FORECAST OF METEOROLOGICAL DATA UTILIZING STATE-SPACE MODEL UTILIZING METRIC-MULTI DIMENSIONAL SCALING

Atsushi Asami[1], Yohei Saika[2]

[1]Department of Advanced Engineering Course,Gunma National College of Technology, 580 Toriba, Maebashi 371-8530 Japan
[2]Department of Information and Computer Engineering, National Institute of Technology, Gunma College, 580 Toriba, Maebashi 371-8530, Japan)

Abstract

Based on Bayesian statistics using the expected a posterior (EAP) estimation, we forecast time evolution of meteorological data, such as the temperature, at a target point via information on a set of time-series of the temperatures at sampling points selected by the metric multi-dimensional scaling (metric-MDS), without using information on that of the target point. Using numerical calculations with respect to the climate statistics in Kanto district, we clarify that the metric-MDS can select a set of sampling points whose data are similar to that of the target. Then, we clarify that the EAP estimation succeeds in predicting time evolution on the temperature in Maebashi using the set of time-series of the temperatures at the selected sampling points around Maebashi. Also, we find that the EAP estimation predicts the time evolution of the temperature more accurately than the conventional auto-regressive model.

Keywords: *forecast of meteorological data, Bayesian statistcs, space-state model, Ising spin, mult-dimensional scaling*

--***--

1. INTRODUCTION

For a long time, a lot of researchers have been investigating data-driven information technology [1]-[9] due to the development of computer technology. In this field, the researchers have applied various data-driven information techniques, such as Bayesian belief network [1]-[3] and visualization techniques [4]-[6] for various problems, such as probabilistic reasoning and decision making. Especially, in order to predict time evolution of large-scale data, many researchers have proposed a large variety of techniques [7]-[9] such as the auto- regression (AR) model and its variants. Also, various techniques in the Bayesian statistics have been studied for these problems. In recent years, theoretical physicists have applied statistical mechanics to various problems in information science and technology [10],[11], such as image restoration. Some researchers [12]-[14] have constructed time-series prediction for stock markets based on the statistical mechanics of Ising model.

On the other hand, a lot of researchers have studied meteorology, as a typical interdisciplinary field, to clarify meteorological phenomena bounded by earth's atmosphere. Currently, the global scale model (GSM) has been used for daily weather forecast in many countries. However, the weather forecast via the GSM is not so accurate for various industries, such as agricultural industry, due to the shortage of sampling points where meteorological data can be observed periodically. For this problem, researchers forecast meteorological data in local areas, some researchers have proposed techniques for forecasting meteorological data in local areas.

In this study, we constructed a technique of time-series prediction for meteorological information on the target point on the basis of Bayesian inference via the expected a posterior (EAP) estimation corresponding to the statistical mechanics of the Ising model. Here, we predict time evolution of temperature at a target point using the time-series of the temperatures observed at several sampling points selected by the visualization technique which is called as the metric multi-dimensional scaling (metric-MDS) [10], [11]. In this method, we first estimate the similarity between the set of time-series of the temperatures. Then, we investigate the performance of the present method for time-series prediction on the temperature at the target point (Maebashi). In this study, we used the set of the time-series of temperature at the target point (Maebashi) and 22 sampling points around Maebashi. First, in order to select the appropriate set of the time-series of the temperatures, we described the two-dimensional constellation with the use of the metric-MDS which visualized the similarity between the time- series of the temperatures at the sampling points. Numerical simulations with respect to 23 time-series of the temperatures clarified that the set of the temperatures at several sampling points evolved with high degree of similarity to that of Maebashi. Next, with the use of these temperatures selected by the metric-MDS, we carried out the time-series prediction of the temperature at Maebashi based on the Bayesian information processing using the expected a posterior (EAP) estimate. We clarified that the present method succeeded in the time-series prediction of the temperature in Maebashi utilizing the appropriate set of the time-series of the temperatures selected by the metric-MDS, if we set an initial condition on the hyper- parameters appropriately.

