Spatio-Temporal Prediction of Meteorological Time Series Data: An Approach Based on Spatial Bayesian Network (SpaBN)

Monidipa Das[(⊠)] and Soumya K. Ghosh

Department of Computer Science and Engineering,
Indian Institute of Technology Kharagpur, Kharagpur 721302, India
monidipadas@hotmail.com, skg@iitkgp.ac.in

Abstract. This paper proposes a *space-time model* for prediction of meteorological time series data. The proposed prediction model is based on a *spatially extended Bayesian network* (SpaBN), which helps to *efficiently* model the *complex spatio-temporal dependency* among large number of spatially distributed variables. Validation has been made with respect to prediction of daily *temperature, humidity,* and *precipitation rate* around the spatial region of *Kolkata, India.* Comparative study with the benchmark and state-of-the-art prediction techniques demonstrates the superiority of the proposed spatio-temporal prediction model.

Keywords: Space-time model · Time series prediction · Spatial Bayesian network · Meteorology

1 Introduction

Spatio-temporal data are ubiquitous in the nature. The meteorological time series data, as collected from spatially distributed weather stations or from sensor networks, are one of the prominent examples in this respect. Prediction of meteorological time series is essential not only to anticipate the weather condition for taking adequate measures, but also it helps in proper management of energy [6]. However, the major challenge in meteorological time series prediction is the complex spatio-temporal inter-relationships among the variables. Modeling such space-time dependency becomes more complicated when the number of spatially distributed influencing variables becomes considerably large. For the same reason, the effectiveness of many of the existing space-time prediction models, especially those are based on graph-based approaches, are significantly hindered.

In the present work, we have proposed an improved, graph-based, probabilistic model, so as to handle such spatio-temporal prediction scenario. The approach is based on the *spatial Bayesian network* (SpaBN) [3], which can efficiently model the influence from large number of spatially distributed variables. Previously, SpaBN has shown encouraging performance in the domain of hydrology [3]. In the present work, we have applied SpaBN as the base technology behind our proposed space-time model for meteorological time series prediction.

1.1   Problem Statement and Contributions

The relevant prediction problem can be formally stated as follows:

– Given, the historical daily time series data set over n meteorological variables in $M = \{m_1, m_2, \cdots, m_n\}$, corresponding to a set of l spatial locations $L = \{l_1, l_2, \cdots, l_l\}$ for previous t years: $\{y_1, y_2, \cdots, y_t\}$. The problem is to determine the daily times series of the variables in M, for any location $x \in (L \cup Z)$, for future q years $\{y_{(t+1)}, y_{(t+2)}, ..., y_{(t+q)}\}$, where, $(Z \cap L) = \phi$ and q is a positive integer.

In this context, spatio-temporal prediction of any meteorological variable $m \in M$ needs to consider influences of its co-located variables in M, from several spatially distributed locations in L. Though the graphical models, like Bayesian networks, are highly suitable for modeling such influences [2], however, considering separate influencing nodes corresponding to each location $l_i \in L$ makes the graphical structure as well as the analysis process extremely complex.

Present work attempts to address this issue by utilizing the effective modeling ability of the spatial Bayesian network (SpaBN). The major contributions in this regard are as follows:

– Exploring the *spatial Bayesian network* (SpaBN) analysis in *multivariate prediction* of meteorological time series data;
– Proposing SpaBN based *spatio-temporal prediction model* which is capable of *efficiently modeling* complex inter-variable dependency over space and time;
– Validating the proposed space-time model with respect to prediction of daily *temperature, humidity* and *precipitation rate* around Kolkata, India;
– Demonstrating the *effectiveness* of the proposed prediction model in comparison with benchmark and state-of-the-art space-time prediction techniques;

The remainder of the paper is organized as follows. The details of the proposed space-time model has been illustrated in Sect. 2. The results of empirical study have been reported in Sect. 3. Finally, we conclude in Sect. 4.

