Pattern Recognition Letters 93 (2017) 192–201

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

semBnet: A semantic Bayesian network for multivariate prediction of meteorological time series data

Monidipa Das*, Soumya K. Ghosh

Department of Computer Science and Engineering, Indian Institute of Technology, Kharagpur 721302, India

ARTICLE INFO

Article history:
Available online 4 January 2017

Keywords:
Bayesian network
Semantic similarity
Domain knowledge
Spatial semantics
Time series prediction
Meteorology

ABSTRACT

Meteorological time series prediction plays a significant role in short-term and long-term decision making in various disciplines. However, it is a challenging task involving several issues. Sometimes, the available domain knowledge may help in dealing with certain issues in this regard. This work proposes a multivariate prediction approach based on a variant of *semantic Bayesian network*, termed as *semBnet*. The key objective of *semBnet* is to incorporate the spatial semantics as a form of domain knowledge, in standard/classical Bayesian network (SBN), and thereby improving the accuracy of meteorological prediction. It has been shown that compared to SBN, the proposed *semBnet* is less prone to parameter value uncertainty. Empirical studies on multivariate prediction of *Temperature, Humidity, Rainfall* and *Soil moisture* demonstrate the superiority of proposed approach over *linear* statistical models (e.g. ARIMA, *spatio-temporal ordinary kriging* (ST-OK)), and *non-linear* prediction techniques based on ANN, SBN, *hierarchical Bayesian autoregressive* model (HBAR) etc. Most significantly, compared to SBN, the proposed *semBnet* shows average 24% improvement in *mean absolute percentage error* of prediction.

1. Introduction

The prediction of meteorological time series such as *temperature, rainfall, soil moisture, wind speed, relative humidity, atmospheric pressure* etc. plays significant role in various disciplines, including weather control, climate impact assessments, agriculture, water system management, and so on. However, the two major challenges in this regard are: 1) complex spatio-temporal interrelationships among the meteorological variables; and 2) influence of various spatial attributes, like latitude, land cover category, land elevation etc. For example, as investigated by Ding et al. [15] and Bagley et al. [2], precipitation is highly influenced by the *surface elevation* and *land-use land-cover (LULC)* type of a region. In this situation, the spatial semantics of the influencing factors can aid in prediction process by providing some added insights. For instance, the *land surface temperature* (LST) of any two places, one belonging to an *urban area* and the other belonging to a *mining region* are influenced more or less in a similar fashion (assuming all other factors to be constant) as both the locations fall under same LULC category which is *'built-up'*. On the contrary, the LST of two other locations, one at an *urban area* and the other inside an *evergreen forest* are influenced in a considerably different manner [37], since in this case the two locations belong to two different LULC categories, namely: *'built-up'* and *'forest'* respectively. Therefore, the domain knowledge on spatial semantics can play an important role in determining meteorological conditions of any location.

Down through the years, numerous models have been proposed for predicting meteorological time series. Most of these are based on various linear statistical processes such as auto-regressive moving-average (ARMA), AR integrated moving-average (ARIMA) [9,18], spatio-temporal kriging [17] etc., or based on computational intelligence (CI) techniques, like artificial neural network (ANN) [32,36], Bayesian network (BN) [1,12], support vector machine (SVM) [7], chaos theory [13] and so on.

Among these CI techniques, the Bayesian network (BN), that can intuitively represent the relevant dependencies among numerous variables, is very much suitable for multivariate prediction in meteorology [6]. With its directed acyclic graph, BN can automatically capture probabilistic information from data and can reason with uncertain knowledge [14,35]. However, one of the major problems with BN is that a proper learning of the network needs large amount of observed data be available during training. Otherwise, it may lead to strongly biased inference results with full of uncertainty [8]. It has been observed by Luo et al. [24] and Chang et al. [8] that in such case, a prior knowledge, more specifically a prior qualitative semantic knowledge, about the respective domain, may help in many ways to adjust the uncertainty. In this regard, two key objectives in our work are:

* Corresponding author.
E-mail address: monidipadas@hotmail.com (M. Das).

http://dx.doi.org/10.1016/j.patrec.2017.01.002
0167-8655/

1. Incorporation of spatial semantics in the Bayesian network model for improving the Bayesian analyses;
2. Employing this semantically enhanced Bayesian network for better modeling of spatio-temporal inter relationships among meteorological parameters and spatial attributes.

1.1. Existing works on semantically enhanced Bayesian network

Although the Bayesian networks with incorporated semantics have proved their usefulness in a number of applications, it is still not a much explored area. A few notable variants of semantic Bayesian network can be found in the works by Kim et al. [19], Butz et al. [5], Zhou et al. [38], and Madsen and Butz [25] respectively.

Kim et al. [19] have used their proposed semantic Bayesian network (SeBN) in a conversational agent to infer the detailed intentions of the user. In SeBN, the network itself contains both probabilistic and semantic relationships. The inference generation is followed by thresholding process to select the appropriate target value corresponding to the user query. Zhou et al. [38] have used semantic Bayesian network (sBN) for web mashup network construction, where sBN has been used to process all information sources on the semantic web. In order to process a semantic graph structure-based attribute, the authors have defined semantic subgraph template using a SPARQL query. The works by Butz et al. [5], and Madsen and Butz [25], are mainly on exploiting semantics in Bayesian network inference. For that purpose, Butz et al. [5] have proposed a join tree propagation architecture in which inference is conducted in a join tree (JT). Each node in JT posses a local BN that preserves all conditional independencies of the original BN. In order to use semantics in Bayesian network inference, Madsen and Butz [25] have used Lazy Propagation. It basically combines a Shenoy–Shafer propagation [26] and variable elimination scheme for computation of messages and marginals.

On the other hand, our proposed one is a new variant of semantic Bayesian network (*semBnet*) which is novel from both learning and inference generation perspectives. In order to incorporate semantics in the Bayesian analysis, the proposed *semBnet* uses a semantic hierarchy representation of the domain knowledge and some appropriate semantic similarity measures between the various concepts (refer Section 2.2). In our work, the proposed *semBnet* has been applied for better modeling of spatio-temporal inter-relationships among meteorological parameters. To the best of our knowledge this is the first attempt of using semantic Bayesian network for multivariate prediction in meteorology. However, the proposed *semBnet* is a generic model which can be applied to diverse set of applications.

1.2. Problem statement and motivations

The overall problem of meteorological time series prediction, addressed in the present work, can be stated as follows:

- Given, the historical daily time series data set over n meteorological parameters in $M = \{m_1, m_2, \ldots, m_n\}$, corresponding to a set of l locations $L = \{l_1, l_2, \ldots, l_l\}$ for previous t years: $\{y_1, y_2, \cdots, y_t\}$. Also given, the spatial attributes $SA = \left\{sa_1^l, sa_2^l, \ldots, sa_p^l\right\}$ for each location $l \in L$. 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 $\left\{y_{(t+1)}, y_{(t+2)}, \ldots, y_{(t+q)}\right\}$, when the spatial attributes of x is observed as $\left\{sa_1^x, sa_2^x, \ldots, sa_p^x\right\}$. Here, Z is a set of k new locations $\{z_1, z_2, \ldots, z_k\}$, such that $z_i \notin L$, for $i = 1$ to k, and q is a positive integer, i.e. $q \in \{1, 2, 3, \cdots\}$.

