and what it means to be a researcher from him. I have decided to spend the rest of my life in research solving the problems. I thank Neil to shape my "purpose of life".

I also express deep regards to my two other supervisors, Raja Jurdak and Branislav Kusy. Both of them are brilliant scientists and I have learnt so much from them. I must say these three researchers have challenged and changed me altogether. I wish other students like me get a chance to work with these brilliant people and find their passion in the field.

I would like to thank my fellow friends Vikram (India), Saheed (Pakistan) Dina (Malaysia), Ida (Indonesia), Ahmed (Lebanon), Mahjad (Iran), Weitao (China), Saju (Bangladesh) for their kind support. I will never forget our endless discussion about power, politics, religion, society and what not. Having friends from more than ten countries and working with them together was a remarkable opportunity for me.

Along with research questions, I also needed to work on a real-life question. I was diagnosed with Hodgkin's Lymphoma and that led to a nervous breakdown. It was painful but my wife, parents and my one-year-old son were there to help me in getting going. I took this as an opportunity to learn what it means to face physical and mental trauma at once. I learned to persevere and found the meaning of life in those difficult times. No problem will ever be a problem for me.

I would like to thank Australian government and CSIRO for the support of my Ph.D. scholarship. I will contribute back to the society as much as I can. In addition, I would like to thank people from the UQ and CSIRO who helped me directly and indirectly during my research. Everyone I met were kind and helpful. Thank you all.

Financial support

This research was supported by an Australian Government Research Training Program Scholarship.

This research was also supported by a CSIRO student scholarship.

Keywords

wireless sensor networks, time series analysis, interpolation, forecasting, environmental monitoring

Australian and New Zealand Standard Research Classifications (ANZSRC)

ANZSRC code: 080504, Ubiquitous Computing, 100%

Fields of Research (FoR) Classification

FoR code 0805 Distributed Computing, 100%

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ADF	Augmented Dicky-Fuller
AIC	Akaike Information Criteria
AR	Auto Regressive
ARIMA	Auto Regressive, Integrated, Moving Average
AUV	Autonomous Underwater Vehicles
CDF	Cumulative Distribution Function
IDW	Inverse distance weighting
LSE	Least Square
MA	Moving Average
MAE	Mean Absolute Error
MAQUMON	Mobile Air Quality Monitoring Network
MESSAGE	Mobile Environmental Sensor System across GRID Environments
ML	Maximum Likelihood
PAQ	Probabilistic Adaptable Query
PEIR	Personal Environmental Impact Report
RMSE	Root Mean Square Error
RV	Random Variables
WSNs	Wireless Sensor Networks

Chapter 1 Introduction

Environmental phenomena are dynamic processes that operate and cycle naturally around us. Air temperature, pressure, humidity, soil moisture are a few examples. Understanding the complete spatiotemporal behaviour of these processes is very important to pinpoint how they are evolving in space and time and impacting the surrounding ecosystem. One example of such monitoring is the Springbrook rainforest monitoring system in South-East Queensland where various environmental parameters are being observed to discover the impact of environmental phenomena on rainforest biodiversity [1].

Understanding the detailed spatiotemporal behaviour of environmental phenomena requires development of an effective observation system. Historically, weather stations have been one widely used environmental monitoring system. Weather stations have a wide range of high precision environmental sensors and capture good quality of environmental data. Being spatially sparse, weather stations only capture large-scale environmental variations. However, meteorological parameters such as surface temperature, wind speed, and humidity can vary at very small spatiotemporal scales [2, 3].

Recently wireless sensor networks (WSNs) have begun being used to observe environmental phenomena at varying spatiotemporal scales. As their costs reduce, WSNs can economically be deployed for comprehensive environmental sensing and monitoring [4]. Sensor networks have been used in various environmental observation including personal environment monitoring [5], building environment monitoring [6, 7], city centre heat monitoring [3], soil moisture measurements [8], volcano monitoring [9], ocean exploration [10], harsh mountain environment monitoring [11, 12] and many more which are listed in several review papers [13-15]. In most of the application scenarios listed, sensor networks are deployed with fixed positions. Mobile nodes, on the other hand, move around the area to be monitored. Small numbers of them may cover the larger area, failed nodes can be replaced by moving working nodes and nodes can change their location in a flexible manner [16]. This thesis deals primarily with static nodes, and how best to choose their spatial positions and their sampling frequency.

This research work considers some techniques to improve the configuration of wireless sensor networks to sense and monitor spatiotemporal environmental processes. For this work, just one environmental parameter – air temperature – is considered.

In particular, the thesis addresses two problems.

The first issue is how frequently sensor nodes should be sampled to give good temporal resolution. Because sensor nodes are typically powered by batteries, often with solar cell energy harvesting, they are energy limited. Nodes sleep, and then periodically wake, sense, record and transmit their data. The sensing frequency has a direct effect on the sensor energy use. If data is sensed less frequently, then intermediate values of temperature can be estimated. This investigation develops a methodology for deciding the best sampling period which maintains good accuracy for interpolated points. For the particular deployment scenario, this technique shows that reasonable accuracy (Root Mean Square Error (RMSE) of 0.2ºC) can be maintained while increasing the sampling period from once every 5 minutes to once every 60 minutes.

The second issue is how many sensors are needed, and where they should be placed spatially. Again in the spatial domain, temperature can be estimated by using the information from nearby locations to estimate temperature at a position without a sensor. This investigation develops a technique which starts with a dense deployment of sensors and uses statistical correlations between sensors to identify a minimum set of sensors which maintains good accuracy. For the particular deployment scenario, this technique shows that reasonable accuracy (RMSE of 0.5ºC) can be maintained with only 25% (3 out of 12) sensors.

The two investigations have been reported in two journal papers, [17, 18] which form the body of this thesis, and each of which contain some relevant background and literature review.

The organization of the thesis is as follows. Chapter 2 presents some broader background and literature review. Chapter 3 is a reformatted version of paper [17]. Chapter 4 is a reformatted version of paper [18]. Chapter 5 presents the conclusion sections of the two papers and some directions for future work.

Chapter 2
Background, Literature Review, and Research Questions

2.1. Background for the research

This section provides some background information about spatiotemporal environmental sensing and estimating. In the approach taken in this thesis, environmental phenomena are considered to be non-deterministic processes and they are normally modelled as random processes. This research follows similar approaches in the literature on environmental monitoring and models environmental phenomena as random processes [19-21]. This research models spatiotemporal variability and utilizes spatiotemporal estimation techniques to estimate environmental phenomena at unobserved locations and times.

2.1.1. Mathematical formulation of the stochastic random process

Consider a finite space and a time domain D and T where $D \subseteq R^d$ and $T \subseteq R^1$, with $d = 2$ for a planar measurement field, and d=3 if a three-dimensional sensing volume is considered. A monitored phenomenon is modelled as a stochastic random process Z that can be characterized as a collection of random variables $(RV)\ \mathrm{Z(u,t)}$ varying in space and time, i.e., D and T.

The domain $\mathrm{D}\times\mathrm{T}$ can have an infinite size. Complete characterization requires observation of the phenomena at each spatiotemporal point. Any realistic sampling strategy, however, samples a few realizations of the random process Z as a sequence of $\mathrm{z}(u_i, t_i)$ and those sparse observations are used to model the statistical behaviour of the phenomenon across the domain $\mathrm{D}\times\mathrm{T}$.

This research considers deployment of wireless sensor networks to sample the environmental phenomena. Deploying sensor network covering the whole spatiotemporal domain $\mathrm{D}\times\mathrm{T}$ may require dense sensor deployment. This research investigates how to characterize the spatiotemporal process Z at desired spatiotemporal scales.

An observation obtained from a sensor network is a realization of the random variable $(RV)\,\mathrm{Z(u,t)}$ at that particular point in space and time. Complete characterization of the random process Z requires the cumulative distribution function (CDF) of all possible random variable $(RV)\,\mathrm{Z(u,t)}$, i.e., $F(u,t;z) = \mathrm{Prob}\{\mathrm{Z(u,t)} \le \mathrm{z}\}, \forall \mathrm{z}, (\mathrm{u,t}) \in \mathrm{D}\times\mathrm{T}$.

2.1.1.1. Stationary assumption of random process

Because fully characterizing phenomena requires an infinite collection of realizations of random variables $(RV)\,\mathrm{Z(u,t)}$, stochastic modelling of environmental phenomena often makes the assumption of the process being stationary. A process is called stationary in the spatiotemporal domain if its behaviour remains statistically consistent in space and time. There are basically two forms of stationarity, strict and weak sense stationary. If the behaviour of the process remains consistent at any order then it is called strict sense stationary. However, behaviours of the process up to second order are considered sufficient for its characterization. That is why much of the research modelling environmental processes assumes second order stationary of the process [22]. Second order stationary process specifies:

1. Mean of the random variables $(RV)\,\mathrm{Z(u,t)}$ remains the same, i.e. $E\{\mathrm{Z(u,t))}\} = \mathrm{m}\,\forall(\mathrm{u,t}) \subseteq \mathrm{D}\times\mathrm{T}$
2. Second order moment i.e., covariance among random variables depends only on the spatiotemporal distance in $\mathrm{D}\times\mathrm{T}$, i.e. $E\{[z(u,t)-m][z(u',t')-m]\} = C_z(h,\tau)$

In the case of a variable like temperature, the average temperature is clearly not constant across a sensing region, so the raw temperature variable will not be second order stationary. Instead, as will be shown later in chapter 3, it is necessary to transform the data to make it stationary to be able to use some common modelling techniques.

2.1.2. Spatiotemporal variability modelling

Variability modelling characterizes the spatiotemporal structural behaviour of the environmental phenomenon. Characterising the variability structure from observed spatiotemporal locations allows us to estimate spatiotemporal observations at unobserved locations[23].

Variability of the observed phenomenon is first captured using a sample variogram:

$$\tilde{\gamma} = \frac{1}{2N(h,\tau)} \sum_{(i,j)\in(h,\tau)}^{n} \left(\mathrm{Z}(u_i,\tau_i) - \mathrm{Z}(u_j,\tau_j)\right)^2 \qquad (2\text{-}1)$$

where $N(\mathrm{u},\tau) = (i, j) : (\mathrm{u}_i - u_j) = h; (\tau_i - \tau_j) = \tau$

The sample variogram, $\tilde{\gamma}$ is then fitted to some standard variogram models. Some standard variogram models are linear, spherical, Gaussian and Matern.

2.1.3. Environmental estimation

Environmental estimation techniques are used in order to estimate the value of environmental phenomenon at unobserved locations. There are two types of estimation techniques in spatiotemporal estimation: deterministic and stochastic [24]. Deterministic estimation techniques use some parameters and estimate the spatiotemporal value at an unobserved location as a deterministic value. Stochastic estimation techniques use the statistical behaviour of the available observations to estimate the value at unobserved locations.

If there are n observations near an unobserved location $\tilde{z}$, any linear deterministic or stochastic estimation approach calculates the value at $\tilde{z}$ weighting each of the nearby observations based on their specific weighting methods.

$$\tilde{z} = \sum_{i=1}^{n} w_i z_i \qquad (2\text{-}2)$$

Weight w_i depends on the estimation approach that is used, as described subsequently.

2.1.3.1. Simple averaging

Simple averaging is one possible approach to estimate environmental phenomena at unobserved locations. It basically uses observations from nearby sample points and estimates values at unobserved locations. It does not consider variability, neither does it consider weighting neighbouring nodes differently.

If there are n observations near an unobserved location $\tilde{z}$ simple averaging estimates the value of $\tilde{z}$ weighting each of the nearby observations equally.

$$\tilde{z} = \sum_{i=1}^{n} w_i z_i, \qquad w_i = \frac{1}{n} \qquad (2\text{-}3)$$

2.1.3.2. Inverse distance weighting

Inverse distance weighting (IDW) is a deterministic estimation method. It estimates environmental phenomenon at unobserved locations giving higher weights to nearby observations compared to observations that are farther away [25]. This technique is simple and computationally very efficient. However, it does not incorporate variability of the phenomenon in the region and so sometimes it has high estimation error.

If there are n observations near an unobserved location $\tilde{z}$, inverse distance weighting estimates the value at $\tilde{z}$ weighting each of the nearby observations by their distance.

$$\tilde{z} = \sum_{i=1}^{n} w_i z_i, \qquad w_i = \frac{1}{d^p} \tag{2-4}$$

Weight w_i depends on the Euclidean distance d between the location to be estimated and nearby observation *i*. The relative weight of the neighbouring observation also depends on the power p in the weighting factor. Selecting higher values of p emphasises the closest neighbouring points. p can either be selected based on previous experience (e.g. p=1 is a common choice) or else the best value of p can be estimated based on detailed analysis of the sensor data.

2.1.3.3. Kriging

Kriging is a stochastic estimation approach. It is unbiased linear estimator and minimizes estimation variance [23].

If there are n observations near an unobserved location, $\tilde{z}$, kriging estimates the value at $\tilde{z}$ weighting each of the nearby observations by their spatiotemporal variability.

$$\tilde{z} = \sum_{i=1}^{n} w_i z_i, \qquad w_i = Cov(z, z')^{-1} Cov(z, \tilde{z}) \tag{2-5}$$

Factor $Cov(z, z')^{-1}$ represents the inverse of the covariance among all the available spatiotemporal samples, and $Cov(z, \tilde{z})$ is the covariance between all sample locations and the location where the estimation is to be performed.

2.1.3.4. Regression Kernels

The above approaches all estimate an unknown value as a linear combination of other known values in the neighbourhood of the unknown value. A more generalized approach, called regression kernels, allows a value to be estimated as a more complex function of the neighbouring values [26].

First a kernel function, such as a zero-mean gaussian is chosen, where the value of the kernel decreases with the distance, d_i, between the location of the unknown value and the observation z_i, at a rate determined by a scaling constant σ:

$$k_i = e^{\frac{-d_i^2}{2\sigma^2}} \tag{2-6}$$

Then each neighbouring observation is weighted by the value of the kernel function (normalised by the sum of all the kernel function values in the summation):

$$\tilde{z} = \sum_{i=1}^{n} w_i z_i, \; w_i = \frac{k_i}{\sum k_j} \tag{2-7}$$

2.1.3.5. Estimation error

In spatiotemporal estimation problems, estimation error performance is commonly measured using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and estimation error variance [24].

Sensors are not perfect. The reading from a temperature sensor is affected by the inherent sensor accuracy, its resolution including digital quantisation, compensation for effects of other variables (such as humidity, pressure or wind) and its calibration history. In this work, we are not aiming to determine the error between the estimated temperature and the actual temperature. Instead, we are aiming to measure the difference between the reading from a sensor at that position, and a reading estimated from nearby temporal or spatial readings. This is what we call estimation error.

2.1.3.6. Data Sources

The data used in Chapter 3 is from real spatiotemporal data traces obtained from Springbrook sensor network situated in southeast Queensland. The network is described in detail in Chapter

3. The data in chapter 4 is based on sensor data from the Meandu mine site rehabilitation and that data is described in more detail in that chapter.

2.2. Literature Review

Environmental monitoring has a long history. As mentioned in [19], the Australian Bureau of Meteorology has been monitoring climatic variables such as temperature, pressure, solar radiation and rainfall since 1957. However, only 4600 monitoring stations are installed to cover the whole of Australia as manufacturing and operating costs of weather stations are very high [19]. They are observing environmental parameters at a spatial separation of more than ten kilometres. Research, however, shows that meteorological parameters fluctuate at a very small spatiotemporal scales [2, 3]. As mentioned in [3], less than 100 metres distance in a city sees different temperature measurements. Such variations can have heat-related effects on the people living or working around the environment. Similar effects of small-scale temperature variations in plant and animal growth may be observed in rainforest environments [1]. Thus, such sparse monitoring system is inadequate to capture small-scale spatiotemporal behaviour of environmental phenomena.

Recent development in the field of sensor technology has enabled a new possibility for environmental sensing and monitoring[4]. Wirelessly communicating groups of sensor nodes are being deployed for in-situ sensing and monitoring of a wide range of environmental phenomena. Cost of the deployment can be high if the spatial area to be covered is large. If nodes fail during observation, parts of the area remain uncovered. Also, adding and removing nodes during operation may not be possible.

Mobile sensor networks are also an area of growing interest. As nodes move around the area of deployment a smaller number of mobile nodes may be sufficient to cover larger spatial regions. Also, development of various mobile communication platforms, such as smart phones and any other portable mobile devices, and the inclusion of various sensors on them has created opportunities for opportunistic sensing of the environment [27]. Sensors can be carried by any mobile entity such as people moving around or animal such as flying fox.

Considering the flexibility provided by a mobile sensor network, researchers have started deploying mobile sensor network in environmental sensing and monitoring. In [28] authors have developed an environmental pollution monitoring vehicular sensor network. Mobile Environmental Sensor System across GRID Environments (MESSAGE) [29, 30] was a large project deployed in the United Kingdom and Europe where static and mobile sensor networks

were deployed in environmental sensing and modelling. In [31], authors from UCLA have developed a system called Personal Environmental Impact Report (PEIR) that senses environmental parameters exploiting location tagged data observed using mobile phones to estimate personalized environmental exposure and its impact. CitiSense, an air quality monitoring platform has been developed in [32] envisioning a "citizen infrastructure" to monitor pollution and environmental conditions where users get exposed in their daily life. Mobile Air Quality Monitoring Network (MAQUMON) has been proposed in [33] that can provide real-time air quality information to the public using it Sensor Map visualization interface. In [34, 35] authors have proposed participatory pollution monitoring using smart phones and discussed their real-time experiments conducted in Zurich Switzerland. In the HazeWatch project at The University of New South Wales,[36] researchers monitored environmental pollutant concentration in Sydney area with the help of sensor mounted in vehicles. In[37], researchers from CSIRO used small and low-cost Autonomous Underwater Vehicles (AUV) to record spatial data between fixed sensors deployed on the surface of the water and the sea bed.

The purpose of sensor networks is to observe environmental phenomena covering as much as possible spatial and temporal domains. However, covering whole spatiotemporal regions can be difficult. As a solution, researchers have employed various spatiotemporal estimation techniques. Spatiotemporal estimations help researchers estimate measurements of observed phenomena at desired spatiotemporal scales. This section reviews some wireless sensor network based spatiotemporal estimation techniques, their focus, results obtained and their relationship to this research proposal. In [23, 38-40], the authors have used a kriging based estimation technique in coverage hole reduction. Sparse data obtained from sensor nodes are spatially interpolated to other locations. In [41] theoretical work on spatiotemporal estimation of sensor networks is explored merely suggesting spatiotemporal characteristics can be exploited in reducing sensor network energy consumption. In [39], the author proposed to use mobile nodes in spatiotemporal estimation and proposed a recursive estimation approach. In [22], authors proposed to use mobile sensor networks with spatiotemporal kriging. Spatiotemporal estimation of environmental phenomena was performed in [42]. A decentralized data fusion approach was proposed in [43] to explore road networks. This research also aspires to estimate sensor observations at unobserved locations and times.

One significant difference from many other projects is that this work is validating results with real-time environmental dataset compared to simulation and theoretical based approach. This

ground truth verification with real platform deployment provides strong evidence of the usefulness of the new techniques that will be developed.

