International Journal of Advanced Smart Convergence Vol.4 No.2 46-53 (2015)
http://dx.doi.org/10.7236/IJASC.2015.4.2.46

IJASC 15-2-8

Efficient Anomaly Detection Through Confidence Interval Estimation Based on Time Series Analysis

Yeong-Ju Kim, Min-A Jeong*

**Department of Computer Engineering, Mokpo National University, Muan-gun, Korea*
{xfile7, majung*}@mokpo.ac.kr

Abstract

This paper suggests a method of real time confidence interval estimation to detect abnormal states of sensor data. For real time confidence interval estimation, the mean square errors of the exponential smoothing method and moving average method, two of the time series analysis method, were compared, and the moving average method with less errors was applied. When the sensor data passes the bounds of the confidence interval estimation, the administrator is notified through alarms. As the suggested method is for real time anomaly detection in a ship, an Android terminal was adopted for better communication between the wireless sensor network and users. For safe navigation, an administrator can make decisions promptly and accurately upon emergency situation in a ship by referring to the anomaly detection information through real time confidence interval estimation.

Keywords: *nautical safety, time series analysis, moving average method, exponential smoothing*

1. Introduction

The development of a service that supports vessel safe navigation was performed many years before; however, the integration of IT technologies into the vessel industry has spurred the development of such service [1].

The purpose of this study is to install wireless sensors in a vessel for safe navigation, build the environment for a wireless sensor network (WSN), notify the vessel administrator of any anomaly in the vessel in real time through an Android terminal, and let the administrator be aware of the anomaly in advance for smooth navigation.

To test anomaly detection, first, the PC in which the sensor data (temperature, humidity, Light and Voltage) collected from the WSN is stored is configured for a web server; then, an Android terminal connects to said web server, and collects the data. Second, the data collected from ServerPC are used to estimate the confidence interval for which an abnormal state in the vessel is to be detected in real time. Third, the estimation of a confidence interval is compared and assessed with time series analysis methods—the exponential smoothing method and the moving average method; then, the current data are collected by applying the moving average method, which has a less mean squared error (MSE), and both the error from

Manuscript Received: Sept. 8, 2015 / Revised: Oct. 9, 2015 / Accepted: Oct. 19, 2015
Corresponding Author: majung@mokpo.ac.kr
Tel: +82-61-450-2774, Fax: +82-61-452-4793
Department of Computer Engineering, Mokpo National University

the method and the data confidence interval are estimated in real time. Fourth, if the data measured in the vessel deviate from the estimated confidence interval, an anomaly has occurred, and the administrator is notified with an alarm. Last, using the data provided to the administrator, he/she can make an immediate and accurate decision during an emergency in the vessel and continue navigating safely.

This study includes the following sections: Section 2 explains related work; Section 3 explains anomaly detection algorithm; Section 4 explains performance evaluation by time series analysis and last, Section 5 describes the conclusion and future study.

2. Related Work

2.1 Time series analysis

Time series estimation evaluates demand with the assumption that a demand pattern found over past time series data at certain time intervals can also be applied to future time intervals. This time series estimation method is good when there are several years of past data with a clear and stable pattern. Such time series data are a set of data with matching time and quantity, where the time is the independent variable and the quantity of load is the dependent variable. Because this estimation method tends to be inaccurate when the environmental condition or quantity changes significantly, it is appropriate for short or mid-term estimation [2, 3]. Therefore, time series estimation is performed under the assumption that a certain pattern is identified in past data on which the estimating variable is based, and is repeated without losing its characteristics [4].

Because this study uses time series sensor data collected at given regular intervals through a wireless sensor network, among many time series estimation methods, we calculated estimated values of the sensor data using the exponential smoothing method in which weight is given to the recent data, and the moving average method that moves by calculating the average at each time series interval. The confidence interval is estimated in real time using the calculated estimated values.

Exponential smoothing method.