As per the definition stated above, this problem is a kind of spatio-temporal prediction that needs to consider spatial as well as temporal aspects of change in inter-relationships among the meteorological variables. Therefore, a Bayesian modeling of the problem may appear as an appropriate solution. However, challenge arises when a spatial attribute $sa \in SA$ has qualitative values with different semantic interpretations. In that case, treating such variable in a conventional manner, without utilizing the available spatial semantics, may results in improper Bayesian learning and inference. For example, consider the example scenario illustrated in Fig. 1.

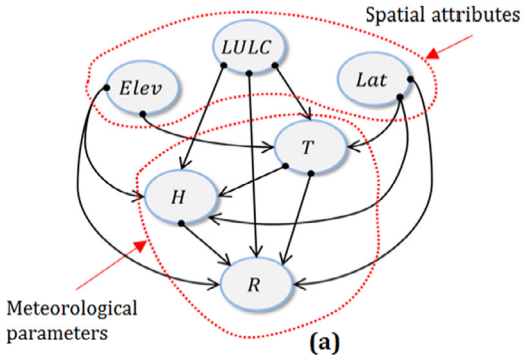
Fig. 1(a) shows a causal dependency graph among three meteorological variables (*Temperature (T), Relative Humidity (H), Rainfall (R)*) and three spatial attributes (*Latitude (Lat), Elevation (Elev), LULC*), which significantly influence these meteorological variables. This graph is basically the directed acyclic graph (DAG) that forms the structure of the Bayesian network. Possible values for each of the quantitative variables (i.e. *T, H, R, Lat,* and *Elev*) are provided in terms of some discrete ranges (refer Fig. 1(b)). On the other side, *LULC* (land-use land-cover) is qualitative variable, which may take the values from its domain: {.'Urban', 'Mining', 'Forest', 'Wetland'.}. Now, suppose, for the variable LULC, some domain knowledge is also available that basically provides insights on the semantic relationships (in this case inheritance) among these domain values of LULC. This knowledge has been represented in terms of a semantic hierarchy [30] in the Fig. 2. Here, it must be made clear that this hierarchy is only the representation of the knowledge; it is not a part of the network/ causal dependency graph in Fig. 1(a). A toy data set over eight separate locations are also provided (refer Fig. 1(c)) for the variable *Temperature* (T).

In this scenario, the standard Bayesian network analyses are performed without using the domain knowledge i.e. without using the semantic relationships expressed through the hierarchy (refer Fig. 2). Therefore, as per the principles of standard/ classical BN, the probability of $T = T_3$, given $Lat = Y_1$, $Elev = E_1$, and $LULC =$ '*Urban*' becomes $P(T_3|Y_1, E_1, \text{'}Urban\text{'}) = \frac{1}{2} = 0.5$, which considers the record $< Y_1, E_1,$ '*Urban*', $T_3 >$ out of $\{< Y_1, E_1,$ '*Urban*', $T_2 >$, $< Y_1, E_1,$ '*Urban*', $T_3 >\}$. Thus, the standard Bayesian network treats all the domain values of LULC, like 'Urban', 'Mining' etc. as separate categorical values. However, as per the hierarchy, 'Urban' is a sub-category of LULC type 'Builtup'. Therefore, 'Urban' is somehow semantically related to 'Mining' as well. (Since the toy data set does not contain any entry on LULC type 'Rural', we have not considered it.) That means, the temperature of 'Urban' and 'Mining' area are influenced more or less in similar manner than the temperature of 'Wetland', 'Forest' etc. So, while measuring $P(T_3|Y_1, E_1,$ '*Urban*'), the effect of two more records: $< Y_1, E_1,$ '*Mining*', $T_3 >$ (corresponding to $location_1$), and $< Y_1, E_1,$ '*Mining*', $T_3 >$ (corresponding to $location_5$) in the data set should also be considered. In order to overcome such limitation in a standard/ classical Bayesian network, we've proposed a variation of semantic Bayesian network, termed as *semBnet*. The *semBnet* provides a mechanism to utilize the domain knowledge, expressed in terms of semantic hierarchical relationships, and incorporate such semantics in a standard Bayesian Analysis.

1.3. Contributions

The key contributions in the present paper can be summarized as follows:

- Defining a new variant of semantically influenced Bayesian network, termed as *semBnet*, that incorporates semantic information during probabilistic learning and inference generation
- Theoretical performance analyses of *semBnet* in comparison with the standard/ classical Bayesian network (SBN)



Quantitative Variable	Range-1	Range-2	Range-3
Temperature (T)	T_1	T_2	T_3
Elevation(Elev)	E_1	E_2	E_3
Lattitude (Lat)	Y_1	Y_2	Y_3
Humidity (H)	H_1	H_2	H_3
Rainfall (R)	R_1	R_2	R_3
Qualitative Variable	Domain values		
Land-use land-cover (LULC)	'Urban', 'Mining', 'Forest', 'Wetland'		

(b)

Location	Lat	Elev	LULC	T
$Location_1$	Y_1	E_1	Mining	T_3
$Location_2$	Y_2	E_3	Forest	T_2
$Location_3$	Y_1	E_2	Wetland	T_1
$Location_4$	Y_2	E_2	Wetland	T_1
$Location_5$	Y_1	E_1	Mining	T_3
$Location_6$	Y_1	E_1	Urban	T_2
$Location_7$	Y_3	E_3	Forest	T_2
$Location_8$	Y_1	E_1	Urban	T_3

(c)

Fig. 1. An example scenario for illustrating the need of semantic knowledge in meteorological prediction: (a) directed acyclic graph (DAG) for Bayesian and semantic Bayesian analysis, (b) values for quantitative and qualitative variables in the graph, (c) a toy data set on the variable *Temperature* (T).