From our literature survey, we observe that many of the sensor network based environmental monitoring projects focused their research on network, MAC and physical layer related networking challenges. As shown in detail in the comparison table below, most of the research works have deployed a limited number of nodes; covered limited spatial region and sampled the environment for a short period of time. This research has demonstrated that sensor networks can be deployed for environmental sensing and monitoring, but not necessarily how they can be best deployed.

A number of approaches have been proposed in the literature for determining the best sampling interval for time series. Alippi et al [44] summarise different adaptive sampling techniques. In many cases, these methods compare the sample with a model and do not transmit data if the data fits the model. However, the node still needs to wake to take the test sample.

Harb et al [45] compare three techniques to optimally set the sampling interval for an industrial process monitoring application. One method uses statistical analysis of data variances to estimate a good sampling interval, another method is based on set-similarity functions which can use past history to inform future readings, and the third technique uses distance-functions to estimate when estimates are stale and new readings are needed.

There has similarly been substantial research into determining the optimal spatial resolution required for sensing. Statistical techniques like those proposed by Marceau [46] look at the spatial frequency at which various phenomena change, and then use Nyquist sampling approaches to decide upon the optimal sampling interval.

Budi et al [47] have very recently proposed using a mobile platform to explore an area prior to sensor deployment, and using those readings to design an optimal sensor placement. Jin et al [48] also propose a robot-based sensing system for indoor air quality, and investigate techniques for interpolating spatio-temporal values from sparse robot readings.

The current state of the works related to spatiotemporal estimation using wireless sensor networks has been reviewed briefly above. Much of the current work on spatiotemporal estimations are limited in answering questions related to the comprehensive spatiotemporal estimation of environmental phenomena. We observe that many questions related to wireless

sensor network based spatiotemporal modelling of environmental are still open, such as the following. How effectively is the sensor network capturing the spatiotemporal behaviour of environmental phenomena compared to reality? What approaches can be used to estimate spatiotemporal dataset at uncovered spatiotemporal locations? How many static nodes would be enough to cover certain regions and capture fine-grained spatiotemporal behaviours?

In short how effective sensor network can be in fine-grained spatiotemporal sensing and monitoring of environmental phenomena still has significant research gaps.

Table 2-1 on the following pages summarize aspects of some previously reported environmental sensing projects.

Table 2-1. Summary of some typical projects involved in environmental sensing and monitoring using wireless sensor network and their environmental analysis

Project name	**Year**	**Observed phenomena**	**sensor nodes (static/mobile)**	**Spatiotemporal Coverage area (analysis)**	**Estimation at all locations**	**Ground truth evaluation**	**Error analysis**
Abbreviations used: ***NP*** = Not performed; ***NS*** = Not specified							
VSN Singapore [28]	2009	Environmental pollution	Single mobile node (car)	Selected routes for experimental period	***NP***	***NP***	***NP***
Citisense San Diego [32]	2012	Environmental pollution	16 smartphones for two weeks	Selected paths that trial users visited	***NP***	***NP***	***NP.***
MESSAGE[29] UK	2008	Air pollution	Mobile and static sensors	Certain traffic routes	***NP***	***NP***	***NP***
Sensorscope [11] Switzerland	2007	High Swiss Alps environmental monitoring	23 sensors deployed for a month and half	only a small area	***NP***	***NP***	***NP***
N-SMART[49]	2008	Air pollution	6 taxis, 4 personal, two weeks experiment	Taxi ways covering some	***NP***	***NP***	***NP***

				parts of Accra, Ghana			
MAQM USA[50]	2008	CO, O3, NO2	Sensor mounted cars (no exact numbers and duration are provided	Covered the road network of the Nashville,	***NP***	***NP***	***NP***
MoDisNet [51] London	2008	Air pollution	12 Static and, 6 mobile sensor nodes	Covered some sections of London	***NP***	***NP***	***NP***
UScan Tokyo [3]	2010	Temperature , vibration, illumination	200 sensor nodes deployed for 2 months (1800 nodes/Km^2)	Small section of tokyo	***NP***	***NP***	***NP***
Tungurahva Equador [9]		Earthquake	16 nodes, sampling at 100 Hz, deployed for 19 days	Coverage of 3 KM area	***NP***	***NP***	No variability, uncertainty and error analysis
MEM Taiwan [52]	2011	Pollution	9 sensor nodes for April 22 to May 3 2011, sampling at every 2 minutes	Points where sensors are deployed	***NP***	***NP***	***NP***

PermaSense Switzerland [12]	2007	sensor nodes monitoring permafrost	10 sensor nodes for months	Only fixed points are observed	***NP***	***NP***	***NP***
Haze Watch Sydney[36]	2012	Environmental pollution	Sensor mounted in Cars	Covered only certain sections of the roads in Sydney	IDW, Kriging at Map	Compared result with Government installed fixed stations	***NP***
Commonsense India[53]	2005-2006	Rain fall, temperature, pressure, soil moisture	10 nodes, sampling at every 5 minutes	Indian Institute of Science campus area	***NP***	Observations are compared with measurements from fixed stations	***NP***
Wannengrat Switzerland[54]	2009	snow monitoring sensors	7 sensors	Covers only deployed area	***NP***	No ground truth verification	***NP***
Opensense Switzerland [35]	2010	Air pollution	***NS***	tram ways and the region covered by the fixed sensor	***NP***	Sensors are calibrated with high quality fixed station based observations	***NP***

PEIR USA[29]	2009	Air pollution	30 users for 6 months	Roads and specific locations	***NP***	publicly available meteorological services are used	***NP***
Participatory air pollution monitoring ETH Zurich [32]	2012	Air quality (O3) measurement system (GasMobile)	Several bikes are used for two months	Only bicycle paths are covered	***NP***	Uses static stations to improve sensor calibration	Analyse effect of mobility on the accuracy of the sensor.
Springbrook[13]	2008	Environmental phenomena	175	Sensors are deployed strategically measuring rainforest regeneration	***NP***	***NP***	***NP***
Airy Notes[50] Shinjuku Gyoen Garden, Japan	2005	uPart sensor (temperature, light, movement)	160 sensors, (May, 25 to June 12 2005), sampling in every 10 seconds	Different regions, business area, border area, forest area, garden field.	***NP***	***NP***	***NP***

2.3. Research Gaps and Research Questions

In the previous section, some typical examples of environmental sensor networks were described. The cost of a wireless sensor network deployment depends on at least two design decisions.

Firstly, the energy requirements of the sensor nodes determine the size of energy storage (batteries) and energy harvesting (e.g. solar cell area). Energy requirements grow with more frequent sampling of environmental parameters and the more frequent transmission of the results. While the existing literature provides some statistical and heuristic methods for determining the sampling period based on the nature of the data time series, the availability of long-term real-world sensor data provides an opportunity to explore this question in more detail.

Secondly, the cost of a deployment depends on the spatial density of the sensor nodes, i.e. how many sensor nodes are deployed. Again, while the existing literature provides some methods for determining the spatial sampling interval based on the nature of the data, the availability of long-term real-world sensor data provides an opportunity to explore this question in more detail.

This research investigates these two issues through a dense spatio-temporal deployment of sensors (i.e. many sensors recording parameters often) to develop data-driven methodologies for determining appropriate density of nodes, node locations, and node sensing duty cycles. Fairly standard time-series analysis techniques are used as the basis for these methodologies.

This leads to two research questions:

Research Question 1: ***Based on time series analysis, can high-frequency sensor data be used to determine appropriate long-term sampling intervals for environment sensor data?***

Research question 1 is answered through the paper "Time Series Data Analysis of Wireless Sensor Network Measurements of Temperature" [17] which forms the basis for chapter 3. Chapter 3 presents the background, literature review, experimental methodology, results and Chapter 5 presents the conclusions for this research question.

Research Question 2: ***Based on time series analysis, can high spatial density temporary sensor deployments be used to determine appropriate long-term spatial density and sensor node locations for environment sensor data?***

Research question 2 is answered through the paper “Time Series Analysis for Spatial Node Selection in Environment Monitoring Sensor Networks” [18] which forms the basis for chapter 4. Chapter 4 presents the background, literature review, experimental methodology, results and Chapter 5 presents the conclusions for this research question.

Chapter 3 incorporates the following paper,
with the conclusions section as part of Chapter 5:

S. Bhandari, N. Bergmann, R. Jurdak, and B. Kusy, "Time Series Data Analysis of Wireless Sensor Network Measurements of Temperature," *Sensors*, vol. 17, no. 6, pp. 21, 2017.

Contributor	Statement of contribution
Siddhartha Bhandari (Candidate)	Conception and design (85%) Analysis and interpretation (85%) Drafting and production (80%)
Neil Bergmann	Conception and design (5%) Analysis and interpretation (5 %) Drafting and production (10%)
Raja Jurdak	Conception and design (5%) Analysis and interpretation (5 %) Drafting and production (5%)
Brano Kusy	Conception and design (5%) Analysis and interpretation (5 %) Drafting and production (5%)

Chapter 3 Temporal Interpolation

3.1. Introduction

Wireless Sensor Networks (WSNs) allow dense spatiotemporal measurement of environmental phenomena such as temperature, humidity, solar radiation and rainfall [13] which in turn can be used to better understand local environmental conditions and processes. However, low-cost WSNs are also characterized by the resource-constrained nature of the WSN hardware. Limited available energy for data sensing, storage and transmission is a common constraint in WSNs in remote areas where mains power in unavailable or uneconomical to access. Sensor nodes are typically battery powered, where node lifetime is determined by battery lifetime. Indefinite operation can be achieved with energy harvesting using technologies such as solar cells, but energy efficiency is still a key factor in determining the cost of deployment since more energy use means larger and more expensive rechargeable batteries and solar cells.

The spatial extent, spatial density and sensing frequency of the WSN nodes is partially determined by the scientific purpose of the deployment, but they will also be determined by the ability to model the processes which generate the environmental data in sufficient detail to be able to interpolate data values between sensed readings, both in time and space. If data can be accurately estimated between readings, then the frequency of making readings can be reduced, which in turn reduces the energy requirements and the deployment cost of the system, while increasing its lifetime. Previous work has not investigated the quantitative effects of reducing sampling frequency on the accuracy of both interpolated and predicted values. The optimal sampling interval will depend on the parameters being sensed, the environment in which they are sensed, the specific features of the sensors, and the scientific requirements for accuracy. This paper demonstrates the use of a data-driven method for determining sufficient sampling intervals through analysis of several specific sensor deployments. While we use temperature as a use case, many features of our approach are generalizable to other sensing modalities.

This paper first investigates the nature of temperature readings in a large scale WSN deployment in Springbrook, Australia [1]. Around 175 microclimate sensor nodes have been deployed for more than 5 years, and they have recorded temperature readings (as well as other environmental phenomena) every 5 min during this time. This provides a rich source of data for further analysis. For this paper, just one week of data has been explored, since there is a significant cost involved in data cleaning and checking prior to statistical analysis. The robustness of results would be improved if the analysis was applied to a larger portion of the data.

In this paper, the temporal dynamics of the temperature recorded by the WSN is analyzed in detail, with a view to answering two questions. Firstly, if the interval between sensing events is increased, how accurately can temperature be interpolated between the sensor readings. Longer sensing intervals will reduce the consumed energy, and hence reduce deployment cost or extend deployment lifetime. Secondly, if real-time readings of temperature are needed, for how long can future values of temperature be accurately extrapolated without needing instantaneous data transmission.

This paper addresses two research questions. Firstly, it analyzes the reduction in measurement accuracy if the sampling interval is extended with temperature interpolated between these values. Also different interpolation methods are compared.

Secondly, we model the temperature phenomenon as a stochastic process and analyse it using a time series modelling framework [55], and use this analysis to determine how the short-term predictability of future temperature is affected by sampling interval, and extrapolation technique.

The rest of the paper is organized as follows: Section 2 reviews the related literature. Section 3 explains the data used, Section 4 examines the first research question about the effect of sampling interval on temperature measurement accuracy, Section 5 repeats the analysis for a different data set, Section 6 explains time series modelling as background for the second research question, Section 7 answers this research question about future temperature prediction, and Section 8 concludes the paper.

3.2. Previous Work

WSNs have the potential to revolutionize environmental sensing, providing high spatial and temporal resolution data [4]. Recent deployments include personal environment monitoring

[5], city monitoring [3], building monitoring [6], ocean exploration [10] and toxic gas monitoring[7].

However, the nature of the measured phenomena is not always well understood. Environmental phenomena can vary at very small spatiotemporal scales [2, 3]. Exhaustive spatiotemporal study of the behaviors of such dynamic phenomena requires the deployment of an adequate number of sensor nodes and effective collection of data.

In terms of temporal resolution, various ad hoc schemes have been proposed to optimize sampling frequency, e.g., in [8] soil moisture is sampled more frequently near rain events to give more useful data, however, such techniques have not considered the detailed statistical nature of the signals.

Techniques have been proposed for spatially interpolating values within a sensor field [22, 23, 56, 57] but these generally assume a smooth gradient across the sensor deployment area, and the techniques have not been well verified in real deployments. Most of the aforementioned references did not consider the statistical behavior of the environmental phenomena or they assume process stationarity [22]. Liu et al. [58] also investigate spatially clustering nodes and reducing sampling interval by having only one sample report from a cluster each sample interval. The same effect could be achieved by simply reducing each cluster to a single node. Also, their spatial redundancy techniques have not been tested on real data, only on synthesized data.

Use of formal time series analysis in sensor networks has been reported by several researchers. Law et al. [59] use time-series modelling to decide the confidence levels for future samples, and skip the future readings if the values are likely to be accurate enough. However, this requires substantial processing, and adjusting time series models continuously for each new reduces the number of required samples by less than 50%.

In [60], Le Borgne et al. use time series prediction for future estimation of samples, so that some data transmission can be suppressed. They present a useful algorithm for selecting a suitable time series, but savings are only achieved for data transmission. The sensors still need to sample data at the full rate. Miranda et al. [61] use autoregressive models to predict samples based on spatially nearby sensors, however, their work does not investigate how to decide upon the optimum sample rate. Liu et al. [62] also present a method for suppressing the transmission of data samples if the receiver is able to accurately forecast samples based on time series models. Sensors are still required to sample data regularly. This method does not allow

sampling intervals to be increased. Recently, Aderohunmu et al. [63] have also used similar time-series modelling for forecasting future sample values so that data transmission can be suppressed. Amidi [64] has used ARIMA modelling for the smoothing of noisy data and for interpolating missing data samples in a series, but again has not analysed the best sample rate to provide accurate data interpolation.

Pardo et al. [65] investigate a neural network model for predicting the future temperature in an indoor environment for use with intelligent air-conditioning. Their neural network predictors perform considerably worse than Bayesian predictors (although the authors claim there is little practical difference), but their work does not investigate the effect of different sampling intervals.

Liu et al. [58] propose on-sensor temporal compression of data by only transmitting a dynamically computed subset of data (with linear interpolation between these). This reduces the quantity of transmitted samples, but not the sampling interval of the sensors, and also increases the latency before receiving measurements.

Tulone and Madden [66] propose a system called Probabilistic Adaptable Query (PAQ) system which develops an Auto Regressive (AR) time series model for every node for predicting future values. If the future predictions based on past transmitted values are below some threshold, then no new data is transmitted. Once this threshold is exceeded, new data is transmitted. Data still needs to be sampled at high temporal resolution, and there is no investigation of what the best sampling interval should be. They also propose round-robin scheduling on sensors in spatial clusters.

In general, these previous works have used time series analysis to model the statistical behavior of the data. They have been used for outlier and anomaly detection, and for separating the underlying trends from noisy signals. They have been used for suppressing data transmissions when forecast values are close to the measured values. However, with such systems, there has been no reduction in the sampling interval, just in the transmitted data. Energy use consists of three main components. Firstly every time data needs to be sampled, the sensor node needs to wake up, wait for the sensor node and sensing transducer to stabilize, undertake any computational tasks (such as calibrating readings, or comparing against predicted estimates of values), and possibly transmitting data to the data sink. Previous work still requires the sensor to wake up, stabilize and compute at high sampling frequency. Even if the energy to wake up, stabilize and compute is relatively small compared to transmission costs, as would be the case

for a temperature sensor, reducing the sensing frequency, and hence the number of wake up times will have a direct impact on sensor lifetime. Substantially more energy can be saved in the sensor sampling interval can be extended without compromising the scientific usefulness of the collected data. Previous work has not used time series analysis to analyse the accuracy of both interpolated and extrapolated data values as the sampling period is varied. This analysis can help a sensor network designer to set a sampling rate that satisfies the required error limit whilst reducing energy consumption.

In this work, no behavioral assumptions of the process are made and all analyses are validated with proper statistical tests. This analysis will allow insights into the required sampling intervals for long-term deployments with moderate accuracy requirements.

It is worth noting that several papers, e.g., [58, 66], reduce sampling intervals by round-robin scheduling of nodes with a spatial cluster of highly correlated nodes. In this paper, only sampling within a single time series is investigated, although we expect to address spatial redundancy in our future work.

3.3. Temperature Data from Springbrook WSN Deployment

This section describes one set of temperature data that used for this study and presents some simple empirical observations. Situated in southeast Queensland, the Springbrook WSN deployment consists of 175 sensor nodes, covering one square kilometre of area, monitoring temperature, pressure, humidity, wind, and several other environmental parameters with a sampling period of 5 min, and it has been operating since 2008[1].

An aerial photograph of the site is shown below in Figure 3-1. The nodes used in Figure 3-2 (nodes 2,3,4 and 5) are shown with blue circles and larger numbered labels beside them. The data from node 2 is used in subsequent data analysis.