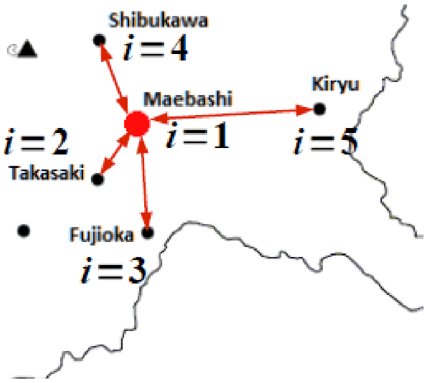


Fig. 1. An example of sampling points

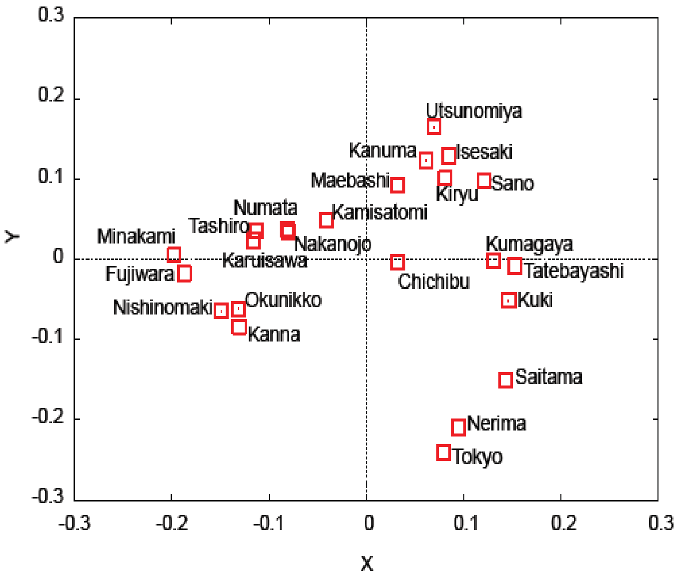
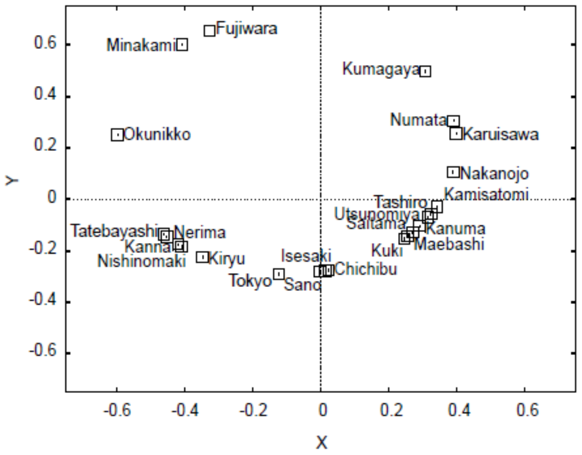
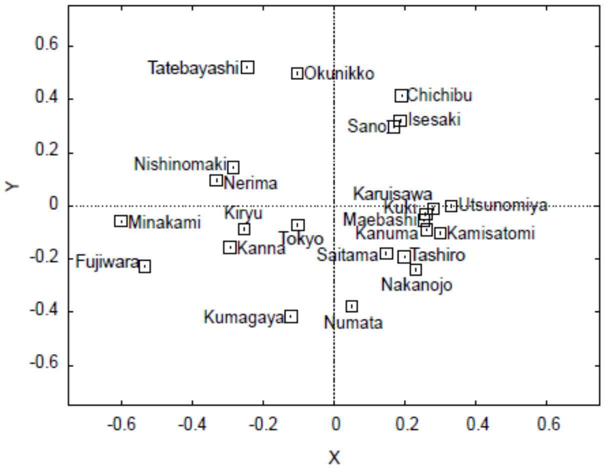