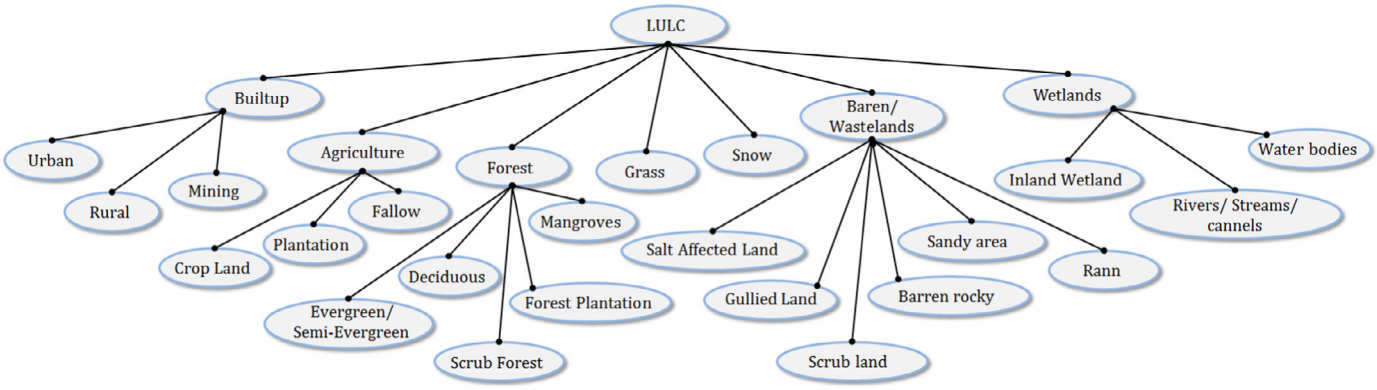


Fig. 2. Hierarchical representation of the domain knowledge on land-use land-cover (LULC).

- Proposing a general purpose forecasting approach based on *semBnet* for multivariate prediction of meteorological time series data
- Validating the proposed approach through case studies on prediction of meteorological variables in two separate spatial regions (*Jharkhand* and *West Bengal*) in India
- Evaluating the semantic Bayesian network (*semBnet*)-based proposed prediction approach in comparison with the traditional statistical model (ARIMA), spatio-temporal ordinary kriging (ST-OK), standard BN (SBN), ANN, and *hierarchical Bayesian autoregressive* model (HBAR).

The rest of the paper is organized in the following manner. Section 2 illustrates the proposed prediction approach along with the details of proposed semantic Bayesian network (*semBnet*). The theoretical performance analyses for *semBnet* has been presented in Section 3. Section 4 reports our experimentation on multivariate prediction of meteorological variables using *semBnet*-based prediction framework. A brief description of the study area and used data sets have been provided in Section 4.1. The experimental set up along with the performance evaluation criteria have been described in Section 4.2. The comparative study of the prediction results has been discussed in Section 4.3. Finally, we conclude in Section 5.

2. Proposed semantic Bayesian network (*semBnet*) based multivariate prediction approach

This section provides a detailed description of the proposed prediction approach that utilizes our newly defined semantic Bayesian network (*semBnet*) for modeling the spatio-temporal inter-relationships among different meteorological parameters.

The overall flow of the proposed prediction approach has been depicted in Fig. 3. The approach consists of two major steps, corresponding to: *1) Data Preprocessing*, and *2) Semantic Bayesian Analysis*.

2.1. Data preprocessing

The proposed prediction method starts with the pre-processing of historical data set, so as to make it suitable for semantic Bayesian analyses in the following steps. Two main objectives in the data pre-processing step are: *1) Data Discretization*, and *2) Capturing short-term variation*. During data discretization, the whole range of values of a quantitative variable is divided into a number of sub-ranges or intervals. According to Uusitalo [34] more dependencies can be achieved even with fewer data by discretizing the data into fewer intervals. On the other side, while capturing the short-term variation of the considered meteorological parameters, the entire historical data is analyzed to find out whether the change is over year, or month, or week, or on daily basis. For example, land surface temperature typically shows a daily variation, however, the precipitation mostly shows a monthly variation.

2.2. Semantic Bayesian analysis

After preprocessing, the proposed prediction system starts analyzing the data using the proposed *semantic Bayesian network* or *semBnet*. The *semBnet* extends standard/ classical Bayesian network to incorporate domain knowledge in terms of some semantic hierarchy representation [27]. Marszalek and Schmid have used semantic hierarchy of discriminative classifiers to perform object detection, whereas in our work, we have used semantic hierarchy of

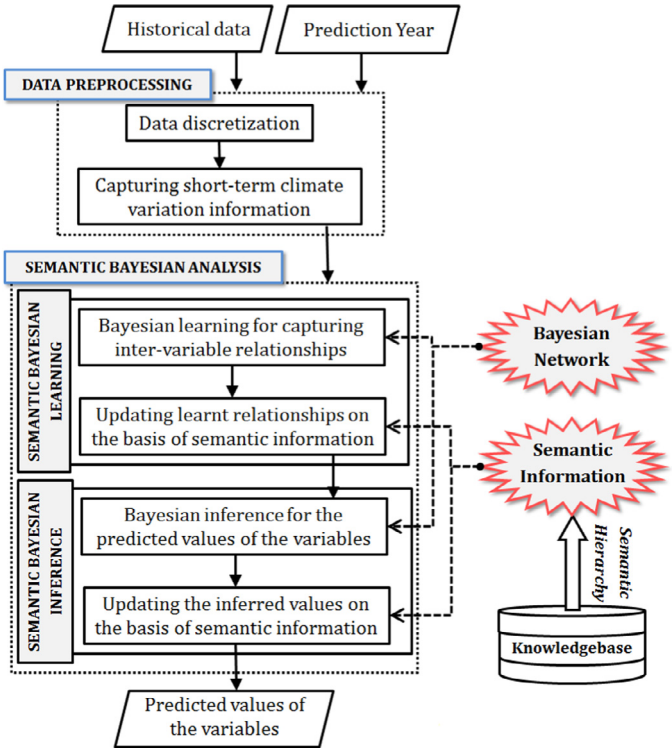


Fig. 3. Proposed prediction system using semantic Bayesian network (*semBnet*).

spatial features as a knowledge base to incorporate semantics (domain knowledge) in standard Bayesian network analysis.

The *semBnet* consists of a *qualitative* and a *quantitative* component. The qualitative component is composed of a causal dependency graph (CDG), containing a set of nodes and edges among themselves. Following is a formal definition for the qualitative component in semBnet.

Definition 2.1. The qualitative component of *semBnet* consists of a directed acyclic graph $G(V_N, V_S, E)$, where V_N is the set of nodes representing random variables with no semantic information available for themselves, V_S is the set of nodes representing random variables having semantic information available for themselves, and E is the set of edges between a pair of nodes in $(V_N \cup V_S)$. An edge $e \in E$ from $v_i \in (V_N \cup V_S)$ to $v_j \in (V_N \cup V_S)$ indicates that variable v_i influences variable v_j.

The quantitative component in *semBnet* is composed of conditional probability distributions (CPDs) associated with the nodes in the causal dependency graph. It can be formally defined as follows:

Definition 2.2. Let $G(V_N, V_S, E)$ be the causal dependency graph in semBnet, where V_N is the set of nodes representing random variables having no semantic information available, V_S is the set of nodes representing random variables having semantic information available for themselves, and E is the set of edges between a pair of nodes in $(V_N \cup V_S)$. Then the conditional probability of a node V_x is denoted as $P^{\dagger}(V_x|Parent(V_x))$ if either $V_x \in V_S$, or $(Parent(V_x) \cap V_S) \neq \emptyset$, or both are true. Otherwise, the conditional probability is denoted as $P(V_x|Parent(V_x))$

Here, $Parent(X)$ denotes the set of all the influencing nodes or parents of a node X.

