

## ARIMA based Value Estimation in Wireless Sensor Networks

A. Amidi

Faculty of Geo-Information Science and Earth Observation, University of Twente, Enschede, The Netherlands.  
Amidi29183@itc.nl

**KEY WORDS:** Temporal correlation modelling, data smoothing, sample size, life span, outlier detection, energy efficiency

### ABSTRACT:

Due to the widespread inaccuracy of wireless sensor networks (WSNs) data, it is essential to ensure that the data is as complete, clean and precise as possible. To address data gaps and replace erroneous data, temporal correlation modelling can be applied, which takes advantage of temporal correlation and is also energy efficient. In this research, the suitability of adapting the ARIMA model into a WSN context is scrutinized, as technological requirements demand special considerations. The necessity of applying a smoothing technique is explored and the selection of an appropriate method is determined. Additionally, the available options with regards to ARIMA set-up are discussed, in terms of achieving accurate and energy friendly predictions. The effect of sufficient historical data and the importance of predictions' life span on the estimation accuracy are additionally investigated. Finally, an adaptive, online and energy efficient system is proposed for maintaining the accuracy of the model that simultaneously detects outliers and events as well as substitutes any missing or erroneous data with estimated values.

### 1. INTRODUCTION

A wireless sensor network (WSN) typically consists of wireless devices that are able to sense a wide range of attributes and variables. These instrumental readings are then transmitted over a wireless channel. These instruments can potentially provide information with high spatial and temporal resolution, which is a key feature of their existence. The quality of the data, however, may be affected by noise and error, missing values, duplicated data or inconstant data. Due to harsh condition of a deployment environment, packet loss, collisions and low sensor battery levels (Elnahrawy and Nath, 2003), not all sensor readings can be successfully gathered simultaneously and some readings are lost altogether.

The communication capability in wireless sensor networks is limited due to energy and cost considerations. The impact of the surrounding environments, such as mountains and obstacles, may cause temporary isolation, which can result in a loss of data. Additionally, the sensors' communication quality may be affected via natural events including rain, thunder, lightning and so on (Amidi et al., 2013). Consequently, the transmission links among sensor nodes may connect and disconnect. WSNs are known as low-power systems, but when a sensor's power is low one may expect unstable records. The low level of power may additionally lead to data loss or/and data abnormalities and errors.

The quality of data provided by WSNs is highly critical, while raw data may be of a lower quality and less reliable, due to the nature of the sensors (Amidi et al., 2013). Limited numbers and low quality of WSN resources, as well as harsh deployment environments, lack of memory capability, computational capacity, and computational bandwidth can all cause unstable data (Zhang, 2010). Thus, observations may include absolute errors, clustered absolute errors, random errors, long-term errors (Zhang, 2010), dead band errors and systematic errors. Consequently, the low quality data should be replaced with adequate estimations when possible. A sensor node, regularly

sensing local observations, can fit a prediction model to the real data-set and then apply the model to estimate the missing or future values. On the one hand, the model should be up to date in order to represent precise and accurate predictions; on the other hand, careful attention should be paid to the model's energy efficiency. The model can be updated if the current measurement differs from the predicted measurement by more than a pre-defined tolerance, thus avoiding unnecessary energy consumption. Typically, the applied model is fixed in advance, since the model parameters are estimated on the basis of incoming data (Santini and Romer, 2006; Tulone and Madden, 2006).

The abovementioned problems are somewhat inevitable due to the inherent characteristics of WSNs. Thus, in an effort to ensure a high quality of service for a WSN, techniques should be made to withstand and combat said undesirable incidents, which will also improve the overall quality of the information.

In this paper an appropriate statistical model, the autoregressive integrated moving average (ARIMA) model, is applied, using real WSN data-set that consists of Grand-St-Bernard nodes. The current research is conducted to investigate the effects of a number of parameters on ARIMA modelling in the context of WSNs including: the necessity of applying smoothing techniques for ARIMA modelling in WSNs, assigning proper methods of smoothing by considering the technological requirements, adequate settings of ARIMA model parameters, the role of sample size in WSNs modelling and optimum prediction life span (prediction age) with respect to specifications of technology and application in WSNs.