2   Proposed Space-Time Model

As depicted in the Fig. 1, the proposed prediction model comprises of three key steps: (i) *data pre-processing*, (ii) *spatial weight calculation*, and (iii) *spatio-temporal prediction*. Each of these steps are illustrated below:

2.1   Data Pre-processing

The main objective of the data pre-processing step is to discretize the continuous meteorological variables, so as to make these suitable for the discrete Bayesian analysis of SpaBN in the subsequent steps. The discretization is performed by considering maximum and minimum observed value ($max(m_i)$ and $min(m_i)$) in the historical time series for each continuous variable $m_i \in M$, and then dividing the whole range into appropriate number (R) of bins or sub-ranges of desired size $size(subRange) = \frac{max(m_i) - min(m_i)}{R}$. The value of R is determined empirically.
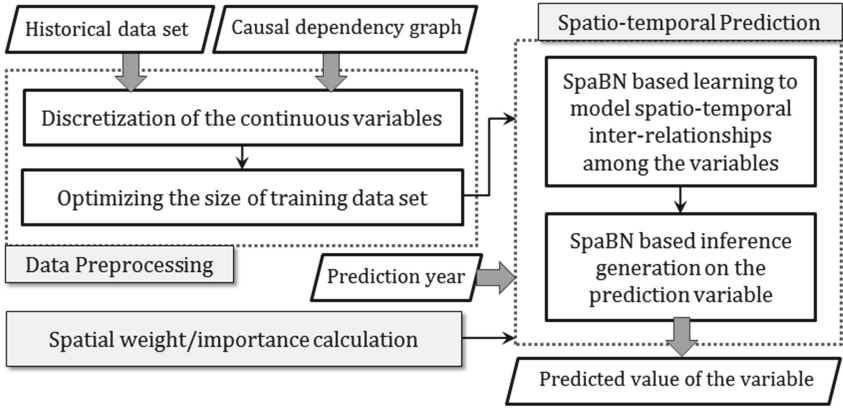
Fig. 1. Framework of the proposed spatio-temporal prediction model

Moreover, in the pre-processing step we also optimize the size of training data set with consideration to the temporal variability of the meteorological variables within short period of time. For example, in general, rainfall shows monthly variation. Therefore, in order to make prediction for a particular day, the rainfall data of the associated month is considered rather than considering the rainfall data of the whole year.

2.2 Spatial Weight/Importance Calculation

This step aims at determining the *spatial importance* or *spatial weight* (SW_i) of each location $l_i \in L$, with respect to the prediction location. The spatial weight SW_i is measured based on the *spatial distance* (SD_i) and the *correlation between the time series of each variable* in the neighborhood locations and that in the prediction location. Suppose, $NCorr^i_{m_j}$ is the *normalized correlation* value between the time series of variable m_j in the i-th neighborhood location and that in the prediction location, such that $NCorr^i_{m_j} \in [0,1]$. Then, the spatial weight of the location is determined as follows:

$$SW_i = \frac{\sum_{j=1}^{|M|} NCorr^i_{m_j} + NISD_i}{\sum_{k=1}^{|L|}(\sum_{j=1}^{|M|} NCorr^k_{m_j} + NISD_k)} \qquad (1)$$

where, $NISD_i$ is the normalized *inverse spatial distance* of i-th location from the prediction location, such that $NISD_i \in [0,1]$.

2.3 Spatio-Temporal Prediction

During the prediction process in the proposed space-time model, the effect of spatial influence of the meteorological variables are learnt with the help of SpaBN analysis. In order to describe the learning process, let's consider an example

scenario, where M_1, M_2 and M_3 are three arbitrary meteorological variables. Also let M_1 is independent, M_2 is influenced by M_1, and M_3 is influenced by M_1 and M_2. Now, because of the inherent spatio-temporal inter-relationships among these variables, variable at one location is also influenced by the variables in its neighborhood locations. Therefore, considering a causal dependency graph, comprising of the representative variables from all the neighboring locations, will ultimately lead to a complex graphical structure for capturing spatio-temporal inter-relationships among the variables (refer Fig. 2a).