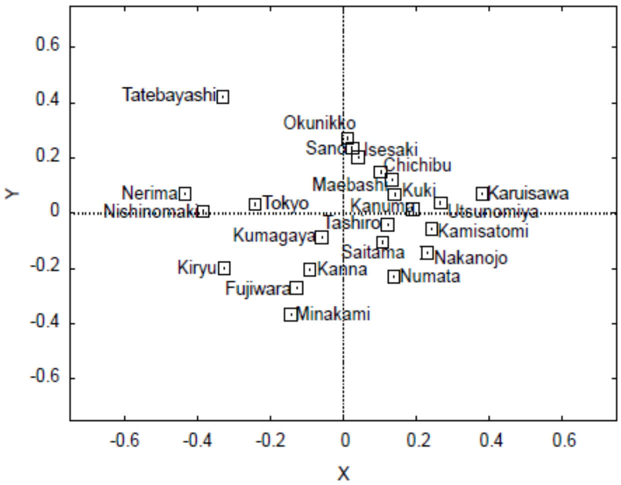
Fig.2 An example of constellation in two dimensions

Also, we found that the present method works more accurately than the conventional AR model under the appropriate initial condition. Next, we estimated the robustness of the present method with respect to the tuning of initial condition of the hyper-parameters. We first clarified that tuning the probabilistic fluctuations around the MAP solution are important to realize time-series prediction with high degree of accuracy. Also, we find that the ferromagnetic interactions between neighboring Ising spins are useful for improving the accuracy of the present method, if we appropriately tune the hyper- parameter J respective of the parameters h_1, h and β. Then, in order to clarify the availability of the time-series of the temperatures at neighboring sampling points for the present method, we evaluated how the mean squared error depended both on the number of sampling points and hyper-parameters. Here, we found that the accuracy in time-series prediction was improved by selecting the neighboring sampling points selected by the metric-MDS appropriately.

The content of this paper is as follows. First, we show the present method for time-series prediction on the temperature on the basis of the Bayesian inference using statistical mechanics of Ising model. Then, we show the visualization technique using the metric-MDS for the set of the time-series of the temperatures of the sampling points named as Maebashi and 22 sampling points around Maebashi. After that, we show the technique of the time-







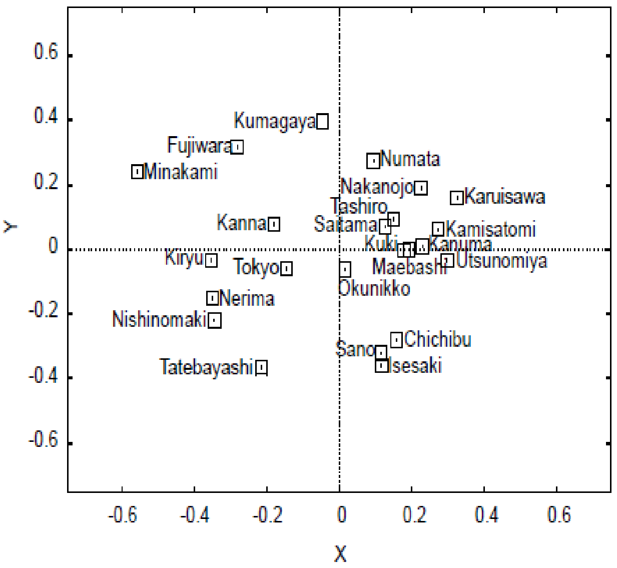


Fig. 3 (a) a constellation for the set of the time-series of the temperatures at the sampling points in Kanto district from Fig. 2 (a) a constellation of time-series of temperatures at Maebashi and its 22 neighbors from 12:00 to 15:00 11th July 2013, (b) a constellation of time-series of temperatures at Maebashi and its 22 neighbors from 11:00 to 15:00 11th July 2013, (c) a constellation of time-series of temperatures at Maebashi and its 22 neighbors from 10:00 to 15:00 11th July 2013, (d) a constellation of time-series of temperatures at Maebashi and its 22 neighbors from 10:00 to 15:00 11th July 2013.

series of the temperature in Maebashi using the set of the time-series of the temperatures of the sampling points selected among the 22 sampling points due to the metric-MDS in two dimensions. Further, we show the performance of the present method using numerical calculations for the previous data. The last part is developed to summary and discussion.