In the exponential smoothing method, the weight of a value, which is given to a time period for an estimated value, decreases exponentially as it reverts in time. Because the exponential smoothing method provides the largest weight to the most recent time period, this method provides several advantages: high model accuracy, easy model determination, easy understanding of the model, less number of calculations, and easy model accuracy test. In addition, this method does not require large memory space. Because of such advantages, this method, among the time series analysis methods, is the most widely used for a short-term estimation [5].

An estimated value by the exponential smoothing method is calculated using Equation (1).

$$\mathrm{F_t} = (1-\alpha)\mathrm{F_{t-1}} + \alpha \mathrm{D_{t-1}} \tag{1}$$

F_t : estimated value of period t

F_{t-1} : estimated value of period t-1

D_{t-1} : measured value of period t-1

α : coefficient of exponential smoothing

Moving average method.

The moving average method is the easiest time series estimation method, and it is useful when the time series data have no abrupt changes or trend changes, no cycle variation, and no seasonal change, but unpredicted variations. In addition, the moving average method is used for calculating a trend or seasonal index when there are time series variation factors. If a set of past time series data show a certain trend pattern, the moving average method estimation tends to be less accurate; however, the application of this method is generally simple and easy so that it is used for estimating a small-sized demand or sales forecast, such as tourist demand or amount of catch [6].

Equation(2) shows an estimation using the moving average method, where N is a certain period determined by the quantity of a time series measurement unit and where x_{t-i} is an observed value at the time of t-i in the collected time series data.

$$\mathrm{MA} = \frac{x_{t-(n-1)} + x_{t-(n-2)} + x_{t-(n-3)} + \cdots + x_{t-(i+1)} + x_{t-1}}{n} + \frac{x_{t-(i-1)} + x_{t-(i-2)} + \cdots + x_{t-(n-n-1)} + x_{t(n-n)}}{n} \tag{2}$$

Where the moving average period N can be divided into Periods 3, 4, or 5 depending on the count of the moving average periods, and it is determined to be period N whose MSE(Mean square error) of an estimated error within an observed period is the smallest. In this study, the moving average period Ns are set to 3, 4, 5, and 6 s; from these, the 5 s period has the smallest MSE.

Mean squared error.

Mean squared error(MSE) is used to square an estimated error, and it is important for calculating the size of an error to ensure an accurate estimate. The smaller the size of an error, the closer is an estimate is to the actual value [7]. MSE is calculated by Equation (3).

$$\mathrm{MSE} = \frac{\sum_{t=1}^{n}(e_t)^2}{n} = \frac{\sum_{t=1}^{n}(y_t - \widehat{y_t})}{n} \tag{3}$$

Where t is the time, $y_t$is the actual measured value, and the estimated value of y_t is $\widehat{y_t}$. Equation (4) presents the difference between an estimated value $\widehat{y_t}$ and an actual value y_t i.e., the estimated error value e_t.

$$e_t = y_t - \widehat{y_t} \tag{4}$$

2.2 Confidence interval

Interval estimation and point estimation scheme is branch of statistical estimation. The interval estimation method for estimating the unknown parameters that is within a specific real number range [a, b]. Here, the greater the range the greater the possibility parameter is in the interval. A standard that determines the size of the possibility is referred to as a confidence level. The range to be obtained under the respective confidence level is referred to as the confidence interval. Generally, the confidence level uses 90%, 95%, 99%[8].

2.3 Research for anomaly detection

There have been a variety of preceding studies that detect sensor values called anomalies, defects, and noise in wireless sensor networks or embedded systems.

Buonadonna et al. considered the convenience and factors for end-users when designing the arrangement of sensors to build a sensor network environment. In their study, data transmission success rates were measured depending on the distance between sending and receiving sensors. The lowest transmission success rate was 22%, whereas the highest rate was 75%. In a sensor network, missing values and outliers occur because of various factors, including sensor equipment, environment, and system, hindering the acquisition of good sensor signal information [9].