2.2.1. Semantic Bayesian learning

As per the principle of proposed *semBnet*, the marginal probability $P(v_x)$ of a node $V_x \in V_N$ is calculated in a same fashion as that of a classical Bayesian network. However, that for a node $V_y \in V_S$ is estimated with consideration to the available semantic information (refer Eq. (1)).

$$P^{\dagger}(v_y) = \gamma \cdot \left[P(v_y) + \sum_{v_{yc}} SS(v_y, v_{yc}).P(v_{yc}) \right], \tag{1}$$

where, v_y and v_{yc} are any two values in the domain of variable $V_y \in V_S$, such that $v_y \neq v_{yc}$; $P(v_y)$ is the classical probability of v_y; γ is the normalization constant; $SS(v_y, v_{yc})$ is the *semantic similarity* between v_y and v_{yc}.

Assuming a hierarchical representation of the semantic knowledge base corresponding to a variable V, the *semantic similarity* can be defined as per Li et al. [22] in the following manner.

Definition 2.3. [Semantic Similarity]. Let H be the hierarchical representation corresponding to the semantic knowledge base of a variable $V \in V_S$ in semBnet. Then the semantic similarity $SS(v_{c_1}, v_{c_2})$ between any two domain values (or concepts) v_{c_1} and v_{c_2} of V in H, can be defined as follows:

$$SS(v_{c_1}, v_{c_2}) = e^{-\delta l} \cdot \frac{e^{\lambda d} - e^{-\lambda d}}{e^{\lambda d} + e^{-\lambda d}} \tag{2}$$

where, l is the length of shortest path between v_{c_1} and v_{c_2}; d is the depth of subsumer in the hierarchy H; $\delta \geq 0$ and $\lambda > 0$ are parameters, scaling the contribution of l and d, respectively.

According to Li et al. [22], the optimal value for δ and λ are 0.2 and 0.6 respectively.

Similarly, as per the principle of *semBnet*, the conditional probabilities $P(V_x|Parent(V_x))$ of the a node $V_x \in V_N$ is calculated in a same fashion as that of a classical Bayesian network, when $(Parent(V_x) \cap V_S) = \emptyset$. Otherwise, the conditional probabilities are estimated with consideration to the available semantic information. There can be three such cases:

(i) $V_x \in V_S$ and $(Parent(V_x) \cap V_S) = \emptyset$:
In this case, the conditional probability of V_x becomes:

$$P^{\dagger}(v_x|parent(V_x)) = \gamma \cdot [P(v_x|parent(V_x)) + \sum_{v_{xc}, v_x \neq v_{xc}} SS(v_x, v_{xc}).P(v_{xc}|parent(V_x))] \tag{3}$$

where, γ is a normalization constant, $parent(V_x)$ is a particular combination of values for $Parent(V_x)$; v_x and v_{xc} are any two particular values for variable V_x; and $SS(v_x, v_{xc})$ is the semantic similarity between v_x and v_{xc}

(ii) $V_x \in V_N$ and $(Parent(V_x) \cap V_S) \neq \emptyset$:
In this case, let $Parent(V_x) = Parent_N(V_x) \cup Parent_S(V_x)$, where $Parent_N(V_x) \subseteq V_N$, and $Parent_S(V_x) \subseteq V_S$. Also let $parent_S(V_x) =$ a particular combination of values for $Parent_S(V_x) = \{v_1, v_2, \ldots, v_{p_x}\}$, where $p_x = |Parent_S(V_x)|$ is the total number of parents of V_x that belong to V_S.
Then the conditional probability of V_x becomes:

$$P^{\dagger}(v_x|parent(V_x)) = \gamma \cdot [P(v_x|(parent_N(V_x) \cup parent_S(V_x))) + \sum_{k=1}^{p_x} \sum_{v_{kc}, v_k \neq v_{kc}} SS(v_k, v_{kc}) \cdot P(v_x|(parent_N(V_x) \cup (parent_S(V_x) - \{v_k\}) \cup \{v_{kc}\}))] \tag{4}$$

where, v_x is a particular value of variable V_x; γ is a normalization constant, $parent_N(V_x)$ is a particular combination of values for $Parent_N(V_x)$; v_k and v_{kc} are any two particular values for variable V_k; and $SS(v_k, v_{kc})$ is the semantic similarity between v_k and v_{kc}

(iii) $V_x \in V_S$ and $(Parent(V_x) \cap V_S) \neq \emptyset$:
In this case, let $Parent(V_x) = Parent_N(V_x) \cup Parent_S(V_x)$, where $Parent_N(V_x) \subseteq V_N$, and $Parent_S(V_x) \subseteq V_S$. Also let $parent_S(V_x) =$ a particular combination of values for $Parent_S(V_x) = \{v_1, v_2, \ldots, v_{p_x}\}$, where $p_x = |Parent_S(V_x)|$ is the total number of parents of V_x that belong to V_S.
Then the conditional probability of V_x becomes:

$$P^{\dagger}(v_x|parent(V_x)) = \gamma \cdot [P(v_x|(parent_N(V_x) \cup parent_S(V_x))) + \sum_{k=1}^{p_x} \sum_{v_{kc}, v_k \neq v_{kc}} \sum_{v_{xc}} [SS(v_k, v_{kc}) + SS(v_x, v_{xc})] \cdot P(v_{xc}|(parent_N(V_x) \cup (parent_S(V_x) - \{v_k\}) \cup \{v_{kc}\}))] \quad (5)$$

where, v_x is a particular value of variable V_x; γ is a normalization constant, $parent_N(V_x)$ is a particular combination of values for $Parent_N(V_x)$; v_k and v_{kc} are any two particular values for variable V_k; $SS(v_k, v_{kc})$ is the semantic similarity between v_k and v_{kc}; v_{xc} is a particular value for variable V_x such that $v_x \neq v_{xc}$; and $SS(v_x, v_{xc})$ is the semantic similarity between v_x and v_{xc}

Example 1. In order to explain the learning principle of *semBnet*, let's consider a similar causal dependency graph $G(V_N, V_S, E)$ as shown in Fig. 1(a). In this case, $V_N = \{Elev, Lat, T, H, R\}$, $V_S = \{LULC\}$, and $E = \{Elev \rightarrow T, Elev \rightarrow H, Elev \rightarrow R, LULC \rightarrow T, LULC \rightarrow H, LULC \rightarrow R, Lat \rightarrow T, Lat \rightarrow H, Lat \rightarrow R, T \rightarrow H, T \rightarrow R, H \rightarrow R\}$

