REAL-TIME URBAN MICROCLIMATE ANALYSIS USING INTERNET OF THINGS

Seminar Report

Submitted in partial fulfillment of the requirements for the award of degree of

BACHELOR OF TECHNOLOGY

In

COMPUTER SCIENCE AND ENGINEERING

of

APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY

Submitted By

GANESH S

Department of Computer Science & Engineering
Mar Athanasius College Of Engineering
Kothamangalam

REAL-TIME URBAN MICROCLIMATE ANALYSIS USING INTERNET OF THINGS

Seminar Report

Submitted in partial fulfillment of the requirements for the award of degree of

BACHELOR OF TECHNOLOGY

In

COMPUTER SCIENCE AND ENGINEERING

of

APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY

Submitted By

GANESH S

Department of Computer Science & Engineering
Mar Athanasius College Of Engineering
Kothamangalam

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
MAR ATHANASIUS COLLEGE OF ENGINEERING
KOTHAMANGALAM

CERTIFICATE

This is to certify that the report entitled **Real-time urban microclimate analysis using internet of things** *submitted by* **Mr. GANESH S, Reg. No. MAC15CS028** *towards partial fulfillment of the requirement for the award of Degree of Bachelor of Technology in Computer science and Engineering Engineering from APJ Abdul Kalam Technological University for December 2018 is a bonafide record of the seminar carried out by him under our supervision and guidance.*

.....	.....	.....
Prof. Joby George	**Prof. Neethu Subash**	**Dr. Surekha Mariam Varghese**
Faculty Guide	***Faculty Guide***	***Head Of Department***

Date:

Dept. Seal

ACKNOWLEDGEMENT

First and foremost, I sincerely thank the 'God Almighty' for his grace for the successful and timely completion of the seminar.

I express my sincere gratitude and thanks to Dr. Solly George, Principal and Dr. Surekha Mariam Varghese, Head Of the Department for providing the necessary facilities and their encouragement and support.

I owe special thanks to the staff-in-charge Prof. Joby george, Prof. Neethu Subash and Prof. Joby Anu Mathew for their corrections, suggestions and sincere efforts to co-ordinate the seminar under a tight schedule.

I express my sincere thanks to staff members in the Department of Computer Science and Engineering who have taken sincere efforts in helping me to conduct this seminar.

Finally, I would like to acknowledge the heartfelt efforts, comments, criticisms, co-operation and tremendous support given to me by my dear friends during the preparation of the seminar and also during the presentation without whose support this work would have been all the more difficult to accomplish.

ABSTRACT

Real-time environment monitoring and analysis is an important research area of Internet of Things (IoT). This research is about the study of tree canopy cover over the microclimate environment using heterogeneous sensor data to reduce the Urban Heat effect. There are several challenges that are addressed, such as obtaining reliable and detailed observations over monitoring area, detecting unusual events from data, and visualizing events in real-time in a way that is easily understandable by the end users. The research proposes an integrated geovisualization framework, built for real-time wireless sensor network data on the synergy of computational intelligence and visual methods, to analyze complex patterns of urban microclimate. A Bayesian maximum entropy-based method and a hyperellipsoidal model-based algorithm have been developed in the integrated framework to address above challenges and all the study was based at Melbourne, Australia as a test site.

Contents

List of Figures

List of Abbrevations

IoT	Internet of Things
BME	Byesian Maximum Entropy
ENOF	Ellipsoidal Neighbourhood Outlier Factor
WSS	Wide Sense Stationary
IBRL	Intel Berkeley Resarch Laboratory
RMSE	Root Mean Square Error

Introduction

Rapid adaptations of Smart City and Internet of Things (IoT) technologies are assisting in urban planning to ensure sustainable cities and lifestyles [1], [2]. Wireless sensor network (WSN) is one of the most important elements of the IoT paradigm which behaves as a digital skin and provides flexible platform to collect data for environmental modeling. In particular, monitoring Urban Heat Island [3] (UHI) effect is important for city councils and government agencies to plan and maintain a healthy Smart City environment.

Increase in human activities, modern urbanisation and subsequent loss of vegetation in the urban landscape have been contributing to the increase of temperature in cities by several degrees higher than the surrounding suburbs, particularly at night. This phenomenon is known as Urban Heat Island effect [3]. In cities, the heat is stored in non-homogenous proportions based on the characteristics of the surrounding environments, whether considering buildings, parks, public places, and other infrastructure [4]. The result is a considerable variability of meteorological parameters such as temperature in the immediate surrounding suburbs, raising the challenge of how best to capture these variabilities in fine details. Increasing the number of trees to reduce the UHI effect is a preferred solution [5], but achieving cost-effective solution and better environmental health benefits require analysis of how different trees, buildings, and parks affect their microclimate. For example, the Melbourne city council's urban forest team recognizes the need of tree species selection based on their cooling benefits, water status, soil conditions, and other parameters [6].

There are a few experimental studies [7]–[10] that used modelling based methods to investigate the ecosystem provided by tress and urban forest. Urban microclimate is described by various parameters, such as humidity, temperature, daylight levels, and wind speed. Modelling of any complex ecosystem, which is impacted by several parameters, is a challenging task and requires the domain-specific knowledge. Thus, it triggers the need for experimental tools to analyze the effects of several factors on microclimates, and use them in model adjustment[2] accordingly. Most city councils and health agencies use geographic information system (GIS)

based visualization tools to analyze the urban climate. The visualization enabled tools[3], [4] for climate analysis are offline and limited to scientific data obtained from meteorological station. They are univariate, i.e., they show only one variable at a time. Such approaches are useful in information presentation, but have severe limitations in finding complex patterns from data that span across multiple dimensions.

There are three main challenges in understanding the relationship between environmental parameters and sensor data: First, is the lack of detailed observation of environmental data in real-time. Real-time observations of environment under different conditions and at higher spatial and temporal resolutions[5] are required for detailed analysis. In many IoT applications, either a small number of high precision sensors (in low spatial resolution) or a large number of inexpensive low precision sensors (in high spatial resolution) are deployed to reduce the overall deployment cost. Due to unavailability of data at locations where sensors are not deployed or due to measurement errors/uncertainties present in low precision sensors, obtaining reliable and detailed observations over a monitoring area in real time is a challenging task. The second challenge is to identify unusual events as patterns from the environmental data. A wide variety of impacts and interactions are possible at spatio-temporal domain in an urban environment. Analyzing voluminous real-time data to identify unusual events automatically is another challenging task. The third challenge is the joint multi-sensor and multivariate visualization that conveys real-time microclimate information including geolocation and time varying sensor data. As more and more data are collected, the analysis such as pattern findings and decision making from the data may become difficult and cumbersome. Therefore, a real-time visualization tool is required that can present multi-sensor, multivariate data as well as the outputs of any computational intelligence algorithm in a unified manner so that patterns can be identified quickly and especially when data volumes are very large.