In such scenario, *spatial Bayesian network* (SpaBN) can be an appropriate tool for modeling these spatio-temporal inter-relationships in an efficient manner. As shown in the SpaBN structure (refer Fig. 2b), all the standard/classical nodes associated with the same but spatially distributed variable have been replaced with *composite nodes*, denoted by double lined circles. The replacement of all such standard nodes with a single composite node reduces both the structural and the algorithmic complexity in Bayesian analysis to a great extent.
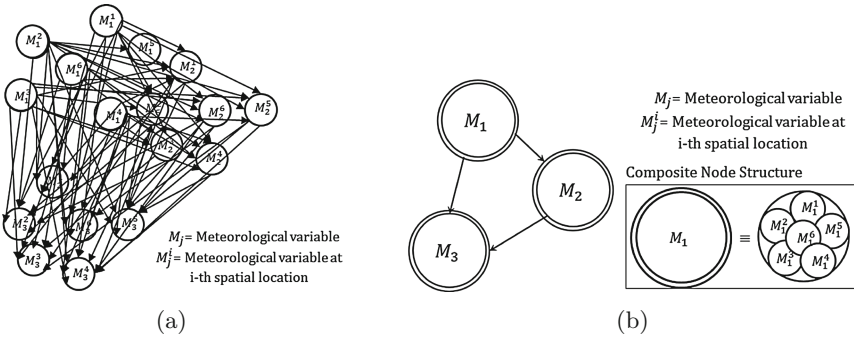

(a)   (b)

Fig. 2. Graph based modeling of spatio-temporal inter-relationships among the three meteorological variables: (a) a complex causal dependency graph of standard BN, (b) equivalent SpaBN structure [considering no. of locations $|L| = 6$]

Let $|L|$ be the number of neighboring locations considered. Then, according to the principle of SpaBN, the marginal and conditional probabilities of the variables in Fig. 2b are estimated in following fashion:

$$P(M_1) = \gamma \cdot \left[\sum_{i=1}^{|L|} P(M_1^i) \cdot SW_i \right] \tag{2}$$

$$P(M_2) = \gamma \cdot \left[\sum_{i=1}^{|L|} P(M_2^i) \cdot SW_i \right] \tag{3}$$

$$P(M_3) = \gamma \cdot \left[\sum_{i=1}^{|L|} P(M_3^i) \cdot SW_i \right] \tag{4}$$

$$P(M_2|M_1) = \gamma \cdot \left[\sum_{i=1}^{|L|} \frac{n(M_1^i, M_2^i)}{n(M_1^i)} \cdot SW_i \right] \tag{5}$$

$$P(M_3|M_1, M_2) = \gamma \cdot \left[\sum_{i=1}^{|L|} \frac{n(M_1^i, M_2^i, M_3^i)}{n(M_1^i, M_2^i)} \cdot SW_i \right] \tag{6}$$

where, SW_i is the spatial weight/importance of the i-th neighboring location with respect to the prediction location; γ is the normalization constant; and $n(< \cdot >)$ is the count of observation for the variable value combination $< \cdot >$.

Now, in order to capture the *temporal evolution* of the inter-variable dependencies, the SpaBN based learning process (as described above) is performed with historical data of each year separately, and finally the probabilistic estimates are combined in a weighted manner, so as to achieve the probability distributions corresponding to the inter-variable dependencies in the prediction year. Higher temporal weight (tw) is assigned to a year which is temporally nearer to the prediction year. The overall process of space-time learning is summarized through Algorithm 1, considering immediately next prediction year $y_{(t+1)}$.

Algorithm 1. Space-time learning using SpaBN

1: **Input:** Historical data set $H = \{ H_{y_1}, \cdots, H_{y_t} \}$ for the past t years: $\{y_1, \cdots, y_t\}$; Causal dependency graph over all meteorological variables $m \in M$

2: **Output:** Probabilistic relationships among the meteorological variables for the prediction year.

3: P_F^m = Final probability estimate for a variable m.

4: P_i^m = Probability estimate corresponding to the the variable m, for the training year y_i

5: **for** each training year y_i, $(0 \leq i \leq t)$ **do**

6:     Apply SpaBN analysis to learn the probabilistic relationships among the variables using H_{y_i}.

7:     **for** each considered meteorological variable m **do**

8:         **for** each probability estimate P_i^m in the probability table of m **do**

9:         $d_i = [y_{(t+1)} - y_i]$; /* calculating temporal distance of y_i from prediction year */

10:         $tw_i = \left(\frac{\frac{1}{d_i}}{\sum_{j=1}^{t} \frac{1}{d_j}} \right)$; /* Estimating temporal weight/importance for the year y_i */

11:         $P_F^m = P_F^m + (tw_i \times P_i^m)$; /* probability updating for the prediction year */

12:     **end for**

13:     **end for**

14: **end for**

15: return P_F^m for all m in the causal dependency graph of SpaBN

Once the parameter learning is over, the inference is generated as per SpaBN by utilizing the spatial weights (SW_i). For example, let the observed/ evidence variables are: M_1 and M_2, from which the value of M_3 is to be inferred.