Then, considering Fig. 2 as the hierarchical representation of the knowledge base corresponding to the variable $LULC \in V_S$, and using the data set as provided in Fig. 1(c), the marginal probability for $LULC =$ '*Urban*', as per Eq. (1), becomes:

$$\begin{aligned} P^{\dagger}('Urban') = \gamma \cdot [&P('Urban') \\ &+ SS('Urban', 'Mining') \cdot P('Mining') \\ &+ SS('Urban', 'Forest') \cdot P('Forest') \\ &+ SS('Urban', 'Wetland') \cdot P('Wetland')] \end{aligned} \quad (6)$$

Since the toy data set in Fig. 1(c) does not contain any entry on LULC type 'Rural' etc., we have not considered these. Thus, from Eq. (6), we get:

$$\begin{aligned} P^{\dagger}('Urban') = \gamma \cdot [&(2/8) \\ &+ e^{-0.2\times 2} \cdot \left(\frac{e^{0.6\times 1} - e^{-0.6\times 1}}{e^{0.6\times 1} + e^{-0.6\times 1}}\right) \cdot (2/8) \\ &+ 0 \cdot (2/8) + 0 \cdot (2/8)] = 0.34\gamma \end{aligned}$$

Here, γ, the normalization constant, is such that the sum of marginal probabilities corresponding to all possible domain values of *LULC* becomes 1. In a similar fashion, it can be determined that $P^{\dagger}('Mining') = 0.34\gamma$, $P^{\dagger}('Forest') = 0.25\gamma$, and $P^{\dagger}('Wetland') = 0.25\gamma$. Therefore, in the present scenario, the value of γ is 0.8474; and thus, $P^{\dagger}('Urban')$ becomes ≈ 0.29

On the other hand, considering the same example scenario, the classical marginal probability (using standard BN or SBN) for $LULC =$ '*Urban*' becomes $P('Urban') = (2/8) = 0.25$

Now, in order to explain conditional probability estimation, let's consider the calculation for probability of $T = T_3$, given $Lat = Y_1$, $Elev = E_1$, and $LULC =$ '*Urban*'. Since, $T \in V_N$, and $\{Lat, Elev, LULC\} \cap V_S = \{LULC\} \neq \emptyset$, therefore, using the given data set, as per Eq. (4) we get:

$$\begin{aligned} &P^{\dagger}(T_3|Y_1, E_1, 'Urban') \\ &= \gamma \cdot \big[P(T_3|Y_1, E_1, 'Urban') \\ &\quad + SS('Urban', 'Mining') \cdot P(T_3|Y_1, E_1, 'Mining') \\ &\quad + SS('Urban', 'Forest') \cdot P(T_3|Y_1, E_1, 'Forest') \\ &\quad + SS('Urban', 'Wetland') \cdot P(T_3|Y_1, E_1, 'Wetland')\big] \\ &= \gamma \cdot \left[0.5 + e^{-0.2\times 2} \cdot \left(\frac{e^{0.6\times 1} - e^{-0.6\times 1}}{e^{0.6\times 1} + e^{-0.6\times 1}}\right).1 + 0.0 + 0.0\right] \\ &= 0.86\gamma \end{aligned}$$

In this case, γ is such that the sum of same conditional probabilities corresponding to all possible available values of T becomes 1. Now, in the similar fashion, it can be determined that $P^{\dagger}(T_1|Y_1, E_1, 'Urban') = 0.0\gamma$, and $P^{\dagger}(T_2|Y_1, E_1, 'Urban') = 0.5\gamma$. Therefore, in the present calculation, the value of γ is 0.7353; and hence $P^{\dagger}(T_3|Y_1, E_1, 'Urban')$ becomes ≈ 0.6324

On the other side, considering the same example scenario, the classical conditional probability for $T = T_3$, given $Lat = Y_1$, $Elev = E_1$, and $LULC =$ '*Urban*', would become $P(T_3|Y_1, E_1, 'Urban') = (1/2) = 0.5$

In order to capture the large scale temporal change in variable characteristics, the learning is performed for each year y_i, and the final probability (whether marginal or conditional) is generated by honoring the temporal auto-correlation property [12], so that the final probability $P_f^{\dagger}$ becomes:

$$P_f^{\dagger} = \sum_{i=1}^{t} \left(P_{y_i}^{\dagger} \times \frac{1/d_i}{\sum_{j=1}^{t} 1/d_j}\right) \quad (7)$$

where, d_i is the temporal distance of the time instant y_i from the prediction time y_p; t is the total number of time instant considered for training; and $P_{y_i}^{\dagger}$ is the estimated marginal/ conditional probability of any variable, for the year y_i.

2.2.2. Semantic Bayesian inference and prediction

In this step, the value for any variable, given the evidence regarding some other variables, can be determined by consulting the conditional and marginal probabilities as estimated during the semantic Bayesian learning. The inferred value becomes the one, associated with the highest probability.

Now, since in data pre-processing step the values for each of the quantitative variables are discretized into certain number of ranges, the inferred value of a quantitative prediction variable is also obtained in the form of a range. In order to obtain a single continuous value, the mean of the range can be assigned to the predicted variable. Following is an example with respect to the scenario depicted in Fig. 1.

Example 2. To demonstrate the inference generation and prediction processes in the proposed *semBnet* based approach, let's consider a case, where the observed/ evidence variables are: *LULC*, *Elev*, and *Lat*, from which the value of *Rainfall* (R) is to be inferred.

In the given scenario, ith range value of the variable R is R_i, where $i = 1, 2, 3$. Now if $IR^{semBnet} \in \{R_1, R_2, R_3\}$ be the inferred range of R as obtained using *semBnet*, then

$$P(IR^{semBnet}) = \underbrace{max}_{\forall i} \left\{P^{\dagger}(R_i|LULC, Elev, Lat)\right\}$$

where, $P^{\dagger}(R_i|LULC, Elev, Lat) = \gamma \cdot \sum_T \sum_H \{P^{\dagger}(LULC).P(Elev).P(Lat).P^{\dagger}(T|Lat, Elev, LULC).P^{\dagger}(H|T, Lat, Elev, LULC).P^{\dagger}(R_i|T, H, Lat, Elev, LULC)\}$, which can be determined by the estimated probabilities from *semBnet* learning phase.

Similarly, if $IR^{standardBN} \in \{R_1, R_2, R_3\}$ be the inferred range of R as obtained using standard Bayesian network, then

$$P(IR^{standardBN}) = \underbrace{max}_{\forall i} \{P(R_i|LULC, Elev, Lat)\}$$

Therefore, the predicted value of *Rainfall* (R) becomes:

$$R_{predict} = \left[\frac{cval(IR^{semBnet}) + cval(IR^{standardBN})}{2}\right]$$

where, $cval(x)$ is the *central* or *mean* value of a range x.