2. GENERAL FORMULATION

For a long time, a lot of researchers have been investigating data-driven information technology [1]-[9] due to the In this chapter, we show our formulation for forecast time evolution of the meteorological data, such as the temperature at the target point based on the expected a posterior (EAP) estimation by utilizing the time-series of temperatures at several sampling points which are selected using the MDS among the sampling points around the target one. First, we outline our general formulation for forecasting the time evolution of the temperature at the target point on the basis of the EAP estimate corresponding to statistical mechanics of the Ising model in two dimensions. In this formulation, we carry out the metric-MDS to select several sampling points whose temperatures synchronously change with that in Maebashi among the set of the sampling points in Kanto district. Next, with the use of information on these sampling points, we predict the time evolution of the temperature at the target sampling point. Here, we show the technique for forecasting the time evolution of the meteorological data at the target point by making use of the temperatures at the set of the sampling points including the target one on the basis of the Bayesian inference using the EAP estimation.

In this formulation, let us suppose a set of time-series on the temperatures $\{y_i(t)\}$. Here $y_i(t)$ denotes the time-series of the temperature at the i-th sampling point as

$$y_i(t) = y_i(t-T) + c\Delta y_i(t-T). \qquad (1)$$

Here, c is a scaling parameter we should set appropriately so that the absolute value of Δyi(t−T) is less than unity. Then, T denotes a sampling interval. Then, we forecast the time evolution of the temperature with the use of the set of the temperature differences $\varDelta y_i(t-T)$ (i=1,…,N) observed at the N sampling points. Then, in order to infer the time evolution of the temperature at the target point, we consider a set of Ising spins $\{S_i(t)\}$ in two dimensions. Here, $S_i(t)$ (=±1) is the Ising spin at the i-th sampling point at t. By making use of these Ising spins $\{S_i(t)\}$, we infer the difference of the temperature $\varDelta y_i$ (t) as

$$\hat{y}_i(t+T) = \hat{y}_i(t) + \Delta\hat{y}_i(t), \qquad (2)$$

$$\varDelta\hat{y}_i(t) = \sum_{\{Si\}(t)\}} \Pr(\{S_i^{(t)}\}) S_i^{(t)} \qquad (3)$$

which is averaged over the probability expressed as the Boltzmann distribution:

$$\Pr(\{S_i^{(t)}\}) = \frac{1}{Z_m} \exp\left[-\beta E(\{S_i^{(t)}\})\right] \qquad (4)$$

Where

$$E(\{S_i^{(t)}\}) = -\frac{J}{N}\sum_{j=2}^{N} S_1^{(t)} S_j^{(t)} - h_1 \sigma_\tau^{(1)} S_1^{(t)} - h\overline{\sigma}_\tau^{(t)} \sum_{j=2}^{N} S_j^{(t)}, \qquad (5)$$

$$\sigma_\tau^{(t)} = \frac{1}{\tau T}(y_i(t) - y_i(t-\tau T)), \qquad (6)$$

$$\overline{\sigma}_\tau^{(t)} = \frac{1}{\tau T}(\overline{y}_i(t) - \overline{y}_i(t-\tau T)), \qquad (7)$$

$$\overline{y}(t) = \frac{1}{N-1}\sum_{j=2}^{N} y_j(t). \qquad (8)$$

Here, J, h_1 and h are hyper-parameters. Then, the first term in the right hand side of eq. (5) is the prior information that time evolution of temperatures changes synchronously between at the target point and its neighbors, if these are located within the same plain. Then, the 2nd and 3rd terms in the right hand side of the eq. (5) are the likelihood which represents the long-term changes in the temperatures at those sampling points. Here, we note that we predict the time-series of the temperature at the target point without using the observed information on the target point, if we set to h_1=0, and then that we predict the time-series of the temperature without using any observed information, if we set to *h*1=0 and h=0.

Here, we show the technique called as the metric-MDS which is one of the techniques for visualization with respect to large-scale data. Especially, we use this technique to visualize the similarity of the time-series of the temperatures between the target and other sampling points by making use of the constellation of these data of the sampling points in two or three dimensions. Here, we show how to draw the constellation in two dimensions due to the metric MDS.