### 2. RELATED WORKS

Many efforts have been made to recover missing data by using spatial information of neighbour nodes. For example Sheikhhasan (2006) and Collins (1995) discussed temperature interpolation with the help of spatial correlations. Generally speaking, spatial correlation for data interpolation and missing

data recovery can be investigated by applying inverse distance weighted averaging (IDWA) and Kriging (Guo et al., 2011; Umer et al., 2008, 2010; Zhang et al., 2012). Additionally, the majority of outlier detection methods that take the advantage of spatial information can be employed to fill in missing values (Cheng, 2008; Wang and Cheng, 2008; Wang and Yu, 2005; Wu et al., 2007; Zhang et al., 2008).

Generally, assuming spatially correlated attributes in real world application of adjacent nodes is oversimplified, especially in the context of WSNs, as they are deployed in harsh environments. Moreover, spatially correlated nodes may lead to errors that are also correlated, and thus utilizing adjacent information is not always a safe option. Alternatively, the main source of energy consumption in WSNs is communication. Utilizing neighbour information requires effective data communication, which may lead to a considerable decrease of network life and subsequent errors. In addition, they typically consider data within a single span and as a result do not benefit from information in the sequential data span.

Several attempts have been made to adopt temporal data for data estimation. Liu et al. (2005) proposed a method that employed the ARIMA model as a way to construct a prediction model for sampled data. Specifically, their proposed method ran both on sensor nodes and at the base station. While the difference between the values sampled on the sensor nodes and those values predicted by the ARIMA model were smaller than a pre-defined tolerance, the values were not transmittable to the network at the base station. While the base station was running the same model, it was using the predicted values as actual values. Singh et al. (2011) utilized ARIMA modelling to locate anomalies within a stream of data for a single node and to correct anomalous data by appropriating forecast values.

### 3. DATA

The Grand-St-Bernard WSN deployment was utilized as a real data-set to perform the experiments in this research. The frequency of the sampling for the deployment was two minutes. The proposed methodology was developed and evaluated on the basis of an ambient temperature, for which the precision of the correspondent sensor was  $\pm 0.3^\circ\text{C}$ . The period of 23:54:59 to 16:02:00 on 2007-09-28 was used to build the temporal model. Moreover, the temperature on 16:04:00 to 18:00:00 from the same day was used as reference data for validation of the methodology.

### 4. BASIC IDEA

In order to estimate and replace the missing values and error related outliers, temporal correlation modelling was performed. An inadequate choice of a prediction model naturally results in poor prediction performance (Le Borgne et al., 2007). An existing correlation among an attribute sensed by WSNs is acknowledged by applying temporal correlation based methods for value estimation.

#### 4.1 ARIMA models

A time series is a chronological sequence of measurements of a particular attribute. Auto-regressive integrated moving average (ARIMA) models establish a powerful class of models, which can be applied to many real time series. ARIMA models are based on three parts: (1) an autoregressive component, (2) a

contribution from a moving average and (3) an element involving the first derivative of the time series.

The auto-regressive (AR) component of the model originates in the theory that individual time series values can be described by linear models based on preceding observations. The general formula for describing AR models is represented by Equation 1, where the order of the model is determined by  $p$ :

$$X(t) = \sum_{i=1}^p \alpha_i X(t-i) \quad (1)$$

The fact that time series values can be expressed as dependent on the preceding estimation errors, leads to moving average models (MA models). Past estimations or forecasting errors are taken into account when estimating subsequent time series values. The difference between the estimation  $\hat{x}(t)$  and the actually observed value  $x(t)$  is denoted  $\varepsilon(t)$ . The general description of MA models is shown by Equation 2.

$$X(t) = - \sum_{i=1}^q \beta_i \varepsilon(t-i) \quad (2)$$

Combining both AR and MA models forms ARMA models. In general, forecasting with an ARMA ( $p,q$ ) model can be described through Equation 3.

$$X(t) = \sum_{i=1}^p \alpha_i X(t-i) - \sum_{i=1}^q \beta_i \varepsilon(t-i) \quad (3)$$

Chatfield (2013) identified the three major steps of time series analysis as: (1) removing the trend and seasonality, (2) fitting an auto-regressive moving average (ARMA) model to the time series and (3) predicating future values using the ARMA. Seasonal effects represent systematic and calendar related properties of the variable. Thus adjustments are made by estimating seasonal effects and then removing them from a given time series. The data need to be seasonally adjusted to uncover the substantive underlying movement in the series, as well as to identify certain non-seasonal characteristics, which may be of interest to analysts (Australian Bureau of Statistics, 2005). When a time series is dominated by a trend or irregular components, it is almost impossible to identify and remove what little seasonality is present, thus seasonally adjusting a non-seasonal series is impractical and will often introduce an artificial seasonal element (Australian Bureau of Statistics, 2005). Alternatively, the trend can be defined as “long term” movement in a time series and can prove to represent the underlying inclination. To achieve a stationary series, trends and seasonality must be accounted for and subsequently removed.

