Environmental Data Recovery using Polynomial Regression for Large-scale Wireless Sensor Networks

Kohei Ohba[1], Yoshihiro Yoneda[1], Koji Kurihara[2], Takashi Suganuma[1], Hiroyuki Ito[1],
Noboru Ishihara[1], Kunihiko Gotoh[1], Koichiro Yamashita[2] and Kazuya Masu[1]
[1]*ICE Cube Center, Tokyo Institute of Technology, Nagatsutacho 4259, Midori-ku, Kanagawa, 226–8503, Japan*
[2]*Network Systems Laboratory, Fujitsu Laboratories Ltd., Kamikodanaka 4–1–1,*
Kawasaki Nakahara-ku, Kanagawa, 211–8588, Japan

Keywords: Wireless Sensor Networks, Polynomial Regression, Data Recovery, Environment Monitoring.

Abstract: In the near feature, large-scale wireless sensor networks will play an important role in our lives by monitoring our environment with large numbers of sensors. However, data loss owing to data collision between the sensor nodes and electromagnetic noise need to be addressed. As the interval of aggregate data is not fixed, digital signal processing is not possible and noise degrades the data accuracy. To overcome these problems, we have researched an environmental data recovery technique using polynomial regression based on the correlations among environmental data. The reliability of the recovered data is discussed in the time, space and frequency domains. The relation between the accuracy of the recovered characteristics and the polynomial regression order is clarified. The effects of noise, data loss and number of sensor nodes are quantified. Clearly, polynomial regression offers the advantage of low-pass filtering and enhances the signal-to-noise ratio of the environmental data. Furthermore, the polynomial regression can recover arbitrary environmental characteristics.

1 INTRODUCTION

Large-scale wireless sensor networks (WSNs) use wireless sensor nodes to monitor environmental parameters such as temperature, humidity, pH, light and air pressure. WSNs have many possible applications, ranging from structual health monitoring to field monitoring. Thanks to the progress in microelectronics based on the integrated circuit technology, small wireless sensor nodes with low power consumption have been achieved. However, problems exist with data loss owing to data collision between the sensor nodes and electromagnetic noise. As the interval of aggregate data is not fixed in the time and space domains, digital signal processing using Fourier or wavelet transforms cannot be applied directly to the aggregated data. Moreover, noise degrades the data accuracy. Because the environmental characteristics have various waveforms, data reliability cannot evaluate by signal analysis. To overcome these problems, various techniques, such as data collection timing (Sivrikaya et al., 2004), redundant system (Yamashita et al., 2014) and data recovery (Doherty et al., 2000), have been used to increase data reliability.

We apply polynomial regression to environmental data recovery based on the correlations among the environmental data. Environmental characteristics are recovered as aggregated data from the sensor nodes using polynomial regression. Thus, data loss is minimized, and the data can be analysed easily. Basic sinusoidal environmental variations are assumed to evaluate the data recovery with polynomial regression. If the sinusoidal characteristics can be modeled appropriately, arbitrary waveform characteristics, such as single-shot, periodic and non-periodic waveforms, can also be modelled. The recovered data accuracy is evaluated by comparing the recovered and source characteristics.

We have also proposed a data reliability evaluation flowchart that does not rely on signal analysis (Yoneda et al., 2014). We also clarify the relation between the accuracy of the recovered characteristics and the polynomial regression order, and the effects of data loss and number of sensor nodes is analysed. Furthermore, we show that the use of polynomial regression has the advantage of low-pass filtering that enhances the signal-to-noise ratio (SNR) of the environmental characteristics. In addition, we show that polynomial regression can recover arbitrary environmental characteristics.

In section 2, we introduce the environmental data recovery technique based on polynomial regression.

Ohba, K., Yoneda, Y., Kurihara, K., Suganuma, T., Ito, H., Ishihara, N., Gotoh, K., Yamashita, K. and Masu, K.
Environmental Data Recovery using Polynomial Regression for Large-scale Wireless Sensor Networks.
DOI: 10.5220/0005636901610168
In *Proceedings of the 5th International Confereence on Sensor Networks (SENSORNETS 2016)*, pages 161-168
ISBN: 978-989-758-169-4

