Wireless Sensor Networks and the Statistical Sciences

Toscas, Peter J.
CSIRO Mathematics, Informatics and Statistics
Private Bag 33
South Clayton, VIC 3169, Australia
E-mail: Peter.Toscas@csiro.au

Garcia-Flores, Rodolfo
CSIRO Mathematics, Informatics and Statistics
Private Bag 33
South Clayton, VIC 3169, Australia

Lee, Dae-Jin
CSIRO Mathematics, Informatics and Statistics
Private Bag 33
South Clayton, VIC 3169, Australia

Abstract

Wireless sensor networks (WSNs) are increasingly being used in many application areas. These can range from environmental monitoring of outdoor natural processes to indoor environmental monitoring of buildings. WSNs are attractive because they are relatively cheap to deploy, but this comes at the price of some sever constraints. These constraints throw up a number of interesting statistical challenges. Here we give an overview of some of the challenges that WSNs pose for statisticians.

1. Introduction

Wireless sensor networks are increasingly being deployed for monitoring purposes for many different applications. Yick *et al.* (2008) breakdown the applications of sensor networks into two broad areas, tracking and monitoring. Tracking applications includes enemy tracking for military purposes, animals for habitat purposes, and traffic congestion. Monitoring applications include inventory monitoring for business purposes, security detection for military and security reasons, animal monitoring for the identification of changes in habitat, patient monitoring for health purposes, structural monitoring for the structural soundness of buildings, bridges and other important infrastructure, machine monitoring for preventative maintenance, and for environmental monitoring. In many of these application areas the big attraction for using sensors is the opportunity to greatly increase the temporal monitoring rate and the spatial extent of monitoring (Benson *et al.*, 2010).

One environmental monitoring example is the Springbrook study in South-East Queensland, where up to 200 sensor nodes will eventually be deployed to monitor the regeneration of the rainforest at an old winery farm site. The nodes are placed in three different areas: old forest, regenerating forest and open grassland. All the nodes monitor a number of variables such as air humidity, air temperature, leaf wetness, and soil moisture. Some of the nodes also monitor other variables such as accumulative hail, accumulative rain, air pressure, rain duration, rain intensity, soil water potential, solar power, wind direction, and wind speed. In addition, some nodes have been fitted out with audio and video sensors for the identification of animals in the local environment (Sensornets CSIRO, 2011). The purpose of the deployment of the network is to undertake a long-term study to identify the factors that contribute to rainforest regeneration and the impact on biodiversity (DERM, 2011). In Figure 1 an image of the Springbrook study area is displayed as are the locations of the currently deployed sensor nodes.

Another environmental example of a sensor network is the one in the Lake Wivenhoe area in South-East Queensland, where over 200 water, land and mobile sensor nodes have been deployed to monitor the movement of cattle, the local environment and weather conditions, and water quality (CSIRO ICT Centre, 2009). The sensors measure a number of variables, including air and water temperature, air humidity, turbidity, wind direction and speed, and the intensity of the sunlight. The aim of the project is to ensure that the drinking water quality for South-East Queensland meets the required standard.

The main difference between WSNs and traditional sensor network designs is that they suffer a number of key resource and design constraints that do not afflict the latter. These include limited storage and processing capability at the sensor nodes, low bandwidth for transmission of data, short communication range often necessitating multi-hop transmission of data to get to server for storage, sever limits on available energy for the node to run (Akyildiz *et al*., 2002, Garcia-Hernandez *et al.*, 2007, and Yick *et al.*, 2008).

These constraints raise a number of challenges for the statistical analysis of data collected using WSNs. In this paper we outline some of these challenges. In the next section we look at some of the constraints on WSNs and what that can mean for the data collection process for WSNs. In the following section we discuss some of the statistical research opportunities that the large volumes of data coming from WSNs provide. In the fourth section we look at some of the statistical issues around sensor network designs. We end the paper with some concluding remarks.