The ARMA or any set of it can be used to predict the future values. ARMA modelling functions through a series of well-defined steps. The first step involves identifying the model. Identification consists of specifying the appropriate structure (AR, MA or ARMA), as well as the order of model. The second step involves estimating the coefficients of the model.

## 4.2 Identifying the numbers of AR and/or MA terms

After a time series has been stationarized, the next step in appropriating an ARIMA model is to determine whether AR or MA terms are needed to correct any autocorrelation that remains in the stationarized series. By examining the autocorrelation function (ACF) and the partial autocorrelation (PACF) plots of the series, one can empirically identify the AR and/or MA terms. An ACF plot is merely a bar chart of the coefficients of correlation between a time series and the time lags. The PACF plot is an illustration of the partial correlation coefficients between the series and the time lags.

## 4.3 Evaluation methods

To evaluate the proposed methodology, an algorithm was implemented in R software (Ihaka and Gentleman, 1996) and the performance was assessed using real sensor data across various scenarios. In the context of WSNs, energy efficiency and accuracy are critical for performance evaluation, given the technological requirements. The metric that is used to evaluate the method performance is leave-one-out cross validation.

For the time-series model, each observation was predicted using its ARMA model. The differences between measured and predicted values are called errors. The mean prediction error (MPE) and the root mean square error (RMSE) are the two main metrics of cross validation that are utilized in this study. The MPE is an indication of bias and RMSE is a measure for accuracy.

## 5. RESULTS

Temporal correlation among consecutive records of the ambient temperature was used for value estimation to replace missing values and error based outliers.

### 5.1 Data smoothing

Figure 1 illustrates the errors and imprecise observations that are represented by small fluctuations. Imprecise properties of WSN data may result in an unreliable temporal model and consequently, incorrect future value predictions. Additionally, those small fluctuations that do not represent informative information can be ignored using smoothing techniques on the original time series. Figure 1 explicates the effects of smoothing with exaggeration for better understanding via the highlighted line. Additionally, the choice of smoothing window size is discussed in experiments.

Commonly used smoothing techniques such as median smoothing (Basu and Meckesheimer, 2007) and average smoothing (Lohninger, 2012), which replace potential outliers using median and mean values within each smoothing window, were utilized. The choice of smoothing window size was determined by performing experiments, as is illustrated by Figure 2 and Figure 3.

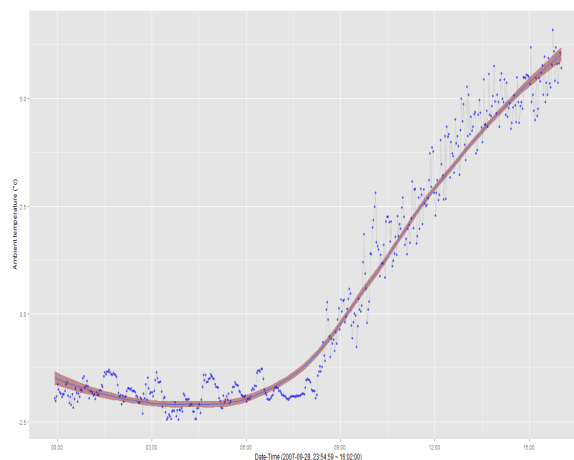


Figure 1 Example of errourness and imprecise data and fluctuations in a real time series data and corresponding data after performing smoothing.