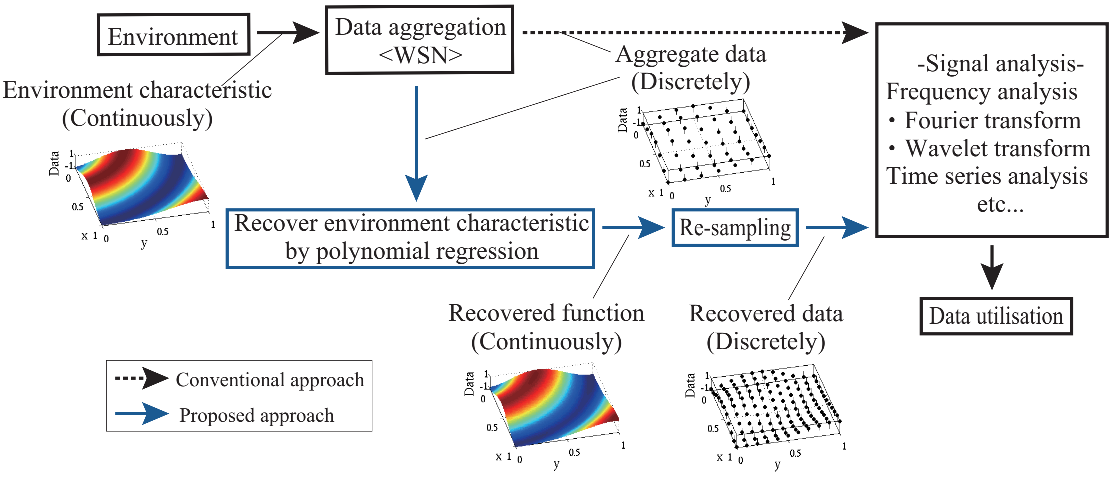


Figure 1: WSN system using polynomial regression.

In section 3, the reliability of the recovered data is discussed. The frequency domain characteristics are evaluated in section 4. In section 5, we confirm the ability of polynomial regression to recover arbitrary environmental characteristics and present the conclusions in section 6.

2 ENVIRONMENTAL DATA RECOVERY USING POLYNOMIAL REGRESSION

The aggregated data analysis is shown in Fig. 1. If the interval of the aggregated sampled data is fixed, the environmental data characteristics can be analysed directly using Fourier or wavelet transforms. However, when the interval of the data is not fixed, the data cannot be directly analysed. Therefore, continuous environmental characteristics are recovered from the aggregated data, and then, the fixed interval data are resampled from the recovered characteristics.

2.1 Polynomial Regression

There are several ways of expressing the recovered characteristics, e.g., Fourier series expansion, polynomial regression, interpolation and so on. Polynomial regression is simple and suitable for expressing continuous characteristics as it tolerates data loss. However, polynomial regression is not good at expressing characteristics with many inflection points. If the frequency band is limited, the environmental characteristics at the limited bandwidth can be expressed using polynomial expressions. Therefore, polynomial regression is used in environmental data recovery. When one-dimensional data are $\boldsymbol{t} = [t_1 \quad \cdots \quad t_N]^{\mathrm{T}}$ and the environmental data are $\boldsymbol{d} = [d_1, \cdots, d_N]^{\mathrm{T}}$, the environmental source characteristics function $f_{\mathrm{W}}(t)$ can be recovered and recovered function $f_{\mathrm{R}}(t)$ is

$$f_{\mathrm{R}}(t) = \sum_{i=0}^{m} a_i t^i \qquad (1)$$

Where $\boldsymbol{a} = [a_0 \quad \cdots \quad a_m]^{\mathrm{T}}$ is the coefficient vector. The value of $\boldsymbol{a}$ is obtained using at least two multiplication methods. $\boldsymbol{m}$ is the order of polynomial equation. For two-dimensional data obtained by sensor nodes arranged in coordinates $(x_1,y_1),\cdots,(x_N,y_N)$ and coordinates $\boldsymbol{x} = [x_1 \quad \cdots \quad x_N]^{\mathrm{T}}$ and $\boldsymbol{y} = [y_1 \quad \cdots \quad y_N]^{\mathrm{T}}$, the recovered function $f_{\mathrm{R}}(x,y)$ is

$$f_{\mathrm{R}}(x,y) = \sum_{\substack{j,k\geq 0\\ j+k\leq m}} a_{jk} x^j y^k \qquad (2)$$

Where the coefficient vector $\boldsymbol{a}$ is the column vector of size $\frac{(m+1)(m+2)}{2} \times 1$.