To address these challenges, we introduce an interactive, multivariate geovisualization framework, built on the synergy of computational intelligence methods and IoT technologies, for real-time monitoring and analysis of complex patterns in urban microclimate data. Our major contributions in this paper are as follows: An interactive, real-time geovisualization framework has been developed for visualization of joint multisensor and multivariate, real-time urban

microclimate data. This framework is built on the following two analytical methods: Spatio-temporal estimation: A novel estimation model, based on the Bayesian maximum entropy (BME) method, is build in both centralized and distributed manner to obtain detailed observation of environmental data in real-time. Pattern detection: An anomaly detection method, based on the hyperellipsoidal models, is utilized to identify unusual patterns from environmental data. We demonstrate the performance of the two analytical algorithms on real-time and historical data, obtained from an indoor deployment (IBRL [6]) and two outdoor deployments using IoT sensors deployed at two strategically selected locations in Melbourne, Australia. A small number of low cost sensors were deployed at both locations to investigate the effectiveness of BME based estimation technique to obtain reliable and detailed observation using limited number of sensors' measurements. We also demonstrate the usefulness of our geovisualization framework for identifying interesting patterns, which are verified by urban forest team of Melbourne city council. To the best of our knowledge, this is the first integrated visualization framework, built on the real-time analytical methods as backbone, to analyze complex patterns of urban microclimate. This analytical tool is intended to be used in future urban design and planning interventions, specially for the selection and positioning of trees in urban spaces. The proposed integrated framework is generic and can also be used for other Smart City applications.

We present an integrated visualization framework for realtime urban microclimate monitoring and analysis, as shown in Fig 1.1 In the proposed framework, the real-time data are collected from our IoT deployment. The details of IoT deployment is provided in Chapter 2. The integrated visualization framework addresses three main challenges as follows: The first challenge is the lack of detailed observation of environmental data. To overcome this problem, a Bayesian Maximum Entropy (BME)[7], [8] based spatio-temporal estimation model is implemented.

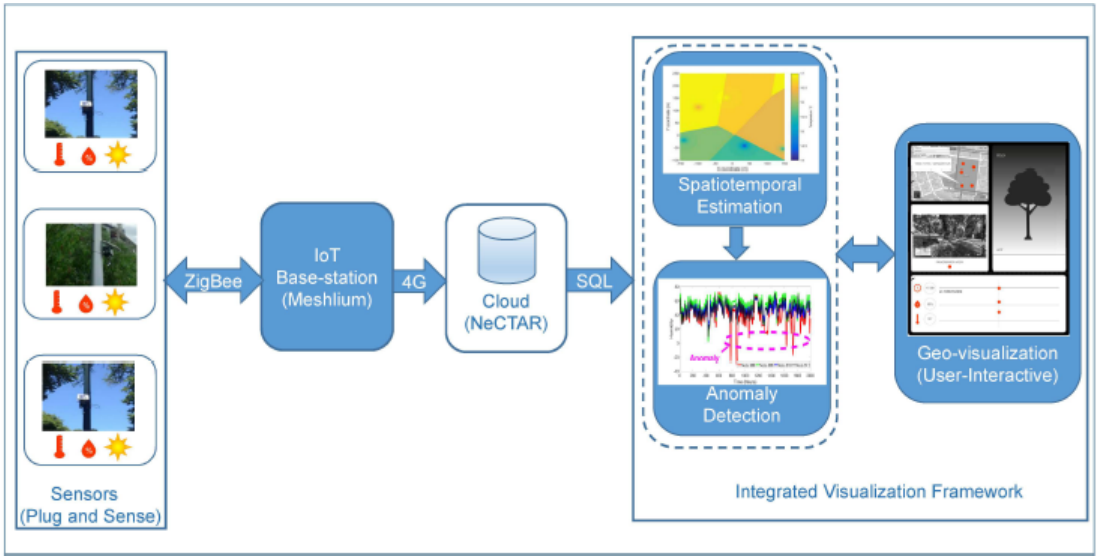


Fig. 1.1: System overview: Communication flow from Sensor nodes to IoT base-station through ZigBee.

This technique can incorporate the measurement errors/uncertainties of low cost, low precision sensors to yield reliable estimates. The detailed explanations of BME method is provided in Chapter 3. The second challenge is to identify the spatio-temporal patterns (unusual) from the observed data automatically. This problem is addressed by utilizing hyperellipsoidal models [9] based distributed anomaly detection algorithm, which is discussed in Chapter 4. In Chapter 5, we examine the utility of both algorithms on historical and real-time data obtained from our IoT deployment, followed by their computational complexities in Chapter 6. To address the visualization (third) challenge, a new geovisualization framework with real-time and userinteractive capabilities is developed that incorporates multisensor and multivariate data. In addition to real-time data visualization, it also incorporates spatio-temporal estimation and anomaly detection algorithm as backbone, and facilitates visualization of their outputs. The geovisualization framework is discussed in Chapter 7, followed by conclusion and future work in Section Chapter 8.

Related works

Fig. 2.1: (a) Fitzroy Gardens with five sensor (b) Dockland Library with four sensor.

For this study, an Internet of things (IoT) based networked devices were deployed in Melbourne, Australia, for real-time urban microclimate monitoring [10]. To study the long term micro-scale relation between canopy coverage and environmental parameters, WSNs were deployed at two strategically selected locations in Melbourne. Five sensor nodes were deployed at Fitzroy Gardens (where different types of tree canopies exist) and four sensor nodes were deployed at Docklands Library (where the location is surrounded by buildings, parks, and Yarra river) to study different urban vegetation. The tree canopies considered include multiple types — established tree canopy, mature tree canopy and minimal canopy cover — based on their canopy cover percentage. The sensor data collected include temperature, humidity and luminosity measurements collected at 10 minute intervals since December 2014. Fig. 2.1 shows the deployment locations of the sensor nodes along with the base stations at Fitzroy Gardens (a) and Docklands Library (b) in Melbourne