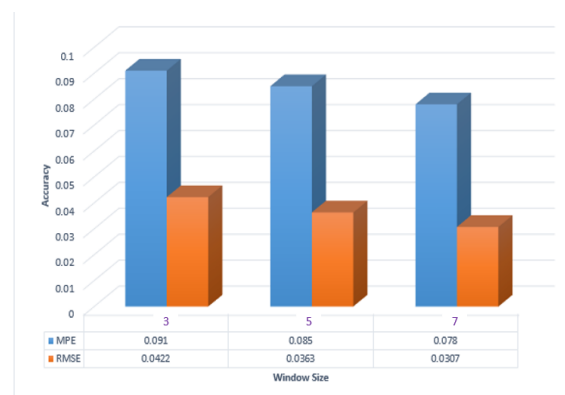


Figure 2 The effect of different window sizes using the moving median smoothing technique.

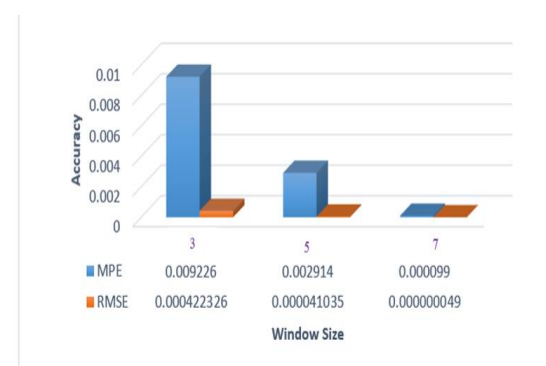


Figure 3 The effect of different window sizes in using the moving average smoothing technique.

Both of the techniques reduce the effect of imprecise observations. Implementing the median smoothing method leads to the omission of infrequent outliers. In the case of average smoothing, however, outliers are not completely disregarded from the data. Importantly, outliers are not always considered errors and may in fact include useful information, to which they are then known as events. As is presented in Figure 2 and Figure 3, larger window sizes represent lower MPE and RMSE, which are translated to greater accuracy. While both techniques for the following data-set represent accurate results,

the moving average method demonstrates superior performance. Furthermore, the existence of outliers was studied using a tolerance of  $\pm 0.3^\circ\text{C}$ , where the measured values exceeded the predicted values. On the basis of the quantity of detected outliers, the moving average identified a larger number of abnormalities compared to the moving median. Thus, the moving average was selected as the smoothing technique for further analyses, since it represented a higher performance in accuracy and outlier detection rate. Additionally, the effects of different window sizes are assessed. Identifying adequate window size is a function of data frequency; whereas high accuracy for this data-set occurred in the presence of a larger window size, too large a window size could change the structure of the data.

## 5.2 Removing trends and Seasonality

Achieving stationary is a prerequisite for building a temporal correlation model, as a result, trends and seasonality need to be accounted for and subsequently removed. Seasonality analysis involves harvesting energy and thus requires a large amount of historical data. The current data-set, which was collected over two months, implies that no significant seasonal behaviour exists. An efficient and non-parametric first order differencing method was applied, due to its low cost of computation and lack of complexity with regards to removing trends.

## 5.3 Identifying the AR and/or MA numerical terms

The selection of an adequate model depends on the nature of the time series, prior knowledge about the data structure, the required accuracy of predictions and the available computational resources. To identify the optimal ARIMA model, ACF and PACF were investigated. Figure 2 illustrates the ACF of the data, before any differencing is performed. The autocorrelations are significant for a large number of time lags. However, the autocorrelations at lags two and above are caused due to the propagation of the autocorrelation at lag one. The evidence is provided by the PACF plot demonstrated in Figure 3. The PACF plot shows a significant spike only at lag one, which is evidenced given all the other higher-order autocorrelations that are effectively explained by lag one autocorrelation. The PACF represents a sharp cutoff, whereas the ACF decays more slowly. Thus, it can be concluded that the stationarized series displays an AR signature, indicating that the autocorrelation pattern can be better explained by adding AR terms as opposed to adding MA terms. Additionally, if the PACF of the differenced series displays a sharp cutoff and/or the lag one autocorrelation is positive, adding one or more AR terms to the model appears to make it more feasible (Nau, 2014). The lag beyond the PACF cutoff reveals the number of AR terms (Nau, 2014). AR models are both theoretically and experimentally prime candidates for making time series predictions (G. Box and G. Jenkins, 1976; Makridakis et al., 2008). Moreover, model parameters can be adopted to the underlying time series in an virtual (i.e. online) manner, without the need to store large amounts of physical historical data (Le Borgne et al., 2007). Subsequently, the ARMA model is simplified to the AR(p) model.

