BESTED: An Exponentially Smoothed Spatial Bayesian Analysis Model for Spatio-temporal Prediction of Daily Precipitation

Monidipa Das*
Department of Computer Science and Engineering
Indian Institute of Technology Kharagpur, India
monidipadas@hotmail.com

Soumya K. Ghosh
Department of Computer Science and Engineering
Indian Institute of Technology Kharagpur, India
skg@iitkgp.ac.in

ABSTRACT

This paper proposes a novel data-driven model (BESTED), based on *spatial Bayesian network* with *incorporated exponential smoothing mechanism*, for predicting precipitation time series on daily basis. In BESTED, the spatial Bayesian network helps to *efficiently model the influence of spatially distributed variables*. Moreover, the incorporated exponential smoothing mechanism aids in tuning the network inferred values to *compensate for the unknown factors*, influencing the precipitation rate. Empirical study has been carried out to predict the daily precipitation in *West Bengal*, *India*, for the year 2015. The experimental result demonstrates the superiority of the proposed BESTED model, compared to the other benchmarks and state-of-the-art techniques.

CCS CONCEPTS

• **Information systems** → **Spatio-temporal systems**; *Geographic information systems*;

KEYWORDS

Spatial Bayesian network, Spatio-temporal prediction, Time series, Exponential smoothing, Precipitation

ACM Reference format:
Monidipa Das and Soumya K. Ghosh. 2017. BESTED: An Exponentially Smoothed Spatial Bayesian Analysis Model for Spatio-temporal Prediction of Daily Precipitation. In *Proceedings of SIGSPATIAL'17, Los Angeles Area, CA, USA, November 7–10, 2017,* 4 pages.
https://doi.org/10.1145/3139958.3140040

1 INTRODUCTION

Prediction of environmental precipitation plays crucial role in various disciplines, including climatology, hydrology, agriculture, transportation and so on. An accurate prediction of precipitation rate can facilitate in improving the outcomes of flooding model, runoff model, crop-growth model etc. and eventually may help in socioeconomic development of a region. However, precipitation is a complex spatio-temporal phenomena. In order to properly model this process, significant efforts have been made till date.

The proposed BESTED (*spatial Bayesian network with Exponential smoothing mechanism for Spatio-Tem-poral prEDiction*) is a probabilistic, data-driven computational approach, which aims at overcoming some of the major challenges, faced by probabilistic graphical models (like BNs) in spatio-temporal prediction. The overall prediction problem, associated challenges, and the contributions of the present work are discussed in the subsequent subsections.

1.1 Problem Statement and Challenges

The prediction problem in the present context can be formally defined as follows:

- Given, historical daily time series data set over precipitation and other influencing meteorological factors, 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 time series of precipitation for any location $x \in (L \cup Z)$ for future m years $\{y_{(t+1)}, ..., y_{(t+m)}\}$. Here, Z is a set of k new locations $\{z_1, z_2, ..., z_k\}$, such that $z_i \notin L$, for $i = 1$ to k, and m is a positive integer, i.e. $m \in \{1, 2, 3, \cdots\}$.

The major challenges in such prediction problem arise mainly due to the spatio-temporal nature of the data involved. Two of these key challenges are described below.

1) *Challenge due to influence from spatially distributed environmental variables:* Being a spatio-temporal variable, the precipitation at a new location and/or at a new time-instant can be predicted based on the historical data of precipitation and other known influencing factors from neighborhood locations. Though the graphical models, like Bayesian networks, are highly suitable for representing such inter-variable influences, yet, for each such influencing variable, introducing representative nodes corresponding to each spatial location leads to extremely high structural and algorithmic complexity of these models (refer Figure 2 [left]).