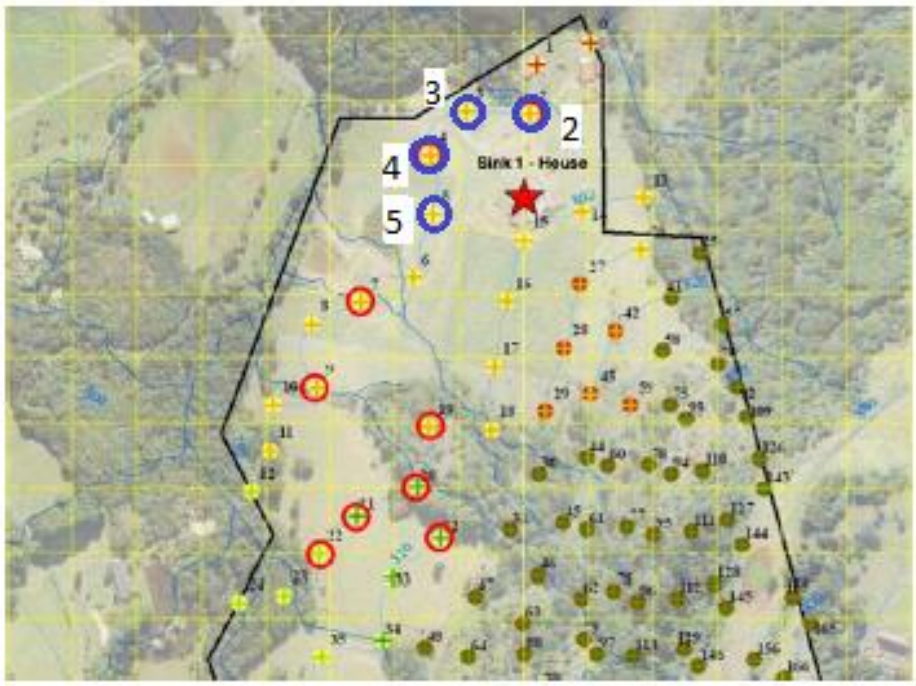
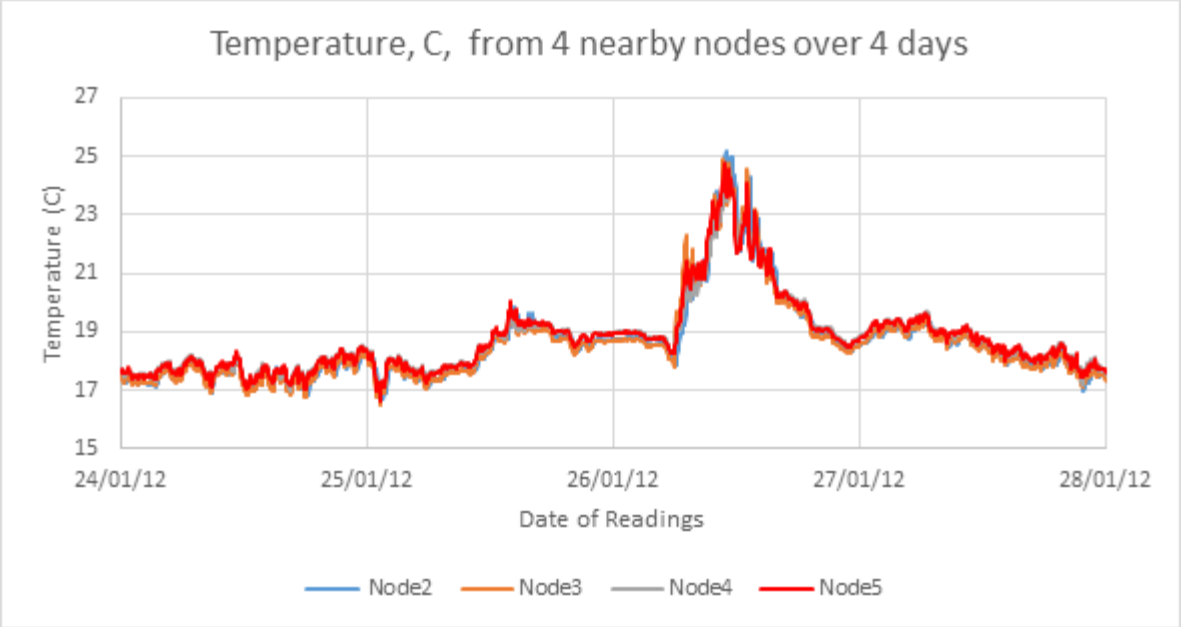


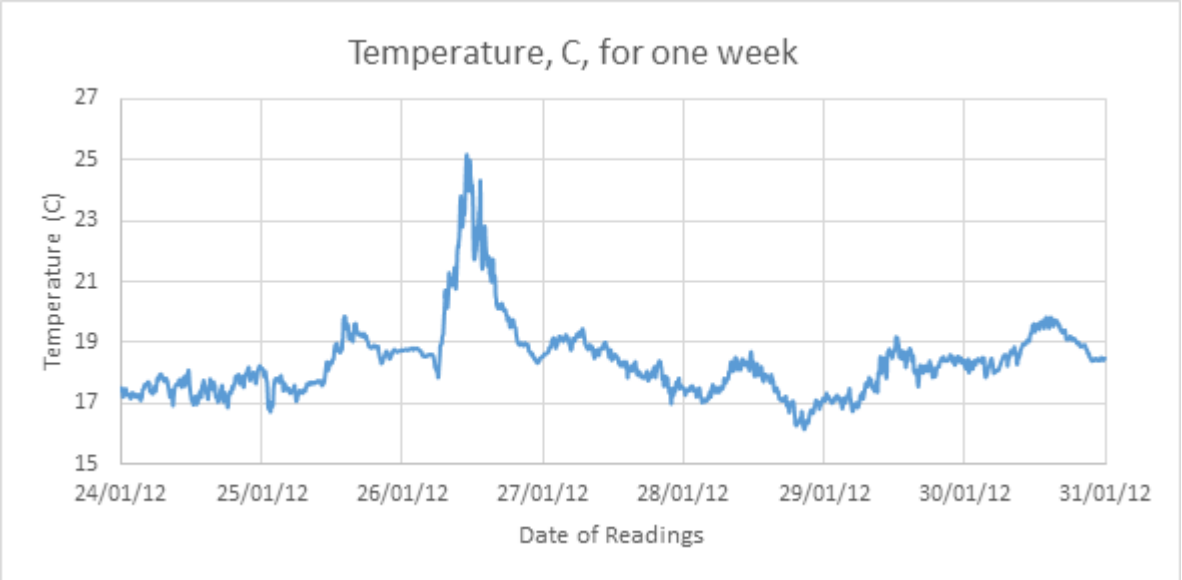
Figure 3-1: Aerial photograph of Springbrook site

Figure 3-2 (a) shows four days of data from four sensors in the deployment. This shows that generally the temperature patterns are highly correlated between nearby sensors, since the values are largely superimposed. This means that interpolation and prediction results from one sensor node should be representative of results from all nodes in that deployment. However, the temporal pattern over the week does not always show a clear daily pattern. This shorter section has been shown (rather than the whole week that is used for subsequent analysis) to more clearly illustrate that temperatures are less highly correlated when temperature changes are rapid such as during the temperature changes on 26/1/2012, and more highly correlated on days with smaller changes, such as 24/1/2012.

Figure 3-2 (b) shows the readings of one sensor over one week which shows that the temperature does not rise and fall smoothly over the course of a day but has a significant component of noise. This data from node 2 will be used for subsequent analysis.



(a)



(b)

Figure 3- 2. (a) Four day time series plot of four nearby sensors. (b) One week of samples from one sensor (node 2).

Figure 3-3 shows a version of the signal, based on differences between consecutive signals, as given by Equation (3-1):

$$Y'(t) = Y(t) - Y(t-1) \tag{3-1}$$

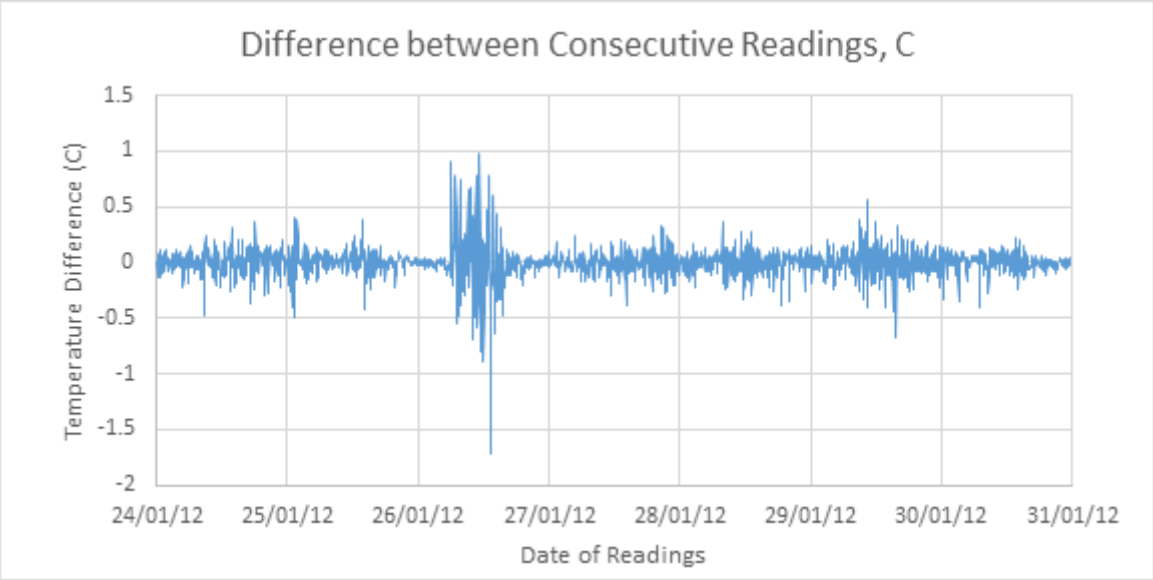


Figure 3-3. One week of difference values.

On first observation, this differenced signal does not have any clear structure, but appears largely random. Simple statistical analysis shows a mean close to zero and a standard deviation of 0.14 °C.

3.4. Accuracy versus Sampling Interval

As mentioned earlier, energy can be saved and sensor lifetime extended if the interval between sensor readings is extended. In this first experiment, the sensing interval is extended from the existing 5 min intervals to intervals of 10 min, 15 min, 20 min, 30 min, 45 min, 60 min, 90 min and 120 min by selecting appropriately spaced samples from the 5-min data for one sensor over one week. Values at the intervening 5 min intervals are then interpolated, and the RMSE (root-mean-square error) and MAE (mean absolute error) of the interpolated values are calculated. Two different interpolation algorithms are chosen. The first method uses linear interpolation between the sampled points, and the second method uses a cubic spline between the sample points. Table 3-1 shows the RMSE and MAE of interpolated values, and the 99th percentile absolute error when the various interpolation methods are applied to the one week sequence shown in Figure3- 2.

Table 3-1. Interpolation Error for Different Sampling Intervals (in °C).

Sampling Interval (Mins)	RMSE Linear	MAE Linear	RMSE Cubic	MAE Cubic	99% Linear	99% Cubic
10	0.0884	0.0528	0.0852	0.0519	0.3250	0.2893
15	0.1097	0.0664	0.1088	0.0669	0.4000	0.4037
20	0.1166	0.0755	0.1228	0.0793	0.4200	0.4496
30	0.1527	0.0937	0.1531	0.0962	0.5800	0.5709
45	0.1865	0.1152	0.1921	0.1190	0.6867	0.7410
60	0.2224	0.1335	0.2330	0.1430	0.8425	0.8753
90	0.2439	0.1566	0.2507	0.1629	0.9133	0.8774
120	0.2646	0.1720	0.2893	0.1882	0.9425	1.0206
240	0.3297	0.2161	0.3290	0.2215	1.2758	1.2189

Figure 3-4 shows the growth of error with increasing sample intervals. The 95% confidence interval for the RMSE of linear interpolation is also shown in Figure 4, and the difference between linear and cubic interpolation is not significant within these confidence intervals. Except at smaller sampling intervals, cubic spline interpolation gives poorer results, and so linear interpolation is preferred.

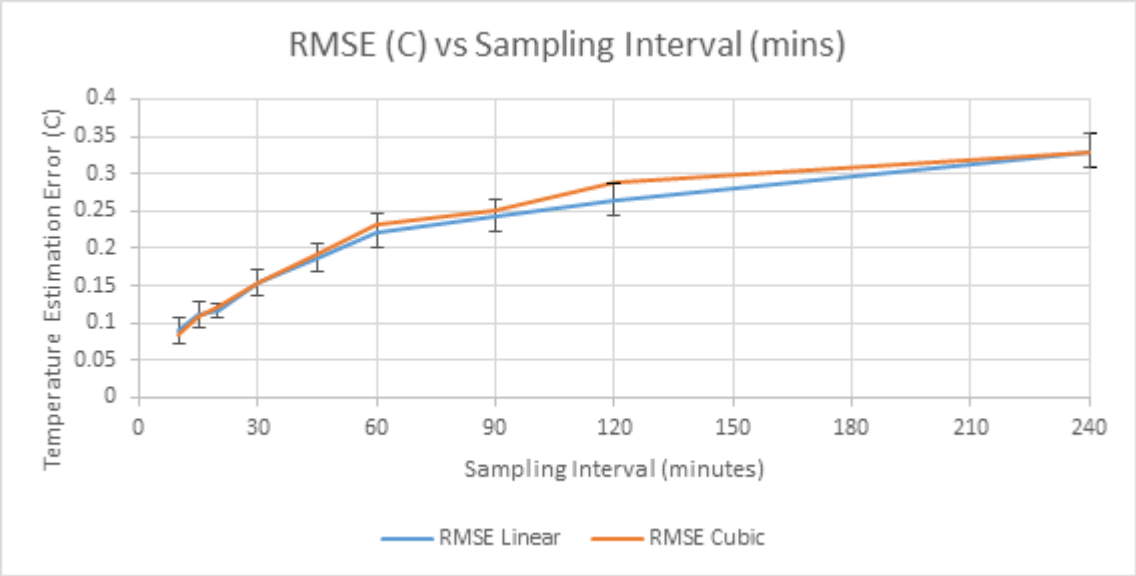


Figure 3-4. RMSE of linear and cubic interpolation showing 95% confidence interval of RMSE Linear.

These results show that with linear interpolation, the MAE remains below the standard deviation of the difference signal (0.14 °C) in Figure 3 when the sampling interval is extended to 60 min. Alternatively, if the accuracy requirement was that 99% of interpolation errors have an absolute magnitude of less than 0.5 °C then the sampling interval can be extended to 20 min.

It should be stressed that these results apply to this particular deployment. The general result, however, is that statistical analysis of sampled data over an initial deployment at relatively high

sampling rate can give insights into a lower long-term sampling rate which does not significantly sacrifice accuracy.

3.5. Repeating for Another Data Series

The analysis above is repeated for another temperature data set using a different set of sensor hardware, a different physical location (a mine rehabilitation and revegetation site) and a different time of year (December 2013), again with samples every 5 min[67]. Figure 3-5 below shows four adjacent sensors over a one week period[1], Figure 3-6 shows one signal, Node 5, in detail, which has a clear cyclic pattern. Figure 3-7 shows the differences between consecutive signals over 7 days. The signal appears mostly like a random noise signal, centred on zero. The variance of the noise is not constant, but also varies cyclically with higher variances in the middle of the day. The standard deviation of the temperature difference is around 0.3 °C.

Table 3-2 repeats the analysis of how well linear interpolation and cubic spline interpolation can estimate intermediate temperatures if the sampling interval is reduced to 10 min, 15 min, 20 min, 30 min, 60 min, 690 min, 120 min or 240 min.

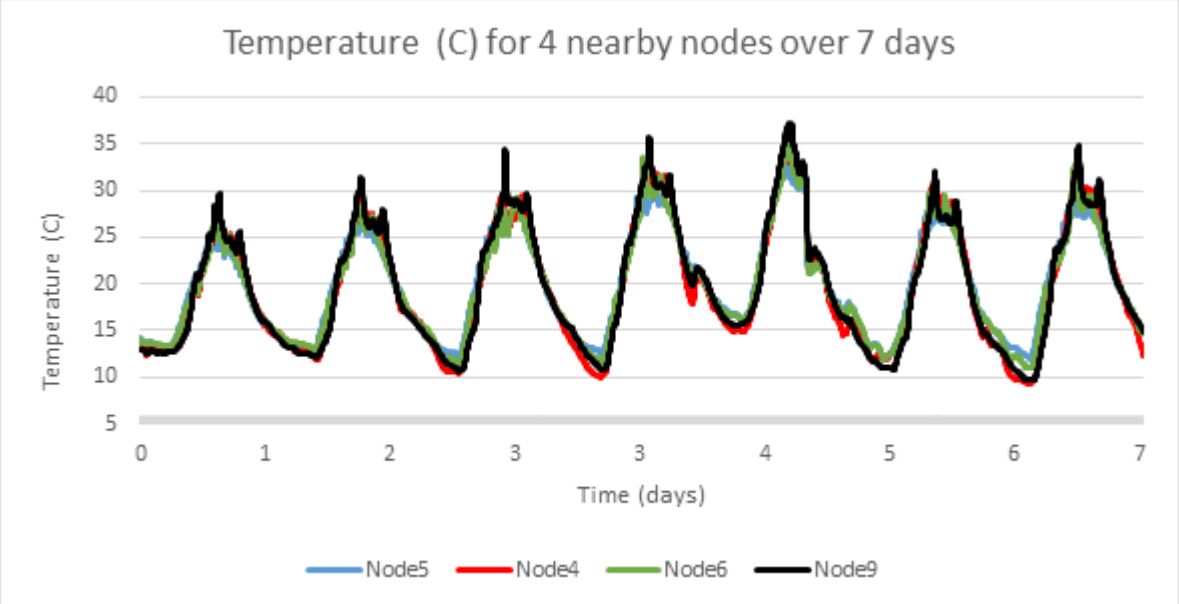


Figure 3-5. Adjacent sensor readings for a second experiment.

[1] These nodes in Figure 3-5 are labelled 4,5,6,9 in the lower left corner of Figure 4-1 in the next chapter.

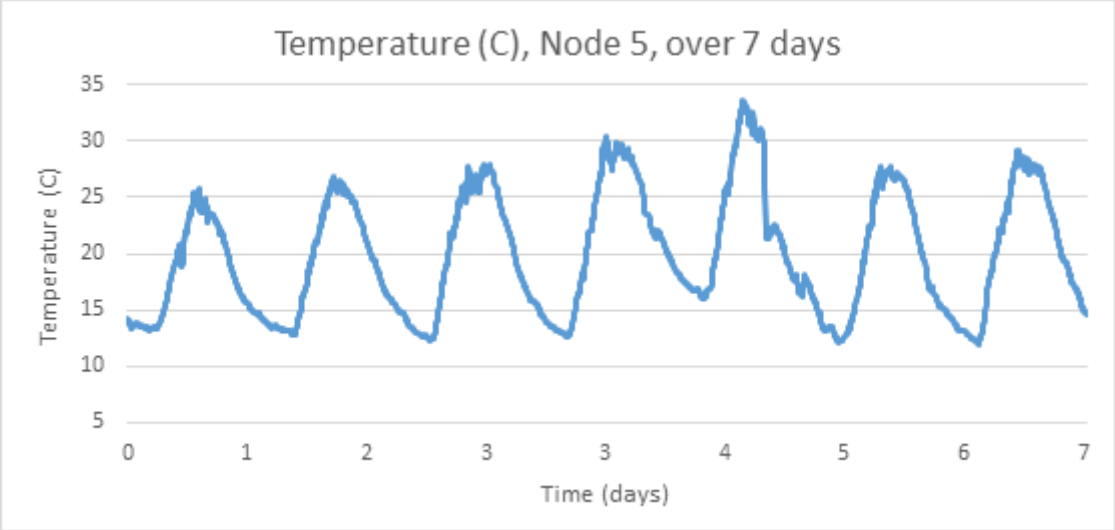


Figure 3-6. Detailed Readings for Node 5.