Werner-Allen et al. suggested the need for a data refining process in order to use sensor networks as scientific tools that require accurate data because they often generate abnormal values [10].

The defects that occur in sensor networks or the features used for detection are divided into environmental, systematic, and data-related factors. Environmental factors include the location of sensors and their surrounding factors, systematic factors include sensor hardware specifications and sensor network systems, and data-related factors include sensor-measured values [11].

Two main types of errors that occur in sensor networks are mechanical errors and random noise found in sensor signals. Elnahrawy et al. focused on random noise and used the Bayesian approach to refine the sensor data, reducing uncertainty in noise [12].

Jeffery et al. proposed a framework called extensible sensor stream processing (ESP) for refining sensor signals. ESP refines data using the range of sensor values, measured temporal interval values, and measured sensor values in the neighboring space. The range of sensor values is a range of the measured values that can actually occur, and any value beyond this range is considered an observed abnormal value. Measured temporal interval values are used to complete missing values and the values measured by sensors for given periods. Missing values are substituted with the average of the values measured by sensors for given periods. Measured sensor values are used to determine whether a signal is reliable. ESP designates neighboring sensors to a group, and calculates the average and deviation of the measured values [13]. However, the above studies require many calculations and preceding processes, which increases cost.

3. Efficient Anomaly Detection Algorithm

3.1 Anomaly detection algorithm

First, the PC in which the sensor data (temperature, humidity, Light and noise) collected from the WSN is stored is configured for a web server; then, an Android terminal connects to said web server, and collects the data. Second, the data collected from ServerPC are used to estimate the confidence interval for which an abnormal state in the vessel is to be detected in real time. Third, the estimation of a confidence interval is compared and assessed with time series analysis methods—the exponential smoothing method and the moving average method; then, the current data are collected by applying the moving average method, which has a less mean squared error (MSE), and the data confidence interval are estimated in real time. Fourth, if the data measured in the vessel deviate from the estimated confidence interval, an anomaly has occurred, and the administrator is notified with an alarm. Last, using the data provided to the administrator, he/she can make an immediate and accurate decision.

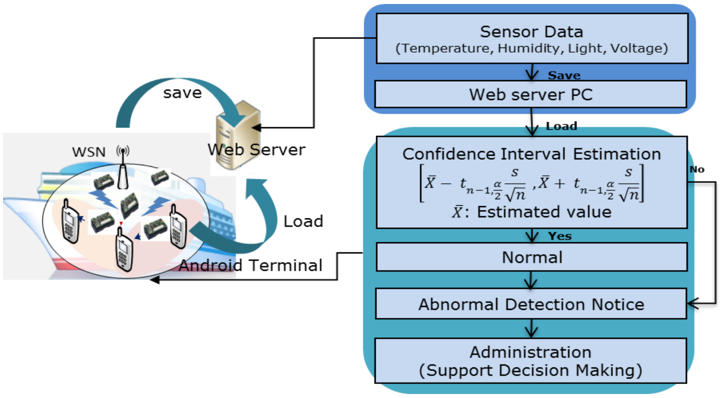

save
WSN
Web Server
Load
Android Terminal
Sensor Data
(Temperature, Humidity, Light, Voltage)
Save
Web server PC
Load
Confidence Interval Estimation
$\left[\bar{X} - t_{n-1,\frac{\alpha}{2}}\frac{s}{\sqrt{n}}, \bar{X} + t_{n-1,\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right]$
$\bar{X}$: Estimated value
No
Yes
Normal
Abnormal Detection Notice
Administration
(Support Decision Making)

Figure 1. Algorithm of anomaly detection

3.2 Confidence interval

σ^2Suppose $\{X_1, \ldots ,X_n\}$is an independent sample from a normally distribution population with mean μ and variance σ^2. Let where $\overline{X}$ is the sample mean, and s^2 is the sample variance. Then has a Student's t-distribution with n-1 degree of freedom [8].

On this basis, sensor data estimated by t-distribution that confidence interval of population mean. In this paper, we estimated 95% confidence interval of the population mean of whole sensor data.