2.2 Data Reliability Evaluation Flow

2.2.1 Evaluation flow

The reliability evaluation flowchart is shown in Fig. 2. Two-dimensional data are assumed in the evaluation. The environmental source characteristic function is $f_{\mathrm{W}}(x_i,y_i)$. To consider the effect of noise and data loss, we define the sensor node model. When the noise is expressed as $f_{\mathrm{N}}(x_i,y_i)$, the sampled data with noise $f_{\mathrm{S}}(x_i,y_i)$ can be expressed as

$$f_{\mathrm{S}}(x_i,y_i) = f_{\mathrm{W}}(x_i,y_i) + f_{\mathrm{N}}(x_i,y_i) \qquad (3)$$

To evaluate the effect of data loss, the following function is added

$$f_{\mathrm{O}}(x_i,y_i) = \begin{cases} f_{\mathrm{S}}(x_i,y_i) & (\mathit{Without\ data\ loss}) \\ \emptyset & (\mathit{With\ data\ loss}) \end{cases} \qquad (4)$$

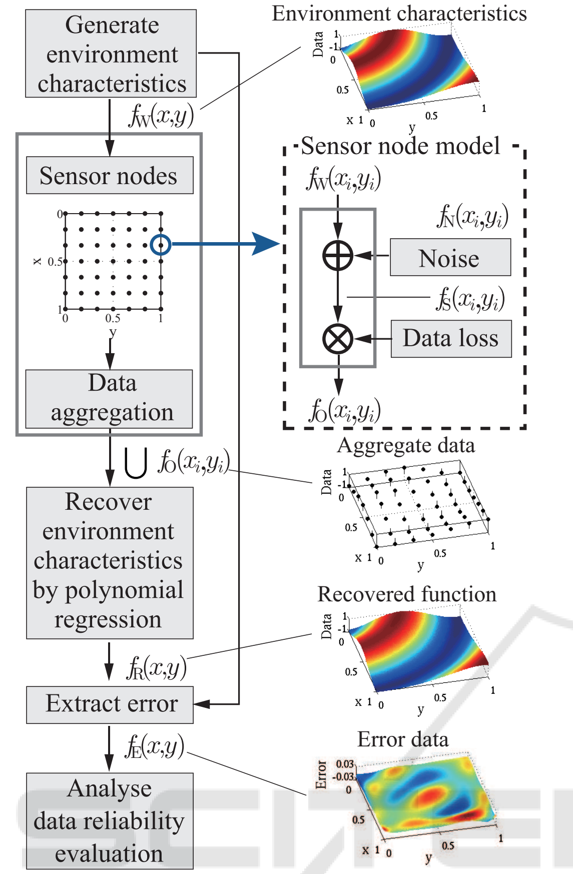


Figure 2: Data reliability evaluation flow (based on Yoneda et al., 2014).

Where $f_O(x_i,y_i)$ represents the sampled data considering the effect of noise and data loss. The amount of data in $f_O(x_i,y_i)$ decreases compared with the number of $f_S(x_i,y_i)$. The error of the recovered data is defined as

$$f_E(x,y) = f_R(x,y) - f_W(x,y) \tag{5}$$

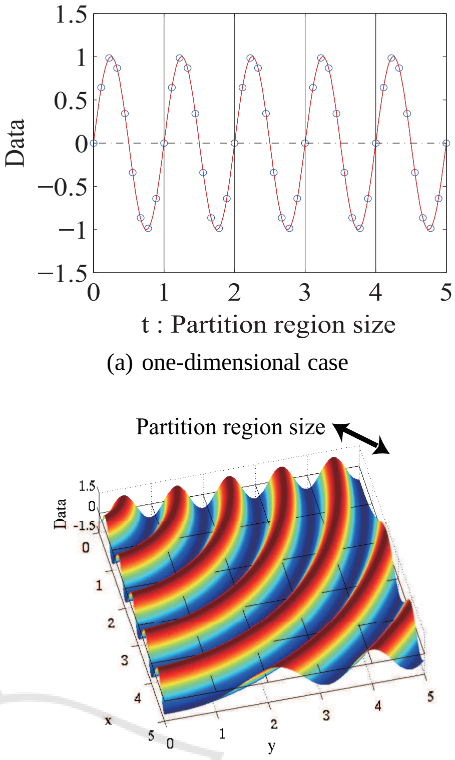
The data accuracy that are sampled at fixed intervals using the above continuous functions are compared with the accuracy of the evaluated data. The root-mean-square error (RMSE) at each comparison point is defined as data reliability. RMSE is given

$$\mathrm{RMSE} = \frac{1}{\sigma_W}\sqrt{\mathrm{mean}\left(f_E(x,y)^2\right)} \times 100(\%) \tag{6}$$

In the evaluation, we carry out 1000 iterations to minimize the effect of noise variation and data loss.

2.2.2 Parameter Setting

To evaluate the data recovery reliability, the following conditions are considered.