In order to draw the constellation of the time-series of the temperatures, we utilize the metric MDS. For this purpose, we first consider a set of the correlation coefficients:

$$\rho_{i,j}^{(t)} = \frac{\overline{r_i^{(t)} r_j^{(t)}} - (\overline{r_i^{(t)}})(\overline{r_j^{(t)}})}{\sqrt{\{\overline{(r_i^{(t)})^2} - \overline{(r_i^{(t)})}^2\}\{\overline{(r_j^{(t)})^2} - \overline{(r_j^{(t)})}^2\}}}, \qquad (9)$$

Where

$$\overline{r_i^{(t)}} = \frac{1}{W}\sum_{k=t-W+1}^{t} r_i^{(k)}, \qquad (10)$$

$$\overline{r_i^{(t)} r_j^{(t)}} = \frac{1}{W}\sum_{k=t-W+1}^{t} r_i^{(k)} r_j^{(k)}, \qquad (11)$$

$$r_i^{(t)} = y_i(t) - y_i(t-T). \qquad (12)$$

In above equations, W denotes the range of the time window. Then, we transform the correlation coefficients into a set of distance as

$$d_{i,j}^{(t)} = \sqrt{\frac{1-\rho_{i,j}^{(t)}}{2}}. \qquad (13)$$

Next, in order to describe the constellation of the set of the time-series of the temperatures, we map the set of the time-series of the temperatures {yi} onto a set of the points {x_i} (xi = (xi1, xi2,…, xiP), i=1,…,N) in the P-dimensional space. Here, the set of the points xi = (xi1, xi2,…, xiP) are satisfied with the relations:

$$d_{i,j} = \sqrt{\sum_{m=1}^{P}(x_{im} - x_{jm})^2} \qquad (14)$$

Then, if we consider three points x_i, x_j and x_k among the set of the points {x_i}, then we define an inner product between position vectors x_j and x_k as

$$x_j \cdot x_k = \sum_{m=1}^{P} x_{jm} x_{km} = \frac{1}{2}(d_{i,j}^2 + d_{i,k}^2 - d_{j,k}^2) \qquad (15)$$

with the use of the position vector x_i according to the law of cosines. Then, because the set of the position vectors {x_i} is translational invariant, so we may rewrite {x_i} into {x_i} as

$$\overline{x}_{jm} = x_{jm} - \frac{1}{N}\sum_{i=1}^{N} x_{im} \qquad (16)$$

so that the mass center of the set of time-series is located at the origin. In this representation, the inner product between x_j and x_k is then expressed as

$$\overline{x}_j \cdot \overline{x}_k = \sum_{m=1}^{P} \overline{x}_{jm}\overline{x}_{km} = \frac{1}{2}\left(\sum_{j=1}^{N}\frac{d_{jk}^2}{N} + \sum_{k=1}^{N}\frac{d_{jk}^2}{N} - d_{jk}^2 - \sum\sum\frac{d_{jk}^2}{N^2}\right). \qquad (17)$$

Further, we consider the N×N matrix B whose entry (*j*,*k*) is an inner product between x_j and x_k. Then we can describe the realistic symmetric matrix:

$$B = \overline{X}\,\overline{X}^T \qquad (18)$$

using P×N matrix:

$$\overline{X} = (\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_N)^T. \qquad (19)$$

As the matrix B is real symmetric, we carry out the eigenvalue decomposition using an orthogonal matrix Y as

$$B = \overline{X}\,\overline{X}^T = Y\Lambda Y^T \qquad (20)$$

where Y is the matrix composed by arranging eigenvectors of the matrix B perpendicularly. Then, Λ is the diagonal matrix that each entry is eigenvalues of the matrix B. The constellation of N points is obtained as

$$\overline{X} = Y\sqrt{\Lambda}\ . \qquad (21)$$

Here, is the matrix whose eigenvalues are the square root of those of the matrix B.

In order to clarify the performance of the present method for time-series prediction, we estimate the root mean square (RMSE):

$$RMSE = \sqrt{\frac{1}{t}\sum_{l=1}^{t}(y_1(l) - y_1(l))^2} \qquad (22)$$

which denotes the hour-wise accuracy between the observed and predicted values.