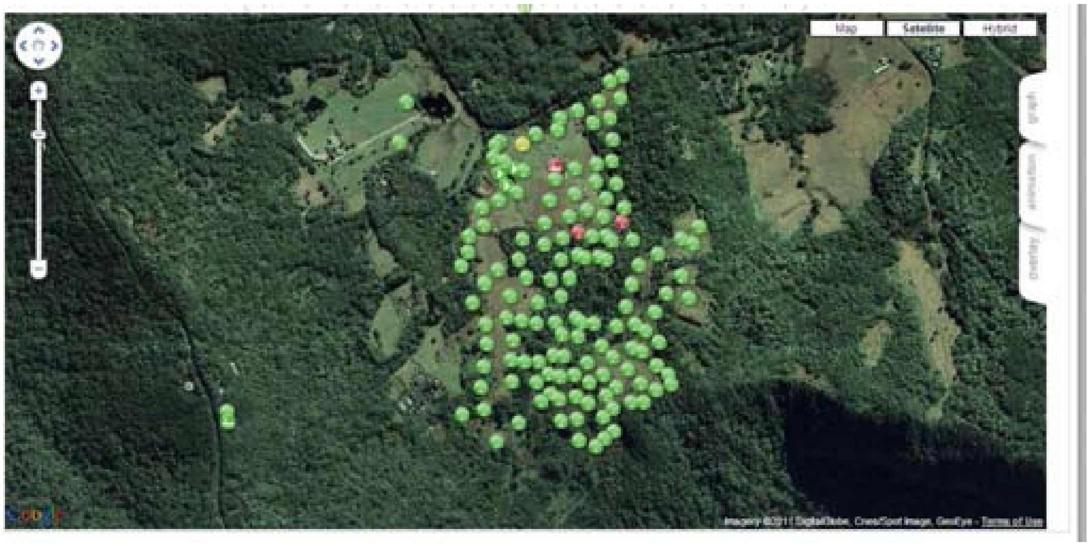
Map
Satellite
Hybrid
Imagery ©2011 DigitalGlobe, Cnes/Spot Image, GeoEye - Terms of Use

Figure 1. Google map image of the Springbrook sensor network with sensor node locations highlighted.

2. Do you know what type of data you are analyzing?

Each node in a WSN has to execute many tasks to collect data and to then transmit the data towards the sink for storage in a server. The node often has to be available to help data packets from nodes further out from the sink reach the sink by receiving the data and then forwarding it to sensors closer to the sink, or processing the received data with the data collected at the node and then forwarding the updated data to sensors closer to the sink. This often has to be done on limited energy resources because the nodes in WSNs are not connected to a power source and have to rely on battery packs with limited power or energy harvesting methods such as solar energy to recharge depleted batteries. In addition, to conserve battery energy the bandwidth for transmission of data is low, limiting the amount of data that can be sent through. The computer science and engineering communities are actively researching methods for optimizing the operation of sensor nodes to minimizing energy usage and thus extend the life of the node, and reliably deliver data packets to the sink with minimal error and lose of data (see e.g. Akyildiz *et al*., 2002, Garcia-Hernandez *et al.*, 2007, and Yick *et al.*, 2008, Buratti *et al*., 2009, and Rosberg *et al*., 2010). These computer science and engineering issues will not be directly discussed further in this paper other than to say that it is a fertile field for those interested in operations research type problems, especially around joint optimization of different components of the operation of sensor nodes and the sensor network.

From a statistical perspective the main interest is how these constraints on WSNs affect the data collection process and what is eventually returned to the sink for storage and data analysis. One approach that can be used is to forward to the sink all the raw data collected. This is the case in the Springbrook study, where sensor nodes take readings every 15 minutes and then forward these to the sink via multi-hop communication. In such a case, data analysis is relatively straight forward from the perspective that so long as there is no major breakdown in the network the data collected comes in regularly and most of the sensor nodes are providing data from their immediate vicinity, thus giving both spatial and temporal coverage. This, however, is not always the case. In many sensor networks data forwarded to the sink may actually be aggregated data (Greenwald and Khanna, 2004 and Yick *et al.*, 2008). This could be in the form of the maximum, minimum, mean, sum or some other summary measure of a number of observations at one sensor node or at multiple sensor nodes, resulting in energy savings since fewer data packets are forwarded to the sink. In such a case supplementary information would be helpful in analyzing the data. For example, the number of observations that make up the aggregated summary values (and the mean and variance, both within a sensor and between sensors, if these are not summary statistics being forwarded back to the sink), and from which sensor nodes the observations came from if the aggregated values summarize observations from more than one sensor. (The mean and variance, both within a sensor and between sensors, for the aggregated values should also be forwarded to the sink if these are not in the set of default summary values being transmitted to the sink.) This information may not always be collected or a query has to be sent to the network to ensure the information is collected. In a self-organizing network the pattern of which sensor nodes “cluster” together to form aggregate values may change as nodes fail, or as the energy availability at each node changes the network may re-organize to minimize energy loss (Culler *et al.*, 2004).