Then, as per the principle of SpaBN,

$$Inferred\,value\,of\,M_3 = \sum_{i=1}^{|L|} P(M_3^i|M_1^i, M_2^i) \cdot SW_i \tag{7}$$

where the value for $P(M_3^i|M_1^i, M_2^i)$ can be determined from the conditional probability table for the variable M_3. Among these inferred values, the predicted value becomes the one corresponding to the maximum probability.

3 Experimental Evaluation

This section describes the empirical study carried out for evaluating our proposed space-time prediction model.

3.1 Data Set and Study Area

The proposed prediction model has been validated by forecasting daily time series of three primary meteorological variables, namely *Temperature, relative Humidity*, and *Precipitation rate*, around *Kolkata [22.57° N, 88.36° E], India*. The corresponding historical daily time series data have been collected from the FetchClimate Explorer [4] for a span of 10 years (2006–2015). Predictions have been carried out for two locations (*Loc-1* [22.93°N, 87.25°E] and *Loc-2* [22.82°N, 88.29°E]), for the year 2016 (refer Fig. 3).
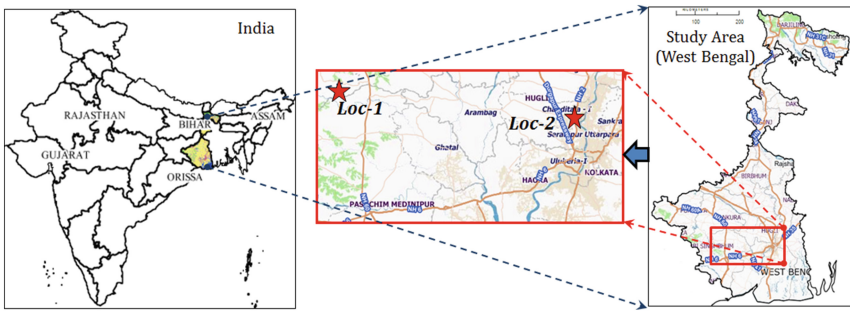

Fig. 3. *Study Area* around *Kolkata* [22.57°N, 88.36°E] in *West Bengal (India)*

3.2 Results

The comparative study has been made with benchmark prediction techniques, like *automated ARIMA* (R-Tool), *standard BN* (SBN), ANN (MATLAB nnTool) etc. and state-of-the-art space-time models, namely *hierarchical Bayesian autoregressive* (HBAR) model [5], and *spatio-temporal ordinary Kriging* (ST-OK) [1]. The prediction performance has been measured in terms of *four* popular statistical measures, namely *normalized root mean square deviation* (NRMSD) [3], *Pearson's correlation coefficient* (CC), *mean absolute error* (MAE), and *mean absolute percentage error* (MAPE) [2]. The best-fit between the observed and predicted value yields NRMSD = 0, CC = 1, MAE = 0 and MAPE = 0. The results of prediction have been summarized in Tables 1, 2 and 3.

Discussions: On analyzing the results, presented in Tables 1, 2 and 3, the following inferences can be drawn:

Table 1. Comparative study of *Temperature* prediction

Prediction approaches	Loc-1				Loc-2			
	Prediction metrics				Prediction metrics			
	NRMSD	MAE	MAPE	CC	NRMSD	MAE	MAPE	CC
A-ARIMA	0.329	3.219	7.463	0.098	0.211	3.064	7.278	0.137
ANN	0.212	1.771	3.698	0.711	0.569	0.761	2.529	0.929
SBN	0.099	1.035	0.147	0.935	0.951	1.028	0.085	0.929
HBAR	0.203	2.245	2.336	0.756	0.221	2.371	3.899	0.687
ST-OK	0.191	2.175	0.179	0.778	0.248	2.300	0.179	0.805
Proposed approach	**0.098**	**1.003**	**0.038**	**0.937**	**0.101**	**0.997**	**0.072**	**0.936**