(a) one-dimensional case

(b) two-dimensional case

Figure 3: Examples of the recovery function.

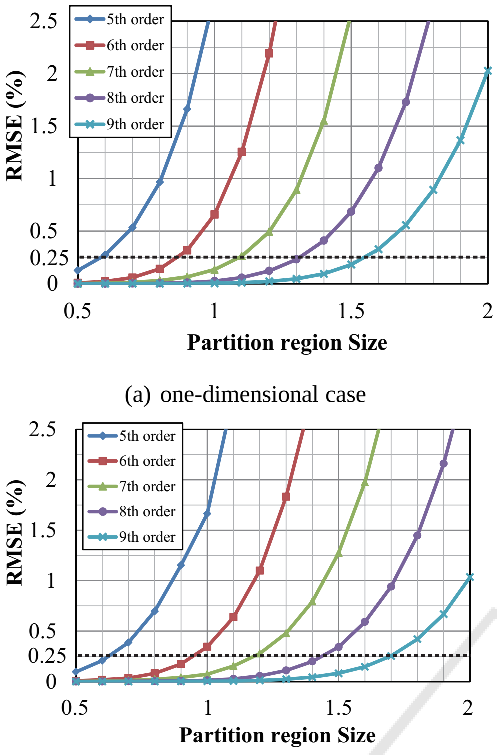
Sinusoidal Environmental Characteristics. The correlations among the actual environmental data are complex. However, to determine a generalized index, it is preferable to use simple data characteristics. In this study, a sinusoidal wave is assumed as the basic environmental characteristic because any arbitrary characteristic can be expressed as a linear combination of a sinusoidal wave. The following equations are the sinusoidal functions used for one- and two-dimensional data.

$$f_W(t) = \frac{A_{pp}}{2}\sin\left(2\pi\frac{t}{L}+\theta\right) \tag{7}$$

$$f_W(x,y) = \frac{A_{pp}}{2}\sin\left(2\pi\frac{\sqrt{(x-x_0)^2+(y-y_0)^2}}{L}+\theta\right) \tag{8}$$

Where (x_0,y_0) is showing the position of the wave generation source, A_{pp} is the peak-to-peak amplitude, L is the wavelength and θ is the phase.

Observation Region. The observation region is the region where the polynomial regression is applied. The observation region is partitioned and then polynomial regression is applied to each partition. Each data recovery function is joined to express the characteristics of the observation region. The partitions of the region are determined by the cycle



(a) one-dimensional case

(b) two-dimensional case

Figure 4: Precision of polynomial regression.

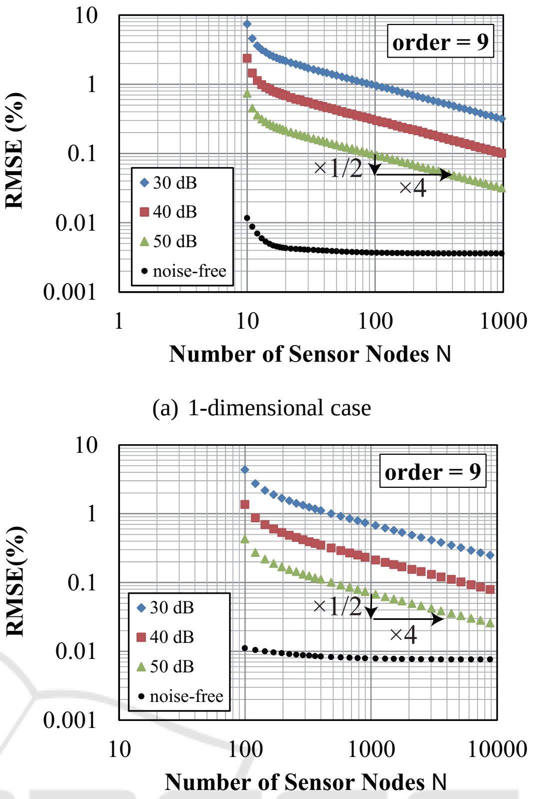
(wavelength) of the highest frequency of the environmental characteristics.

Number of Sensor Nodes. The sensor nodes are arranged at equal intervals in the observation region, including the upper boundary. In the case of two-dimensional structures, the sensor nodes are set on a grid. The density of the sensor nodes is represented by the number of sensor nodes N in the observation region. When the analysis is in the time domain, the one-dimensional coordinate axis is evaluated with respect to the time axis. In this case, the number of sensor nodes in the observation region represents the number of sampled data.