Aggregation raises the issue of quality assurance and quality control (QA/QC) for the data collected. When all the raw data are forwarded back to the sink methodologies for indentifying erroneous or suspect observations can be used to examine the data. Due to the spatio-temporal nature of the data this could involve looking for a combination of temporal, spatial, and multivariate consistency in the data (Sparks and Okugami, 2011). Temporal consistency checks involve examining the time-series of data for a variable from one node. Spatial consistency checks involve examining the data for a variable from a number of geographically neighboring sites, and multivariate consistency checks is when a number of variables collected at the same node and at the same times are examined for consistency. Joint temporal, spatial and multivariate consistency checks are also possible (Sparks and Okugami, 2011). For aggregated data being sent to the sink, the impact of the inclusion of a small number of erroneous observations in the aggregated statistic will be mitigated by the non-erroneous observations in the aggregated statistic, but if the aggregated value being sent to the sink is not a robust measure such as a median or trimmed mean, the aggregated value is likely to be corrupted by the inclusion of the erroneous observations. This means that it is important for QA/QC capabilities to be deployed at the sensor nodes when aggregated data is sent back to the sink. This capability has to be simple but robust as the computational capability at the sensor node is limited and energy needs to be conserved.

Other energy and communication savings approaches can also impact the quality of the data. Although the Sprinkbrook sensors nodes forward the raw data, to save energy the time that the observations are recorded is not forwarded. The time stamp in the database where the data are ultimately stored is the time that the data arrived at the sink. Often this is not a serious problem when the network is working well as the data collected arrives at the sink within seconds, but if there are delays in the transmission of data to the sink due to node failures or other problems in the network, the time stamp recorded in the database may be very different to the time that the observations were actually recorded. One simple solution is for the sensor node to increment by one an integer variable every time it is scheduled to take recordings. This will help with chronologically ordering the observations if delays in the system results in a breakdown in the sequence in which observations arrive at the sink. This approach, however, does not handle clock drift at the sensor node.

3. Large data sets

Sensor networks can generate lots of data in a short time. For example, in the Springbrook study sensor nodes record every 15 minutes, which means that if there are 175 sensor nodes finally deployed, with at least five variables recorded at each sensor node every time measurements are taken, then at least 84,000 observations will be recorded and transmitted to the sink each day or over 30 million in a year (assuming no network problems). One approach to handling such large data is to reduce the data size by aggregating as discussed in the previous section. Another approach is to use compressive sampling (or sensing) techniques such as wavelets to reduce the volume of data needed to be sent to the sink for the signal to be recreated (Masiero *et al.*, 2009, and Yick *et al.*, 2008). It would be helpful to build on the work of Fuentes *et al.* (2006) to develop spatio-temporal wavelets methodology for the analysis of sensor network data compressed using wavelets. The attraction of spatio-temporal modelling using wavelets is that it can naturally handle non-stationarity.

The recording of audio sound or video can dramatically increase the data volumes making it infeasible for most WSNs to transmit this information to the sink due to energy and communication bandwidth constraints. The Sprinkbrook study is exploring the use of bio-acoustic and video monitoring methodology for the identification of specific animal species (DERM, 2011). This can be done by providing greater electronic storage space at each sensor node for the audio or video recording to be stored until it is physically downloaded onto laptops at regular intervals. For other applications, such as security and intrusion detection, this is unsatisfactory. For these applications it is more desirable to have real-time event detection and pattern recognition methodology at the sensor node or distributed across a number of sensor nodes. This will offer the best chance of quickly identifying any threats. This entails the development of event detection and pattern recognition methods that can work in highly constrained computational environments and can work in a distributed way to garner more computational resources from surrounding sensor nodes.

The need for real-time monitoring at the node or distributed across a number of nodes could be useful for environmental studies as well. For example, in the Springbrook study continuous audio or video recording will quickly consume the electronic storage space and deplete the energy resources of the sensor nodes. There is a need for event detection to identify if there is an animal in the vicinity, and then pattern recognition to identify if it is a species that is of interest. If so, then recording can begin. Another reason for real-time monitoring in environmental studies is to monitor the energy stores in sensor nodes and in the network, to help regulate its operation by minimizing energy usage (Basha *et al.*, 2011).