Low cost sensors, namely the Waspmote[1] wireless sensor nodes, which are on open source wireless sensor platform developed by Libelium, were deployed to measure temperature, humidity and luminosity parameters. Fig. 1.1 shows the end-to-end system overview of the deployment. The Plug and Sense module creates a packet at every 10 minute intervals using the sampled sensor data, and transmits it to the base station via ZigBee protocol. Meshlium was used as the base station to collect the data from all the sensor nodes and save them in a Cloud

server. Data were transmitted to an external Cloud database using a 3G/4G wireless modem. We used the Cloud infrastructure provided by the NeCTAR (National eResearch Collaboration Tools and Resources[2]) for this purpose. This research cloud provides high capacity storage for the real-time (streaming) data feeds. In order to make the real time data collected from the deployment accessible to the general public via the Internet, the aggregated data is periodically copied from the NeCTAR Cloud to the City of Melbourne's (CoM) Open Data platform servers[3]. The real-time data from both the deployments have been used in evaluating the spatio-temporal estimation and anomaly detection algorithms for microclimate analysis. In the next section, we discuss the Bayesian maximum entropy based estimation method in detail.

2.1 Detailed observation using estimation

The detailed observations of the monitored phenomena over an area of interest can be obtained in the form of overall spatial map. In continuous monitoring applications, such as environmental monitoring[4], [5], inexpensive IoT sensors are used in order to reduce the deployment cost. Generally, these low-cost, low-precision sensor nodes have limited memory and processing power. Obtaining a reliable and detailed observation of a geographical area in the form of overall spatial map (estimates), using measurements of these inexpensive sensors, is a challenging task. Several interpolation techniques have been used to obtain spatial map of urban microclimate data. Jeganathan et al[6] used Inverse Distance Weighting method for offline spatial interpolation of weather station data. This method does not incorporate the spatial covariance information [7] effectively. Sánchez et al[8] used Kriging method for offline spatial interpolation of urban transect data. Both the studies have been done for offline analysis of weather station data. In addition, both algorithms do not consider the measurement errors of low-precision sensors in estimation. We employ a Bayesian maximum entropy (BME) based spatio-temporal estimation method to estimate the value of the microclimate parameters at unobserved locations. BME is a spatio-temporal analysis and mapping technique, which can incorporate the measurements errors of low precision. sensors in the form of interval or probabilistic data. Below we discuss the BME technique and its ability to integrate low precision sensor measurements as soft data.

In many IoT applications, either a small number of high precision sensors (in low spatial resolution) or a large number of inexpensive low precision sensors (in high spatial resolution) are deployed to reduce the overall deployment cost. Environmental data obtained from various sensor nodes are spatio-temporal in nature, thus each point p in this continuum can be represented by space and its time information as data p = (s,t), where s is the spatial location represented by longitude and latitude, and t is the time. Given a set of N sensor nodes at locations S=[s1,s2,..,sN], realization of random variable X (e.g., temperature) at these locations can be represented as Xdata=[X1,X2,X3,..,XN]. For real-time visualization of spatio-temporal map within the geographical area of interest, we use the BME based scheme for estimating the values at unobserved location. In particular, we estimate the realization XE of a random variable at a location set , where it is a set of locations where estimation is to be performed), and then a spatio-temporal (ST) map is generated using the realizations Xmap=[X1,X2,X3,..,XN ,XE] at locations Smap = [s1,s2,..,sN ,sE]. The total physical knowledge K regarding a natural process, used by BME to estimate the values, comprises two prime knowledge bases: general knowledge KG, such as law of sciences, structured patterns, summary statistics; specificatory knowledge, KS, obtained through experience with specific situations and associated with physical data points. Physical data points may consist of hard data points Xhard, which are exact measurements of natural process with probability one such as high precision sensor measurements; and soft data points Xsoft , which can be an interval or a probabilistic type of data, that capture uncertain knowledge, intuition or low precision sensor outputs etc., such that Xdata is a collection of set Xsoft and Xhard. BME technique uses three stages for knowledge acquisition and processing, as follows: Prior stage, which starts with the basic set of assumptions and general knowledge, KG, with the goal of prior information maximization; Pre-posterior stage, which uses specificatory knowledge, KS, including hard and soft data; ; and Posterior stage, in which the knowledge from prior and pre-posterior stages are integrated and used with the goal of posterior probability maximization. In the prior stage, the expected information contained in the prior probability distribution function (pdf) is defined using the Shannon's information measure as follows,

$$\sum(InfoG[Xmap]) = -\int G(Xmap)logG(Xmap)dXmap, \qquad (Equ:1)$$

here G(Xma p) is the prior pdf model, which refers to knowledge KG before any specific knowledge base, KS, has been taken into consideration, and E(n) denotes the expectation operator. The shape of the prior pdf is derived by maximizing the expected information which takes the following constraints into consideration :

$$\sum(g[Xmap]) = \int gn(Xmap)G(Xmap)dXmap, \qquad (Equ:2)$$

where n= 0,1,2...Nc and gn are known functions of Xmap with E(g(Xmap)) = 1, and Nc is such that stochastic moments, that involve all p = (s,t) points, are included.

In this work, mean Xmap(p) and covariance functions of sensor measurements were used as general knowledge. The space-time variability of X is described in terms of a centered covariance function as: Cmap= E[(Xmap(p)] Space and time lags respectively, and show that this covariance is spatially isotropic and stationary in time. Hence, gn is adapted such that expectation E(g(Xmap)) defines ST mean and covariance functions throughout the ST domain of interest. In this work, a nugget-exponential function [9] (discussed in Section V-A) is used to model spatio-temporal covariance structure known as variogram. Variogram (or semivariogram) is an experimental function, which is used to determine spatial correlations in observations measured at sample locations and time. The optimization for maximization is performed using the Lagrange multipliers n. Hence, G represents the operator processing the general knowledge KG, and is given by P which encapsulates E(Xmap). Similarly, specificatory knowledge KS is considered in the pre-posterior stage, and the prior pdf G is updated by means of Bayesian conditionalization. In this paper, we use interval I = [l,u] for expressing the soft data, where interval ranges are defined using measurement error/uncertainty of low precision sensors. At the posterior stage, the updated pdf is a conditional pdf, and can be expressed in terms of the prior pdf and specificatory knowledge considered at pre-posterior stage as K(Xdata) which is represented as 10S[G(Xma p),Xsoft], where S[G] represents the posterior operator that incorporates the soft data. Estimate at E can be obtained by maximizing the posterior pdf with respect to XE such that BME estimate minimizes the root mean square error (RMSE).