Noise. Electromagnetic noise generated at the sensor interface of an amplifier and an analogue-to-digital converter and electromagnetic noise in the environment mainly contribute to data noise. The SNR is defined by following equation.

$$\mathrm{SNR} = 10\log_{10}\frac{\mathrm{var}\,(f_S)}{\mathrm{var}\,(f_N)} \tag{9}$$

White noise (Gaussian noise) is added in the reliability evaluation.



(a) 1-dimensional case

(b) 2-dimensional case

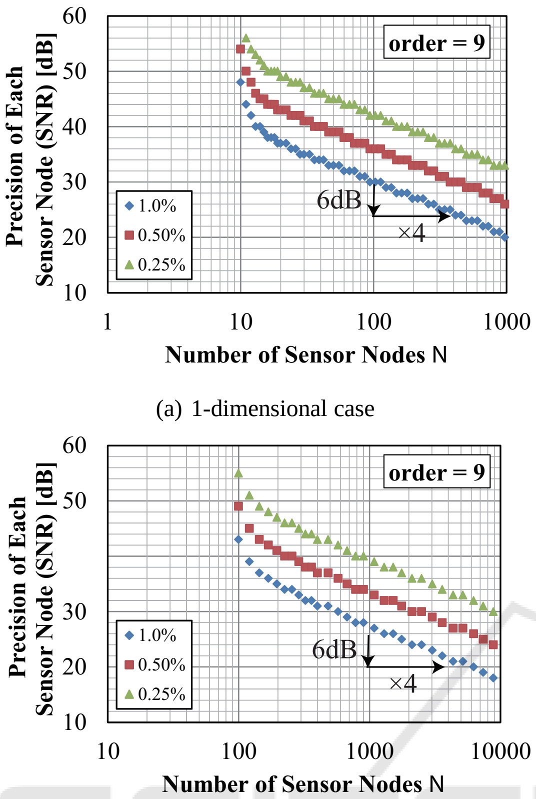
Figure 5: Required number of sensor nodes.

Data Loss. Data loss occurs because of data collisions or intermittent failures in the wireless communication. To simulate the effect of data loss, we use a pseudorandom data generation technique.

3 DATA RELIABILITY OF PERIODIC CHARACTERISTICS

The reliability of recovered data that are re-sampled from the recovered function is evaluated by comparing with the environmental source characteristics function. The data reliability is evaluated using the conditions described in the previous section. Figure 3 shows the sinusoidal signal that is assumed as the environmental characteristics of one- and two-dimensional conditions.

The SNR at each sensor node is set at 40 dB. Therefore, the reference position of the RMSE is determined at 1.0%. When the number of sensor nodes is increased, the reference position of the RMSE is 0.5%. Without sensor node noise, the reference position of the RMSE is 0.25%.



(a) 1-dimensional case

(b) 2-dimensional case

Figure 6: Sensor accuracy (based on Yoneda et al., 2014).

3.1 Application Range of the Polynomial Regression

Firstly, the relation between the partition region cycle and RMSE was analysed by changing the order of the polynomial without the sensor node noise. The number of sensor nodes is ten for the one-cycle partition region in the one-dimensional case and 10 × 10 for the one-cycle partition region in the two-dimensional case. We also examined the five-cycle partition region and monitored the maximum error. Results for the one- and two-dimensional sinusoidal signals (Fig. 3) are shown in Fig. 4. For 0.25% error and one-cycle partition region, the order of the polynomial should be higher than seven for one- and two-dimensional signals.

3.2 Effect of Sensor Node Number

The number of sensor nodes is thought to strongly affect the data evaluation reliability. The relation between the number of sensor nodes and RMSE was analysed for SNR of 50, 40 and 30 dB at each sensor node. And ninth-order polynomial was used in the analysis. The results for the one- and two-dimensional cases are shown in Fig. 5. Obviously, the errors reduced with the number of sensor nodes. The increasing number of sensor nodes reduced the RMSE owing to noise. Figure 6 shows the results for the required SNR at each sensor node when the RMSE is 0.25%, 0.5% and 1%. The precision of each sensor node is improved by increasing the number of sensor nodes.

There are two ways to improve the data evaluation reliability. The first is to increase the number of sensor nodes and the second is to use high SNR. If the number of sensor nodes is increased four times, the RMSE decreases 50%. If the SNR of each sensor node is improved by 6 dB, the RMSE decreases by 50%.

3.3 Effect of Data Loss

The relation between data loss rate and the RMSE was analysed. A ninth-order polynomial and 40-dB SNR at each sensor node was assumed. The number of sensor nodes was selected to satisfy the RMSE of 0.5% and 0.25%. In the one-dimensional case, 36 and 149 nodes were selected for the analysis. In the two-dimensional cases, 225 and 841 were selected.