Real-time monitoring in sensor networks offers opportunities for research into spatio-temporal data assimilation methodologies. One of the tasks of the sensor network in the Lake Wivenhoe project is to validate hydrodynamic models that model the 3-D movement of water in the system (CSIRO Smart Sensors, 2010). In a major flood event, or if contaminants enter the system either by accident or deliberately due to terrorist or criminal activity, data assimilation methodology such as the ensemble Kalman filter (EnKF) (Evensen, 2003) or particle filter (PF) (Doucet *et al.*, 2001) can be used to combine observational data from the sensor network with the hydrodynamic model to forecast the 3-D movement of the pollutants or contaminants in the water system, thus helping decision makers make better informed decisions. In studies where the sensor network covers a very large area, assimilating sensor network data with remote sensing data and process models may be beneficial, with the remote sensing data providing the spatial observation coverage while the sensor network provides the temporal observation coverage. The sensor network observations can be used to validate the remote sensing information.

4. Sensor network design

As noted previously, WSNs are constrained by energy, communication bandwidth, transmission range, and storage capacity at the nodes. This means that when designing a monitoring sensor network these constraints need to be taken into account, in addition to the normal statistical issues considered in designing monitoring networks. It is not enough to base sensor network designs on optimizing spatial covariance estimation (Zhu and Stein, 2005) or some function of the spatial prediction variance (Sacks and Schiller, 1988, and Cressie, 1993). Any optimization needs to account for these constraints to minimize energy consumption, otherwise repair and energy source replacement costs can escalate, and the probability of successfully transferring data to the sink may be too low thus compromising the data collection process and hence statistical inference. Krause *et al.* (2011) have developed methodology for near optimal selection of sensor node locations based on communication costs and on prediction uncertainty. They use Gaussian processes for spatial prediction and modelling the spatial variability of the communications link quality. This work needs to be extended to take account of multiple criteria, such as available energy resources, communication bandwidth and transmission range, which may be dependent on the environment in which the network is to be deployed. For example, dense foliage or a hill near a sensor node may restrict the transmission distance in certain directions. Not accounting for this in the sensor network design may result in little data from this sensor node getting back to the sink. Extending this work further for the prediction and estimation of multivariate data adds another layer of complexity.

Accounting for the energy constraints will require network designs that can tolerate adaptation in the network. As the energy reserves in some sensor nodes deplete over time, to prolong the life of the network, these sensor nodes will have to sample and transmit information less regularly. This will impact the spatial sampling coverage at any given time and the temporal coverage at sensor node locations. There will be a need for network design methodology that can adaptively change the configuration of which sensor nodes sample and when they sample to minimize prediction or estimation uncertainty under a changing regime of energy resource availability in the network. An associated issue is the addition to, or removal from the network of sensor nodes. Recently research effort has gone into developing methods for the addition or removal of sites from a monitoring network (e.g. Arbia and Lafratta, 1997, Fuentes *et* al., 2007, and Ainslie *et al.*, 2009). These methods will need to be extended or new methods developed to account for the various constraints on the operation of WSNs.

Related to the need for the sensor network design to be adaptive is the issue of sensor network designs being robust to changes in the environment. Krause *et al.* (2011) look at a simplified version of this design robustness issue in which it is assumed that after a period of monitoring, the environment being monitored changes. They propose optimizing the sensor network design over a number of environmental scenarios, so that the sensor network design is robust to changing environments. Research is needed in identifying sensor networks designs that are robust to environmental scenarios not considered previously in scenarios during the planning phase of the sensor network.

WSNs may also be characterised by the existence of mobile sensors in the network (Yick *et al*., 2008). The Lake Wivenhoe study above is an example of a sensor network that uses a combination of stationary and mobile sensors. An autonomous catamaran is deployed to undertake a number of tasks, including calibrating the stationary sensors, ascertaining if anomalous readings from stationary sensors are valid, and measuring relevant environmental variables (CSIRO Smart Sensors, 2010). The measuring equipment on the catamaran is more accurate and precise than those on the stationary sensors. Developing algorithms for optimising sensor network designs where stationary and mobile sensors exist poses a number of interesting research challenges. These include choosing the spatio-temporal path of the mobile sensor that will maximize information gain (Singh *et al*., 2010), but doing this in a way that accounts for the less precise and less accurate stationary sensor network, and factors in the maintenance and repair duties the mobile sensor may also be required to perform. This is a complex multiple criteria optimisation problem.