3. PERFORMANCE

In this chapter, we investigate the performance of the present method for the time-series prediction of the temperature at Maebashi using the EAP estimate by utilizing the time-series of the temperatures at Maebashi and 22 sampling points around Maebashi in Kanto district, such as Kusatsu and Tokyo. Here, as shown in Fig. 1, we use the set of time-series of the temperatures both at Maebahsi and 22 sampling points around Maebashi in Kanto district form 1st July 2013 to 30th July 2013. First, we examine the efficiency of the metric-MDS which selects the time-series of the temperatures at sampling points with high degree of similarity to Maebashi. Then, we examine the accuracy of the EAP estimate for the time-series prediction of the temperature at Maebashi.

Here, we investigate the performance of the metric-MDS to select the set of the sampling points whose time-series of the temperatures are similar to that of the target point. Here, we select the set of the sampling points among 22 sampling points located around Maebashi in Kanto district. Here, as shown in Fig. 2 and Figs. 3(a)-(d), we draw the constellations of the set of the time-series on the temperatures of 23 sampling points. Here, we set the range of time window to W=743 (3,4,5,6).

First, we examine which points are close to that in Maebashi from the constellation in Fig. 2. As shown in this figure, we find that the set of the points was distributed around the origin. We find that the sampling point closest to the target point (Maebashi) is Kanuma. Here, we note that the minimum distance from Maebashi to Kanuma is 0.3566 and the correlation coefficient between these sampling points is 0.7457. Then, we find that the time-series of the temperature at Maebashi is similar to those at Kanuma, kiryu, Kamisatomi, Utsunomiya Isezaki and so on.

Then, we examine how the accuracy of the time-series prediction depends on the range of the time window W, if we set the time window as W=3 (from t=12:00AM to 15:00), 4 (from t=11:00AM to 15:00), 5 (from t=11:00AM to 15:00), 6 (from t=10:00AM to 15:00). As seen from Figs. 3 (a)-(d), we find that the accuracy for the time-series prediction is dmin=0.0138 (0.0232, 0.0571, 0.1292), if we set to W = 3(4, 5, 6). These results show that the minimum distance dmin from Maebashi becomes longer with the increase in the time window from W=3 to W=6. These figures indicate that the similarity of the sampling point with minimum distance to the target point (Maebashi) depends on the choice of the time window W. For instance, the sampling point nearest to the target point (Maebashi) is Kuki (Chichibu), if we set to W=3, 4, 5 (W=6) in the pattern of the constellations.

In this chapter, we investigate how the accuracy of the EAP estimation for the time-series prediction on the temperature at Maebashi from 30th July 2013 to 1st, July, 2013. In this simulation, we use the sampling points selected by the metric- MDS in two dimensions among 22 sampling points in Kanto district.

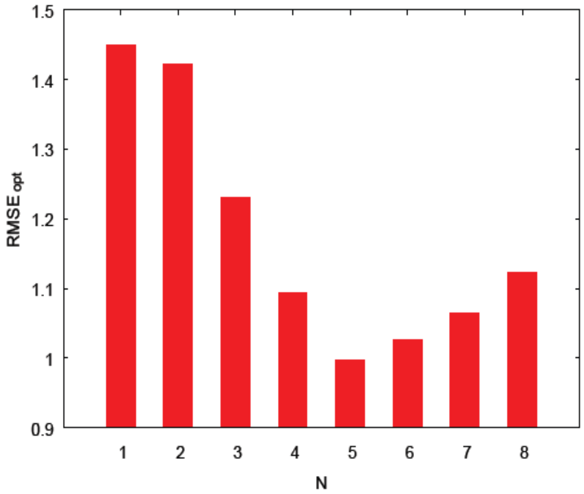


Fig. 4. Root mean square error as a function of the number of the sampling points composed of Maebashi and its neighbors. Each value is estimated under an appropriate condition.