3. Theoretical analysis of *semBnet*

In this section, we analyze the time and space complexities of learning the proposed *semBnet*. We also show that the *semBnet* is less susceptible to parameter value uncertainty than the standard/ classical Bayesian network.

Let $G(V_N, V_S, E)$ be a causal dependency graph of *semBnet* containing n number of *nodes*, where V_N is the set of nodes without semantic information and V_S is the set of nodes with semantic information. Moreover, let the *maximum number of parents* of any node in G is C and the *maximum domain size* of any variable $\in (V_N \cup V_S)$ is D.

Lemma 3.1. *The proposed semantic Bayesian network (semBnet) has a complexity of $O(n.n_s.D^{C+3})$ in terms of computational time requirement, where n_s is the maximum number of parents (for any variable) having semantic knowledge base.*

Proof. As per the network learning for a standard Bayesian network, the total number of iterations required for learning/updating a node, having i number of parents, is at most $(D-1).D^i$. Now, if we consider that the number of nodes having i $(0 \le i \le C)$ number of parents is n_i such that $\sum_{i=0}^{C} n_i = n$, then classical computational cost for learning parameters of all the n nodes is:

$$\begin{aligned} TC_{standardBN}(G) &\le \sum_{i=0}^{C} (D-1).n_i.D^i \\ &= O(n.D^{C+1}) \end{aligned} \tag{8}$$

Now, once the classical probabilities (marginal and conditional) are available, the computation of conditional probabilities involving semantic information needs maximum $n_{s_i}.D^2$ time (refer Eq. (5)), where n_{s_i} is the maximum number of parents from within the i parents of a variable, such that these parents have semantic information available with them. Therefore, the computational cost for learning parameters of all the nodes in *semBnet* is:

$$\begin{aligned} TC_{semBnet}(G) &\le \sum_{i=0}^{C} \left[\left((D-1).n_i.D^i\right) \times \left(n_{s_i}.D^2\right)\right] \\ &= D^2.(D-1)\sum_{i=0}^{C} n_i.D^i.n_{s_i} \\ &= O(n.n_s.D^{C+3}) \end{aligned} \tag{9}$$

In worst case, $n_s = (n-1)$. Therefore, the worst case time complexity of *semBnet* becomes $O(n^2.D^{C+3})$. On the other side, the best case occurs when the semantic information is available for neither of the variables, i.e. n_s becomes 0, which suppresses the term $\left(n_{s_i}.D^2\right)$ to 0. Therefore, the best case time complexity of *semBnet* becomes $O(n.D^{C+1})$, which is similar to that of the standard Bayesian network (standard BN or SBN). □

Lemma 3.2. *The proposed semantic Bayesian network is not more complex than standard/ classical Bayesian network in terms of computational space requirement.*

Proof. For any Bayesian network, the minimum amount of space requirement for any node $x = (|D(x)| - 1).\prod_i |D(Parent_i)|$, where $|D(Parent_i)|$ denotes the domain size of the ith parent $Parent_i$ of x, and $|D(x)|$ is the domain size for the variable x. Therefore, space requirement for classical learning/updating a of node, having i number of parents, is $\le (D-1).D^i$.

Now, if we consider that the number of nodes having i $(0 \le i \le C)$ number of parents is n_i such that $\sum_{i=0}^{C} n_i = n$, then the space requirement for classically learning parameters of all the nodes becomes:

$$SC_{standardBN}(G) \le \sum_{i=0}^{C} (D-1).n_i.D^i = O(n.D^{C+1}) \tag{10}$$

Now, once the classical probabilities (marginal and conditional) are available, the computation of conditional probabilities involving semantic information needs constant space to determine *semantic similarity* (refer Eqs. (3)–(5)) between any two pair of domain value/ concept of a variable $V \in V_S$. Therefore, the computational space requirement for learning all parameters in *semBnet* is:

$$SC_{semBnet}(G) \le 2 \cdot \sum_{i=0}^{C} (D-1).n_i.D^i + c_0 = O(n.D^{C+1}) \tag{11}$$

where, c_0 is a constant.

Therefore, the space complexity of *semBnet* becomes $SC_{semBnet}(G) = O(n.D^{C+1})$, which is similar to that of standard/classical Bayesian network learning involving no semantic information. This proves the lemma. □

It is evident from *Lemma 3.1* and *Lemma 3.2* that even with the embedded process of knowledge incorporation, the overall complexity of semBnet analysis remains equivalent to that of the classical/standard Bayesian network (SBN) analysis. Further, it can also be noted that due to the knowledge incorporation capability, the *parameter value uncertainty* in semBnet analysis is more likely to be *lesser* than that in SBN analysis.

The *Parameter value uncertainty*, which is observed during the process of learning/developing specific value(s) or parameter(s), appears due to lack of knowledge [23]. One of the major sources of parameter uncertainty is the linguistic imprecision and vague concepts [16]. In general, the fuzzy set theoretic approach is used to describe such imprecise concepts. However, sometimes it becomes difficult to determine the fuzzy membership functions associated with the vague concepts. For example, though the concept like 'mining area', 'urban area', 'rural area' etc. are somehow relevant to the concept of 'built-up area', we cannot properly define a fuzzy membership function in this respect. Therefore, the recent research has focused on utilizing semantic similarity relations among the linguistic labels/ concepts [20,21]. An analogous idea has been adopted in the present work. However, instead of using fuzzy relation, as used by Lawry [20] and Lawry and Tang [21], in this work we have used *semantic hierarchy* to measure the semantic similarity between any pair of concepts.

As discussed in *Example-1*, because of lack of domain knowledge, SBN treats 'mining', 'urban' etc. as non-related concepts. Therefore, the parameter learning in SBN does not consider a number of training records which could have significant relevance in the process involved. This also leads to draw inferences based on limited samples/ records, and thereby further increases the parameter uncertainty. On the other hand, during parameter learning, the proposed *semBnet* considers every relevant record along with the associated similarity value, measured based on the supplied *domain knowledge* (represented in terms of *semantic hierarchy*). Hence, with the help of available semantic information, the proposed *semBnet* is able to tune the parameter values by considering the relevance of each record. This also helps *semBnet* to