5. Concluding remarks

In this paper we have discussed some of the statistical research issues posed by WSNs. WNSs are characterized by a number of constraints on their operation that need to be accounted for when designing the placement of the nodes in the WSN. Accounting for these constraints while minimizing estimation or prediction uncertainty, raises a number of interesting optimization challenges. The collection of large data and the need to transmit all or some summary of this data by WSNs necessitates a number of compromises that require fast and efficient statistical methodology that can be executed in computing environments with limited electronic storage and processing capability.

In this paper we have predominantly focused on the immediate needs for collecting and analysing sensor network data, particularly around univariate spatial and spatio-temporal analyses. Given that sensor nodes often will be collecting data on multi variables, the challenges listed previously become more so as one thinks about undertaking real-time multivariate spatio-temporal analyses of sensor network data for real-time or near real-time decision making. Using a batch approach to analyse the data is not suitable, as new data will be arriving at regular intervals (Domingos and Hulten, 2003). The extensions of data assimilation methods such as the EnKF and PF to the multivariate spatio-temporal data setting will be important, as will be evaluating the performance of the EnKF and PF as other data assimilation methods may need to be developed if the EnKF and PF do not work well for the multivariate spatio-temporal setting.

Research into data streaming tools will be required to help quickly quantify and visualise changing spatial and multivariate associations over time. Current spatio-temporal modelling approaches for estimating associations do not work in real-time, at least not at the time scales at which some sensor networks can record observations and transmit them for analysis. To work in real-time new multivariate spatio-temporal tools will need to efficiently exploit sparsity in multivariate space. New methods for estimating partially missing data, such as censored observations, will be required, as the Expectation-Maximisation and Markov chain Monte Carlo algorithms are too slow to work in real-time. Recent advances in the development of approximate Bayesian methods (Rue and Martino, 2007 and Eidsvik *et al*., 2009) and the use of sparse matrix methodology may offer computational advantages to considerably speed up convergence times compared to MCMC analyses. The extension of approximate Bayesian approaches to dynamical systems (Toni *et al*., 2009) offers opportunities for extending these methods for data assimilation purposes.

Acknowledgement

The authors would like to thank Dr Ross Sparks for reviewing the paper and providing useful insights that have helped to improve it.

REFERENCES

[1] Ainslie, B., Reuten, C., Steyn, D.G., Le, N.D. and Zidek, J.V. (2009). Application of an entropy-based Bayesian optimization technique to the redesign of an existing monitoring network for single air pollutants. *Journal of Environmental Management*, **90**, 2715 – 2729.

[2] Akyildiz, I.F., Su, W., Sankarasubramaniam, Y. and Cayirci, E. (2002). Wireless sensor networks: a survey. *Computer Networks*, **38**, 393 – 422.

[3] Arbia, G. and Lafratta, G. (1997). Evaluating and updating the sample design in repeated environmental surveys: monitoring air quality in Padua. *Journal of Agricultural, Biological, and Environmental Statistics*, **2**, 451 – 466.

[4] Basha, E., Cotillon, A., Jurdak, R. and Rus, D. (2011). Scalable solar current prediction. Submitted to *ACM Sensys'11*.

[5] Benson, B.J., Bond, B.J., Hamilton, M.P., Monson, R.K., and Han, R. (2010). Perspectives on next-generation technology for environmental sensor networks. *Frontiers in Ecology and the Environment*, **8**, 193 – 200, doi:10.1890/080130.

[6] Buratti, C., Conti, A., Dardari, D, and Verdone, R. (2009). An overview on wireless sensor networks technology and evolution. *Sensor*, **9**, 6869 – 6896; doi:10.3390/s90906869.

[7] Cressie, N.A.C. (1993). *Statistics for Spatial Data*. Revised edition. Wiley-Interscience: New York.

[8] CSIRO ICT Centre (2009). http://research.ict.csiro.au/research/labs/information-engineering/ie-lab-projects/envoronmental-sensing. Last accessed 5 May, 2011, at 12:38 pm.

[9] CSIRO Smart Sensors (2010). http://www.csiro.au/science/smart-sensors-monitoring-water-quality.html. Last accessed 9 May, 2011, at 12:39 pm.

[10] Culler, D., Estrin, D. and Srivastava, M. (2004). Overview of sensor networks. *Computer*, **37**, 41 – 49, doi:10.1109/MC.2004.93.

[11] Department of Environment and Resource Management (DERM) (2011). Springbrook Wireless Sensor Network: Information Sheet. http://www.derm.qld.gov.au/register/p02588aa.pdf. Last accessed 5 May, 2011, at 11:59 am.