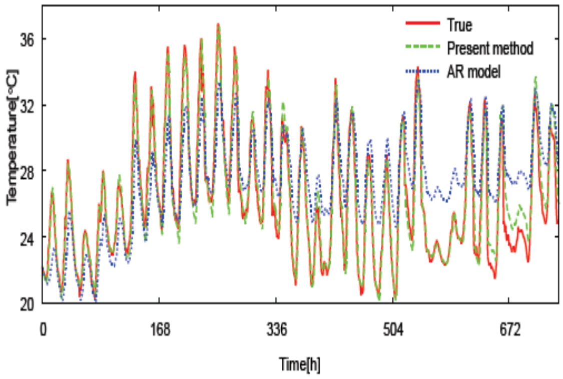


Fig. 5. Time-series of the temperatures in Maebashi (true), time-series of the temperatures predicted due to the EAP estimation and that due to the conventional AR model.

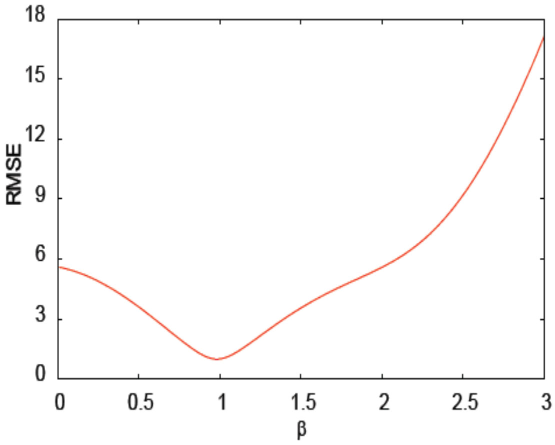


Fig. 6. RMSE as a function of the parameter *β*, if we set to h_1=0.05, *h*=0.05, *J*=0.35.

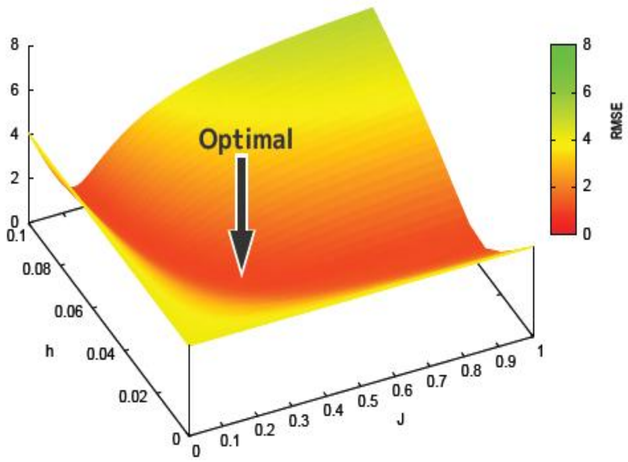


Fig. 7. RMSE as a function of the parameters h and J, if we set to β=0.98, h_1=0.05.

First, we examine how the accuracy of the present method depends on the number of the sampling points around Maebashi. In order to realize high performance, we utilize the set of time-series on the temperatures. Here, we select several sampling points around Maebashi by making use of the two-dimensional constellation in Fig.2. As shown in Fig. 4, we find that the accuracy for time-series prediction is improved with the increase in the number of the sampling points from N=1 to N=4, and that the optimal performance is achieved by using 4 time-series of the temperatures of sampling points, i.e. Kanuma, Isezaki, Kamisatomi and Sano. However, we find that the accuracy is suppressed with the increase in the number of the time-series of the temperatures, if N>6. This means that the time-series of the temperatures with high degree of similarity to that in Maebashi are available of improving the accuracy of the time-series prediction on the temperature in Maebashi.

Then, we first show in Fig. 4 the time evolution of the temperature in Maebashi from 1st July to 30th July in 2013 predicted both by Bayesian inference using the EAP estimation and the conventional auto regression (AR) model. This figure shows that the EAP estimation achieves the time-series prediction on the temperature with high degree of accuracy in comparison with the conventional AR model, if we use the set of the time-series of the temperatures of the sampling points, i.e. Kanuma, Kiryu, Utsunomiya, Sano and Maebashi, which are appropriately selected by the metric-MDS and set the hyper-parameters to h1=0.05, h=0.05, J =0.35 and β =0.98. Next, we evaluate how the RMSE depends on the parameters, such as β, J and h. As shown in Fig. 6, we find that the RMSE is minimized (RMSEopt= 0.9888) at β=0.98, if we set to J=0.35, h1=0.05 and h=0.05. Also, we find that the accuracy of the present method becomes worse in the limit of β→0 (β→∞). These results indicate that tuning the probabilitic fluctuations around the MAP solution of the assumed cost function in eq. (5). Then, as shown in Fig. 7, we evaluate how the RMSE depends on the hyper- parameters J, h1 and h, if we set to β=0.98. This figure indicate that the optimal performance is achieved, if we set to J=0.35, h1=0.05, h=0.05 and β=0.98. This indicates that the ferromagnetic interactions between Ising spins are available of improving accuracy for time-series prediction on the temperature in Maebashi, and that the external fields h1 and h, both of which enhance the observed information are also effective for improving the accuracy for this problem.