Proposed work

Most of the unusual patterns appear as anomaly or outlier in the spatio-temporal data. Several methods, have been proposed for anomaly detection in sensor network, but most of them are unable to detect anomalies in real-time with low computational and communication complexity. It is the only visualization framework which uses clustering for classification of urban transects. However, it is limited for weather station data and suitable for offline analysis. One of the challenges in analyzing voluminous real time data is to identify the unusual events, called as ”anomalies”, automatically, and in a timely manner. The proposed framework uses an energy efficient and distributed anomaly detection algorithm, developed using multiple hyperellipsoidal models. This anomaly detection algorithm has significantly less communication overhead, memory and computation complexity, which makes it suitable for resource-constrained WSNs. Consider a set of N sensor nodes V = Vj, j = 1...n, having realization of random variable X (e.g., temperature). At every time interval i, each sensor node Vj measures data vector xj After a window of n sensor measurements, each sensor has collected a set of measurements as Xj = x j, j : i = 1...n. The aim is to find the local (each sensor measurements) and global (measurements from multiple nodes) anomalies in the collected data. A summary of the proposed distributed approach is presented below (a network of sensors is considered to have a hierarchical topology, where a parent and child relationship exists).

3.1 Local anomaly detection

Each node Vj performs hyperellipsoidal clustering on its data using our HyCARCE (Hyperellipsoidal Clustering algorithm for Resource Constrained Environments) algorithm. This results in a set of multiple hyperellipsoids Ej for each sensor node (Note that the number of clusters is determined algorithmically). Then the anomaly detection algorithm (presented below) is applied to the hyperellipsoidal clusters Ej to identify the locally anomalous hyperellipsoids, and subsequently the locally anomalous data vectors of that node.

3.2 Global anomaly detection

Following steps are performed to detect global anomalies. Each node Vj sends a summary of its hyperellipsoids to its immediate parent node (in the form of tuple (m,A,ID), where m, A, and ID are centroid, (inverse) covariance matrix (positive definite), and cluster ID of ellipsoid respectively. Geometrically, a hyperellipsoid in h space is represented by e(A,m,r) whose all points are constant A distance(r) from its center m, where r is also called as effective radius of ellipsoid. In our experiment, we consider h = 2 i.e., shape is ellipse. The parent node combines all hyperellipsoidal clusters from its children with its own cluster. Then, this parent node sends summaries to its immediate parent. This process continues up to the gateway (Base-station), where anomaly detection algorithm is performed to detect globally anomalous hyperellipsoids. Summary of anomalous global hyperellipsoidal clusters are communicated back to all the nodes in network. Then, each node identifies corresponding global anomalous data using this summary information. The anomaly detection algorithm consists of two components: similarity measures for pairs of ellipsoids, and a mechanism to score the ”outlierness” of the hyperellipsoids.

3.3 Similarity measure

An ellipse e = (A,m,r) can be constructed by tracing the curve whose distance from foci f1 and f2 is a positive constant. Let D be the similarity based on focal distance between two ellipsoids ea = (Aa,ma,ra), and eb = (Ab,mb,rb). If (m, M) are the minimum and maximum eigenvalues of A corresponding to orthogonal eigen vectors (m,M), the focal segment f1 and f2 is given by performing the distance vector operation on the segment. Euclidean distance between x and y. For one or more of the values from above distances, if the orthogonal projection of f for each foci f falls on the opposing focal segment then default distance can be replaced by this distance, otherwise minimum distance between f and two opposing foci can be used in distance calculation. The focal distance between ea and eb is the average of these four distances.

3.4 Ellipsoidal neighbourhood outlier factor

Here the outlier scoring mechanism is presented to identify outlying ellipsoids. Consider an ellipsoid ea E, where E is a set of hyperellipsoids, has a set of k nearest neighbour ellipsoids denoted by NNk(ea). A reachability distance RDk(ea,eb) of the ellipsoid ea from eb is defined as the maximum of the focal distance d(ea,eb) and the kd(eb), where kd(eb) is generalized focal distance of ea to the k nearest neighbour ellipsoid of eb. It means that the ellipsoids belonging to the k nearest neighbours of eb are considered to be at the same distance. Neighbourhood reachability density NRD(ea) of the ellipsoid ea can be defined as the distance at which ea can be reached from its neighbours i.e., reciprocal of the average reachability distance of ellipsoid ea from its neighbours, denoted as, N RD(ea) = 1/ 1 N Nk (ea) P eb N Nk (ea) RDk(ea,eb). By comparing the neighbouring reachability densities with those of the neighbours, the ellipsoidal neighbourhood outlier factor ENOF(ea) can be obtained as:

$$ENOF(ea) = \frac{\sum(ebNNk(ea)NRD(eb))}{NRD(ea) * NNk(ea)} \qquad (Equ : 3)$$

This is a ratio between the average neighbourhood reachability density of the neighbours and the ellipsoid's own neighbourhood reachability density. It can be inferred that ENOF(ea) becomes 1 when ea becomes comparable to its neighbouring ellipsoids, thus it is not an anomaly. This ratio becomes less than 1 when ea lies in a denser region. ENO(ea) becomes significantly higher than 1 for anomalous ellipsoids. Generally, a threshold of some higher value than 1 is required to declare an ellipsoid as anomalous depending on the dataset. For ENOF(ea) greater than TH, an ellipsoid ea will be considered as 'anomalous', and 'normal' otherwise, where TH greater than radius of shorter side NOF is the threshold, SDE NOF is the standard deviation of the ENOF scores for the set of ellipsoid E, and the parameter z= 1,2,3 determines the sensitivity of the detector. In particular, an ellipsoid that belongs to a dense group of ellipsoids, has a small outlier score than an ellipsoid that is far from this group of ellipsoids.

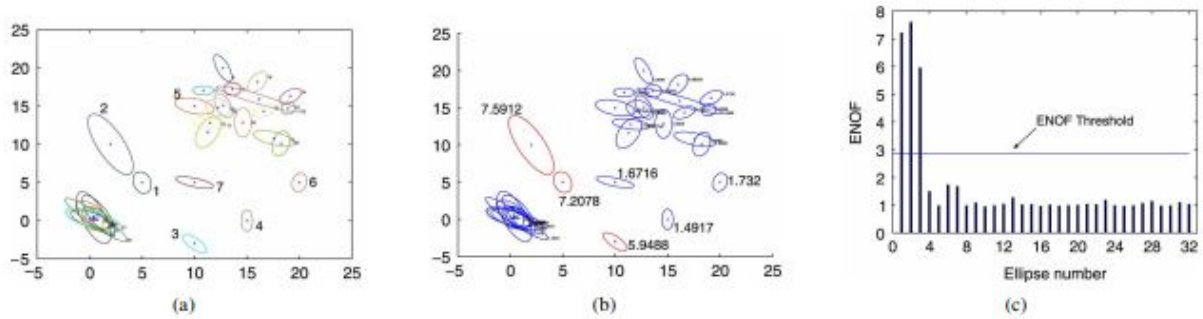


Fig. 3.1: (a) Ellipses with numbers, (b) Ellipses with ENOF values, (c) ENOF values: Threshold= 2.8653, for k = 5, z = 1.