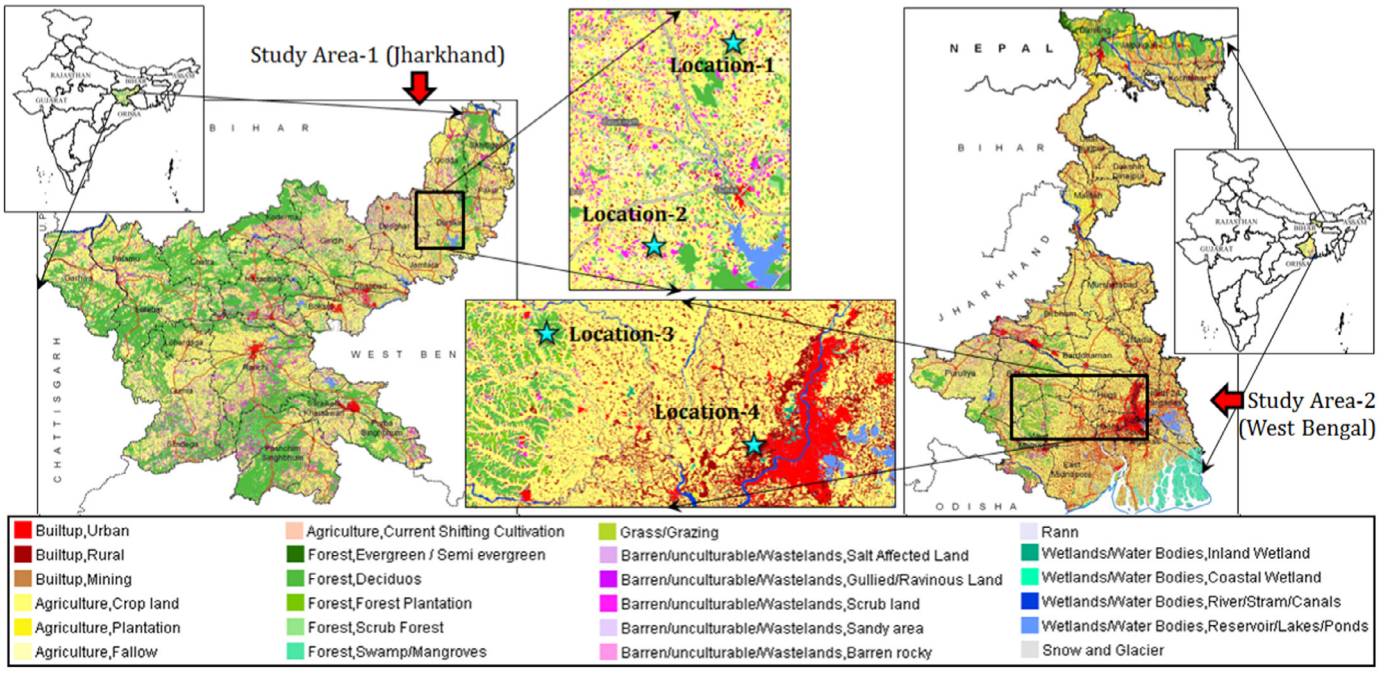


Fig. 4. *Study Area-1* in Jharkhand (India) and *Study Area-2* in West Bengal (India).

draw inferences based on comparatively more samples/ records than SBN. Thus, the parameter value uncertainty due to lack of knowledge is reduced in our proposed *semBnet*.

4. Experimental evaluation

This section provides the details of the experimental evaluation for our proposed *semBnet* scheme, in comparison with existing linear and non-linear forecasting techniques. As the linear forecasting approaches, we have selected *statistical ARIMA models*, and *spatio-temporal ordinary kriging* (ST-OK) [10]; whereas the *artificial neural network* (ANN), *standard Bayesian network* (SBN) and *hierarchical Bayesian autoregressive model* (HBAR) [31] have been chosen as the non-linear prediction models. The overall results of comparative study are found to be encouraging.

4.1. Data set and study area

The experimentation has been carried out in two spatial regions, one in the state of Jharkhand (India), comprising of around 3625 km^2 area, and the other in the state of West Bengal (India), comprising of 12,390 km^2 area (refer Fig. 4). Both the study areas are full of diverse categories of land-use land-cover (LULC) which include: *agricultural crop-land, agricultural fallow-land, forest-scrub, forest plantation, urban area, rural area, mining area, waste-land, waterbodies* and so on. However, the *Study Area-2* (in West Bengal) has more homogeneous distribution of LULC categories than the *Study Area-1* (in Jharkhand). The hierarchical representation of these LULC categories, as obtained from the Bhuvan portal [4,30], is depicted in Fig. 2 in terms of a semantic graph. The raw data on LULC have been collected from the *National Bureau of Soil Survey and Land Use Planning, Govt of India*. In this experimentation, four variables, namely *Temperature, Humidity, Rainfall*, and *Soil moisture*, have been chosen as the meteorological parameter of study. The corresponding historical daily time series data have been collected from the *FetchClimate Explorer* [29] for a span of 14 years (2001–2014). The prediction has been made for two locations (refer Fig. 4) in each of the study areas, for the year 2015. The details of the prediction locations are given in Table 1.

Table 1
LULC details of the prediction locations.

Location-ID	Study area	State	LULC category
Location-1	Study Area-1	Jharkhand	Agricultural crop-land
Location-2	Study Area-1	Jharkhand	Agricultural fallow-land
Location-3	Study Area-2	West Bengal	Forest plantation
Location-4	Study Area-2	West Bengal	Rural area

4.2. Experimental set-up and performance metrics

The overall experimental study has been carried out using *MATLAB* 7.12.0 (R2011a) and *R-tool* version 3.2.3 (64 bit) in Windows 7 (64-bit Operating System, 3.10 GHz CPU, 4.00GB RAM). MATLAB has been utilized to implement the proposed semantic Bayesian network (*semBnet*) based prediction approach. The same system configuration and MATLAB version have been used to experiment with *standard* BN and ANN (*feed forward back propagation network*, NNTool). On the other hand, the R-tool has been used to study with *hierarchical Bayesian autoregressive model* (HBAR) [31], and the linear statistical models: *Holt-Winters Approach*/(HW method) [18], *Automated ARIMA* (A-ARIMA) and *spatio-temporal ordinary kriging* (ST-OK) [10].

The performance of prediction using proposed *semBnet* has been evaluated in terms of four popular statistical measures, namely *root mean square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE)* [28], and the ratio of the variance of estimated values to the variance of the observed values (RVAR) [33]. The formal definition for each of these metrics are given below:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(M_o - M_p)^2} \tag{12}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|M_o - M_p| \tag{13}$$

$$MAPE = \frac{|M^s_{mo} - M^s_{mp}|}{|M^s_{mo}|} \times 100 \tag{14}$$

Table 2
Comparative study of proposed *semBnet* based multivariate prediction approach.