The above results indicate that the present method is useful for improving the accuracy of the time-series prediction by making use of several time-series of the temperatures around Maebashi.

4. SUMMARY AND DISCUSSION

In the previous chapters, based on statistical mechanics of the Ising model, we have constructed the technique for the time-series prediction on the temperature at the target point (Maebashi) by utilizing the set of the time-series of the temperatures around the target point. Then, we have estimated the performance of the present method for the time-series prediction with respect to the set of the time-series of the temperatures at the target point and 22 sampling points around the target point. Using numerical calculations for the set of time-series of the temperatures, we have found that the present method achieved the time-series prediction on the temperature at the target point by tuning the probabilitic fluctuations around the MAP solution, and also that the accuracy of the present method was improved by utilizing the set of the time-series of the temperatures at the sampling points appropriately selected by using the metric-MDS. Also, we find that the present method succeeded in time-series prediction at the point where no information was observed, if we utilized the set of the time-series of the temperatures appropriately selected by the metric-MDS under the appropriately condition.

As a future problem, we are going to apply this technique to the time-series prediction on the electricity power used in the small-scale organization.

REFERENCES

[1] MacKay, D. J. (2003) *Information Theory, Inference and Learning Algorithm*, U. K., Cambridge University Press.

[2] Bishop, C. M. (2006) Pattern Recognition and Machine Learning, New York, Springer-Verlag.

[3] Barber, D. (2012) *Bayesian Reasoning and Machine Learning*, U. K., Cambridge University Press.

[4] Borg, I, Groenen, P. (2005) Modern Multidimensional Scaling: theory and applications (2^{nd} ed.), New York, Springer-Verlag.

[5] Cox, T. F. Cox, M. A. A. (2001) *Multidimensional Scaling*, New York, Chapman and Hall.

[6] Borg, I, Groenen, P. J. F, Mair, P. *Applied Multidimensional Scaling (SpingerBriefs in Statistics)*, New York, Springer-Verlag, 2012.

[7] Kitagawa, G. (1987) Non-Gaussian state-space modeling of nonstationary time series", Journal of American Statistics Association, vol. 82, pp. 1032-1063.

[8] M. West, J. Harrison, *Bayesian Forecasting and Dynamic Models*, New York, Springer-Verlag, 1996.

[9] A. Gelman, J. B. Carlin, H. S. Stern and D. B. Rubin, *Text in Statistical Science*, *Bayesian Data Analysis (Second Edition)*, New York, Chapman & Hall., 2003.

[10] Nishimori, H. (2001) Theory of spin glasses and information; An introduction, Oxford, London.

[11] K. Tanaka, "Statistical mechanical approach to image processing," *Journal of Physics A:Mathematical and General*, vol. 35, no. 37, pp. R31-R150, 2002.

[12] T. Mohri, H. Tanaka, "Weather Forecasting by Memory-Based Reasoning", *Journal of Japan Society of Artificial Intelligence*, pp. 798-805, 1995.

[13] T. Hotaka, M. Nakagawa, "Learning and Prediction of Environmental Time Series with Chaos Recurrent Neural Network", *ICICE Technical Report, NLP2006-75*, pp. 51-56, 2006.

[14] Y. Michihiro, Y. Sato, Y. Suzuki, "Development and Application of Climate Change Information Database", Prev. Res. Inst., Kyoto Univ., No. 54B, pp. 747-755, 2011.