2) *Challenge due to unknown influencing factors:* Another common challenge faced by many prediction models is that the information about all factors influencing the prediction variable is not known always. For example, precipitation is not only dependent on the level of humidity, wind speed, temperature, latitude, altitude etc., but also on several other factors, like atmospheric current, ocean current and many

*Corresponding Author, Tel: +91 3222-281440

SIGSPATIAL'17, November 7–10, 2017, Los Angeles Area, CA, USA

ACM ISBN 978-1-4503-5490-5/17/11.
https://doi.org/10.1145/3139958.3140040

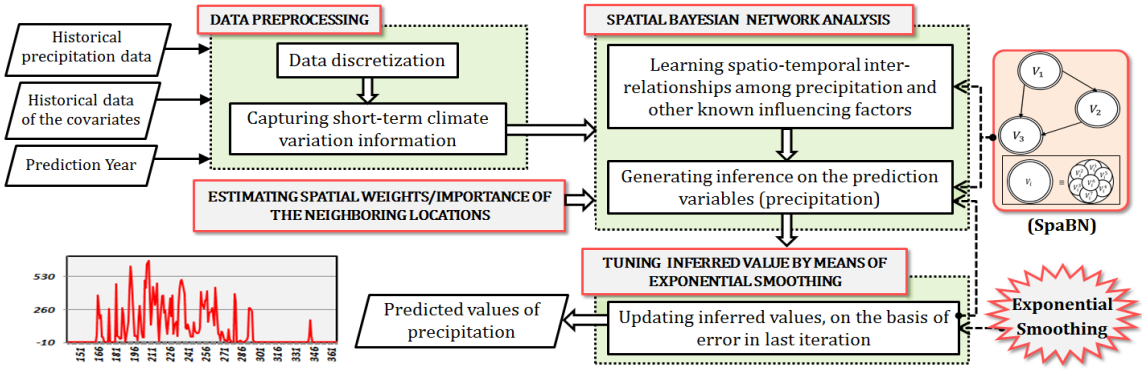


Figure 1: Workflow of the proposed BESTED model

more, which may be even unknown. Therefore, training of a prediction model in absence of these information always leads to some imperfection in the prediction process.

1.2 Contributions

Our proposed prediction model addresses these two challenges by using *spatial Bayesian network* (SpaBN) [2] with *incorporated exponential smoothing mechanism*. The key contributions in this regard are as follows:

- proposing a data-driven model (BESTED) for spatio-temporal prediction of daily precipitation time series;
- deriving an improved spatial BN analysis method with integrated exponential smoothing mechanism;
- validating the efficacy of the BESTED model with an empirical study on predicting daily precipitation in the state of *West Bengal, India*;
- performing comparative study with the well-known benchmarks (e.g. ARIMA, VARMA etc.), and state-of-the-art data-driven prediction approaches, based on ANN, SVM, standard BN, spatial BN, hierarchical Bayesian auto-regressive (HBAR) model, spatio-temporal kriging (ST-OK) etc.

2 BESTED: A SPATIO-TEMPORAL PREDICTION MODEL

As shown in the Figure 1, the flow of the BESTED model consists of four major steps: 1) Data pre-processing, 2) Spatial weight/importance calculation, 3) Spatial Bayesian network analysis, and 4) Tuned inference generation.

2.1 Data pre-processing

The step of data pre-processing comprises of data discretization and capturing short-term variation from the historical time series of each influencing meteorological factors.

Discretization makes the historical data suitable for discrete spatial Bayesian analysis in the subsequent step. If v_{min} and v_{max} are the minimum and maximum value observed in the historical data of variable v, then the length/size (S) of the discretized range is calculated as: $S = \frac{v_{max} - v_{min}}{R_c}$, where, R_c is the number of discretized ranges for the variable v.

On the other side, the step of capturing short-term climatic variation helps to prepare an optimal size data set for training purpose. For example, in general, the precipitation shows monthly variation, and therefore, in case of prediction for a particular day, the historical data of corresponding month is more insightful, compared to the data of the whole year.

2.2 Spatial weight/importance calculation

This step aims at assigning some appropriate weight to each of the locations l_i at the neighborhood of the prediction location x. The value of the weight reflects the strength of spatial influence of the location l_i on x. In the proposed BESTED model, the spatial weight is estimated as follows:

Let SW_i be the spatial weight for neighboring location l_i, SD_i be the *spatial distance* between l_i and prediction location (x), and $NCorr_v^i$ be the normalized correlation between the meteorological variable v at the location l_i and that at the prediction location x. Then the spatial weight (SW_i) for location l_i is estimated as follows:

$$SW_i = \frac{\sum_v NCorr_v^i + NISD_i}{\sum_{j=1}^{K}(\sum_v NCorr_v^j + NISD_j)} \quad (1)$$

where, $NISD_i$ is the normalized inverse spatial distance between l_i and x, such that $NISD_i \in [0, 1]$; K is the total number of neighboring locations considered.