In order to demonstrate the anomaly detection procedure, a synthetic dataset consisting of anomalous ellipses in between two groups of normal ellipses is created as depicted in the Fig. 3.1(a). A total of 32 ellipsoids are created, and the ENOF values obtained using k = 5 and z = 1 are shown in Figures 3.1 (b) and (c). Ellipses which have ENOF values more than threshold value of 2.87 (obtained using T H calculation equation) are marked in red.

3.5 Spatio-temporal estimation on historical and real-time data

We first evaluate the performance of our scheme on a large real-life sensor network (IBRL) dataset, and then demonstrate its applicability for real IoT dataset obtained from our deployment. Both centralized and distributed approaches of BME are implemented in our framework. In the centralized approach, all the computations are performed at the central location, whereas, in distributed (decentralized) approach, each node estimate its value using neighbourhood sensor measurement. For the evaluation of the estimation algorithms, we used the distributed approach of BME and Kriging at each node.

This is a publicly available WSN dataset consisting of around 2.5 million readings collected between February 28th 2004 and April 5th, 2004. This deployment consists of 54 sensor nodes measuring humidity, temperature, luminosity and battery level at a sampling interval of 31 seconds. The measurements from sensor node IDs 5,15,18 were noisy, hence were not considered in experiment. As temperature and humidity are slowly varying phenomenon, we

resample them at 10-minute interval by averaging them in 10-minute window. The experiments were carried out on a total of 1200 samples (around 8 days) of temperature values. As, we do not have the classification of high and low precision sensors for this dataset, we considered 10 sensor nodes distributed at even spatial locations as high precision sensors, and the remaining as low precision sensors. The spatial arrangements of hard and soft sensors for IBRL WSN deployment is shown in Fig. 3.2 (a). In order to make remaining sensor nodes as low-precision sensors, we preprocessed their measurements in such a way that their measurement lies within some interval I = [l,u], which means the measured values are of low precision and lies within the lower l and upper u limits with probability 1. A larger interval value corresponds to a higher measurement error, i.e., less precise measurement. Since the IBRL data do not have the information about the measurement errors, we added some uncertainty levels into the original measurements to make them as low-precision measurements, i.e., soft data. These levels were determined by the widths of the interval of data recordings lining soft sensors.

In this work, mean and covariance functions (as mentioned in Section III) were used as the general or the prior knowledge. The mean was computed assuming a uniform distribution. The next step is to compute the experimental variogram in order to obtain the covariance estimates at sample locations.

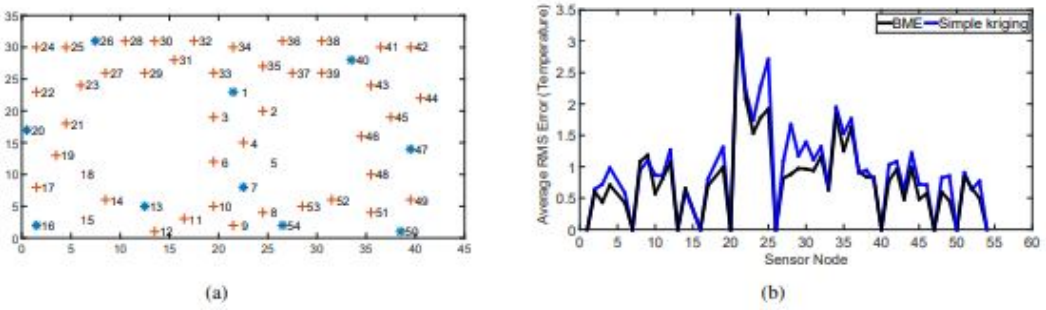


Fig. 3.2: (a) High-precision and low-precision data. (b) Average RMS error of all sensor nodes (IBRL dataset) for BME and simple kriging.

This process includes fitting a wide sense stationary (WSS), spatially isotropic, spatio-temporal covariance model to the real data in order to compute the variogram of the sensor measurements. Since we performed the estimation iteratively in time domain, only spatial

covariance model was used to fit the data. Based on the best fitting, the nugget-exponential model was chosen as the optimal variogram model, as depicted by the equation Cmap, where C0 is a constant (nugget) due to nugget effect which raises the whole theoretical semivariogram by C0 units, as is the distance at which samples become independent of each another, called as the range of a sample, and C is the sill value of Cmap (), at which the semivariogram graph levels off. The variogram parameters obtained from optimal fitting are sill equal to 0.587, range = 24.35 and nugget = 0. This variogram model is used in the BME for estimation. In the second experiment, all measurements from the weather station and five sensor nodes from the Fitzroy Gardens were considered as hard data. Fig. 3.2 (b) shows the comparison of the average RMS error for each sensor node over the entire duration of the experiment, for BME and simple kriging. It can be clearly seen that BME method provides the lower RMS error value for almost every sensor node. The average RMS error for BME and kriging based methods over all the sensor nodes for the entire duration of the experiment was 0.80 and 0.95 respectively. One point to note is that though the BME based method outperforms the Kriging based scheme, the difference between the average RMS error is small. The main reason was that this experiment was conducted in a controlled environment (office), hence the readings of all the sensors are pretty close to each other, which do not cause much difference in performance for both the methods. For outdoor IoT deployment, BME based algorithm is expected to perform much better than the Kriging method, where a higher variance between the measurements of each sensor is expected. Next, we discuss the experiments with outdoor deployment data.

The data collected from the deployment in the City of Melbourne were used for urban microclimate analysis. Both real-time and historical data of temperature, humidity and luminosity measurements collected between 14th December 2014 and 22nd May 2015 were used in the experiments.