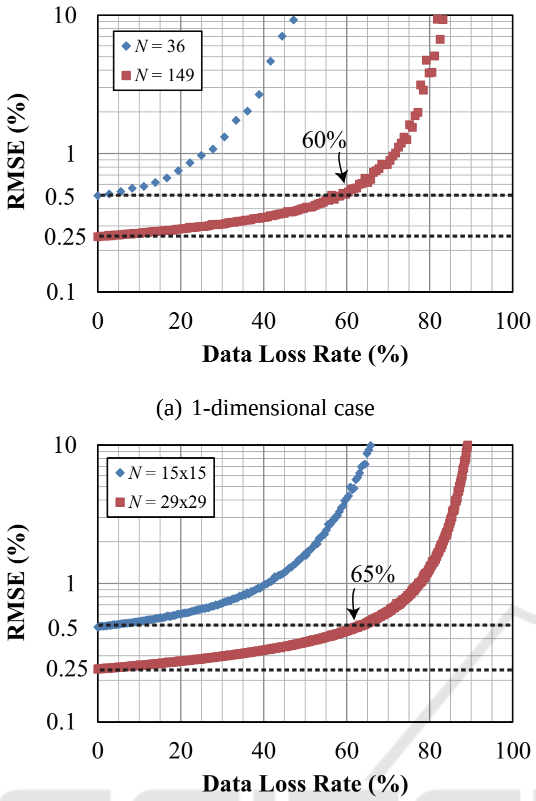
The results for the one- and two-dimensional cases are shown in Fig.7. The RMSE increases with data loss rate, of course. However, by increasing the number of sensor nodes, the RMSE decreases. The number of sensor nodes satisfies the RMSE of 0.25% adequately, whereas the data loss rate is 60% for RMSE of 0.5% in the one-dimensional case and 65% in the two-dimensional case. These results suggest that a redundant system can enhance the data reliability.

4 RELIABILITY IN THE FREQUENCY DOMAIN

The reliability of the recovered data using polynomial regression was also evaluated in the frequency domain (Ohba et al., 2015). The fast Fourier transform (FFT) was applied to the recovered data.

4.1 FFT Analysis

The one-dimensional sinusoidal environmental characteristics are assumed to be the same as in the previous sections. Recovered data at fixed intervals are obtained by sampling the data recovered by polynomial regression. FFT is applied to the recovered data. The signal-to-noise and distortion ratio (SNDR) and spurious-free dynamic range (SFDR) were evaluated.



(a) 1-dimensional case

(b) 2-dimensional case

Figure 7: Data loss robustness (based on Yoneda et al., 2014).

The SFDR is used to discuss the effect of harmonic distortion.

When the fundamental frequency is f_0, the number of FFT points is FFT_{POINT} and the sampling frequency is F_{S}, f_0 is

$$f_0 = \frac{F_{\mathrm{S}}}{FFT_{\mathrm{POINT}}} \tag{10}$$

Therefore, the input signal frequency f_{in} and wavelength of the input signal per division λ_{in} is

$$f_{\mathrm{in}} = m f_0 \tag{11}$$

$$\lambda_{\mathrm{in}} = \frac{m}{D_{\mathrm{n}}} \tag{12}$$

Where D_{n} is the number of divisions and m is an integer number.

4.2 Data Reliability in the FFT Analysis

4.2.1 Effect of Input Signal Cycle (Wavelength)

The relation between signal cycle (wavelength) in the polynomial regression and the evaluation indices of gain, SNDR and SFDR was analysed using FFT. In the analysis, a ninth-order polynomial, 40-dB SNR at each sensor node, 32 divisions dividing FFT points into partition region and 1024 of FFT points are assumed. The results are shown in Fig. 8. This is showing the relation between input frequency cycle (wavelength) for polynomial regression and the evaluation indexes. Decreases of 1 dB are tolerated by the SNDR and SFDR and for wavelength with the maximum partition of 1.6 cycles. Above 1.6 cycles, the partition region signals are filtered out. The gain is flat up to the three-cycle partition region. Thus, the polynomial regression acts as a low-pass filter. This means that the SNDR and SFDR improve because the polynomial regression limits the bandwidth of the environmental signals.