Table 2. Comparative study of *Relative humidity* prediction

Prediction approaches	Loc-1				Loc-2			
	Prediction metrics				Prediction metrics			
	NRMSD	MAE	MAPE	CC	NRMSD	MAE	MAPE	CC
A-ARIMA	0.323	8.333	2.187	0.603	0.325	8.873	2.093	0.606
ANN	0.158	3.038	4.068	0.942	0.252	6.077	6.524	0.791
SBN	0.077	1.976	2.029	0.980	0.171	2.850	4.080	0.922
HBAR	0.156	3.696	2.782	0.910	0.150	3.353	1.357	0.904
ST-OK	0.192	4.970	0.789	0.868	0.288	6.255	4.463	0.790
Proposed approach	**0.069**	**1.328**	**0.669**	**0.989**	**0.064**	**1.315**	**1.150**	**0.985**

- In all the cases, the proposed prediction model produces the least NRMSDs and MAEs, which are significantly lesser than the other prediction techniques considered. This indicates superiority of our SpaBN based prediction, compared to the others.
- The high values of CC ($\approx$1) (refer Tables 1, 2 and 3) reveals that the series predicted by our model have the best match with the observed time series.
- The least MAPE values corresponding to the proposed space-time prediction model also demonstrate its better efficacy in comparison with the others.

Moreover, the improvement in computation time of the proposed model, with respect to standard Bayesian network (SBN) based prediction, has also been studied considering number of neighboring location $|L| = 6$, and variable domain size $R = 3$. The result (refer Table 4) proves the effectiveness of using SpaBN for modeling spatio-temporal dependency among the spatially distributed meteorological variables, as accomplished by our proposed space-time prediction model.

Table 3. Comparative study of *Precipitation rate* prediction

Prediction approaches	Loc-1				Loc-2			
	Prediction metrics				Prediction metrics			
	NRMSD	MAE	MAPE	CC	NRMSD	MAE	MAPE	CC
A-ARIMA	0.401	81.274	37.245	0.388	0.410	86.448	37.718	0.387
ANN	0.202	28.286	20.277	0.912	0.153	24.462	15.913	0.940
SBN	0.217	12.947	19.572	0.985	0.098	21.457	11.917	0.976
HBAR	0.410	92.910	81.257	0.840	0.374	88.518	71.203	0.863
ST-OK	0.260	51.856	5.175	0.797	0.259	51.869	5.346	0.807
Proposed approach	**0.063**	**11.684**	**1.602**	**0.990**	**0.038**	**06.818**	**1.449**	**0.996**

Table 4. Comparative study of computation time (considering $|L| = 6$, $R = 3$, *single* prediction day, and SBN with *spatially distributed* variables)

Prediction techniques	Standard BN (SBN)	Proposed approach
Execution time (s)	1500.254	0.068195

4   Conclusions

The objective of the present work is to address the challenge of handling complex spatio-temporal dependency among the meteorological variables during multivariate time series prediction. For that purpose, we have proposed a *space-time model* based on spatial Bayesian network (SpaBN) which is inherently capable of efficiently modeling the inter-dependency among large number of spatially distributed variables. Experimental study has been carried out in comparison with several benchmarks (ARIMA, SBN, ANN) and state-of-the-art prediction techniques (HBAR, ST-OK). Overall, the proposed space-time model has shown encouraging performance with respect to both *accuracy* and *computational cost* in meteorological time series prediction.

References

1. Cressie, N., Wikle, C.K.: Statistics for Spatio-Temporal Data. Wiley, London (2015)
2. Das, M., Ghosh, S.K.: semBnet: a semantic Bayesian network for multivariate prediction of meteorological time series data. Pattern Recogn. Lett. **93**, 192–201 (2017)
3. Das, M., Ghosh, S.K., Gupta, P., Chowdary, V., Nagaraja, R., Dadhwal, V.: FORWARD: a model for FOrecasting Reservoir WAteR Dynamics using Spatial Bayesian Network (SpaBN). IEEE Trans. Knowl. Data Eng. **29**(4), 842–855 (2017)
4. Microsoft-Research: FetchClimate (2015). http://research.microsoft.com/en-us/um/cambridge/projects/fetchclimate/app/. Accessed Jan 2016
5. Sahu, S.K., Bakar, K.S.: Hierarchical Bayesian autoregressive models for large space-time data with applications to ozone concentration modelling. Appl. Stoch. Models Bus. Ind. **28**(5), 395–415 (2012)
6. Voyant, C., Nivet, M.L., Paoli, C., Muselli, M., Notton, G.: Meteorological time series forecasting based on MLP modelling using heterogeneous transfer functions. J. Phys. Conf. Ser. **574**, 1–6 (2015). IOP Publishing