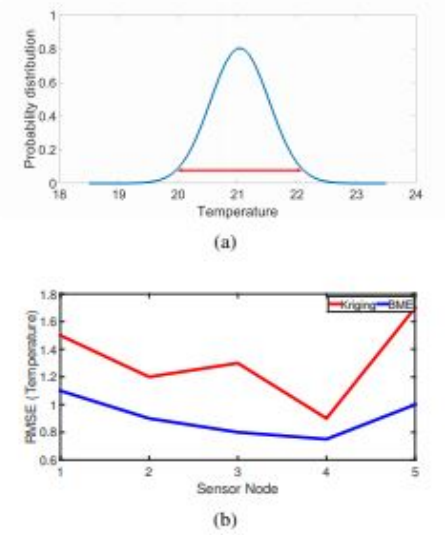


Fig. 3.3: (a) Gaussian distribution of sensor measurements (b) Performance of BME and Kriging data.

. A combined total of approximately 60,000 measurements have been collected during this period. Spatial estimation procedure is performed to generate a real exposure map of climate parameters over the complete study area for appropriate analysis. Two experiments were performed using temperature measurements collected from the sensors deployed in the Fitzroy Gardens during a week period starting from 17th December 2014 to 24th December 2014. In the first experiment, weather station's data (located nearby the Fitzroy Gardens) are considered as the hard data, and the measurements from the five deployed (low precision) sensors are considered as the soft data (of interval type). The measurement errors of (soft) sensor nodes can be used to derive the interval value for each of them. One of the ways to obtain this is to take multiple measurements at any location under the same environmental condition, and then fit a Gaussian distribution model to them. This will provide a means to obtain the measurement error (1.0 in the figure) as shown in Fig. 3.3 (a). Accordingly, the interval value for each soft sensor is obtained as 0.5, and used in subsequent computations. The values of variogram parameters cs and as were chosen as 0.42 and 20.3297 respectively. These variogram parameters were chosen such that they well represent the data within the given geographical area. In the second experiment, all measurements from the weather station and

five sensor nodes from the Fitzroy Gardens were considered as hard data.

As we do not have ground truth information at low-precision sensors' location, we assumed actual measured values as ground truth for evaluation. Then, these actual measurements were preprocessed into interval values (as mentioned for IBRL data) based on their measurement errors (0.5) to make them as soft data. Evaluation of the estimation algorithm was performed by comparing the ground truth value (actually measured) at any sensor location with the estimated value at that location using measurements of other sensor nodes in the neighbouring locations.

The root mean square error (RMSE) was computed to evaluate these estimates for two different scenarios: considering both hard and soft data and considering only hard data. In the later case, the BME scheme reduces to a specialized scheme known as the Kriging. The second approach is less computationally expensive due to only using hard sensor for data processing. Fig. 3.3 (b) shows the RMS values for both BME and Kriging operations. The RMS error is observed to be lower for BME estimation within the region. Thus, the BME was used to produce real-time spatial estimates at any arbitrary location, and incorporated into our visualization framework. When a user clicks at any point within the GIS map in our visualization application, the corresponding spatial location is fed to the estimation algorithm as an estimate location sE to compute the estimate XE using the hard data. i.e., [X1,X2,X3,..,XN]. As shown in Fig. 3.4 (c), the estimated value with variance can be seen on the screen as well as at the place where the mouse pointer is placed, where the sensor nodes are not present. An additional option to view the heat map is also provided in the visualisation framework. In this work, distributed anomaly detection algorithm

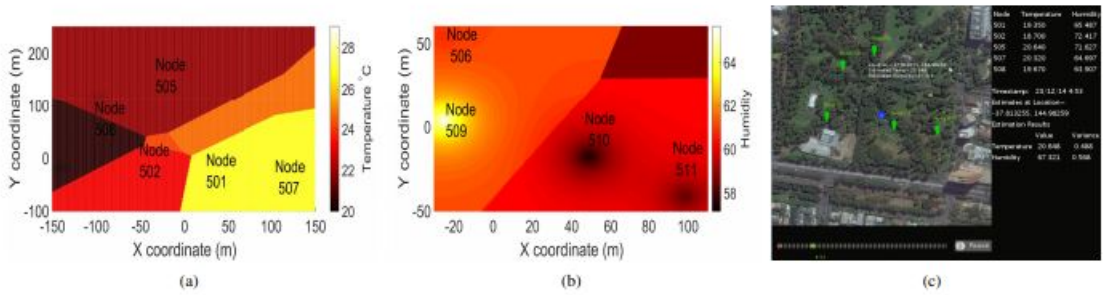


Fig. 3.4: e (a) heatmap of spatial estimat (b) heatmap of spatial estimate (c) an underlying GIS based framework.

among the five sensors nodes, while the region near node 501 and 507 was relatively warm because it is near to road (Clarendon street) and have less vegetation, the region near node 508 was coolest due to its deployment under the thick mature canopy cover (more shade). Fig. 3.4 (a) and (b) shows the estimation heat map of humidity over the region of the Docklands Library. It can be seen that node 509 shows more humidity (in yellow) as it is situated near the Yarra river, while node 510 shows lowest humidity as it is situated near the building and road pavement.

3.6 Anomaly detection

In this work, distributed anomaly detection algorithm (as explained in Section IV) was used on the time series data (temperature and humidity), collected from the sensors in the Docklands Library deployment, to automatically identify any unusual patterns in the data. This helps the Melbourne council's urban forest team to further analyse the microclimate around the canopy at those detected time periods. The data considered for this analysis were collected from nodes with node IDs 509, 510 and 511 between the periods 21st December 2014 and 11th January 2015, sampled at 10 minute intervals. Fig. 3.5 (a) shows the time series plot of the temperature and humidity values for each node. Each node data is first clustered using the hyperellipsoidal clustering scheme. The local anomalies can be found by performing the ENOF on the ellipsoids obtained at each node level. However, there is no local anomalies detected for

these three nodes in this experiment (the results are not shown for brevity).