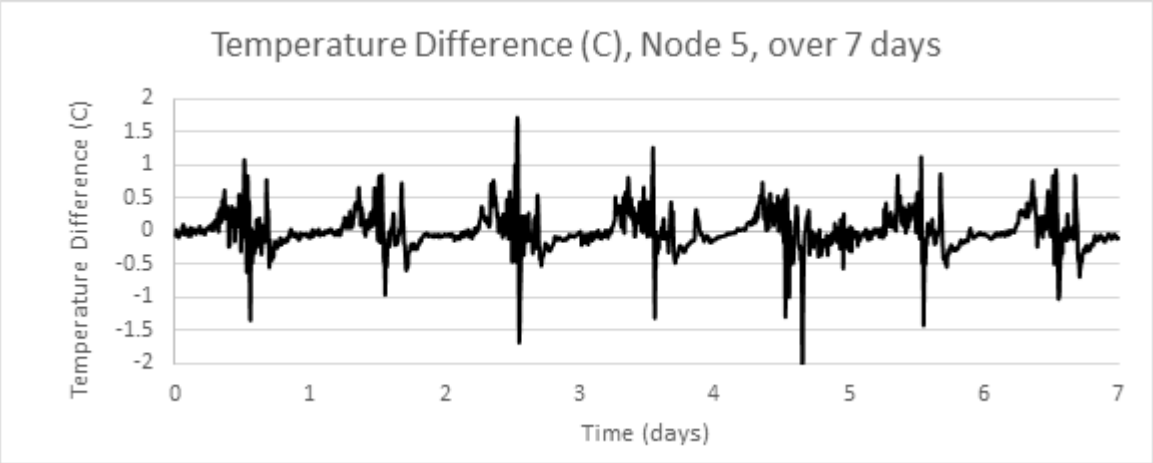


Figure 3-7. Temperature Difference, Node 5 over 7 days.

Table 3-2. Interpolation Error (°C) for Different Sampling Intervals for Mine Data.

Sampling Interval (Mins)	RMSE Linear	MAE Linear	RMSE Cubic	MAE Cubic	99% Linear	99% Cubic
10	0.1746	0.0941	0.1751	0.0960	0.6740	0.6366
15	0.2085	0.1164	0.2185	0.1211	0.7554	0.8286
20	0.2342	0.1360	0.2487	0.1459	0.8862	0.9436
30	0.2723	0.1588	0.2846	0.1693	1.0099	1.0027
45	0.3664	0.2029	0.3694	0.2087	1.2578	1.3131
60	0.4655	0.2498	0.4635	0.2493	1.5781	1.5309
90	0.5837	0.3093	0.5762	0.3033	1.9658	1.8047
120	0.6057	0.3836	0.5840	0.3663	2.1344	2.0859
240	0.9780	0.6687	0.8121	0.5515	3.0073	2.7782

Again linear interpolation gives better estimates at smaller sampling intervals up to 60 min. For sampling intervals over 60 min, there is a small advantage for cubic spline interpolation. The results also show that the sampling interval can be extended to about 60 min without the errors in the interpolated values exceeding 0.3 °C, which is the standard deviation of the difference signal between consecutive samples.

3.6. Time Series Analysis of Random Processes

The next experiment involves forecasting future values of temperature based on past samples. Liu et al. [62]described a system for saving sensor transmission energy when real-time estimates of temperature are needed. Samples are taken at regular intervals, and at each interval both the sender and the receiver calculate an estimated value based on the past time series. If the actual sensed value at the transmitter is within an error margin (say 0.5 °C) then no data is sent, and the receiver uses the forecast estimate. Once the error exceeds the error limit, then the actual current value plus any recent past values needed for future forecasting are sent. Liu et al. show a reduction in transmitted data of 70% with a corresponding reduction in energy use. However, their work uses indoor temperature readings with a very smooth behavior. We are interested if such a forecasting approach also works in a much more variable outdoor environment. Forecasting of future values uses an ARIMA process model and the subsequent sections explain the theoretical background behind such forecasting before such techniques are applied to our data.

3.6.1. Time Series and Stochastic Process

Due to the lack of complete knowledge of the complex underlying physical processes that generate local climate, environmental phenomena are in general modelled as stochastic processes [20]. A stochastic process varying in time is characterized by the sequence of a random variable. Any time sequenced realization of such a process is called a time series. Time

series analysis involves a range of investigations of the behavior of the observed stochastic process. Such analyses reveal structural behavior of the process that can be used to fit a suitable statistical model and understand short-term and long-term behaviour. Time series analysis is widely employed in areas such as signal processing, business processes and economic modelling, and there are many references which explain the concepts in detail [68-70].

Typically, in time series analysis, a process $Y(t)$ is assumed to consist of several sub-components: a trend, $\mu(t)$, a periodicity $P(t)$, seasonality, $P(t)$, and a random shock $e(t)$, as shown in Equation (1). The trend component represents a deterministic tendency such as long-term global warming; a periodicity represents regularly repeating behavior such as diurnal temperature variations; seasonality represents longer-term patterns such as summer and winter, and the random shock captures the effects of local short-term changes which are not explained by the longer term patterns:

$$Y(t) = \mu(t) + P(t) + S(t) + e(t) \tag{3-2}$$

If the properties of a process vary with time, then it is difficult to predict future values from its observed time series $Y(t)$ and such a process is called a non-stationary process. Most environmental phenomena fall in this category. In order to analyze a random process and perform state estimation, some sort of stationarity assumption needs to be made. In general, a second order stationarity assumption is made which assumes that the mean and the variance characteristics of the process do not change over time.

3.6.2. Time Series Model Development Strategy

Time series model development involves estimating a process characterizing components mentioned in Equation (3-2) with several sequential steps as shown in Figure 3-8. This generic time-series analysis framework is also known as Box-Jenkins time series modelling[68]. Structural analyses study the sample autocorrelation function and examine the stationarity property of the process.

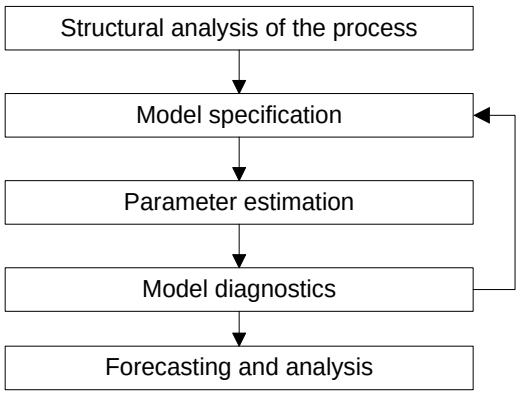


Figure 3-8. Time series model development strategy.

3.6.2.1. Model Specification

In general, the current state of any random process may depend on time, its past states, and some random shocks or a combination of these. Such dependencies of the observed series need to be extracted. Linear or nonlinear regression captures the trend component of the process. Dependencies with previous states can be captured by regression of the current state with the previous state and the effect of random shocks can be captured by involving noise components.

There are many different possible time series modelling approaches, but the most general of these is the Auto Regressive, Integrated, Moving Average (ARIMA) model. Stationarity of time series can be determined from the analysis of sample autocorrelation function and by conducting an Augmented Dickey-Fuller (ADF) unit root test[68]. If the time series is found to be non-stationary, transformation of the series can be performed that makes the series stationary. Logarithmic and power transformation and series differencing are the most commonly used transformation approaches. If the difference is taken to make the time series stationary, then the model is an Integrated model (i.e., ARIMA rather than ARMA). The order of the differencing is represented by a parameter d.

The ARIMA model specification involves finding suitable autoregressive (AR) and moving average (MA) sub-components of the Integrated model. The model represented in Equation (3-2) and can then be specified as in Equation (3-3):

$$Y_t = \mu + \emptyset_1 Y_{t-1} + \cdots + \emptyset_p Y_{t-p} + e_t - \theta_1 e_{t-1} - \cdots - \theta_q e_{t-q} \qquad (3\text{-}3)$$

Parameters specify deterministic (μ), autoregressive (φ), moving average (θ), and error *(e)* components. *p* and *q* represent the orders of AR and MA components which are determined by analyzing sample autocorrelation and extended autocorrelation function of the time series. Overall, the time series is then modelled by an ARIMA (*p*, *d*, *q*) model.

3.6.2.2. Parameter Estimation

After specifying differencing to achieve stationarity and specifying the AR and MA orders, the next step is the estimation of the parameters involved in Equation (3-3). For most random processes, parameters φ_i and θ_i are estimated using a Least Square (LSE) or Maximum Likelihood (ML) estimator. These parameters can then be used to estimate future values of the series

3.6.2.3. Model Diagnostics

Model specification deals with examining the goodness of the fit of the model parameters. Analysis of the residuals and over-parameterized models are two approaches used for validation. If residuals obtained after fitting a model fit a Gaussian noise distribution, then the model is considered to be valid. Over-parameterizing models involve internationally over-fitting the model with higher orders of p and q. If the over-fitted model doesn't show significant improvement in its residuals, the fitted model is considered to be valid.

3.6.2.4. Time Series Forecasting

After fitting a suitable model, the future state of the time series can be forecast. These future values can themselves be used to estimate further future values of the series. The forecasting power of the time series model is based on how many future sample values can be estimated with some desired accuracy.

3.7. Forecasting Experiments

As mentioned earlier, forecasting of future values can reduce the transmission energy for real-time temperature modelling. Analysis of the mine site temperature data (from Figure 6 above) is undertaken to estimate the forecasting accuracy of future samples.

The time series analysis in Section 3.6 uses standard methods to characterize the physical process. This section proposes a mechanism that uses the results of the time series analysis to identify the best sampling interval for a sensor deployment. We also observe what level of prediction improvement is gained by use of ARIMA models.

Environmental time series are usually non-stationary and require data cleaning to deal with missing data due to energy failures or other causes. The non-stationary nature is addressed by applying differencing and checking that the difference signal is stationary, as described in Section 3.7.1. The data used here has been manually checked the series here have been cleaned of any missing or repeated data (which was less than 1% of the data samples).

3.7.1. Structural Analysis of Time Series

Data analyses in this paper are primarily done in R [71], specifically using the package developed in [72]. Microsoft Excel and MATLAB are used for some data formatting and data plotting.

The chosen data series is the one-week sample series shown in Figure 6 above. Stationarity is checked by examining the one-week sample autocorrelation plot of the selected series, as shown in Figure 3-9. This autocorrelation plot has a clear structure which varies with the autocorrelation lag. Temperature patterns in one day are clearly correlated with the pattern the next day. This shows the clear presence of non-stationary (periodic) behavior in the series. After applying differencing, the time series in Figure 3-7 above was obtained. Figure 3-10 shows the autocorrelation of the differenced signal. Compared to the sample autocorrelation of Figure 9, the differenced series has an autocorrelation function which still has some regular structure, but the magnitude of the autocorrelation is less than 0.2 for all lags.

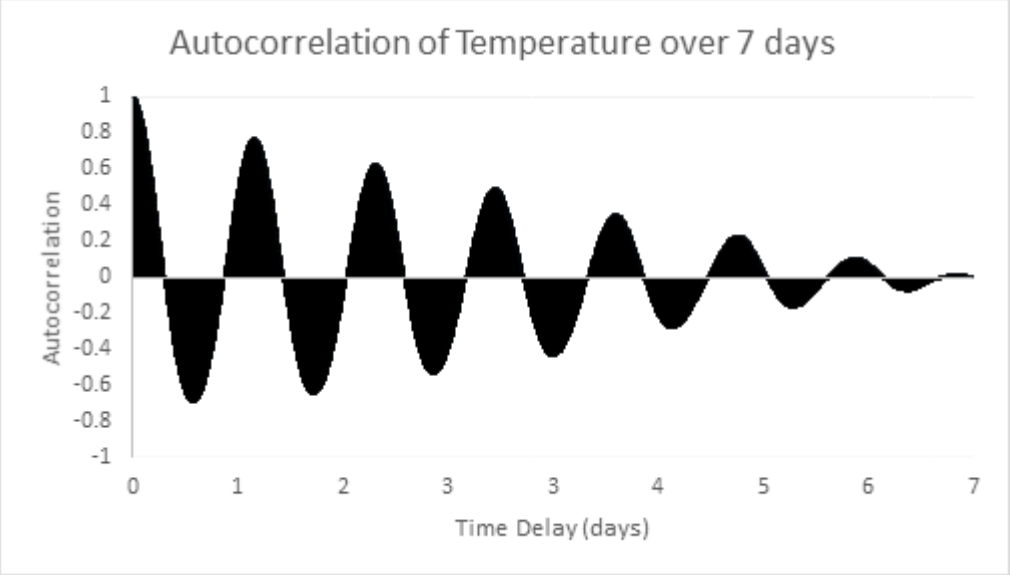


Figure 3-9. Sample autocorrelation of temperature in experimental data (5 min samples over 1 week).

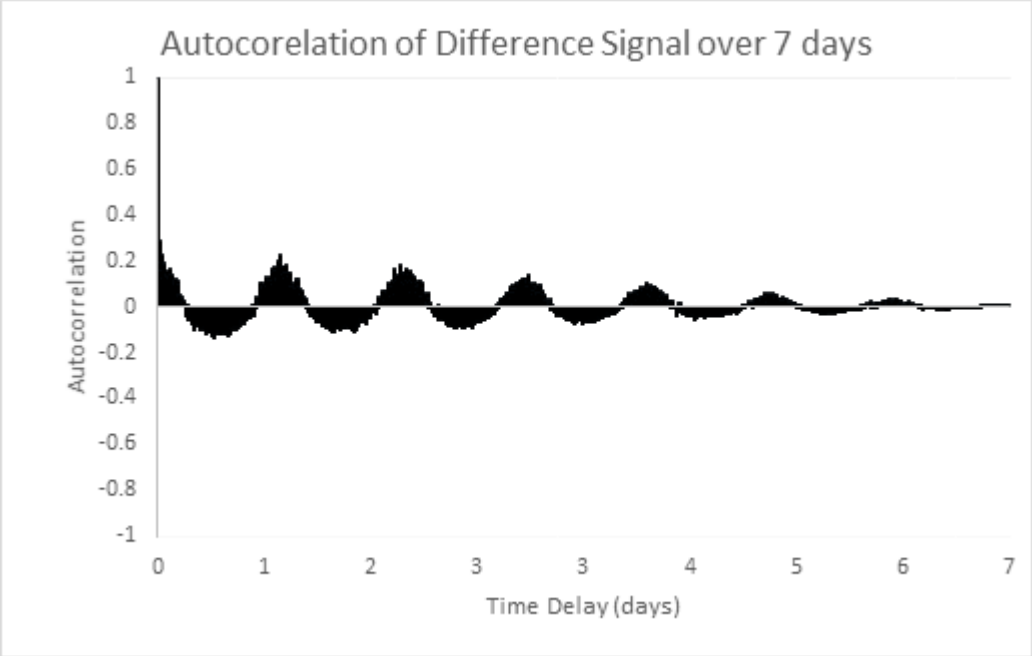


Figure 3-10. Autocorrelation of the differenced sample series over 7 days.

After applying one more round of differencing (a doubly differenced series) the autocorrelation in Figure 3-11 results which shows the double differences are uncorrelated. However, since single differencing gives the low autocorrelation values in Figure 3-10, the singly differenced signals will be used for further analysis.

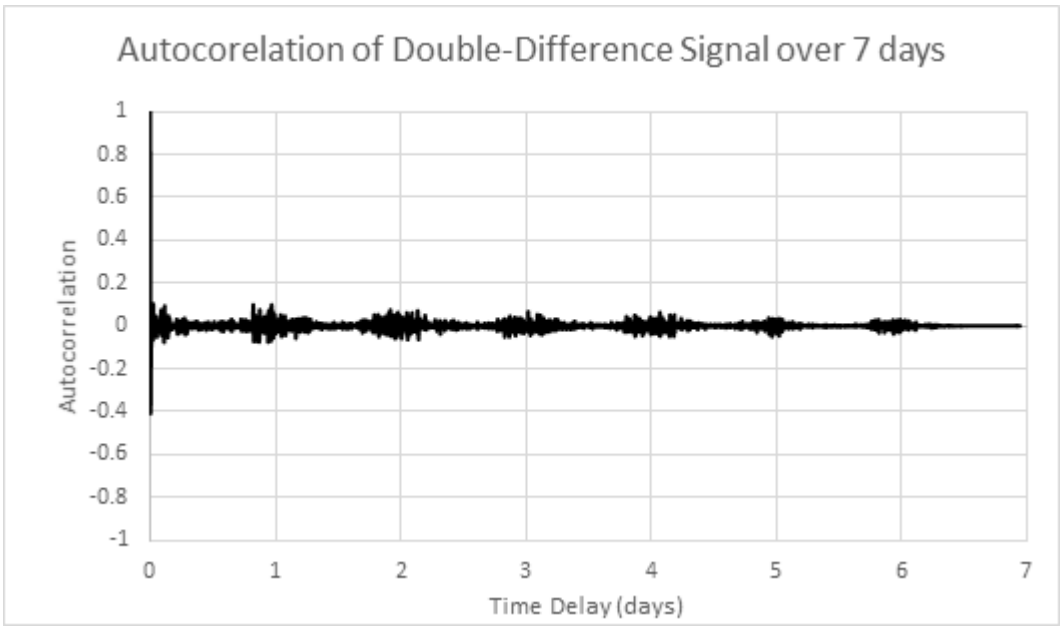


Figure 3-11. Autocorrelation of the doubly differenced sample series.

3.7.2. Model Order Selection

As the series becomes stationary after differencing, an ARIMA model will be used for the time series model. As the average of the differenced series varies about zero, the expected value of

the deterministic trend can be considered to be zero. The next step is to determine the orders of AR and MA components for the most suitable model. The Akaike Information Criteria (AIC) are widely used criteria which trade off the increased accuracy of higher order models with the parsimonious use of fewer model components [21]. Using the "auto.arima" routine from the *forecast* package in R which tests many different models, AR and MA orders of the series are estimated for different sampling rates. Estimation of AR and MA orders for different sampling rates help us to examine how the time series model varies with different sampling rates of the deployed sensors. Table 3-3 shows the models with the best AIC score based on the first three days of data as shown in Figure 3-8 above, for different sampling rates (i.e., for subsampled subsets of the original data). These different sampling rates capture different realizations of the process and specify different orders for the ARIMA models, however, there is not any clear interpretation of how the ARIMA model order varies with the sampling rate, other than the fact that for this data set, 60 min sampling gives the simplest model.

Table 3-3. AR and MA orders for different sampling rates.

Sampling Rate (Minutes)	Fitted Models
5	ARIMA(3,1,1)
10	ARIMA(2,1,2)
15	ARIMA(1,1,3)
20	ARIMA(1,1,3)
30	ARIMA(2,1,1)
60	ARIMA(1,1,0)
120	ARIMA(3,1,1)

Experiments on other data (such as the data shown in Figure 3-2, or on different subsets of the week in Figure 3-8) show that the best ARIMA model order is not very consistent between different deployments or different periods and would need to be revised regularly when used for prediction. Rechecking and updating the best predictive model order once a week for each different sensor (rather than using a single model order for all deployments) would allow seasonal changes in model order to be tracked.

3.7.3. Forecasting

To test the forecasting ability of the time series models, the ARIMA models are used to forecast the remaining four days of data shown in Figure 3-6. In particular, the following procedure is used. For each sampling rate, the ARIMA model of the order shown in Table 3-3 is trained on three days of data, and then used to predict up to two hours forward from that point, e.g., for 5 min sampling, 24 future points are estimated, for 30 min four future points are estimated, and for 120 min, one future point is estimated. Then the 3 day training window is moved forward

by 2 h, the models retrained, and the process repeated for the remainder of the four "testing" days of the sample. For sampling rates greater than 5 min, the future predictions at 5 min intervals are linearly interpolated between the future prediction points. For example, for 30 min sampling, the future prediction at 5 min is linearly interpolated between the last data point and the first predicted point.