2.3 Spatial Bayesian network analysis

In this step, the BESTED model learns the spatio-temporal inter-relationships between precipitation and the other influencing factors. Then, based on given evidences, it infers the precipitation rate value for the prediction day.

The process is carried out on the basis of spatial Bayesian network (SpaBN) analysis, proposed in our earlier work [2]. For any spatially distributed variable, instead of introducing representative node for each spatial location, as depicted in Figure 2 [left], the SpaBN replaces all such nodes with a single *composite node* (refer Figure 2 [right]), and thereby significantly reduces the network complexity from both representational as well as computational perspective.

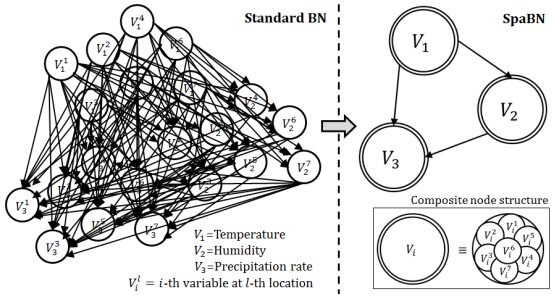


Figure 2: Complex causal dependency graph of standard BN and the equivalent SpaBN structure

2.3.1 Learning spatio-temporal inter-relationships among precipitation and other known influencing factors. In order to explain the SpaBN based learning process in BESTED model, let's consider the SpaBN structure in Figure 2, representing causal dependency between temperature (V_1), humidity (V_2), and precipitation (V_3), spatially distributed in seven locations. Let K be the number of neighboring locations considered. Then, according to the principle of SpaBN, the marginal and conditional probabilities are estimated in following manner (example given with respect to the variable V_3):

$$P(V_3) = \gamma \cdot \left[\sum_{i=1}^{K} P(V_3^i) \cdot SW_i \right] \tag{2}$$

$$P(V_3|V_1, V_2) = \gamma \cdot \left[\sum_{i=1}^{K} \frac{n(V_1^i, V_2^i, V_3^i)}{n(V_1^i, V_2^i)} \cdot SW_i \right] \tag{3}$$

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

In a similar fashion, in order to make prediction for any day, the network is trained with the data of each previous year $(y_1, y_2, \cdots, y_t)$ separately, to learn the associated probabilistic relationships among the variables during each year. At the end of training for each year, the marginal and conditional probability estimates P_p^v (corresponding to each variable v) for the prediction year y_p is generated by honoring the temporal auto-correlation property [3], in following manner:

$$P_p^v = \sum_{i=1}^{t} \left(P_{y_i}^v \times \frac{1/d_i}{\sum_{j=1}^{t} 1/d_j} \right) \tag{4}$$

where, d_i is the temporal distance of training year y_i from prediction year y_p; t is the total number of training years.

2.3.2 Generating inference on precipitation. Once the parameter learning is over, the inference for precipitation is generated as per SpaBN by utilizing the spatial weights (SW_i). For example, let the observed/ evidence variables are: *temperature* (V_1) and *humidity* (V_2), from which the value of *precipitation* (V_3) is to be inferred. Then, as per SpaBN,

$$\text{Inferred value of precip. } (V_3) = \sum_{i=1}^{K} P(V_3^i|V_1^i, V_2^i) \cdot SW_i \tag{5}$$

where the value for $P(V_3^i|V_1^i, V_2^i)$ can be determined from the conditional probability table for precipitation (V_3). Among these inferred values, the predicted value becomes the one corresponding to the maximum probability estimate.