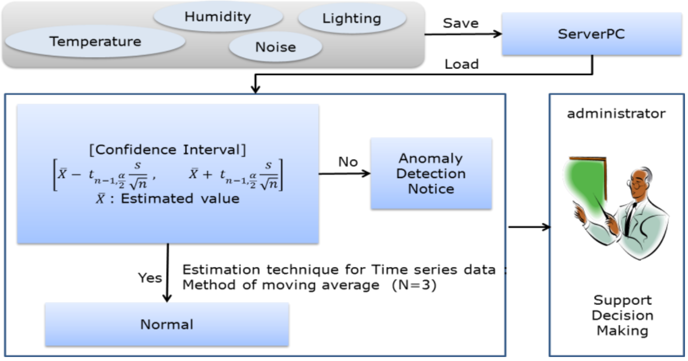

Humidity
Lighting
Temperature
Noise
Save
ServerPC
Load
[Confidence Interval]
$\left[\bar{X} - t_{n-1,\frac{\alpha}{2}}\frac{s}{\sqrt{n}},\quad \bar{X} + t_{n-1,\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right]$
$\bar{X}$: Estimated value
No
Anomaly Detection Notice
administrator
Yes
Estimation technique for Time series data :
Method of moving average (N=3)
Normal
Support Decision Making

Figure 2. Real-time confidence interval estimation

4. Performance Evaluation

4.1 Sensor data cleaning

We performed the time series analysis because of sensor data. Figure 3 shows the Intel Lab Dat. Intel lab data is collected from 54 sensors deployed in the Intel Berkeley Research lab between February 28th and April 5th, 2004. Each sensor collected the temperature, humidity, and light and voltage value every the rod of 31 seconds. We generated time series data(the total of 72) by averaging sensor data (the total of 5,797) collected for 3 days (march 1 day ~3 month 3 day)s.