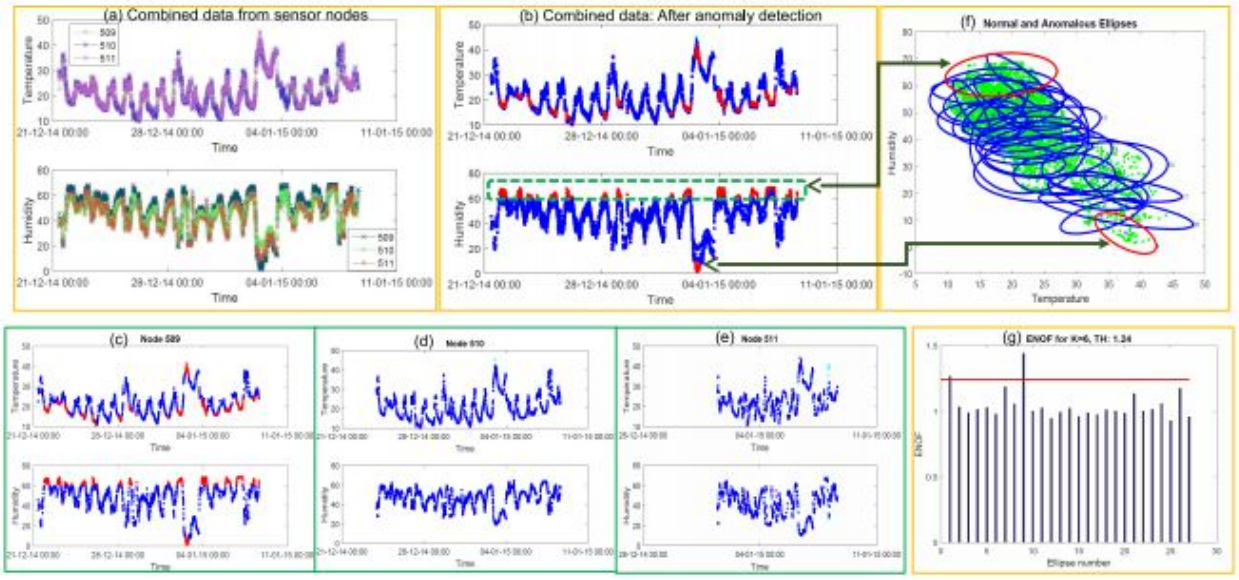


Fig. 3.5: (a) temperature and humidity data from node (b) labelled data of anomaly detection, (c) Node 509 data (d) Node 510 data (e) Node 511 data (f) scatter plot of all the node data (f) ENOF values of each ellipsoid and the Threshold TH (shown with a red horizontal line).

To find the global anomalies, measurements from all the nodes were considered. Hence, the clusters from all the nodes were combined to perform the distributed anomaly detection. The results are shown in Fig. 3.5. Fig. 3.5 (f) shows the scatter (in green) plot of the data from all the nodes and their corresponding ellipsoidal clusters. The blue ellipsoids are the normal ellipsoids and the red ones are the detected anomalous ellipsoids depiction (global anomalies). Fig. 3.5 (g) shows the ENOF values obtained for each of the ellipsoids. The threshold TH=1.24 used for this scheme is shown as a red horizontal line in that plot. The parameters used here are z = 2 and K = 6 (0.2 times of the total no of data points). Fig 3.5 (b) shows the labelled time series data (combined) after the anomaly detection is performed. The normal data vectors are shown in blue and the anomalous vectors are shown in red. Further, the anomalous data vectors that correspond to the anomalous ellipsoids shown in Fig. 3.5 (g) are highlighted using green arrows for easy identification. Figs. 3.5 (c), (d), and (e) show individual node data and their detected anomalies (in red). It can be observed that our algorithm correctly identifies the data vectors that are different from the other nodes'. In here, some of the node 509's

humidity values differ from those of the other two nodes (see Fig. 3.5 (a)). Note that, the same pattern was observed in the BME heat map of Docklands Library region, where the region near node 509 was bright yellow (more humid). This has been identified by the anomaly detection scheme automatically. This demonstrates the algorithm's ability to automatically identify the interesting regions in the data collected from several nodes in the monitored environment.

The estimation heat map and anomalies detected using our framework are being used by the city council's urban team to automatically identify the unusual environmental events, and then perform a focused and detailed analysis about how effectively the type of plant and the canopy cover density are responding to the unusual heating and cooling events in real-time. Further, they study the influence of the surrounding environment, such as the building, waterways and road sides, on the temperature and humidity changes in real time, over different seasons. Moreover, they aim to develop an app to assist pedestrians, community and ecology using these real time data. In the short-term, the detection of anomalous events, such as high temperatures, will assist in providing pedestrians with information such as the distribution of the temperature (or any variable of interest, such as pollution) over the region and help them move to a safer, cooler or shaded regions during extreme heat events. In the long-term, the analysis of the data will be used to make strategic decision about: increasing the canopy cover, increasing the forest diversity, improving the vegetation health, improving the soil moisture and water content and informing the community.

3.7 Complexities

In this section, we briefly analyse the computational complexities of implemented algorithms. There would be a computational overhead of O(n) for Kriging. The BME approach improves the estimation accuracy at the cost of computing multiple integrals for calculating multivariate Gaussian probability distribution function. In BME, first, regression is performed in O(n) to obtain primary estimate from hard data. Second, a multivariate cumulative distribution function (cdf) calculation is performed with a time complexity of O(2k(k1)!) [7] for k random variables (number of soft sensors considered in estimation). The third time consuming part is to find the minima of a single valued function within a range, which uses a

golden section search method with successive parabolic interpolation algorithm. The computational complexity to compute -accurate solution at a linear rate is given by O(log(1/n)). A distributed approach of BME and anomaly detection is implemented in our framework. As only neighbouring node measurements will be used for estimation, computational complexity can be reduced significantly. The distributed anomaly detection algorithm incurs a maximum computational complexity of O(n+h +4h+ k + 1). Furthermore, environmental physical variables do not change rapidly thus it reduces the need for fast sampling rate, and hence will not affect the performance of our framework in real-time IoT applications.

3.8 Multivariate visualization framework

An interactive, real-time geovisualization application was developed using open visualization software called Processing. A distributed approach of BME is designed in Processing to integrate spatio-temporal estimation with this visualization application to facilitate real-time visualization of spatial estimates. This application provides the real-time visualization of sensor data as well as visualization of spatial estimates at the location where sensors are not present physically. In addition, user can visualize the real-time data or historical data within some selected (past) time periods. Furthermore, user can visualize combinations of environmental attributes such as humidity, temperature, and luminosity for any particular sensor or pair of sensors simultaneously (to compare) with changing visual cues (brightness, blur, colour gradient) that depend on the attribute values, to understand and analyze the multivariate patterns and inter-relationships between these climate parameters.