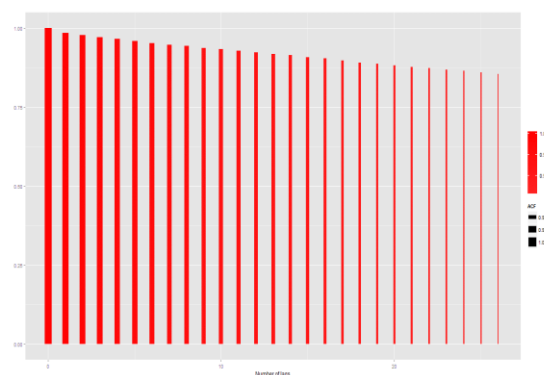


Figure 4 Autocorrelation expolarity plot.

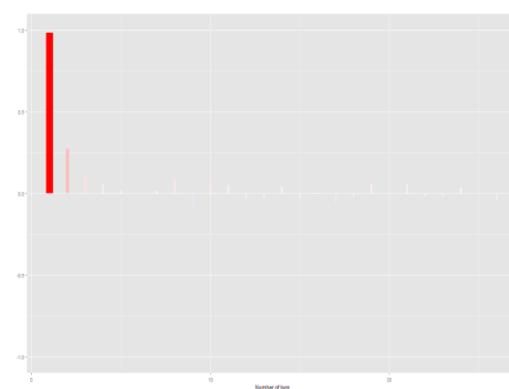


Figure 5 Partial autocorrelation expolarity plot.

The order of the model varied between [1, 2, 3 and 4] to identify the best selection as depicted by Figure 6.

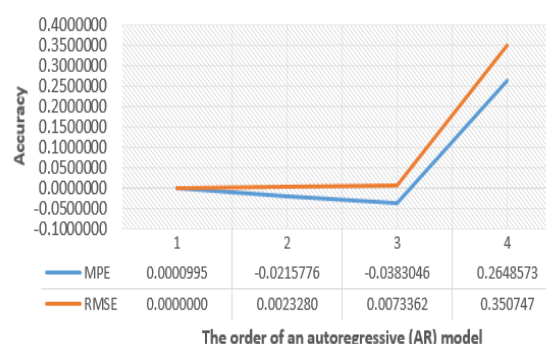


Figure 6 The effect of order variation in AR(p) modelling.

As it is illustrated by Figure 6, the orders of AR (up to order three) result in a quite constant accuracy. Thus, the performances of AR models are essentially equivalent and convincing, regardless the model order. Applying AR with a small order maintains the system's thrifty computation. However, in the context of WSNs, it is not only the accuracy, but also energy concerns that are of considerable importance.

## 5.4 Effectiveness of the historic data size

To identify the structure and patterns that are used for prediction, there needs to be a sufficient amount of historical data available (Hyndman and Kostenko, 2007). Accordingly, this research investigates the importance of historical data.

Analyses of historical data are performed to determine the effects of the sample size of historical data, of which achieve the most robust and accurate model(s). Nowadays, in the presence of cheap data storage and high-speed computing, historical data is made readily available and accessible and thus can and should be used to build statistical models. However, within the WSN's domain, due to temporary deployments and special circumstances, accessing large amounts of historical data is not always feasible. At times, it may be necessary to forecast very short time series, and thus it is helpful to understand the minimum sample size requirements when fitting statistical models to such data. Table 1 displays the accuracy of the predictions, with respect to the different sample sizes.

Sample size	486	300	200	100	50	10
MPE	0.00009953	0.00117117	0.00448142	0.01639968	0.02623359	0.02977228
RMSE	0.00000004	0.00000685	0.00010041	0.00134474	0.00344100	0.00443194

Table 1: The effects of sample size on the accuracy of the predictions.

The results of Table 1 confirm that the larger sample size, including the trend, diurnal and seasonal components, produces the most accurate predictions.

The number of times ahead for which the model is predicting is studied. Table 2 shows the results of predictions in the presence of various prediction life spans.

Predictions age	5	20	30	50	70	90
MPE	0.00009953	0.03366696	0.00136691	-0.66426113	-1.38378312	-2.03018432
RMSE	0.00000004	0.02266929	0.00005605	22.0621421	0.00344100	370.948243

Table 2: The effects of life span on the accuracy of predictions.