4.2.2 Effect of the Number of Sensor Node

The number of sensor nodes per partition region is evaluated. The results are shown in Fig. 9. The FFT results for the source environmental signals were 40-dB SNDR and 59-dB SFDR. For 13 sensor nodes, the SNDR is the same as the result of the source environmental signals. For 25 sensor nodes, the SNDR is the same as the result of source environmental signals. Higher SNDR and SFDR are possible by increasing the number of sensor nodes. By increasing the number of sensor nodes four times, both SNDR and SFDR improved by 6 dB.

4.2.3 Frequency Spectrum

The frequency spectrum is evaluated by FFT. The results are shown in Fig. 10. Based on the results of Figs. 8 and 9, the 1.6-cycle (wavelength) input signal region and 25 sensor nodes per partition region were assumed. Compared with the spectrum of the source environmental signal, the noise level of the high-frequency region is filtered out. Table 1 summarizes the FFT results and Table 2 summarizes the conditions of the FFT analysis. By limiting the observation region in the polynomial regression, both SNDR and SFDR are improved.

Table 1: FFT analysis result.

	SNDR[dB]	SFDR[dBc]
(A) FFT	40.0	59.0
(B) FFT using recovered data	45.2	59.9
(=B-A) Difference	+5.2	+0.9

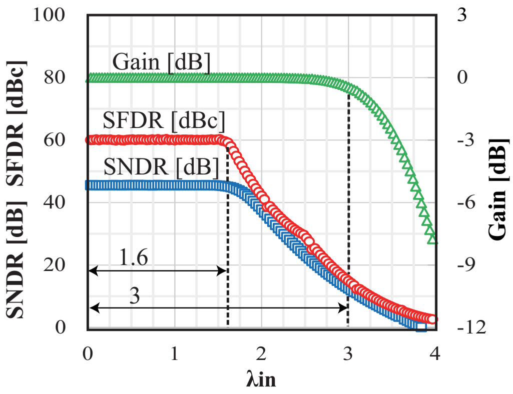


Figure 8: SNDR, SFDR and gain vs. input wavelength (based on Ohba et al., 2015).

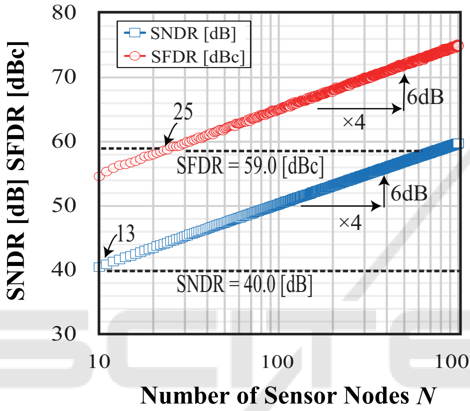


Figure 9: SNDR and SFDR vs. number of sensor nodes(based on K. Ohba, 2015).

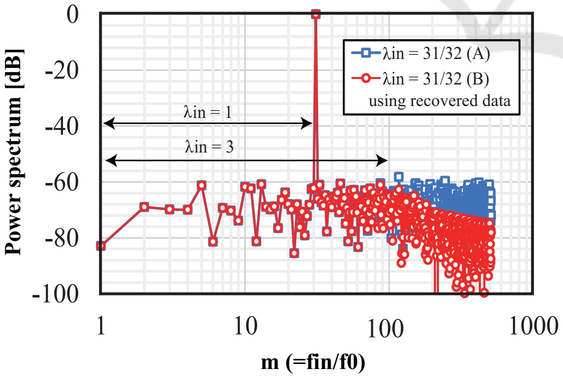


Figure 10: Frequency spectrum with and without recovered data.

Table 2: FFT analysis conditions.

Parameters	Conditions
Noise	40[dB]
Polynomial regression order	9
FFT points	1024
Division	32

5 ARBITRARY CHARACTERISTICS RECOVERY BY POLYNOMIAL REGRESSION

In sections 3 and 4, it was clarified that polynomial regression can recover the sinusoidal environmental characteristics. Polynomial regression for arbitrary characteristics is also validated by selecting the order of the polynomial equation for each partition region. Scale-space filtering (SSF) (Witkin, 1984) and Akaikefs information criterion (AIC) (Akaike, 1974) were used to select the partition region and the order of the polynomial regression based on aggregate data. The SSF detects extreme values by the convolution of the Gaussian function. The partition region is obtained as the region between the extreme points. The AIC is a statistical measure that estimate the quality of the environmental source characteristics from aggregate data, including noise and data loss. The order of the polynomial regression for the partition region is obtained by the SSF. By detecting the extreme values by SSF and estimating the quality of source characteristics using polynomial regression between the appropriately selected extreme points by AIC, arbitrary characteristics can be recovered (Haze, et al., 2012). Figure 11 shows extreme values of arbitrary environmental characteristics with 40-dB SNR detected by SSF. The observation region is divided into partition regions using the extreme values and the order of polynomial regression is thus optimized. The regions divided by the criterion of extreme values are the partition regions, and the order of the polynomial regression is set at each partition region. Figure 12 shows the recovered data from arbitrary characteristics with 40-dB SNR using the SSF, AIC and polynomial regression. The RMSE is under 0.1%; thus, the polynomial regression can obviously recover the arbitrary environmental characteristics.