[12] Domingos, P. and Hulten, G. (2003). A general framework for mining massive data streams. *Journal of Computational and Graphical Statistics*, **12**, 945 – 949, doi:10.1198/1061860032544.

[13] Doucet, A., de Freitas, N. and Gordon, N. (2001). *Sequential Monte Carlo Methods in Practice*. Springer-Verlag: New York.

[14] Eidsvik, J., Martino, S. and Rue, H. (2009). Approximate Bayesian inference in spatial generalized linear mixed models. *Scandinavian Journal of Statistics*, **36**, 1 – 22, doi:10.1111/j.1467-9469.2008.00621.x.

[15] Evensen, G. (2003). The ensemble Kalman filter: theoretical formulation and practical implementation. *Ocean Dynamics*, **53**, 343 – 367, doi:10.1007/s10236-003-0036-9.

[16] Fuentes, M., Chaudhuri, A. and Holland, D.M. (2007). Bayesian entropy for spatial sampling design of environmental data. *Environmental and Ecological Statistics*, **14**, 323 – 340, doi:10.1007/s10651-007-0017-0.

[17] Fuentes M., Guttorp, P. and Sampson, P.D. (2007). Using transforms to analyze space-time processes. In *Statistical Methods for Spatio-Temporal Systems*, B. Finkenstädt, L. Held, and V. Isham, eds. Chapman & Hall/CRC: Boca Raton.

[18] Garcia-Hernández, C.F., Ibargüengoytia-González, P.H., Garcia-Hernández, J. and Pérez-Díaz, J.A. (2007). Wireless sensor networks and applications: a survey. *IJCSNS International Journal of Computer Science and Network Security*, **7**, 264 – 273.

[19] Greenwald, M.B. and Khanna, S. (2004). Power-conserving computation of order-statistics over sensor networks. *PODS* June 14 – 16, 2004, Paris, France.

[20] Krause, A., Guestrin, C., Gupta, A. and Kleinberg, J. (2011). Robust sensor placements at informative and communication-efficient locations. *ACM Transactions on Sensor Networks*, **7**, No. 4, Article 31, doi:10.1145/1921621.1921625.

[21] Masiero, R., Quer, G., Rossi, M. and Zorzi, M. (2009). A Bayesian analysis of compressive sensing data recovery in wireless sensor networks. *Proceedings of the International Conference on Ultra Modern Telecommunications*, October 12 – 14, 2009, St. Petersburg, Russia.

[22] Rosberg, Z., Liu, R.P., Dinh, T.L., Dong, Y.F. and Jha, S. (2010). Statistical reliability for energy efficient data transport in wireless sensor networks. *Wireless Netw*, doi:10.1007/s11276-009-0235-5.

[23] Rue, H. and Martino, S. (2007). Approximate Bayesian inference for hierarchical Gaussian Markov random field models. *Journal of Statistical Planning and Inference*, **137**, 3177 – 3192.

[24] Sacks, J. and Schiller, S. (1988). Spatial designs. In *Statistical Decision Theory and Related Topics IV, Vol. 2*, S.S. Gupta and J.O. Berger, eds. Springer: New York, 385 – 399.

[25] Sensortnets CSIRO (2011). http://150.229.98.68/deployments/63. Last accessed 5 May, 2011 at 11.07 am.

[26] Singh, A., Ramos, F., Whyte, H.D. and Kaiser, W.J. (2010). Modeling and decision making in spatio-temporal processes for environmental surveillance. *IEEE International Conference on Robotics and Automation*, May, 3 – 8, 2010, Anchorage, Alaska, USA.

[27] Sparks, R.S. and Okugami, C. QA/QC of measurements collected on very large scale: rainfall and streamflow datasets. ISI August, 22 – 26, 2011, Dublin, Ireland

[28] Toni, T. Welch, D., Strelkowa, N. Ipsen, A. and Stumpf, M.P.H. (2009). Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, **6**, 187 – 202, doi:10.1098/rsif.2008.0172.

[29] Yick, J., Biswanath, M. and Ghosal, D. (2008). Wireless sensor network survey. *Computer Networks*, **52**, 2292 – 2330.

[30] Zhu, Z. and Stein, M.L. (2005). Spatial sampling design for parameter estimation of the covariance function. *Journal of Statistical Planning and Inference*, **134**, 583 – 603.