2.4 Tuning of inferred values

Significance: One of the major issues in probabilistic graphical prediction model is that it is not always known precisely which variable influences which other. In that case, due to the lack of appropriate influencing nodes in the dependency graph, the modeling of spatio-temporal interrelationships using graphical model becomes a challenging task. The absence of major influencing variables/nodes in the graph may cause inadequate parameter learning, which can eventually lead to poor inference generation from the model. Therefore, the objective in this step is to tune the inferred value of precipitation, as generated by SpaBN, in such a way that the absence of influencing nodes can be recompensed at some level. For that purpose, the proposed BESTED model uses *exponential smoothing mechanism.*

Let, at the end of training with data of past t years, the inferred value of precipitation for a particular day in prediction year $y_{(t+1)}$ is $I_{(t+1)}$. Then, the tuned inferred value becomes:

$$I'_{(t+1)} = I_{(t+1)} + \epsilon_t \tag{6}$$

where, the ϵ_t is termed as the *tuning component* and it is recursively determined as follows:

$$\epsilon_t = (\alpha E_{t-1}) + (1 - \alpha)\epsilon_{t-1} \tag{7}$$

Here, $\alpha \in [0, 1]$ is the *smoothing factor* and E_{t-1} is the error in inference corresponding to the same day in year $y_{(t-1)}$ and is calculated as follows:

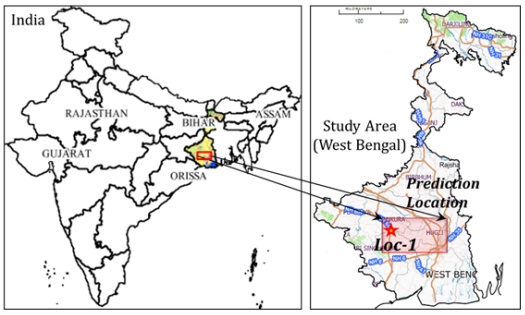
$$E_{t-1} = ActualValue - I'_{t-1} \tag{8}$$

3 EXPERIMENTATION

The experimentation has been carried out in a spatial region in the state of *West Bengal, India* (refer Figure 3). In order to predict *precipitation* (P), two more meteorological parameters, namely, *temperature*(T) and *relative humidity* (H), have been chosen as the influencing co-variates. Historical data[1] of 2010-2014, has been used to predict precipitation rate at the *Loc-1*[22.93°N, 87.31°E], for the year 2015. Nine more locations have been chosen as the neighboring locations from the study area and let these be denoted as *Loc-2*, *Loc-3*, and so on. Estimation of spatial weight/importance with respect to the prediction location *Loc-1* is shown in the Figure 3.

The comparative study has been performed with eight other pure statistical and computational intelligence based techniques, including automated ARIMA, VARMA, HBAR (*Hierarchical Bayesian Auto-Regressive model* [5]), ST-OK (*Spatio-Temporal Ordinary Kriging* [1]), ANN, SVM, standard BN (SBN), and SpaBN. MATLAB (NNTool) has been utilized to perform prediction using ANN, and for implementing the SBN and SpaBN[2] based predictions, whereas

[1] http://research.microsoft.com/en-us/um/cambridge/projects/fetchclimate/app/



Spatial weight/importance estimation with respect to *prediction location* (Loc-1)

	Loc-2	Loc-3	Loc-4	Loc-5	Loc-6	Loc-7	Loc-8	Loc-9	Loc-10
$NCorr_T$	0.5096	0.9872	0.9763	0.9880	0.5424	0.3461	0.9906	0	0.9619
$NCorr_H$	0.3025	0.9968	0.9918	0.9892	0.4392	0.2399	0.9878	0	0.9507
$NCorr_P$	0.7237	0.9928	0.9855	0.9941	0.7412	0.5550	0.9935	0	0.9743
$NISD$	0.0299	0.7734	0.6083	0.7800	0.0383	0.0189	0.7781	0	0.3453
Spatial Weight/ Importance(SW)	0.0695	0.1664	0.1581	0.1665	0.0782	0.0515	0.1664	0	0.1434

Figure 3: *Study area* in *West Bengal* (India)

to predict precipitation using A-ARIMA, VARMA, HBAR, ST-OK and SVM, the R-tool (version 3.2.2) has been used.