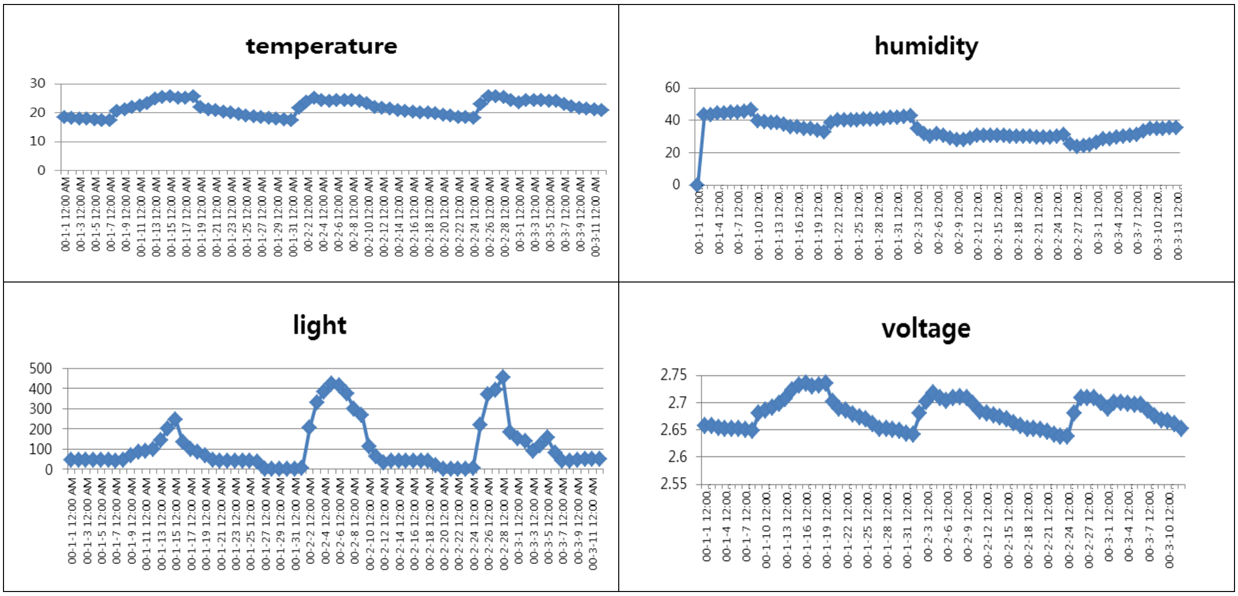


Figure 3. Sensor data of time series

In general, time series data can be explained by a level, trend, seasonal change, and noise. A level is the average of the time series data, and a trend is an overall change of pattern from one time point to another. A seasonal change means a periodic pattern for a short period, and last, noise is a random variation that occurs from an unknown advection. The easiest method for examining time series data elements is through temporal graphs [14]. Figure 3 presents time series graphs that contain 72 refined temperature, humidity, Light and Voltage data, and presents irregular variations. An irregular variation is a horizontal series whose averages remain almost unchanged, regardless of period. In addition to the time series analysis methods, the moving average method or exponential smoothing method can be used to smooth irregular variation and estimate future time series values [15].

4.2 Estimation by the moving average method

The table 1 shows mean square error specified moving average method. In the period 3, the smallest mean square error is shown.

Table 1. Mean square error by each period

Moving average method Period	Mean Square Error(MSE)			
	Temperature	Humidity	Light	Voltage
3 Period	0.776	1.079	36.736	0.008
6 Period	1.693	2.272	84.446	0.018
12 Period	2.764	3.599	136.979	0.029

4.3 Estimation by the exponential smoothing method

The table 2 shows mean square error by smoothing factor of the exponential smoothing method. In the smoothing factor 0.7, the smallest MSE is shown.

Table 2. Mean square error by smoothing factor

Exponential smoothing method smoothing factor	Mean Square Error(MSE)			
	Temperature	Humidity	Light	Voltage
0.1	2.365	3.531	94.666	0.024
0.3	1.731	2.329	76.831	0.018
0.7	1.063	1.495	50.801	0.011

4.4 Result of time series analysis

The graph in figure 4 shows the measured value and estimated value (in moving averages method and exponential smoothing method). In the estimated value, the method of moving average (t=3) is more dependent than the method of exponential smoothing. The exponential smoothing method is used most frequently for short-term estimation of time series data, and it has the advantage of real time monitoring. However, as indicated in Figure 4, this method fails to reflect the variation of real time data, whereas the results of the moving average method reflect the trend of the measured value.

Therefore, our proposed system uses the moving average method to reflect the most recent sensor data for real time anomaly detection and to determine the confidence interval.