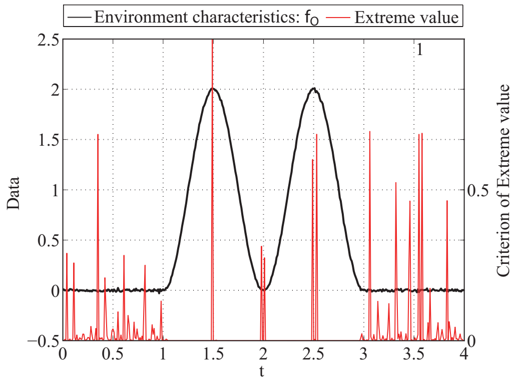


Figure 11: Extreme values of arbitrary characteristics.

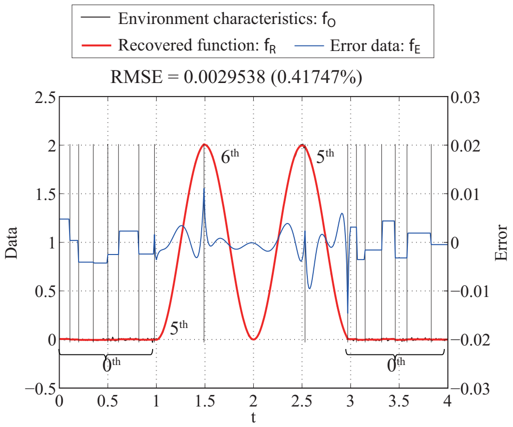


Figure 12: Data recovery function based on arbitrary characteristic.

6 CONCLUSIONS

We used polynomial regression for environmental data recovery from large-scale WSNs.
(1) A data reliability evaluation procedure for WSN was proposed.
(2) The recovered data reliability depends on the order of the polynomial regression, the number of sensor nodes and the effect of noise and data loss were quantified.
(3) From FFT analysis, it is seen that polynomial regression act as a low-pass filter. Data recovery using polynomial regression enhances the SNDR or SFDR in the WSNs system.
(4) Polynomial regression can recover arbitrary environmental characteristics and can be used with SSF and AIC.

In conclusion, environmental data recovery technique using polynomial regression can be applied to large-scale wireless sensor networks.

ACKNOWLEDGEMENTS

Part of this work was a joint research project of Tokyo Institute of Technology and Fujitsu Laboratories Limited. We would like to express our thanks to all who supported this project.

REFERENCES

Sivrikaya, F., et al., 2004. Time synchronization in sensor networks: a survey.h Network, IEEE, 18.4: 45-50.

K. Yamashita, et al., 2014. Implementation and Evaluation of Architecture Search Simulator Including Disturbance for Wide-range Grid Wireless Sensor Network.h Multimedia, Distributed, Cooperative, and Mobile Symposium. 1368-1377.

Doherty, L., et al., 2000. Algorithms for position and data recovery in wireless sensor networks.h Diss. Department of Electrical Engineering and Computer Sciences, University of California at Berkeley.

Y. Yoneda, et al., 2014. A study on Data Reliability Evaluation Index of Wireless Sensor Network for Environmental Monitoring.h The 41st SICE Symposium on Intelligent Systems.

K. Ohba, et al., 2015. A method of recovering environment data using polynomial regression for large-scale wireless sensor networks.h IEICE Technical Report ASN2015-72.

A. P. Witkin, 1984. SCALE-SPACE FILTERING: A New Approach To Multi-Scale Description,h *IEEE International Conference on Acoustics, Speech and Signal Processing*, Vol. 9, pp. 150–153, DOI:10.1109/ICASSP.1984.1172729.

H. Akaike, 1974. A New Look at the Statistical Model Identification,h *IEEE Transactions on Automatic Control*, Vol. AC-19, No. 6, DOI:10.1109/TAC.1974.1100705.

K. Haze, et al., 2012. Modeling home appliance power consumption by interval-based switching Kalman filters.h Technical Report of IEICE, 112.31: 39-44.