Additionally, two other prediction models are used based on the 5 min sampled data. The "zero difference" model uses the last data point in the undifferenced series as the predictor for the next two hours. This is the same as using the mean (zero) of the differenced series as the predictor of the next difference. The "same difference" model linearly extrapolates from the last two data points in the undifferenced series, which is the same as assuming that the next difference value is the same as the current difference value.

The accuracy of the future predictions are measured by the RMSE of the predictions across the four days, and also the MAE of the predictions. Table 3-4 shows the results for RMSE and Table 3-5 shows the results for MAE. Figure 3-12 shows a plot of the RMSE for the different predictors versus the forecast time, where, for example, "ARIMA5" means the ARIMA model with 5 min sampling interval.

Table 3-4. RMSE of Future Temperature Predictions in °C.

Forecast	Simple Models		ARIMA Models Sampling Intervals (Minutes)						
Time (Mins)	Zero Diff	Same Diff	5	10	15	20	30	60	120
5	0.33	0.49	0.33	0.17	0.13	0.13	0.11	0.12	0.12
10	0.48	0.78	0.45	0.45	0.38	0.39	0.35	0.34	0.34
15	0.59	1.07	0.51	0.51	0.51	0.51	0.45	0.44	0.46
20	0.62	1.36	0.53	0.54	0.54	0.63	0.54	0.54	0.57
30	0.91	2.02	0.75	0.76	0.77	0.90	0.86	0.84	0.85
60	1.56	4.05	1.07	1.07	1.06	1.29	1.17	1.39	1.33
120	3.32	8.47	2.50	2.52	2.52	2.71	2.58	2.80	2.48

Table 3- 5. MAE of Future Temperature Predictions in °C.

Forecast	Simple Models		ARIMA Models Sampling Intervals (Minutes)						
Time (Mins)	Zero Diff	Same Diff	5	10	15	20	30	60	120
5	0.24	0.27	0.21	0.11	0.08	0.08	0.07	0.08	0.09
10	0.35	0.49	0.32	0.32	0.26	0.26	0.23	0.21	0.22
15	0.45	0.65	0.38	0.38	0.38	0.36	0.31	0.29	0.31
20	0.52	0.87	0.42	0.42	0.42	0.46	0.40	0.37	0.43
30	0.73	1.31	0.58	0.58	0.59	0.63	0.63	0.55	0.62
60	1.27	2.60	0.82	0.82	0.81	0.85	0.82	0.90	0.96
120	2.71	5.71	1.91	1.94	1.98	2.03	1.97	2.03	1.74

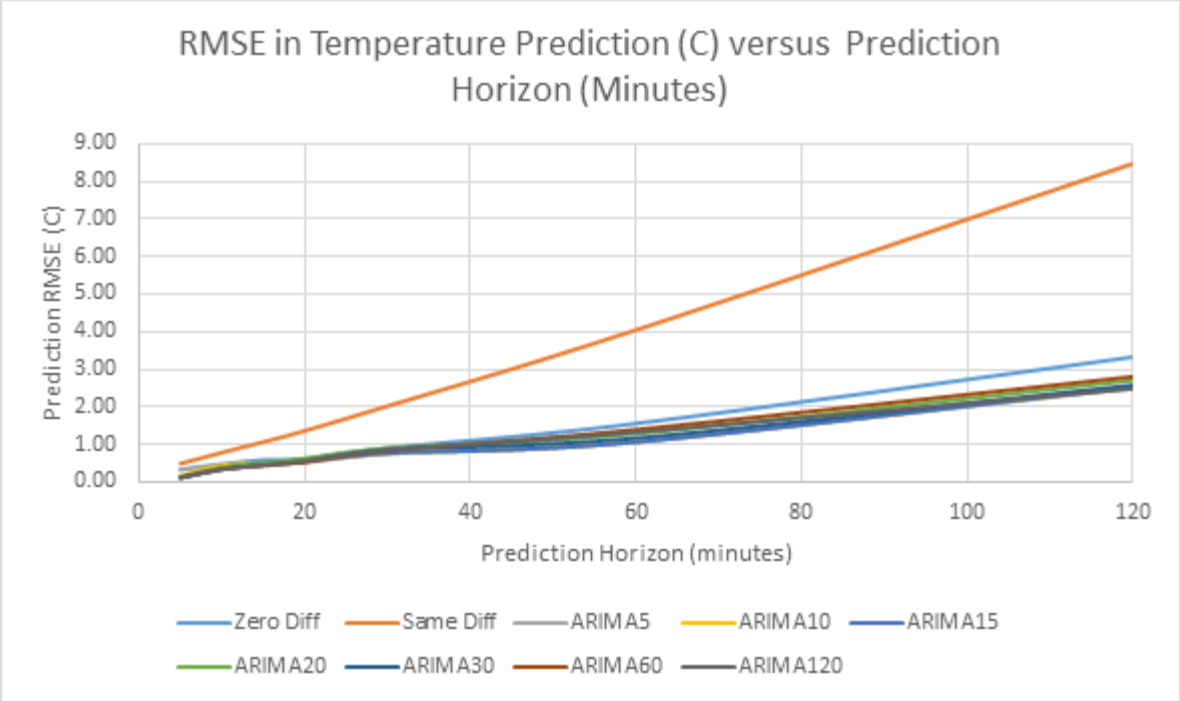


Figure 3-12. RMSE versus Prediction Horizon for Different Predictors.

Because the different predictors are difficult to distinguish in Figure 12, Figure 13 shows an expanded close up of the prediction up to 60 min, with the poorly performing linear extrapolation (Same Difference) excluded. Figure 13 also shows the 95% confidence interval for the ARIMA60 results, showing that the differences between predictors are small compared to the confidence interval.

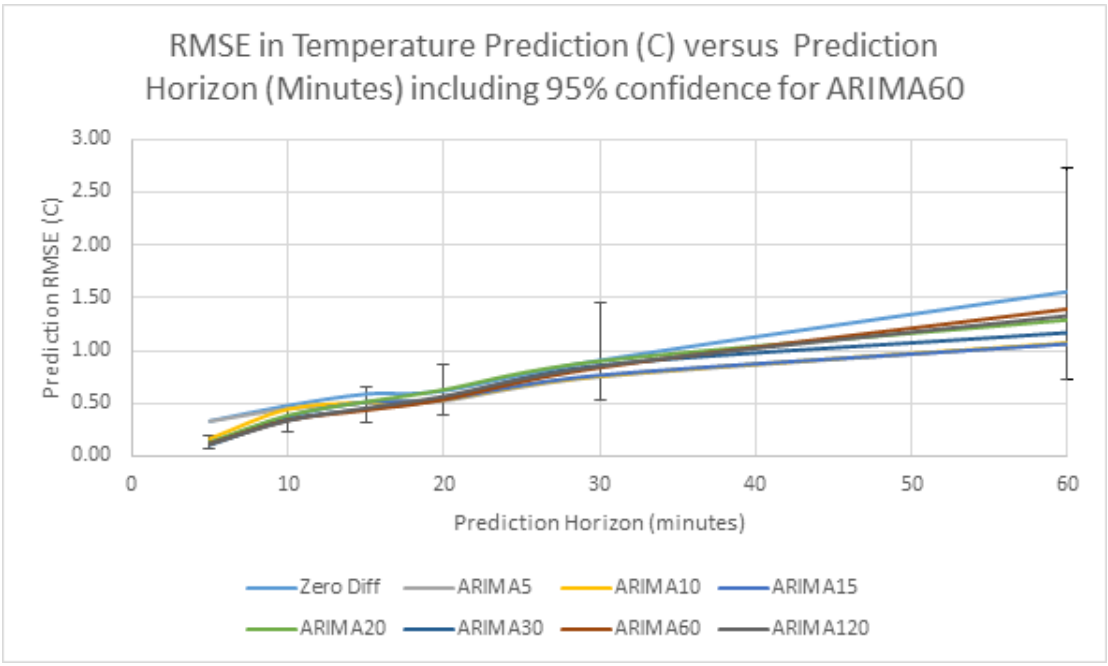


Figure 3-13. Detail of RMSE versus Prediction Horizon for Different Predictors with 95% confidence interval for ARIMA60.

As can be seen from this data, the RMSE in forecasting increases as we forecast further in the future and it exceeds 1°C after about 60 min. This behavior can be explained by the sample autocorrelation function in Figure 3-8. The correlation between samples decreases steadily as the lag increases, and so, as predicted, the prediction error steadily increases. Another interesting observation from Figure 3-13 is that the forecasting error does not change significantly with sampling interval. The “Same Difference” or linear extrapolation method performs very poorly, and the “Zero Difference” method also performs worse than any of the ARIMA models. In this particular example, the ARIMA model prediction with 30 min sampling has the lowest error. The differences between the ARIMA models with different sampling intervals is small, and it is expected that the differences are artifacts of the particular data series. However, a clear message is that prediction accuracy does not depend on high-frequency data sampling.

Chapter 4 incorporates the following paper,
with the conclusions section as part of Chapter 5:

S. Bhandari, N. Bergmann, R. Jurdak, and B. Kusy, "Time Series Analysis for Spatial Node Selection in Environment Monitoring Sensor Networks," *Sensors*, vol. 18, no. 1, pp. 11, 2017.

Contributor	Statement of contribution
Siddhartha Bhandari (Candidate)	Conception and design (85%) Analysis and interpretation (85%) Drafting and production (80%)
Neil Bergmann	Conception and design (5%) Analysis and interpretation (5 %) Drafting and production (10%)
Raja Jurdak	Conception and design (5%) Analysis and interpretation (5 %) Drafting and production (5%)
Brano Kusy	Conception and design (5%) Analysis and interpretation (5 %) Drafting and production (5%)

Chapter 4 Spatial Interpolation

4.1. Introduction

Environmental phenomena such as temperature, pressure, humidity, and soil moisture are dynamic processes. Understanding the spatio-temporal behaviour of these processes is relevant for understanding the surrounding ecosystem's state. Environmental phenomena, in general, vary at a small spatio-temporal scale [2, 3] that impact the local ecosystem. The microclimate (temperature, solar radiation and other phenomena at small scale) affects ecological changes in forests[73], soil characteristics in mine rehabilitation [67], and diseases in agriculture[74]. Thus it is crucial for many application scenarios to monitor environmental phenomena at high spatio-temporal resolution.

Understanding the spatio-temporal behaviour of the environmental phenomena requires the development of an effective monitoring system. In past decades, weather stations have been the widely used for monitoring. However, weather stations are spatially sparse, and they only capture coarse-grained environmental variations, which are not sufficient for monitoring variations in small-scale ecological processes.

Recently, wireless sensor networks have been widely used in small-scale environmental monitoring as they can be economically deployed for fine-grained environmental sensing and monitoring. Example applications include city centre heat monitoring [3], air quality monitoring [5], building environment monitoring [6], soil moisture measurement [8], volcano monitoring [9], ocean exploration [10], and harsh mountain environment monitoring [11]. In most of these sensor network deployments, the number and positions of sensor nodes are selected based on intuition, domain knowledge, or cost constraints. There is currently a lack of an objective method for determining the best number of nodes and their spatial distribution. The challenge is that the optimal node number and locations are dependent on the specific spatiotemporal processes in the monitored environment. The dynamics of these processes are not known a priori, which is, in fact, the motivation for monitoring the environment. Two of the sensor networks deployed by our research lab for rainforest monitoring [1, 13] and mine rehabilitation monitoring [67] are clear examples where the number of nodes that were deployed was not based on any evidence-based understanding of the number that would be

needed. The question of the optimal number and placement of sensor nodes needed for adequate environmental monitoring remains a challenge, and that is the topic that this paper addresses.

In a real application scenario, it is important to know the optimal number of sensor nodes to be deployed and the best position to achieve the project's scientific or business objectives. A large number of sensors incurs high deployment and operational costs. On the other hand, fewer sensors may fail to capture sufficient local details. The design goal should be to achieve the scientific objectives at the most economical cost.

Strategies for determining the target number of deployment nodes vary from analytical to simulation-based approaches. Some of the strategies are theoretically-based where environmental phenomena are modelled as spatio-temporally correlated processes and suitable sampling strategies are developed, such as in [75] where Gaussian process modelling is used. In [76], Monte-Carlo simulation has been used to find the locations of nodes in space that produces the lowest spatial variability. In [42], a geometrical approach is used treating sensor deployment as an area coverage problem. Our approach balances theory with initial experimental evaluation of the sensor deployment area to ensure that the coverage is adequate for the specific deployment scenario.

This work considers a practical application scenario, using the example of a mine rehabilitation monitoring program over an area of several square kilometres [4]. The objective is to monitor small-scale spatio-temporal variations using empirical data from a short-term, high-density deployment to optimize the deployment of a number of long-term sensor nodes. First, a larger number of static sensor nodes are deployed across the sensor area. The observations at each sensor location form a time series while observations at different locations form multiple time series. A time series analysis framework is then applied on each individual series as well as at the multiple series. Co-integration analysis is then used to determine the relationships between series. Co-integration provides information on which time series are most similar to each other. Similar time series are used to determine one location that can be used as an estimate for its co-integrated locations. Redundant sensors can be re-used elsewhere, or alternatively, initial deployments can be with a large number of low-cost, short lifetime sensors that are replaced by fewer yet more robust long-term sensors. Implementing our proposed co-integrated multiple time series analyses for temperature measurement in the mine rehabilitation scenario showed that 75% of the existing sensors are found to be co-integrated with the other 25%. In other words, similar temperature monitoring accuracy could be achieved with only 25% of the

existing deployment. The proposed approach is general enough that it can be utilized in any spatio-temporal monitoring application.

The rest of the paper is organized as follows: Section 2 reviews previous work. Background information on the techniques used is described in Section 3. The analytical approach that is used and the algorithms developed for the approach are discussed in Section 4. Section 5 presents analytical results from the particular mine rehabilitation sensor network. Section 6 concludes the paper.

4.2. Previous Work

In [73], authors have described the association between ecological processes and microclimate (temperature, solar radiation and other phenomena at small scale). Temperature variation up to 8 °C within a small forest patch was reported and linked to ecological changes. The effect of small-scale climatological condition on the development of a fungal disease on a potato crop and forest canopy was observed in [77]. Variation of temperature within a small urban area has been reported in [3] while the microclimate effects on soil characteristics in mine rehabilitation were reported in our previous work [67]. In all scenarios, variations in the environmental phenomena at small scale are observed and linked to environmental changes, motivating the need for accurate understanding of local microclimate conditions in many scenarios.

Environmental monitoring has a long history. As described in [19], The Australian Bureau of Meteorology has been monitoring climatic variables including temperature, pressure, sun radiation, and rainfall since 1957. However, only 4600 monitoring stations are installed to cover the whole 7.7 million square kilometres of Australia since the capital and operating costs of weather stations are very high [19]. Such a coarse-grained spatio-temporal environmental monitoring would not suffice for the small-scale environmental impact analyses needed in mine rehabilitation [67] or rain forest monitoring [1] scenarios.

Significant research has been undertaken in the design of monitoring networks in sensor network applications. In general these works can be divided into three groups: mathematical, geometrical and simulation approaches. A selection is reviewed here.

Environmental phenomena are modelled mathematically as a spatio-temporal random field where the monitoring network design problem becomes the problem of sampling the assumed random field. In [75], the phenomenon is modelled as a Gaussian process and sampling strategies are designed. In [75], the authors also deployed sensor nodes for some time to learn the parameters of the Gaussian process.

Another approach to designing a sampling strategy has been the geometry-based approach. Within a spatial region, various geometrical approaches are used to select the positions of the sensors. Voronoi tessellation, Delaunay triangulation, and cell declustering are some of the examples of these geometric arrangements [75]. In [42], Voronoi tessellation is used to optimize the node positions. The main issue with such approaches is the strong assumption regarding the nature of the process. Environmental phenomena will not have convenient geometrical regions of similarity. The limitation of such an approach in monitoring temperature is shown empirically in [75] where temperature variations among equidistant points are different.

Other work by Chen et al. [78] also addresses geographic sensor node selection, although in their case they select a subset of nodes from a heterogeneous collection of web-connected sensors for a particular application using a web-services approach. In their case, geographical sensor selection is based on proximity and they do not provide a method for interpolating between sensor positions, which is the focus of this work. Wang et al. [79] have described a wide area technique for selecting the site of ground precipitation sensors to complement satellite observations. Their work is based on maximizing the geographical coverage of sensors, sensitive to local terrain conditions. Such techniques could be useful for determining the initial dense deployment of sensors and is complementary to our work which then identifies the best subset of those sensor locations.

In the simulation approach, sensors are placed at selected points and simulated sample measurements are drawn from the expected sensor responses to check the quality of the measurement. In [76], Monte Carlo simulation is used to choose sensor locations. However, this requires the spatio-temporal variability of the data to be estimated before any measurements are made.

Several studies have conducted time-series analysis in sensor networks [59-62]. Some works are based on simulation while others are based on real observed series. One common objective of all the studies has been to identify the nature of the time series from each sensor node and somehow use the knowledge to reduce communication among sensor nodes which is important in energy saving in resource constrained nodes. For example, in [62] sensor data is only transmitted when it cannot be accurately forecast by a time series model of past data. Most works are based on univariate analysis of measurements at one point. Our work considers the correlation of time series across space basing the analysis on multivariate or multiple time series. The main focus of our work is to explore co-integrated time series and exploit their

behaviour to optimize the number of sensors needed to monitor the desired environmental phenomena at the required accuracy.

4.3. Background Information

This section briefly describes some background information required for this research. It includes information on time-series analysis and a technical specification of the environmental sensor network involved in this paper. Mathematical details are kept to a minimum, and readers are referred to [68] for further information.

4.3.1 Theory of Time Series Analysis

Time series analysis is a framework for analysing sequentially observed data in time. It involves analysing the temporal correlation of the observation which can be used for identification of the process model that generates the data. Identification of the model helps in generalizing the nature of the underlying process and estimating past and future values based on available observations. Environmental phenomena that are observed sequentially at regular sampling intervals are best suited for this analysis. Environmental phenomena which form time series include temperature (T), solar radiation (S), soil moisture (M), and rainfall (R). Each variable has an observation at each sampling instant (t). The series of sampling intervals can be numbered (t_0, t_1, ..., t_n). The value of one variable at successive sampling instants forms a time series, e.g., (T_0, T_1, ..., T_n).

4.3.1.1. Univariate and Multivariate Time Series

Univariate time series analysis is concerned with the study of a single time series. A series of temperature readings (T_i) measured at one sensor node is an example of a univariate time series. Most of the environmental phenomena are measured in many locations generating multivariate time series which are correlated among themselves. Multivariate time series analysis is the process of analysing more than one-time series at a time. Time series such as temperature (T_0, T_1, ..., T_n), solar radiation (S_0, S_1, ..., S_n), and soil moisture (SM_0, SM_1, ..., SM_n) have relationships between them that can be analysed under multivariate time series analysis. Similarly, measurements of the same variable at different locations, e.g., temperature from different sensors, can be analysed using multivariate analysis.