Table 1: Comparative study of daily *precipitation rate* prediction in *West Bengal, India*

Prediction Approaches	Performance Metrics			
	NRMSD	**MAE**	**MAPE**	R^2
A-ARIMA	0.377	81.274	37.245	0.388
VARMA	0.406	93.19	55.731	0.387
ANN	0.190	28.286	20.277	0.912
SVM	0.066	14.437	9.651	0.994
HBAR	0.385	92.91	81.257	0.840
ST-OK	0.260	51.856	5.175	0.797
SBN	0.066	12.947	1.957	0.985
SpaBN	0.059	8.762	1.602	0.990
BESTED (proposed)	**0.029**	**7.622**	**0.700**	**0.996**

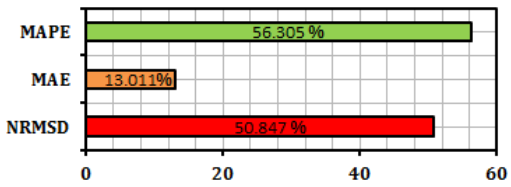


Figure 4: Percentage improvement of BESTED in comparison with pure SpaBN based prediction

3.1 Results

The performance of BESTED model has been measured in terms of four popular statistical measures: *normalized root mean square deviation (NRMSD)* [2], *mean absolute error* (MAE), *mean absolute percentage error* (MAPE)[4], and *Coefficient of determination or R-squared* (R^2). The experimental results have been summarized in Table 1 and in Figure 4.

Discussions: On analyzing the results, summarized in Table 1 and in Figure 4, the following inferences can be drawn:

- In almost all the cases of prediction, the proposed BESTED has outperformed the other prediction techniques by producing least NRMSD and least MAE values. This proves the superiority of the BESTED model in precipitation prediction.
- The MAPE measures also show a significantly lesser value (0.70%) for the BESTED model, demonstrating its efficacy in learning spatio-temporal dependency between precipitation and other influencing factors.
- The high values of R^2 (≈ 1) in every case also indicate that the series predicted by the BESTED model have the best match with the observed precipitation time series of the corresponding year.
- Moreover, it is evident from Figure 4 that, in every case, the proposed BESTED model has attained significant improvement over the performance of pure SpaBN based analysis which does not consider the inference tuning process. The average improvement in NRMSD, MAE, and MAPE are ≈51%, ≈13%, and ≈56%, respectively. This proves the effectiveness of exponentially smoothing the inferred value of precipitation, as adopted by the BESTED model.

4 CONCLUSIONS

The present work proposes BESTED, a data-driven model for spatio-temporal prediction of daily precipitation time series. The novelties in this work are twofold: 1) unlike the existing probabilistic graphical models of precipitation prediction, the proposed BESTED has an incorporated facility of modeling influences from spatially distributed variables in an efficient manner. This is achieved by utilizing the spatial Bayesian network (SpaBN) analysis; 2) the integrated exponential smoothing mechanism in BESTED model offers a superior inference generation ability, which can compensate for the absence of unknown factors influencing precipitation. The experimental results are found to be encouraging. Most significantly, the proposed exponentially smoothed spatial Bayesian analysis in BESTED shows ≈ 56% improvement in mean absolute percentage error of prediction, compared to the standard spatial Bayesian network [2] based analysis.

REFERENCES

[1] Noel Cressie and Christopher K Wikle. 2015. *Statistics for spatio-temporal data*. John Wiley & Sons.
[2] Monidipa Das, Soumya Ghosh, Pramesh Gupta, VM Chowdary, R Nagaraja, and VK Dadhwal. 2017. FORWARD: A Model for FOrecasting Reservoir WAteR Dynamics using Spatial Bayesian Network (SpaBN). *IEEE Transactions on Knowledge and Data Engineering* 29, 4 (2017), 842–855.
[3] Monidipa Das and Soumya K Ghosh. 2014. A probabilistic approach for weather forecast using spatio-temporal interrelationships among climate variables. In *IEEE 9th International Conference on Industrial and Information Systems*. IEEE, 1–6.
[4] Monidipa Das and Soumya K Ghosh. 2017. semBnet: A Semantic Bayesian Network for Multivariate Prediction of Meteorological Time Series Data. *Pattern Recognition Letters* 93 (2017), 192–201.
[5] Sujit Kumar Sahu and Khandoker Shuvo Bakar. 2012. Hierarchical Bayesian autoregressive models for large space–time data with applications to ozone concentration modelling. *Applied Stochastic Models in Business and Industry* 28, 5 (2012), 395–415.