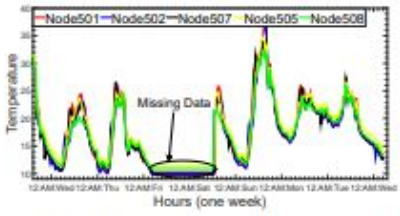
As each sensor node is surrounded by different tree species and different canopy covers, the real panoramic view of a chosen sensor node helps the user to relate the climate data with the knowledge of actual tree. The brightness of panoramic view changes as luminosity changes, which assists users to analyze the tree microclimate under different daylight conditions visually. Canopy cover density is shown visually using shadows of the animated tree, which changes its diameter depending on the tree canopy cover percentage and species. Fig. 8 demonstrates the GIS interactive visualization application for microclimate analysis of tree species. Fig. 3.6 shows the visualization for spatial estimates of multivariate data. A video demonstration of

these two visualization components in action can be seen from.

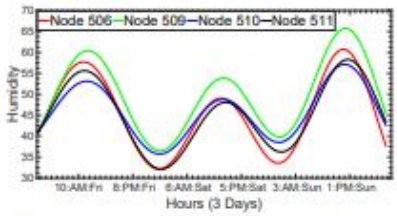
Fig. 3.6: Real-time interactive geovisualization of multivariate data.

The estimation heat map and anomalies detected using the framework are being used by the city council's urban team to automatically identify the unusual environmental events,and then perform a focused and detailed analysis about how effectively the type of plant and the canopy cover density are responding to the unusual heating and cooling events in real-time. Further, they study the influence of the surrounding environment, such as the building,waterways and road sides, on the temperature and humidity changes in real time, over different seasons. Moreover, they aim to develop an app to assist pedestrians, community and ecology using these real time data. In the short-term, the detection of anomalous events, such as high temperatures, will assist in providing pedestrians with information such as the distribution of the temperature (or any variable of interest, such as pollution) over the region and help them move to a safer, cooler or shaded regions during extreme heat events. In the long-term, the analysis of the data will be used to make strategic decision about: increasing the canopy cover,increasing the forest diversity, improving the vegetation health, improving the soil moisture and water content and informing the community. The given mean Xmap(p) and covariance functions of sensor measurements were used as general knowledge. The space-time variability of X is described in terms of a centered covariance function as: Cmap= E[(Xmap(p)] Space and time lags respectively, and show that this covariance is spatially isotropic and stationary in time. Hence, gn is adapted such that expectation E(g(Xmap)) defines ST mean and covariance functions through-

out the ST domain of interest. In this work, a nugget-exponential function [9] (discussed in Section V-A) is used to model spatio-temporal covariance structure known as variogram Some of the unusual temporal patterns identified by urban forest team using this framework are presented here. Interestingly, throughout the summer the minimum temperature during day and night was observed at two different locations. Fig. 3.7(a) shows the one week temperature measurements of all sensor nodes of Fitzroy Gardens. It can be observed that during the afternoon, the node 508 showed the lowest temperature (coolest location), while at night, the node 502 showed the lowest temperature. This is because the node 508 is deployed under a mature tree, which weakens the solar radiation (strong in noon) effectively, while the node 502 is surrounded by minimal canopies, which provide the cool air passages in the night. Similar type of urban microclimate patterns were observed. Similarly, interesting patterns were observed in humidity measurements at Docklands Library as shown in Fig. 3.7(b). Node 510 and 506 showed lowest humidity during the day and night respectively. This is because node 510 does not have any vegetation around it, and situated near the



(a) Temperature measurements of one week



(b) Humidity measurements (smooth) of three days

Fig. 3.7: Measurements of all sensor nodes deployed in (a) Fitzroy Gardens, and (b) Docklands Library

road pavements. Hence, during the day it shows less humidity (as compared to 506), attenuated due to strong solar radiation (high sky view), while it becomes relatively more humid

in the night due to its proximity to the Yarra river. The node 509 shows the highest humidity during day and night as it is very close to the Yarra river and situated near road pavements besides the Library building. The visual observations from our application reveal that the microclimate parameters are significantly influenced by the attributes of urban vegetation.

Conclusion

The analysis presents an integrated framework with detailed implementation of an IoT platform that aids in creating actionable knowledge. Bayesian Maximum Entropy based spatio-temporal estimation and hyperellipsoid based anomaly detection algorithm were used as backbone in the framework to address the three main challenges in urban microclimate analysis. Since, these challenges are same for many Smart City applications, the proposed framework can also be used to analyze other parameters of interest in a Smart City environment. The proposed framework also includes development of an interactive geovisualization tool to visualize spatio-temporal data with integrated algorithm outputs. On micro-level, the visualization assists in observing urban microclimate for different urban canopies under different daylight conditions.

Reference

[1] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of things (iot): A vision, architectural elements, and future directions "Future Generation Computer Systems, vol. 29, no. 7, pp. 1645 –1660, 2013",

[2] IEEE Internet of Things Journal, vol. 1, no. 2, pp. 112–121, 2014,". [Online]. Available: *J. Jin, J. Gubbi, S. Marusic, and M. Palaniswam*

[3] "Qut. Jour. of the Royal Meteorological Society, vol. 108, no. 455, pp. 1–24, 1982."

[4] "H. Akbari, M. Pomerantz, and H. Taha *https://www.w3.org/Submission/wot-model/* "Cool surfaces and shade trees to reduce energy use and improve air quality in urban areas , vol. 70, no. 3, pp. 295–310, 2004

[5] "Internet of things." *https://en.wikipedia.org/wiki/Internet_of_things* Accessed on: July. 18, 2017.

[6] Ashton, K "City of Melbourne (2014). Draft Urban Forest Strategy" 2012- 2032 , vol. 87, no. 3, pp. 210–222, 2008

[7] *"Characterising the urban environment of uk cities and towns: A template for landscape planning," Landscape and Urban Plang.,, vol. 97, no. 3, pp. 168–181, 2010. no. 3* Accessed on: July. 18, 2017.

[8] *A comparison of mesua ferrea l. and hura crepitans l. for shade creation and radiation modification in improving thermal comform* Accessed on: July. 18, 2017., vol. 47, pp. 256–271, 2012.

[9] L. Shashua-Bar, O. Potchter, A. Bitan, D. Boltansky, and Y. Yaakov. [Online]. Available: *"Microclimate modelling of street tree species effects within the varied urban morphology in the mediterranean city of tel aviv, israel," Int. journal of climatology* Accessed on: July. 18, 2017., vol. 47, pp. 186–271, 2012.

[10] "E. Ng, L. Chen, Y. Wang, and C. Yuan: *A study on the cooling effects of greening in a high-density city: an experience from hong kong* Accessed on: July. 18, 2017.vol. 52, no. 7, pp. 3823–3832, 2014.