4.3.1.2. Stationary and Non-Stationary Time Series

A time series is called a stationary if it exhibits a consistent temporal statistical pattern. Such time series are amenable to time series analysis. If the moments of the time series such as mean and variance do not change with time, the series is called stationary to the mean and the variance. (M_0, M_1, ..., M_n) is called stationary of order (1, 2, 3, ..., n) if moments (m_1, m_2, m_3, ..., m_n) remain constant over time. For many applications, a time series is examined for second order stationarity. Second order stationarity is based on the assumption that the underlying phenomena are a Gaussian stochastic process for which first and second order moments (mean and variance) are sufficient to characterize it. A second-order stationary time series whose covariance is such that Cov(Xt_1, Xt_2) can be generalized by Cov(τ) where $\tau = (t_1 - t_2)$ is called weakly stationary. Any time series that doesn't show regularity about its moments is called a non-stationary time series, and simple time-series analysis techniques cannot be used. Temperature (T_0, T_1, ..., T_n) measured at a particular location is a good example of a non-stationary time series. Expected value, correlation, and variance all vary with time. Non-stationarity can occur due to seasonal variation, unknown noise involved or due to the nature of the underlying phenomena.

4.3.1.3. Co-Integrated Time Series

Time series are called co-integrated if they show some similarity amongst themselves. If two-time series are co-integrated, even if they are non-stationary, one can be estimated using the other. Many studies on co-integrated non-stationary time series have been conducted in the field of econometrics where various quantitative and qualitative economical series are analyzed [80, 81]. Linear modelling can be performed among co-integrated series and ordinary least square estimation becomes the best unbiased estimation. Such estimation is mathematically tractable and statistically efficient. Most environmental phenomena are non-stationary in nature, so that linear estimation cannot be performed without the assumption of stationarity or some transformation. Assumptions may lead to invalid conclusions while some transformations render the data difficult to interpret in the transformed scale. If multiple time series exhibit co-integrated characteristics, no assumptions and transformation are needed. Co-integration analysis that has been proposed in econometrics for economic time series modelling is adapted for environmental time series in this work. As co-integration analyses search for similarly behaving series, this can help to determine environmental series which are redundant, and so the sensors generating those redundant time series are not needed.

4.3.1.4. Augmented Dicky-Fuller Test

Before conducting any inferential analysis, the co-integrated nature of the time series needs to be validated. Researchers in [80, 81] provided a framework to validate whether time series are co-integrated. The Augmented Dicky-Fuller (ADF) test is a statistical procedure that tests the stationarity hypothesis of a univariate time series. Given a time series, the ADF test fits varying degrees of autoregressive (AR) models and provides statistics needed for acceptance or the rejection of an initial non-stationarity hypothesis. Equation (1) shows an AR(1) process:

$$y_i = c + \rho y_{i-1} + \varepsilon \tag{4-1}$$

where ε is a Gaussian white noise process with zero mean, and c is a drift constant

The process is non-stationary if $|\rho| \geq 1$ and the process is stationary if $|\rho| < 1$. In the ADF test, non-stationarity is tested for higher degrees of order p using Equation (2) i.e., to check if the time series fits an AR(p) model:

$$\Delta y_i = \rho y_{i-1} + \sum_{j=1}^{p-1} b_j \Delta y_{i-j} + \varepsilon \tag{4-2}$$

where the difference operator Δ is $\Delta y_{i-j} = y_{i-j} - y_{i-j-1}$

The ADF test is available in the libraries of statistical computing platforms like R [82]. The Dickey-Fuller Test Statistic is a statistical measure that is used to confirm that the nodes are co-integrated. It should be less than a critical value determined by the number of observations, and the confidence of decision. The needed critical threshold value and related statistics for various orders of the process and the number of observations are tabulated in [80]. Table 4-1 below, shows the values for different numbers of observations and different confidence levels for an order 1 process. For a confidence level of 99% and more than 100 observations, it is common practice to choose a critical value of the ADF test statistic of −3.5.

Table 4-1. Critical Values for Dickey-Fuller Test Statistic.

Sample Size	99% Confidence Level	95% Confidence Level
50	**−3.58**	**−2.93**
100	**−3.51**	**−2.89**
500	**−3.44**	**−2.87**
Infinity	**−3.43**	**−2.86**

4.3.2. Mine Rehabilitation Monitoring Sensor Network

This study uses environmental sensor network data obtained from the Meandu open cut coal mine situated in a remote location of Queensland, Australia [67]. The industrial site of the mine is fairly large and spread across several sections of the mine site. The mine was established in the 1980s. Mining activity involves removing overburden, then removing the coal, and then replacing the overburden. After the mining is completed in one section, the rehabilitation phase commences. Rehabilitation involves restoring the previous environment, i.e., regenerating soil and re-establishing plants (grass, shrubs, trees) back to the condition of the natural environment. Sensor networks are deployed in rehabilitation sites, as shown in Figure 4-1, to monitor microclimate in order to assist with the timing of operations such as planting, and watering. Air temperature, soil temperature at two levels of depth, solar radiation, soil moisture, rainfall are measured in each rehabilitation site. The coloured outlines on the map show areas where rehabilitation has begun in different years from before 2000 up to 2010. The numbered boxes show the locations of sensor nodes.

Figure 4-1. Meandu mine rehabilitation site and sensor deployment.

The sensor network designed by CSIRO has been deployed in several rehabilitation sections. In the current deployment, there are four sections, 12 sites and 24 transects in which 30 sensor platforms are deployed. For ground truth validation, several sophisticated weather stations are also deployed. Locations of the sensor nodes are selected based on the requirement of the rehabilitation monitoring. A custom sensor network platform using a 900 MHz IEEE 802.15.4 compatible radio was designed. A collection tree-based data collection protocol is used to for

data communication from sensor to the gateway. The gateway station then forwards data to a centralized server using 3G connectivity. The server provides access to the data and further analysis. Technical details of the platform are given in [67].

4.3.3. Limitations and Assumptions

This paper represents a first exploration of using the time-series analysis method of co-integration for improved placement of sensors in an environmental sensing scenario. There are many assumptions and restrictions to the applicability of this model, as follows.

Firstly, the method is only applicable to sensing parameter fields that are spatially correlated, i.e., where values at locations that are close spatially tend to have similar values. Environmental parameters such as air temperature, humidity, wind speed and barometric pressure would be examples of such parameters. There are many parameters, especially in the built environment, which would not be amenable to such analysis, such as smart power meters in one street, or traffic density in nearby streets. Part of the analysis in the next section is to identify if time series data are suitable for this approach.

Another assumption is that spatial correlations between sensor readings persist over the long term. An initial exploration of the estimation error over a whole year based on one week of training data is presented in Section 4.5.4.

In some situations, dense sensor deployments may be intended to detect data anomalies, for example, a sudden increase in temperature due to an approaching forest fire. Again, since the approach here uses a few sensors to interpolate parameters at other locations, it will be less sensitive to local anomalies, and would not be suitable for such applications.

This initial investigation uses temperature as the example environmental variable since it is easy to measure and changes relatively slowly. Our future work plans to extend this work to other sensors.

4.4. Proposed Analytical Methodology and Algorithms

4.4.1. Data Analytic Framework

This section describes the analytical framework used for the analysis of the multivariate time series. Figure 4-2 shows the different steps involved in the analytical process.

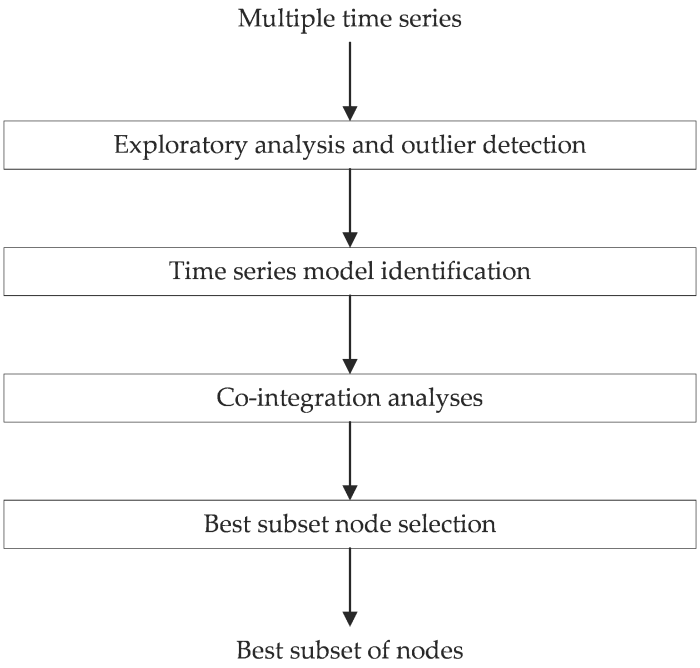


Figure 4-2. Multivariate time series analysis framework.

First, exploratory analysis of time series data looks for any significant inconsistencies. Spatially proximate sensors are plotted together for this. Outlier detection is performed including univariate and multivariate features. The detailed approach to performing outlier detection analysis is available in our previous work [67]. The next step is to identify the time series model. Stationary behavior of the series is analysed using an Augmented Dicky-Fuller test for each sensor. As expected, none of the periodic temperature time series are stationary. Co-integration analysis is then performed for all possible pairs of sensors. The result of the co-integration analysis is the confirmation or failure of the co-integration test of the pairs of the available sensors. After co-integration analysis, the Best Subset Node Selection step is performed that searches for the best possible subset of the sensor nodes that can estimate each of the time series.

4.4.2. Co-Integrated Series Selection Algorithm

Firstly, a decision must be made about which set of nodes are sufficiently close in location to be considered as possible co-integrated nodes. This means identifying a local neighbourhood of nodes. For example, in the experiments we describe here, 12 nodes in the north-east corner of the mine site (numbered 201 to 212 in Figure 4-1 above) are selected. They are within 1 km of each other. It would be less likely that nodes in the south-west corner of the mine would be as closely correlated. Within this neighbourhood, all possible pairs of nodes are examined.

The co-integrated series selection algorithm searches for the best co-integrated node for each sensor node. This algorithm starts fitting a linear model on one node with all the other nodes. After fitting the model each residual series is then evaluated for stationarity using the Augmented Dicky Fuller test. At the end of the run, the algorithm generates the best co-integrated node for each
sensor node.

In the case where the most correlated node has a Dickey-Fuller test statistic which is above the critical value of −3.5, then it cannot be estimated accurately from other nodes, and that node would be one of the critical locations for a permanent sensor node.

Algorithm 1: Co-integrated time series selection.

```
1: TS ← sensor series
2: for each time series i do
3:                  # fit a linear model with each other node j
4:                  lm[i][j] ← linear model TS(i, j)
5:                  resd[i][j] ← residual(lm[i][j])
6: end for
7: for each residual i,j do
8:                  # run Dicky − Fuller test
9:                  DF [i][j] ← ADFtest(resd(i, j))
10: end for
11: for each time series i do
12:                 ts ← maximum(abs(DF(i, j)))
13:                 Cointegrated[i] ← ts
14: end for
```

4.4.3. Best Subset Sensor Nodes Selection Algorithm

After validating that the observed time series are co-integrated, a best subset nodes selection algorithm searches for the best subset of nodes that can be used to estimate the value at each unobserved location. At each location, the proposed algorithm starts searching for the best linear combination of observations at other locations that can reproduce the observed value. It is possible to set the maximum number of nodes to be searched from 1 to N, where N is the total number of available nodes. If the maximum node to be selected is set to 1, the algorithm selects a single best node for the estimation. The searching involves all available series. A linear combination of temperature at a particular location is calculated based on Equation (4-3):

$$Y = \beta X + \varepsilon \tag{4-3}$$

where $\beta = (\beta_0, \beta_1, ..., \beta_N)$ are corresponding linear weights and X is the matrix of variables with each column representing a single series.

The least square cost function to minimize is given by $(Y - \beta X)^T(Y - \beta X)$ which when differentiated with $(\beta_0, \beta_1, ..., \beta_N)$ provides the least squares unbiased estimation of the parameters as given by Equation (4-4):

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{4-4}$$

In each iteration, the algorithm selects one more co-integrated series that has not been previously selected. The selection is based on the node whose addition to the subset most reduces the estimation error. After parameter estimation, the estimated value of this series based on the linear combination of other series can then be calculated for a test set (different from that used to select parameters) using parameters from Equation (4-4).

In each iteration, the algorithm produces the training error for each series. Observing training errors, a suitable number of nodes can be selected which can generate all the series. This suitable number may be determined by operational requirements, e.g., one might have only 4 permanent sensing stations for deployment, and wish to choose the best four locations. Alternatively, this number could be chosen by scientific requirements, such as needing a maximum of 0.5 °C RMSE error at all the estimated positions. Finally, the number could be chosen on a statistical basis, such as identifying when adding an additional node does not significantly reduce the RMSE of estimated readings (using something like the heuristic "elbow" criterion in a graph of RMSE versus the number of nodes). Pseudocode of the algorithm that selects the best subsets is given in Algorithm 2.

Algorithm 2: Best subset selection of M co-integrated nodes from $N - 1$ candidates for each of N nodes.

```
1: # Search for the best subset of M sensors for each individual sensor, i
2: M ← number of sensors in the subset
3: for each sensor i do
4:                  searchspace ← set of all sensors minus sensor i
5:                  bestsubset[i] ← NULL
6:                  for j = 1 to M do  #add one more sensor to best subset for i
7:.                             lowest estimation error ← infinity
8:                              for each sensor k in searchspace
9:                                      fit linear model to sensor i using (k + bestsubset[i])
10:                                     if estimation error from linear model < lowest estimation error
11:                                             lowest estimation error ← estimation error from linear model
12:                                             bestsensor ← k;
13:                                     end if
14:                             end for
15:                             searchspace ← searchspace − bestsensor
16:                             bestsubset[i] ← bestsubset[i] + bestsensor
17:                 end for
18: end for
```

It is useful to estimate the computational complexity of Algorithm 1 and Algorithm 2. Both algorithms basically have the same structure, which is for every pair of nodes, find a least squares estimator for one node from the other, and then calculate the goodness of fit, either by calculating the Dickey-Fuller statistic or the estimation error. The parameters which affect which affect computational complexity are N, the number of nodes, M the size of the best subset, C = 2M, the number of parameters that have to be estimated in the linear model, and S, the number of samples.

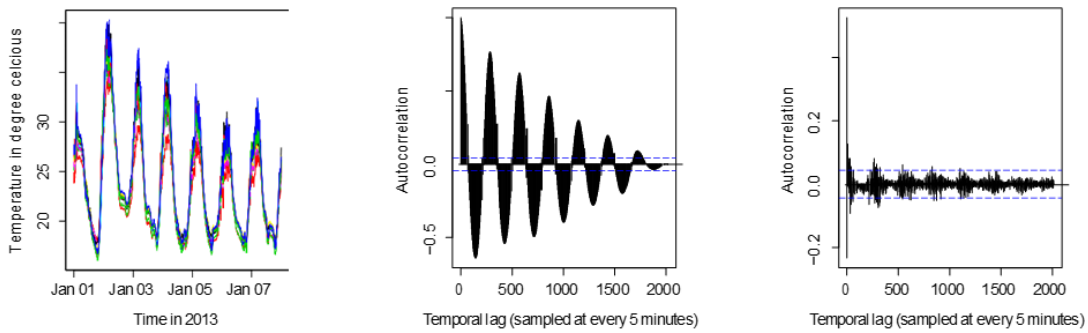
Equation (4) is the basis of fitting a linear model, and in terms of time complexity it consists of a matrix multiplication X^TX which is $O(C^2S)$, a matrix multiplication X^TY which is O(CS) a matrix inverse which is order (C^3), and a final matrix multiply which is $O(C^2)$. The calculation of the error metric or statistic consists of estimating S values from C parameters, O(CS). For the case where M = 1 (using just one estimator node), and therefore C = 2 is a constant, the order of one linear fit is O(S). If this is repeated for every pair of nodes, the total complexity is $O(N^2S)$. The N^2 term suggests that it may be infeasible to apply this method directly to thousands of nodes, instead these nodes should be divided into disjoint neighbourhoods of less than 100 nodes. For M > 1 (i.e., larger subsets of estimators), the complexity grows to $O(N^2M^2S)$, and so for these experiments, we just use M = 1 to reduce the computation time.

4.5. Analysis of Results

This section provides results obtained from implementing the proposed algorithms on the 12 sensors in a 1 km × 1 km area in the north-east of the Meandu mine site, as shown in Figure 4-1. The average distance between neighbouring nodes is about 100 m. Three weeks of temperature time series starting from 1 January 2013 are used for the analyses. The first week of data is used to select three "permanent" nodes from the 12, and to train models to estimate the other nine. Then the temperature is estimated at the nine positions from the three "permanent" nodes for 10 days, and the estimated temperature compared to the actual temperature at those nine positions. Temperature is selected as a representative time series as it has been analysed in other works [2, 3, 75], and is known to be amenable to time series analysis. We hope to investigate other parameters in future work.

4.5.1. Univariate Analysis

Figure 4-3a shows the multiple time series plot of 12 nearby sensors superimposed. It helps to evaluate obvious inconsistencies among the series which is not present in this case. Figure 4-3b shows the temporal autocorrelation of temperature from one of the sensors. From the nature of the correlation, it is obvious that the series is non-stationary. Any series that possesses periodicity in their correlation are non-stationary. The Augmented Dicky-Fuller test is run for each time series to verify that its non-stationarity is of order 1. Also, the time series model identification utility available in R is used for model identification. Figure 4-3c shows that after first order differencing, the autocorrelation is reduced to small values for all lags, and so this differenced sequence is stationary and amenable to analysis.



(a) (b) (c)

Figure 4-3. (**a**) Multiple time series plot for 12 nearby sensors; (**b**) Sample autocorrelation for a univariate temperature series; (**c**) Sample autocorrelation for differenced time series. Horizontal dashed lines indicate the +/−5% bounds normally used to identify stationarity in the ACF.

4.5.2. Co-Integration Analysis

After confirming that all series are first order non-stationary, co-integrated analysis is then performed for each node. The nodes are given ID's ranging from node N1 to N12. Table 4-2 shows the statistics of the ADF test value for each sensor node with the rest of the nodes.

Table 4- 2. ADF-test for time series, Best Match **bold**, NN = Physically Nearest Neighbour.