The accuracy of predictions in Table 2 reveal that greater accuracy is achieved for the more recent predictions. Sample size, frequency of observations, expected accuracy, application domain and the structure of the data should all be taken into account when determining the adequate time to base predictions.

### 5.5 The Dynamic adaption of the AR model

An AR model is unlikely to be an adequate fit for non-linear physical phenomena (Tulone and Madden, 2006). Tulone and Madden (2006) note that WSN data is locally linear, but that it is also periodic non-linear, which cannot be precisely predicted by AR. Thus, to achieve convincing prediction accuracy, the model should be dynamically adapted. The precision of the current data-set is at  $\pm 0.3^{\circ}\text{C}$ ; at the point where the prediction error exceeds the tolerated threshold, the model should be fed by the data being collected. Hence the model can be re-conceptualized using data coefficients, enabling the predicted values to be close to the actual records. This design inherently includes the advantages of being effective, (near) real time and energy efficient in outlier detection, since no communication overhead is needed. Outliers can be an error or, alternatively, can provide potentially useful information, known as an event. While errors happen infrequently, the system should nonetheless be able to distinguish errors from events. If the system updates the model when errors occur then it may misinterpret acceptable data as outliers; to solve this problem, the system should update the model only in cases of events. Consequently, the model is

updated based on new observations whenever the multiple number of successive outliers (events) are recorded, as it assumes that the possibility of experiencing successive errors is low.

### 5.6 Energy

Energy consumption for WSNs is highly dependent on the data sampling rate, communication overhead and, last but not least, the applied prediction model. As was previously stated, larger sample sizes result in more accurate predictions. However, a wide availability of historical data implies more computation and storage energy consumption (Amidi et al., 2013; Li and Wang, 2013). While communication costs require several orders of magnitude higher, extra memory usage and computation costs are essentially negligible. Thus, the proposed methodology is also considered to be an energy friendly system.

## 6. CONCLUSION

The ARIMA model has the ability to capture a wide variety of realistic phenomena and it is efficient in terms of both memory and computational cost. Its applicability has not yet been recognized within the research community of WSNs, specifically with respect to technological requirements. WSN observations are often corrupt, missed or inaccurate due to the inherent imprecise characteristics of WSNs. Applicability of ARIMA modelling was assessed with respect to the nature of sensor data and the specific requirements and limitations of WSNs. Among data smoothing techniques, the moving average method represented better performance since it did not completely exclude outliers from data. The proposed method maintains the abnormalities where they may potentially introduce useful information called events. Moreover, the results demonstrate that a larger window size can improve predictions accuracy, up until the point where it becomes too large to differentiate the structure of the data. The experiments reveal the suitability of simplifying ARIMA to AR, as well as the need to, apply low order AR, which result in greater accuracy and energy efficiency. Furthermore, the results indicate that accurate and precise predictions hinge on sample sizes. Subsequently, larger sample sizes produce more accurate and precise predictions. The life span of the predictions is also relevant and accurate predictions tend to be those more recent ones. The proposed system was furthermore designed to monitor its local model and continually update as needed. Finally, it includes a capability to detect outliers and events in an (near) real time manner.

## REFERENCES

- Amidi, A., Hamm, N.A.S., Meratnia, N., 2013. Wireless Sensor Networks and Fusion of Contextual Information for Weather Outlier Detection. ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 1, 37-41.
- Australian Bureau of Statistics, 2005. Time Series Analysis: The Basics.
- Basu, S., Meckesheimer, M., 2007. Automatic outlier detection for time series: an application to sensor data. Knowl Inf Syst 11, 137-154.