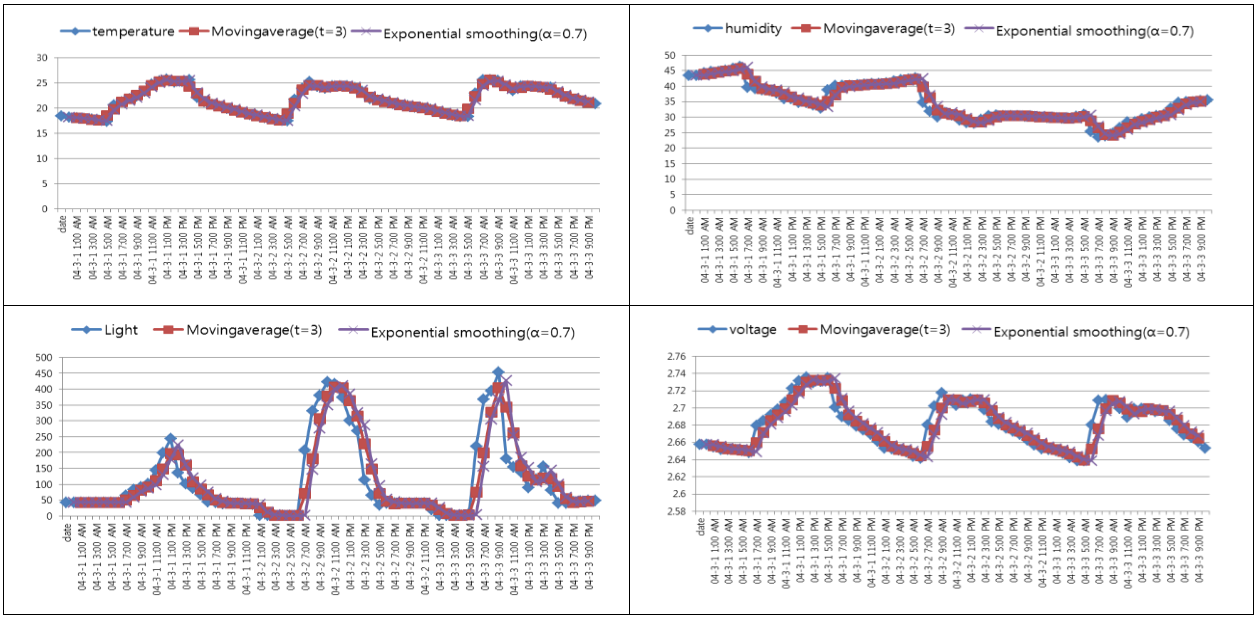
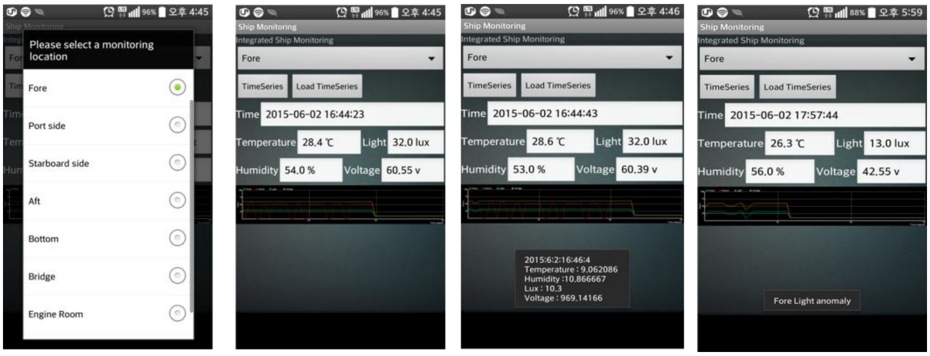


Figure 4. Result of time series analysis

4.5 Implementation



Please select a monitoring location

- ● Fore
- ○ Port side
- ○ Starboard side
- ○ Aft
- ○ Bottom
- ○ Bridge
- ○ Engine Room

5.a

Fore

Time 2015-06-02 16:44:23

Temperature 28.4 ℃ Light 32.0 lux

Humidity 54.0 % Voltage 60.55 v

5.b

Fore

Time 2015-06-02 16:44:43

Temperature 28.6 ℃ Light 32.0 lux

Humidity 53.0 % Voltage 60.39 v

2015:6:2:16:46:4
Temperature : 9.062086
Humidity :10.866667
Lux : 10.3
Voltage : 969.14166

5.c

Fore

Time 2015-06-02 17:57:44

Temperature 26.3 ℃ Light 13.0 lux

Humidity 56.0 % Voltage 42.55 v

Fore Light anomaly

5.d

Figure 5. Mobile display of ship internal anomaly detection

Figure 5 presents screen captures of an Android device that implemented to detected an internal anomaly in the vessel. From Figure 5.a to Figure 5.c is the screen on which the monitoring location, such as the fore or the post side, is selected, and Figure 5.d is an alarm window that displays an alarm when abnormal state is actually detected on the port. The delay-time of the proposed system assumes as 0. Moreover, data received from the sensors a stored in the database every 5 seconds.