	N1	N2	N3	N4	N5	N6	N7	N8	N9	N10	N11	N12
N1	-	**−43.26**	**−35.17**	**−25.90**	**−28.06**	**−24.65**	**−30.53**	**−29.79**	**−3.90**	−30.20	**−3.55**	−7.86
N2	−43.26	-	−45.02	**−28.53**	−29.82	**−26.89**	**−31**	**−30.33**	**−3.53**	**−27.60**	**−3.64**	**−7.02**
N3	**−35.18**	−45.01	-	**−25.36**	**−24.35**	**−25.21**	**−25.92**	**−25.08**	**−3.82**	**−26.42**	**−3.55**	**−6.58**
N4	**−26.07**	**−28.71**	**−25.19**	-	**−25.59**	**−29.65**	**−43.97**	**−42.87**	**−3.82**	**−29.49**	**−3.54**	**6.48**
N5	**−28.16**	**−29.91**	**−24.26**	**−25.67**	-	**−22.60**	**−24.43**	**−25.65**	**−3.91**	**−20.41**	**−3.57**	**−6.63**
N6	**−24.73**	**−26.96**	**−25.12**	**−29.75**	**−22.61**	-	**−30.01**	**−29.86**	**−3.84**	**−22.45**	**−3.56**	**−6.69**
N7	**−30.53**	**−31.13**	**−25.79**	−43.92	**−24.40**	−30.92	-	−49.12	**−3.83**	**−22.78**	**−3.57**	**−6.57**
N8	**−29.96**	**−30.48**	**−24.96**	**−42.05**	**−25.60**	**−29.90**	−49.09	-	**−3.87**	**−22.45**	**−3.56**	**−6.68**
N9	**−3.90**	**−3.93**	**−3.19**	**−3.16**	**−3.40**	**−3.31**	**−3.26**	**−3.37**	-	**−3.52**	−5.16	**−3.88**
N10	**−30.10**	**−27.49**	**−26.51**	**−20.69**	**−20.54**	**−22.59**	**−22.97**	**−22.29**	**−3.79**	-	**−3.57**	**−6.68**
N11	**−3.55**	**−3.55**	**−3.68**	**−3.74**	**−3.94**	**−3.98**	**−3.97**	**−3.02**	−5.13	**−3.44**	-	**−4.48**
N12	**−7.86**	**−7.07**	**−6.82**	**−6.77**	**−6.93**	**−7.01**	**−6.87**	**−7.02**	**−3.49**	**−7.25**	**−3.68**	-
NN	**N2**	**N4**	**N4**	**N2**	**N6**	**N5**	**N8**	**N7**	**N10**	**N9**	**N8**	**N10**
Best	**N2**	**N3**	**N2**	**N7**	**N2**	**N7**	**N8**	**N7**	**N11**	**N1**	**N9**	**N1**

In order for a series to be co-integrated with another, the test statistic should be less than the ADF test threshold which is normally set to −3.5, as described earlier in Section 4.3.1.4. It can be seen that almost all ADF test statistics are less than the critical value which means all series are statistically co-integrated. More negative values of the test statistic indicate a higher co-integration between series. Almost all series have a high degree of co-integration with all other series, with the test statistic for most pairs in Table4- 2 significantly more negative than the −3.5 threshold. The exceptions are nodes 9 and 11 with a test statistic close to the threshold when paired with other series. Among the co-integrated series, some are highly co-integrated with a single series. Node N1, N3, and N5 are highly co-integrated with N2. Similarly, N4, N6, and N8 are most co-integrated with N7. N9 and N11 are less co-integrated with other nodes, but they are co-integrated with each other. Also, N10 and N12 are co-integrated with N1 which in turn is co-integrated with N2. Note that the most co-integrated node is rarely the physically Nearest Neighbour node, shown in the NN row in the table.

This co-integration result shows that three sensor nodes, namely N2, N7, and N11, are co-integrated with all of the rest of the nodes. This indicates that using these three co-integrated series, the remaining series should be able to be accurately estimated by using a linear estimator.

4.5.3. Estimation of Observation at Co-Integrated Nodes

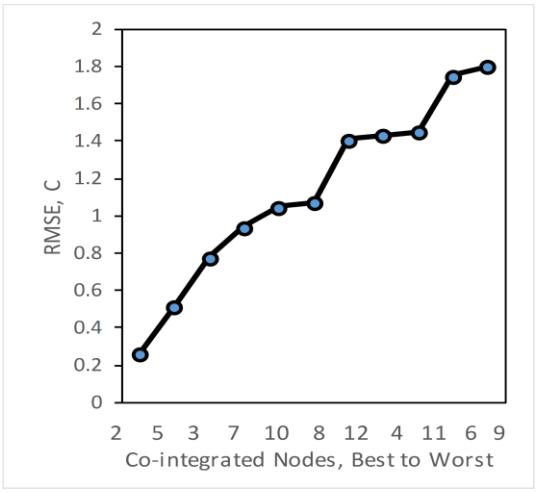
This section analyses results about how co-integrated series can be used for the estimation of the temperature value. The best subset selection algorithm is used to search for the best subset of nodes among co-integrated nodes. The maximum subset to be selected is set to 1 to evaluate how useful the most co-integrated node is for the estimation of temperature at other sensor nodes.

For each node, the most co-integrated node from Table 4-2 is selected as the estimator. Temperature is then estimated during a separate 10 day test period using the linear model learned during the training phase and mean test error is recorded.

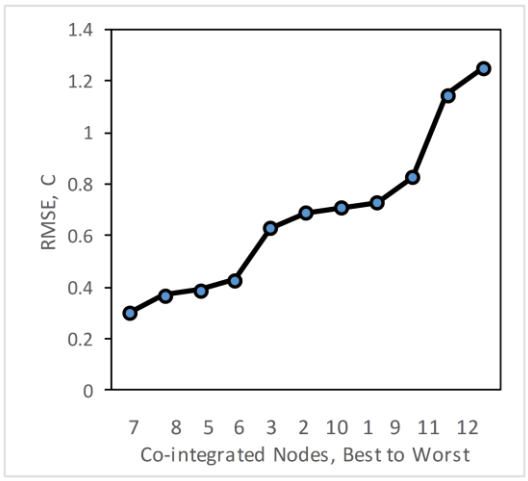
We then also analyse how the estimation varies if other nodes are selected instead of the most-co-integrated node. The RMSE is recorded for each of the other nodes used as an estimator. Figure 4 shows how the root mean squared error (RMSE) varies when different nodes are used for estimation – the order of nodes on the x-axis is from best to worst, left to right. The least RMSE for estimation of node N1 in Figure 4-4a is with the most co-integrated node N2 with an RMSE of 0.26 °C.

Based on the ordering given by RMSE, the quality order (best to worst) of estimators is N2, N5, N3, N7, N10, N8, N12, N4, N6, N11, N9. It is worth noting that this is different to an ordering based on the ADF test statistic as shown in Table 2, where the most co-integrated nodes for N1 are (in order) N2, N3, N7, N10, N8, N5, N4, N6, N12, N9, N11. The ADF test statistic, as shown in Table 2, gives a measure of the confidence that two nodes are co-integrated, rather than a direct measure of the quality of prediction. So, we recommend using Algorithm 1, based on the ADF, to establish where nearby series are sufficiently co-integrated for this method to be valid, and then use algorithm 2 based on RMSE to actually select the best estimator nodes.

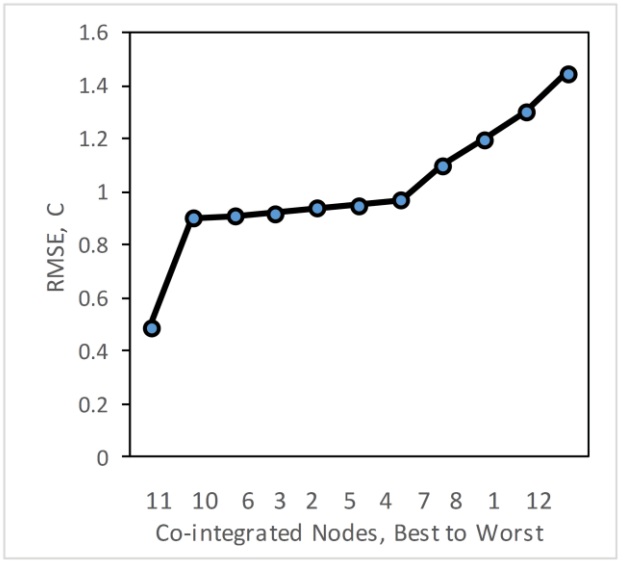
We repeat the analysis at node 4, which is most co-integrated with node 7 as shown in Figure 4-4b. From this figure, it can be seen that RMSE for node 4 is small with mostly co-integrated nodes 7, 8, 5 and 6 while estimation error is higher with node 11 which is less co-integrated. In the case of node 9, the lowest RMSE is obtained with node 11 as shown in Figure 4-4c.



(a) **Node 1**



(b) **Node 4**

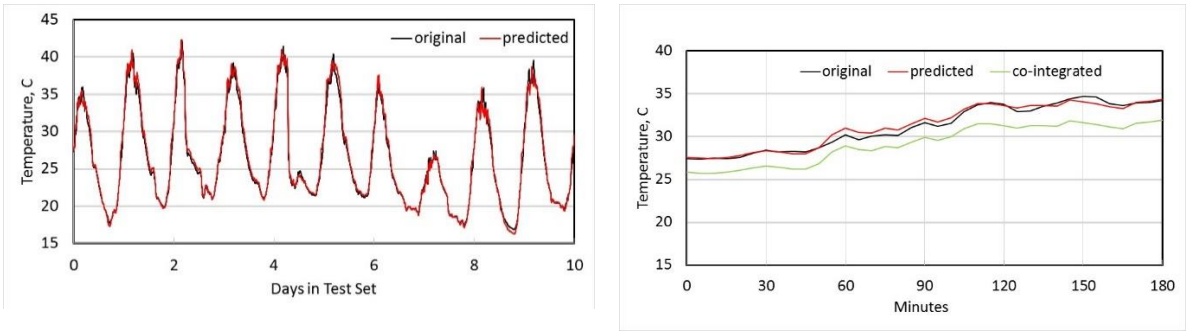


(c) **Node 9**

Figure 4-4. Root Mean-squared estimation error for co-integrated series at (**a**) Node 1, and (**b**) Node 4, and (**c**) Node 9, using all other nodes as estimators.

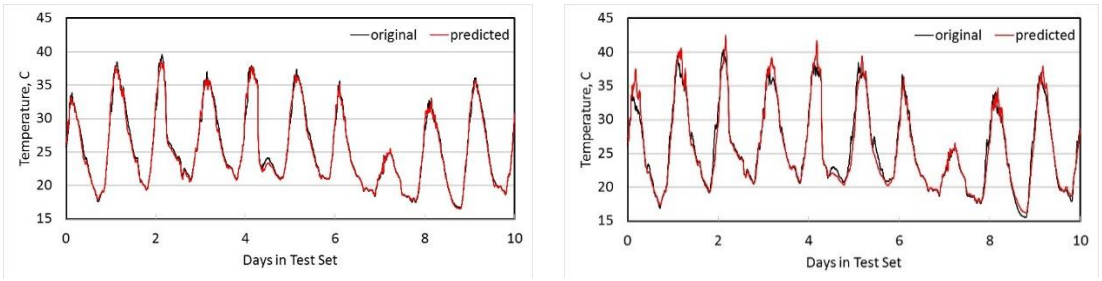
If the RMSE error threshold for temperature measurement in all nodes were set to 0.5 °C, nodes 2, 7 and 11 would be sufficient to estimate all other nodes within the required accuracy. So the number of deployed nodes could be reduced by 75%.

Figure 4-5a shows both the original measured temperature at node N1, and the temperature estimated from using co-integrated node N2 over the 10-day test set. Figure 4-5b shows the detail of these two-time series for the first 3 h, as well as the original measured temperature at N2, and it is clear that a linear estimator is significantly better than simply using N2 directly as an estimate. Figure 6a shows the original measured temperature at N4 and the estimated temperature from its most co-integrated node N7, while Figure 6b shows the original and estimated temperature at node N9. In all cases, the linear estimates from co-integrated nodes give good approximations to the actual measured temperatures.



(a) (b)

Figure 4-5. Estimation of temperature at node N1 using most co-integrated node N2 (**a**) over 10 days; (**b**) detail over first three hours, including the co-integrated baseline used for estimation.



(a) N4 estimated from N7 **(b) N9 estimated from N11**

Figure 4-6. Estimation of temperature nodes N4 and N7.

4.5.4. Discussion

While we have demonstrated the proposed approach on temperature time series, the approach is broadly applicable for determining the minimal set of sensor nodes for monitoring a given area. Since the sensor fields for each area will have unique spatiotemporal dynamics, our approach requires an initial dense deployment of sensor nodes for a short period. Once enough data is collected, we can determine nodes that are highly co-integrated and select the minimal set of nodes that can capture the sensor processes accurately. The deployment can then be reduced to include only the minimal set of nodes, thereby minimizing the monetary cost and network scale, along with its associated bandwidth overheads.

Several issues remain for future work. Firstly, how densely should the initial nodes be deployed? This obviously depends on the nature of the parameter being measured and its spatial variability. For this experiment, we have used temperature sensors that have been deployed at approximately 100m intervals, and we have shown that 75% of sensors can be estimated by spatial interpolation. Our suggestion would, therefore, be to deploy sensors at approximately four times the density of the expected final deployment, with the expectation that 75% are unnecessary, but the remaining 25% will be placed at better positions. This is clearly an area for more future investigation.

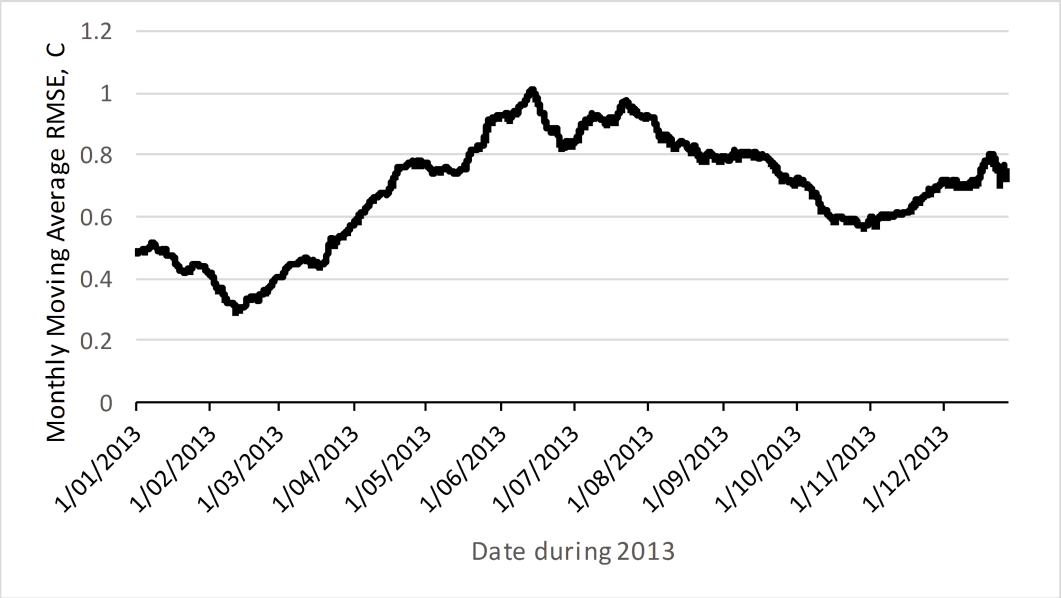


Figure 4-7. RMSE (moving average over 1 month) of prediction error using linear parameters from one week of training data in January.

A second question is whether the co-integrated prediction is reliable into the future, given that the test data in Figures 5 and 6 is immediately after the training data. Figure 4-7 shows how the RMSE changes over the course of the next year, using estimation parameters from just one week of training data. The monthly moving average RMSE error peaks at about 1 °C in the opposite season (winter in July versus training data during summer in January). This suggests that the RMSE error in the opposite season may be twice that close to the training data. If the deployment is planned to be very long term, this suggests temporary deployments that includes summer and winter periods may be useful to get better prediction accuracy. Again, this is a fruitful area for further research. Another area for further research is the use of non-linear models, including more complex machine-learning estimators which could include the season as a prediction input.

Chapter 5 Conclusions and Future Work

5.1. Conclusions on Temporal Interpolation

In chapter 3, univariate time series analysis is performed on an environmental sensor array deployed for monitoring outdoor environmental temperatures. Statistical properties of the phenomenon are observed and a suitable time series model is fitted. After parameter estimation, evaluation of the forecasting error of the future temperature is performed with varying sampling period of the sensor. Interpolation between subsampled series is also performed, and linear interpolation is preferred to more complex cubic spline interpolation. Temperature can be interpolated with an RMSE accuracy of less than 0.2 °C while extending the sampling interval to 60 min. For prediction, an RMSE in prediction of less than 1 °C is possible if the sampling interval is extended to around 60 min.

Altogether, this detailed analysis shows that frequent temperature sampling (every 5 min) provides limited additional information over-sampling at intervals up to 60 min. Such a down-sampling can be helpful in extending the energy-limited lifetime of the sensor and reducing the data storage requirements.

This analysis has shown that it is not possible to state the best sampling interval for all deployments based on experiments from one deployment. Instead, determination of the best sampling intervals would need to be done on a case-by-case basis after some initial high-frequency sampling. Then detailed data analysis using the methods described above can be used to determine a suitable sampling interval for that particular deployment. Subsequent work described in Chapter 4 work moved from the required temporal resolution to look at the required spatial resolution for measuring sensor data across a geographical area.

5.2. Conclusions on Spatial Interpolation

The work in Chapter 4 has proposed a time series-based analytical approach to develop sampling node selection in environmental sensor networks. Co-integration is found to be a useful tool to investigate temporal variation of the monitored phenomena. From the analyses

conducted with temperature series in a mine rehabilitation scenario, a significant number of sensing nodes are found to be redundant. Co-integrated nodes are shown to be capable of estimating observations at their co-integrated neighbour without exceeding a small error threshold. Such an approach of finding the best co-integrated nodes and using them to estimate observations for the rest of the nodes can be useful for developing a long-term environmental monitoring strategy.

To monitor a large spatial area, monitoring can begin with a large number of short-deployment sensors and analysing their co-integrated nature. Where sets of nodes are found to be co-integrated, redundant sensing positions can be removed. Permanent sensors are needed only in the positions of the non-redundant nodes. Alternatively, a small set of nodes can be densely deployed in one part of the area, the best positions can be chosen, then the unused nodes would be moved to another section of the area and this can be continued until the whole spatial region is covered. However, while this approach would provide local optima for sensor positions for each neighbourhood, it is more difficult to guarantee an optimum deployment over a large area. One suggestion would be to start at the centre of the deployment area, and then gradually move outwards. The pool of candidate nodes could include all the already committed permanent nodes from previous areas in the pool of potential co-integrated nodes. The best algorithm for extending this technique to cover a larger area would be an interesting topic for future work.

Currently, this work only focuses on static sensor nodes. Future work could include using mobile nodes to map the co-integrated regions of the sensing field prior to permanent node deployment.

5.3. Future Directions

So far this work has only examined the measurement of temperature. It would be useful to extend this work to other parameters, such as incident radiation, rainfall, soil moisture content, and humidity.

One currently suggested method for the dense deployment of nodes which is used to “train” the spatiotemporal interpolation, is to deploy a large number of low-cost, low-lifetime nodes to decide the position of the long-term nodes. Another option would be to use mobile sensors to make many measurements across the sensing area, perhaps over several weeks. This also has the advantage that the field could be recalibrated in the opposite season, since, as shown in chapter 4, errors are largest about 6 months away from the initial training.

Chapter 6 Bibliography

[1] T. Wark, W. Hu, P. Corke, J. Hodge, A. Keto, B. Mackey, G. Foley, P. Sikka, and M. Brunig, "Springbrook: Challenges in developing a long-term, rainforest wireless sensor network," in International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP 2008), Sydney, Australia, 2008, pp. 599-604.

[2] K. Lengfeld, and F. Ament, "Observing local-scale variability of near-surface temperature and humidity using a wireless sensor network," *Journal of Applied Meteorology and Climatology,* vol. 51, no. 1, pp. 30–41, 2012.