- Chatfield, C., 2013. The analysis of time series: an introduction. CRC press.
- Cheng, T.-Y.W.Q., 2008. collaborative event-region and boundary-region detections in wireless sensor networks. Signal Processing.
- Collins, F.C., 1995. A comparison of spatial interpolation techniques in temperature estimation, IBM Government Systems - Resource Management and Distributed Systems.
- Elnahrawy, E., Nath, B., 2003. Online data cleaning in wireless sensor networks, Proceedings of the 1st International conference on Embedded networked sensor systems. ACM, pp. 294-295.
- G. Box, G. Jenkins, 1976. Time Series Analysis: Forecasting and Control, Holden-Day Series in Time Series Analysis, San Francisco, CA, USA.
- Guo, D., Qu, X., Huang, L., Yao, Y., 2011. Sparsity-based spatial interpolation in wireless sensor networks. Sensors 11, 2385-2407.
- Hyndman, R.J., Kostenko, A.V., 2007. Minimum sample size requirements for seasonal forecasting models. Foresight 6, 12-15.
- Ihaka, R., Gentleman, R., 1996. R: A Language for Data Analysis and Graphics. Journal of Computational and Graphical Statistics 5, 299-314.
- Le Borgne, Y.-A., Santini, S., Bontempi, G., 2007. Adaptive model selection for time series prediction in wireless sensor networks. Signal Processing 87, 3010-3020.
- Li, G., Wang, Y., 2013. Automatic ARIMA modeling-based data aggregation scheme in wireless sensor networks. EURASIP Journal on Wireless Communications and Networking 2013, 1-13.
- Liu, C., Wu, K., Tsao, M., 2005. Energy efficient information collection with the ARIMA model in wireless sensor networks, Global Telecommunications Conference, GLOBECOM'05. IEEE, pp. 5 pp.-2474.
- Lohninger, H., 2012. Fundamentals of Statistics. Epina e-Book Team.
- Makridakis, S., Wheelwright, S.C., Hyndman, R.J., 2008. Forecasting methods and applications. John Wiley & Sons.
- Nau, R., 2014. Lecture notes on forecasting, Summary of rules for identifying ARIMA models. Fuqua School of Business, Duke University.
- Santini, S., Romer, K., 2006. An adaptive strategy for quality-based data reduction in wireless sensor networks, Proceedings of the 3rd International Conference on Networked Sensing Systems (INSS 2006), pp. 29-36.
- Sheikhhasan, H., 2006. A comparison of interpolation techniques for spatial data prediction. Master's Thesis, Department Computer Science, Universiteit van Amsterdam, Amsterdam, The Netherlands.
- Singh, A.K., Giridhar, B., Mandal, P.S., 2011. Fixing Data Anomalies with Prediction Based Algorithm in Wireless Sensor Networks. arXiv preprint arXiv:1111.3334.
- Tulone, D., Madden, S., 2006. PAQ: Time Series Forecasting for Approximate Query Answering in Sensor Networks, in: Römer, K., Karl, H., Mattern, F. (Eds.), Wireless Sensor Networks. Springer Berlin Heidelberg, pp. 21-37.
- Umer, M., Kulik, L., Tanin, E., 2008. Kriging for localized spatial interpolation in sensor networks, Scientific and Statistical Database Management. Springer, pp. 525-532.
- Umer, M., Kulik, L., Tanin, E., 2010. Spatial interpolation in wireless sensor networks: localized algorithms for variogram modeling and Kriging. Geoinformatica 14, 101-134.
- Wang, T.-Y., Cheng, Q., 2008. Collaborative event-region and boundary-region detections in wireless sensor networks. IEEE Transactions on Signal Processing, 56, 2547-2561.
- Wang, T.-Y., Yu, C.-T., 2005. Collaborative event region detection in wireless sensor networks using markov random fields, Wireless Communication Systems, 2005. 2nd International Symposium on. IEEE, pp. 493-497.
- Wu, W., Cheng, X., Ding, M., Xing, K., Liu, F., Deng, P., 2007. Localized outlying and boundary data detection in sensor networks. IEEE Transactions on Knowledge and Data Engineering, 19, 1145-1157.
- Zhang, Y., 2010. Observing the unobservable: distributed online outlier detection in wireless sensor networks. University of Twente, PhD thesis.
- Zhang, Y., Hamm, N.A.S., Meratnia, N., Stein, A., van de Voort, M., Havinga, P.J., 2012. Statistics-based outlier detection for wireless sensor networks. International Journal of Geographical Information Science 26, 1373-1392.
- Zhang, Y., Meratnia, N., Havinga, P., 2008. An online outlier detection technique for wireless sensor networks using unsupervised quarter-sphere support vector machine, International Conference on Intelligent Sensors, Sensor Networks and Information Processing, ISSNIP. IEEE, pp. 151-156.