5. Conclusion

We proposed an anomaly detection method in a vessel. Proposed method is the data received from the sensors are above or below a designated confidence interval, and such an abnormal state is notified to the administrator with an alarm, after which the administrator can form an immediate and accurate decision, and continue safe navigation. Here, to improve the accuracy of anomaly detection estimation, reflecting a characteristic where continuous time series data of sensor data have a changing confidence interval in real time, we performed an analysis through which we confirmed that the moving average method, which reflects the trend of the time series data, is effective. And the confidence interval was estimated by using the estimated value of the moving average method of the interval estimation average. Therefore, proposed method is sensor data received in real time could be anomaly detection in real time.

In the future, the continuous time series data estimation method needs to be complemented by collecting, testing, and analyzing discrete time series data at the same time interval. In addition, a study that can predict malfunction separately, missing values, outliers, and monitoring application of various sensor data based on time series data analysis results is required.

Acknowledgement

This paper was supported by Research Funds of Mokpo National University in 2013.

References

[1] J. H. Park, B. T. Jang, and D. S. Lim, "Safe operation of the shipyard and ship building digital technology developments supported," *Korean Inst. Inf. Sci. Eng.(KIISE)*, Vol. 31, No. 1, pp. 55-63, Jan. 2003.

[1] A. C. Harvey, *Time Series Models*, 2nd Ed., MIT Press, (308), 1993.

[2] H. Zou and Y. H. Yang, "Combining Time Series Model for Forecasting," *Int. J. Forecasting*, Vol. 20, No. 1, pp. 69-84, 2004

[3] K. H. Cho and D. H. Lee, "A Study on Traffic Anomaly Detection Scheme Based Time Series Model," *J. KICS*, Vol. 33, No. 5, pp. 304-309, 2008.

[4] H. G. No, *SPSS / Excel by time series analysis*, HYOSAN, (323), 2008.

[5] Lim, M. Michael, "Time Series Forecasts of International Travel Demand for Australia," *Tourism Management*, Vol. 23, No. 4, pp. 389-396, Aug. 2002.

[6] Y. H. Kim, *Time Series Prediction*, HSPN, (448), 2002.

[7] E. H. Kim, S. H. Lee, Teaching Statistics, KYUNGMOON, (270), 2007.

[8] P. Buonadonna, D. Gay, J. M. Hellerstein, W. Hong, and S. Madden, "Task: Sensor network in a box," in *Proc. 2nd European Workshop on Wirel. Sensor Netw.*, pp. 133-144, Istanbul, Turkey, Feb. 2005.

[9] G. Werner-Allen, K. Lorincz, J. Johnson, J. Lees, and M. Welsh, "Fidelity and yield in a volcano monitoring sensor network," in *Proc. 7th USENIX Symp. Operating System Design and Implementation*, pp. 381-396, Berkeley, USA, Nov. 2006.

[10] K. Ni, N. Ramanathan, M. N. H. Chehade, L. Bal zano, S. Nair, S. Zahedi, E. Kohler, G. Pottie, M. Hansen, and M. Srivastava, "Sensor Network Data Fault Types," *J. ACM Trans. Sensor Netw.*, Vol. 5, No. 3, pp. 1-29, Aug. 2009.

[11] E. Elnahrawy and B. Nath, "Cleaning and Querying Noisy Sensors," in *Proc. Int. Workshop Wirel. Sensor Netw. Appl.*, pp. 78-87, New York, USA, Sept. 2003.

[12] S. R. Jeffery, G. Alonso, M. J. Franklin, W. Hong, and J. Widom. "Declarative support for sensor data cleaning," in *Proc. Int. Conf. Pervasive Computing, Lecture Notes in Comput. Sci.*, Vol. 3968, pp. 83-100, Dublin, Ireland, May 2006.

[13] G. Shmueli, N. R. Patel, and P. C. Bruce, *TData Mining for Business Intelligence*, E&B, (460), 2006.

[14] S. D. Lee and U. R. Lee, *Time series data analysis using SAS*, TAMJI, (319), 2006.