[3] N. Thepvilojanapong, T. Ono, and Y. Tobe, "A deployment of fine-grained sensor network and empirical analysis of urban temperature," *Sensors,* vol. 10, no. 3, pp. 2217–2241, 2010.

[4] J. K. Hart, and K. Martinez, "Environmental sensor networks: A revolution in the earth system science?," *Earth-Science Reviews,* vol. 78, no. 3, pp. 177-191, 2006.

[5] P. Dutta, P. M. Aoki, N. Kumar, A. Mainwaring, C. Myers, W. Willett, and A. Woodruff, "Common sense: participatory urban sensing using a network of handheld air quality monitors," in 7th ACM Conference on Embedded Networked Sensor Systems, Berkeley, CA, USA, 2009, pp. 349-350.

[6] X. Cao, J. Chen, Y. Xiao, and Y. Sun, "Building-environment control with wireless sensor and actuator networks: Centralized versus distributed," *IEEE Transactions on Industrial Electronics,* vol. 57, no. 11, pp. 3596-3605, 2010.

[7] S. De Vito, and G. Fattoruso, "Wireless chemical sensor networks for air quality monitoring," in 14th International Meeting on Chemical Sensors-IMCS 2012, 2012, pp. 641-644.

[8] R. Cardell-Oliver, K. Smettem, M. Kranz, and K. Mayer, "Field testing a wireless sensor network for reactive environmental monitoring," in Intelligent Sensors, Sensor Networks and Information Processing Conference (ISSNIP 2004), Melbourne, Australia, 2004, pp. 7-12.

[9] G. Werner-Allen, J. Johnson, M. Ruiz, J. Lees, and M. Welsh, “Monitoring volcanic eruptions with a wireless sensor network,” in Second European Workshop on Wireless Sensor Networks Istanbul, Turkey, 2005, pp. 108-120.

[10] N. E. Leonard, D. A. Paley, F. Lekien, R. Sepulchre, D. M. Fratantoni, and R. E. Davis, “Collective motion, sensor networks, and ocean sampling,” *Proceedings of the IEEE,* vol. 95, no. 1, pp. 48-74, 2007.

[11] G. Barrenetxea, F. Ingelrest, G. Schaefer, and M. Vetterli, “Wireless sensor networks for environmental monitoring: the sensorscope experience,” in 2008 IEEE International Zurich Seminar on Communications, Zurich, Switzerland, 2008, pp. 98-101.

[12] I. Talzi, A. Hasler, S. Gruber, and C. Tschudin, “PermaSense: investigating permafrost with a WSN in the Swiss Alps,” in 4th Workshop on Embedded Networked Sensors, 2007, pp. 8-12.

[13] P. Corke, T. Wark, R. Jurdak, W. Hu, P. Valencia, and D. Moore, “Environmental wireless sensor networks,” *Proceedings of the IEEE,* vol. 98, no. 11, pp. 1903-1917, 2010.

[14] L. M. Oliveira, and J. J. Rodrigues, “Wireless Sensor Networks: A Survey on Environmental Monitoring,” *Journal of Communications,* vol. 6, no. 2, pp. 143-151, 2011.

[15] M. F. Othman, and K. Shazali, “Wireless sensor network applications: A study in environment monitoring system,” *Procedia Engineering,* vol. 41, pp. 1204-1210, 2012.

[16] M. Di Francesco, S. K. Das, and G. Anastasi, “Data collection in wireless sensor networks with mobile elements: A survey,” *ACM Transactions on Sensor Networks* vol. 8, no. 1, pp. 7, 2011.

[17] S. Bhandari, N. Bergmann, R. Jurdak, and B. Kusy, “Time Series Data Analysis of Wireless Sensor Network Measurements of Temperature,” *Sensors,* vol. 17, no. 6, pp. 21, 2017.

[18] S. Bhandari, N. Bergmann, R. Jurdak, and B. Kusy, “Time Series Analysis for Spatial Node Selection in Environment Monitoring Sensor Networks,” *Sensors,* vol. 18, no. 1, pp. 11, 2017.

[19] S. J. Jeffrey, J. O. Carter, K. B. Moodie, and A. R. Beswick, “Using spatial interpolation to construct a comprehensive archive of Australian climate data,” *Environmental Modelling & Software,* vol. 16, no. 4, pp. 309-330, 2001.

[20] P. C. Kyriakidis, and A. G. Journel, “Geostatistical space–time models: a review,” *Mathematical Geology,* vol. 31, no. 6, pp. 651-684, 1999.

[21] L. Spadavecchia, and M. Williams, “Can spatio-temporal geostatistical methods improve high resolution regionalisation of meteorological variables?,” *Agricultural and Forest Meteorology,* vol. 149, no. 6, pp. 1105-1117, 2009.

[22] J. Cortés, “Distributed Kriged Kalman filter for spatial estimation,” *IEEE Transactions on Automatic Control,* vol. 54, no. 12, pp. 2816-2827, 2009.

[23] M. Umer, L. Kulik, and E. Tanin, “Spatial interpolation in wireless sensor networks: localized algorithms for variogram modeling and Kriging,” *Geoinformatica,* vol. 14, no. 1, pp. 101, 2010.

[24] J. Li, and A. D. Heap, *A Review of Spatial Interpolation Methods for Environmental Scientists,* vol. 2008/23, Geosciences Australia, Canberra, 2008.

[25] C. G. Karydas, I. Z. Gitas, E. Koutsogiannaki, N. Lydakis-Simantiris, and G. Silleos, “Evaluation of spatial interpolation techniques for mapping agricultural topsoil properties in Crete,” *EARSeL eProceedings,* vol. 8, no. 1, pp. 26-39, 2009.

[26] K. Q. Weinberger, and G. Tesauro, “Metric learning for kernel regression,” in International Workshop on Artificial Intelligence and Statistics, 2007, pp. 612-619.

[27] X. Wu, K. N. Brown, and C. J. Sreenan, “Analysis of smartphone user mobility traces for opportunistic data collection in wireless sensor networks,” *Pervasive and Mobile Computing,* vol. 9, no. 6, pp. 881-891, 2013.

[28] K.-J. Wong, C.-C. Chua, and Q. Li, “Environmental monitoring using wireless vehicular sensor networks,” in 5th International Conference on Wireless Communications, Networking and Mobile Computing, 2009. WiCom'09, 2009, pp. 1-4.

[29] R. North, M. Richards, J. Cohen, N. Hoose, J. Hassard, and J. Polak, “A mobile environmental sensing system to manage transportation and urban air quality,” in IEEE International Symposium on Circuits and Systems 2008, pp. 1994-1997.

[30] R. J. North, J. Cohen, S. Wilkins, M. Richards, N. Hoose, J. W. Polak, M. Bell, P. Blythe, B. Sharif, and J. Neasham, “Field deployments of the MESSAGE system for environmental monitoring,” *Traffic Engineering & Control*, vol. 50, no. 11, 2009.

[31] M. Mun, S. Reddy, K. Shilton, N. Yau, J. Burke, D. Estrin, M. Hansen, E. Howard, R. West, and P. Boda, “PEIR, the personal environmental impact report, as a platform for participatory sensing systems research,” in 7th International Conference on Mobile Systems, Applications, and Services, 2009, pp. 55-68.

[32] P. Zappi, E. Bales, J. H. Park, W. Griswold, and T. Š. Rosing, “The Citisense air quality monitoring mobile sensor node,” in 11th ACM/IEEE Conference on Information Processing in Sensor Networks, Beijing, China, 2012.

[33] P. Völgyesi, A. Nádas, X. Koutsoukos, and Á. Lédeczi, “Air quality monitoring with sensormap,” in IEEE International Conference on Information Processing in Sensor Networks, , 2008, pp. 529-530.

[34] D. Hasenfratz, O. Saukh, S. Sturzenegger, and L. Thiele, “Participatory air pollution monitoring using smartphones,” *Mobile Sensing*, vol. 1, pp. 1-5, 2012.

[35] J. J. Li, B. Faltings, O. Saukh, D. Hasenfratz, and J. Beutel, “Sensing the air we breathe-the opensense zurich dataset,” in National Conference on Artificial Intelligence, 2012, pp. 323-325.

[36] V. Sivaraman, J. Carrapetta, K. Hu, and B. G. Luxan, “HazeWatch: A participatory sensor system for monitoring air pollution in Sydney,” in IEEE 38th Conference on Local Computer Networks 2013, pp. 56-64.

[37] P. A. de Souza, G. Timms, A. Davie, B. Howell, and S. Giugni, “Marine monitoring using fixed and mobile sensor nodes,” in IEEE OCEANS Sydney, 2010, pp. 1-4.

[38] G. Hernandez-Penaloza, and B. Beferull-Lozano, “Field estimation in wireless sensor networks using distributed kriging,” in IEEE International Conference on Communications (ICC), , 2012, pp. 724-729.

[39] S. Martinez, “Distributed interpolation schemes for field estimation by mobile sensor networks,” *IEEE Transactions on Control Systems Technology*, vol. 18, no. 2, pp. 491-500, 2010.

[40] R. Tynan, G. O'Hare, D. Marsh, and D. O'Kane, “Interpolation for wireless sensor network coverage,” in Second IEEE Workshop on Embedded Networked Sensors, EmNetS-II. , 2005, pp. 123-131.

[41] M. C. Vuran, Ö. B. Akan, and I. F. Akyildiz, “Spatio-temporal correlation: theory and applications for wireless sensor networks,” *Computer Networks*, vol. 45, no. 3, pp. 245-259, 2004.

[42] B. Lu, J. Oyekan, D. Gu, H. Hu, and H. F. G. Nia, “Mobile sensor networks for modelling environmental pollutant distribution,” *International Journal of Systems Science,* vol. 42, no. 9, pp. 1491-1505, 2011.

[43] R. Ouyang, K. H. Low, J. Chen, and P. Jaillet, “Multi-robot active sensing of non-stationary Gaussian process-based environmental phenomena,” in International Conference on Autonomous Agents and Multi-agent Systems, 2014, pp. 573-580.

[44] C. Alippi, G. Anastasi, M. Di Francesco, and M. Roveri, “Energy management in wireless sensor networks with energy-hungry sensors,” *IEEE Instrumentation Measurement Magazine,* vol. 12, no. 2, 2009.

[45] H. Harb, and A. Makhoul, “Energy-Efficient Sensor Data Collection Approach for Industrial Process Monitoring,” *IEEE Transactions on Industrial Informatics,* vol. 14, no. 2, pp. 661-672, 2018.

[46] D. J. Marceau, D. J. Gratton, R. A. Fournier, and J.-P. Fortin, “Remote sensing and the measurement of geographical entities in a forested environment. 2. The optimal spatial resolution,” *Remote Sensing of Environment,* vol. 49, no. 2, pp. 105-117, 1994.

[47] S. Budi, P. de Souza, G. Timms, F. Susanto, V. Malhotra, and P. Turner, “Mobile platform sampling for designing environmental sensor networks,” *Environmental Monitoring and Assessment,* vol. 190, no. 3, pp. 130, 2018.

[48] M. Jin, S. Liu, S. Schiavon, and C. Spanos, “Automated mobile sensing: Towards high-granularity agile indoor environmental quality monitoring,” *Building and Environment,* vol. 127, pp. 268-276, 2018.

[49] R. Honicky, E. A. Brewer, E. Paulos, and R. White, “N-smarts: networked suite of mobile atmospheric real-time sensors,” in Second ACM SIGCOMM workshop on networked systems for developing regions, 2008, pp. 25-30.

[50] W. Hedgecock, P. Völgyesi, A. Ledeczi, X. Koutsoukos, A. Aldroubi, A. Szalay, and A. Terzis, “Mobile air pollution monitoring network,” in ACM Symposium on Applied Computing, 2010, pp. 795-796.

[51] Y. Ma, M. Richards, M. Ghanem, Y. Guo, and J. Hassard, “Air pollution monitoring and mining based on sensor grid in London,” *Sensors,* vol. 8, no. 6, pp. 3601-3623, 2008.

[52] O. A. M. Popoola, “Studies of urban air quality using electrochemical based sensor instruments,” Department of Chemistry, PhD Thesis, Dept. of Chemsitry, University of Cambridge, 2012.

[53] J. Panchard, T. Prabhakar, J.-P. Hubaux, and H. Jamadagni, “Commonsense net: A wireless sensor network for resource-poor agriculture in the semiarid areas of developing countries,” *Information Technologies & International Development,* vol. 4, no. 1, pp. pp. 51-67, 2007.

[54] S. Michel, A. Salehi, L. Luo, N. Dawes, K. Aberer, G. Barrenetxea, M. Bavay, A. Kansal, K. A. Kumar, and S. Nath, “Environmental Monitoring 2.0,” in IEEE 25th International Conference on Data Engineering, ICDE'09., 2009, pp. 1507-1510.

[55] D. S. G. Pollock, R. C. Green, and T. Nguyen, *Handbook of time series analysis, signal processing, and dynamics*: Academic Press, 1999.

[56] D. Guo, X. Qu, L. Huang, and Y. Yao, “Sparsity-based spatial interpolation in wireless sensor networks,” *Sensors,* vol. 11, no. 3, pp. 2385-2407, 2011.

[57] J. Shinomiya, Y. Teranishi, K. Harumoto, S. Takeuchi, and S. Nishio, “An examination of sensor data collection method for spatial interpolation on hierarchical Delaunay overlay network,” in Eleventh IEEE International Conference on Mobile Data Management (MDM), 2010, pp. 407-412.

[58] C. Liu, K. Wu, and J. Pei, “An energy-efficient data collection framework for wireless sensor networks by exploiting spatiotemporal correlation,” *IEEE Transactions on Parallel and Distributed Systems,* vol. 18, no. 7, 2007.

[59] Y. W. Law, S. Chatterjea, J. Jin, T. Hanselmann, and M. Palaniswami, “Energy-efficient data acquisition by adaptive sampling for wireless sensor networks,” in 5th International Conference on Wireless Communications and Mobile Computing (WiCOM 2009), Beijing, China, 2009, pp. 1146-1151.

[60] Y.-A. Le Borgne, S. Santini, and G. Bontempi, “Adaptive model selection for time series prediction in wireless sensor networks,” *Signal Processing,* vol. 87, no. 12, pp. 3010-3020, 2007.

[61] K. Miranda, and T. Razafindralambo, “Using efficiently autoregressive estimation in wireless sensor networks,” in International Conference on Computer, Information and Telecommunication Systems (CITS 2013), Athens, Greece, 2013, pp. 1-5.

[62] C. Liu, K. Wu, and M. Tsao, “Energy efficient information collection with the ARIMA model in wireless sensor networks,” in IEEE Global Telecommunications Conference (GLOBECOM ’05), St Louis, MO, USA, 2005, pp. 2470-2474.

[63] F. A. Aderohunmu, G. Paci, D. Brunelli, J. D. Deng, L. Benini, and M. Purvis, “An application-specific forecasting algorithm for extending wsn lifetime,” in IEEE International Conference on Distributed Computing in Sensor Systems (DCOSS), 2013, pp. 374-381.

[64] A. Amidi, “ARIMA based value estimation in wireless sensor networks,” *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences,* vol. 40, no. 2, pp. 41, 2014.

[65] J. Pardo, F. Zamora-Martínez, and P. Botella-Rocamora, “Online learning algorithm for time series forecasting suitable for low cost wireless sensor networks nodes,” *Sensors,* vol. 15, no. 4, pp. 9277-9304, 2015.

[66] D. Tulone, and S. Madden, “PAQ: Time series forecasting for approximate query answering in sensor networks,” in European Workshop on Wireless Sensor Networks, 2006, pp. 21-37.

[67] B. Kusy, C. Richter, S. Bhandari, R. Jurdak, V. J. Neldner, and M. R. Ngugi, “Evidence-based landscape rehabilitation through microclimate sensing,” in 12th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON 2015), Seattle, WA, USA, 2015, pp. 372-380.

[68] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time Series Analysis: Forecasting and Control*, Hoboken, NJ, USA: John Wiley & Sons, 2015.

[69] J. D. Cryer, and K.-S. Chan, "Time series regression models," *Time Series Analysis: with Applications in R*, J. D. Cryer, ed., pp. 249-276: Springer, 2008.

[70] J. G. De Gooijer, and R. J. Hyndman, "25 years of time series forecasting," *International Journal of Forecasting,* vol. 22, no. 3, pp. 443-473, 2006.

[71] R. Ihaka, and R. Gentleman, "R: a language for data analysis and graphics," *Journal of Computational and Graphical Statistics,* vol. 5, no. 3, pp. 299-314, 1996.

[72] R. J. Hyndman, and Y. Khandakar, *Automatic time series for forecasting: the forecast package for R*: Monash University, Department of Econometrics and Business Statistics, 2007.

[73] J. Chen, S. C. Saunders, T. R. Crow, R. J. Naiman, K. D. Brosofske, G. D. Mroz, B. L. Brookshire, and J. F. Franklin, "Microclimate in forest ecosystem and landscape ecology variations in local climate can be used to monitor and compare the effects of different management regimes," *BioScience,* vol. 49, no. 4, pp. 288-297, 1999.

[74] K. Langendoen, A. Baggio, and O. Visser, "Murphy loves potatoes: Experiences from a pilot sensor network deployment in precision agriculture," in IPDPS 20th International Parallel and Distributed Processing Symposium, Rhodes Island, Greece, 2006, pp. 1-8

[75] A. Krause, C. Guestrin, A. Gupta, and J. Kleinberg, "Near-optimal sensor placements: Maximizing information while minimizing communication cost," in 5th International Conference on Information Processing in Sensor Networks (IPSN 2006), Nashville, TN, USA, 2006, pp. 2-10.

[76] C. C. Castello, J. Fan, A. Davari, and R.-X. Chen, "Optimal sensor placement strategy for environmental monitoring using wireless sensor networks," in 42nd Southeastern Symposium on System Theory (SSST 2010), Tyler, TX, USA, 2010, pp. 275-279.

[77] Y. Liu, Y. He, M. Li, J. Wang, K. Liu, and X. Li, "Does wireless sensor network scale? A measurement study on GreenOrbs," *IEEE Transactions on Parallel and Distributed Systems,* vol. 24, no. 10, pp. 1983-1993, 2013.

[78] N. Chen, C. Xiao, F. Pu, X. Wang, C. Wang, Z. Wang, and J. Gong, "Cyber-Physical Geographical Information Service-Enabled Control of Diverse In-Situ Sensors," *Sensors,* vol. 15, no. 2, pp. 2565, 2015.

[79] K. Wang, Q. Guan, N. Chen, D. Tong, C. Hu, Y. Peng, X. Dong, and C. Yang, "Optimizing the configuration of precipitation stations in a space-ground integrated

sensor network based on spatial-temporal coverage maximization," *Journal of Hydrology,* vol. 548, no. Supplement C, pp. 625-640, 2017.

[80] D. A. Dickey, and W. A. Fuller, "Likelihood ratio statistics for autoregressive time series with a unit root," *Econometrica,* vol. 49, no. 4, pp. 1057-1072, 1981.

[81] S. Johansen, "Statistical analysis of cointegration vectors," *Journal of Economic Dynamics and Control,* vol. 12, no. 2-3, pp. 231-254, 1988.

[82] *R Core Team. R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2017.