Prediction variable: *Temperature* (T)					
Loc.	Prediction techniques	Prediction year 2015			
		RMSE	MAE	MAPE	RVAR
Location-1	HW	08.724	07.900	42.810	≈ 0
	A-ARIMA	04.577	03.626	09.845	00.199
	ST-OK	02.955	02.167	02.027	01.396
	ANN	01.599	00.851	02.975	00.587
	HBAR	02.249	01.887	02.875	01.706
	SBN	01.343	01.114	00.568	01.094
	Proposed ***semBnet***	**01.289**	**01.085**	**00.533**	**01.081**
Location-2	HW	08.740	07.911	42.873	≈ 0
	A-ARIMA	04.584	03.611	09.666	00.203
	ST-OK	02.979	02.192	02.068	01.372
	ANN	01.609	00.757	02.820	00.573
	HBAR	02.247	01.899	02.527	01.693
	SBN	01.345	01.114	00.611	01.087
	Proposed ***semBnet***	**01.290**	**01.086**	**00.576**	**01.074**
Location-3	HW	08.151	07.313	27.570	≈ 0
	A-ARIMA	04.352	03.219	07.463	00.187
	ST-OK	02.525	02.175	00.178	00.087
	ANN	02.809	01.771	03.698	00.677
	HBAR	02.692	02.245	02.336	00.234
	SBN	01.310	01.035	00.035	00.916
	Proposed ***semBnet***	**01.164**	**00.978**	**00.034**	**00.946**
Location-4	HW	07.595	06.830	25.623	≈ 0
	A-ARIMA	04.049	03.064	07.278	00.211
	ST-OK	03.067	02.299	02.002	01.684
	ANN	01.550	00.761	02.529	00.568
	HBAR	02.731	02.370	03.899	00.303
	SBN	01.279	01.028	00.084	00.949
	Proposed ***semBnet***	**01.155**	**00.967**	**00.025**	**00.979**

Table 3
Comparative study of proposed *semBnet* based multivariate prediction approach.

Prediction variable: *Relative Humidity* (H)					
Loc.	Prediction techniques	Prediction year 2015			
		RMSE	MAE	MAPE	RVAR
Location-1	HW	12.478	10.346	05.257	≈ 0
	A-ARIMA	11.643	09.895	02.101	00.004
	ST-OK	07.630	06.453	02.377	00.980
	ANN	06.652	04.584	04.137	00.394
	HBAR	04.902	04.180	00.726	01.072
	SBN	03.246	02.711	03.334	01.075
	Proposed ***semBnet***	**03.118**	**02.638**	**02.749**	**01.063**
Location-2	HW	12.522	10.385	05.009	≈ 0
	A-ARIMA	11.688	09.939	02.065	00.004
	ST-OK	07.556	06.412	02.875	01.706
	ANN	19.842	13.973	18.467	00.359
	HBAR	04.885	04.195	00.157	01.055
	SBN	03.405	02.853	03.743	01.067
	Proposed ***semBnet***	**03.253**	**02.778**	**03.119**	**01.061**
Location-3	HW	10.764	08.612	05.251	≈ 0
	A-ARIMA	09.769	08.333	02.187	00.007
	ST-OK	05.813	04.970	00.788	00.933
	ANN	04.792	03.038	04.068	00.567
	HBAR	04.728	03.696	02.781	00.652
	SBN	02.325	01.976	02.029	01.199
	Proposed ***semBnet***	**02.196**	**01.862**	**01.634**	**01.103**
Location-4	HW	10.164	08.923	00.601	≈ 0
	A-ARIMA	09.825	08.873	02.093	00.005
	ST-OK	07.589	06.254	04.463	01.272
	ANN	07.636	06.076	06.524	00.606
	HBAR	03.956	03.352	01.357	00.989
	SBN	03.063	02.374	03.239	01.045
	Proposed ***semBnet***	**02.896**	**02.245**	**02.639**	**01.039**

$$RVAR = \frac{var(M_p^s)}{var(M_o^s)} \tag{15}$$

where M_o and M_p denote the observed value and the corresponding predicted value of a meteorological variable. M_{mo}^s is the mean value of the observed series; M_{mp}^s is the mean value of the predicted series; $var(M_o^s)$ and $var(M_p^s)$ are the variation in the observed time series and variation in corresponding predicted time series respectively; and n is the total number of observations in the series.

The best-fit between observed and predicted values under ideal conditions yields $RMSE = 0$, $MAE = 0$, $MAPE = 0$, and $RVAR = 1$.

4.3. Results

Tables 2–5 present the comparative results of predicting the meteorological time series of *Temperature, Relative humidity, Rainfall* and *Soil moisture* respectively, for the target year 2015. Both for standard BN (SBN) and proposed *semBnet*, same causal dependency graph has been used for capturing the spatio-temporal inter relationships among the considered meteorological variables. In the dependency graph, *LULC* has been the only variable having semantic knowledge base. In order to incorporate these semantic information, the proposed *semBnet* has further utilized the semantic hierarchy of *LULC*, as depicted in Fig. 2. Moreover, to model the yearly change in the inter variable dependencies, both SBN and *semBnet* have used a similar learning *framework* as proposed by Das and Ghosh [12], taking care of the temporal auto-correlation property.

4.3.1. Discussion

From the experimental results (refer Tables 2–5), the following inferences can be drawn:

Table 4
Comparative study of proposed *semBnet* based multivariate prediction approach.

Prediction variable: *Rainfall* (R)					
Loc.	Prediction techniques	Prediction year 2015			
		RMSE	MAE	MAPE	RVAR
Location-1	HW	156.78	104.78	2556.7	≈ 0
	A-ARIMA	116.02	084.32	058.21	000.05
	ST-OK	079.13	055.37	006.96	001.12
	ANN	091.37	058.72	034.84	001.15
	HBAR	038.92	025.59	002.47	001.30
	SBN	040.10	025.27	013.17	001.43
	Proposed ***semBnet***	**038.08**	**023.24**	**010.58**	**001.38**
Location-2	HW	157.67	105.28	096.25	≈ 0
	A-ARIMA	116.94	084.81	037.07	000.05
	ST-OK	079.89	055.58	007.36	001.10
	ANN	127.70	078.39	021.58	001.64
	HBAR	038.65	025.09	003.29	001.28
	SBN	039.84	025.02	012.78	001.41
	Proposed ***semBnet***	**038.09**	**023.43**	**010.00**	**001.38**
Location-3	HW	157.25	110.09	096.44	≈ 0
	A-ARIMA	113.04	081.27	037.24	000.06
	ST-OK	073.17	051.86	005.18	001.18
	ANN	057.04	028.28	020.28	001.28
	HBAR	115.40	092.91	081.25	001.25
	SBN	019.79	012.95	009.63	001.16
	Proposed ***semBnet***	**015.34**	**010.29**	**003.27**	**001.06**
Location-4	HW	164.07	115.88	093.44	≈ 0
	A-ARIMA	118.00	086.45	037.72	000.06
	ST-OK	073.01	051.87	005.35	001.17
	ANN	044.17	024.46	015.91	000.90
	HBAR	107.78	088.52	071.20	001.05
	SBN	028.26	021.46	009.42	001.04
	Proposed ***semBnet***	**025.14**	**019.27**	**007.66**	**001.02**

- As shown in the Tables 2–5 (first column), RMSE for the SBN is significantly less than that of the HW, A-ARIMA, ST-OK, ANN and HBAR model. Further, the RMSE corresponding to proposed *semBnet* is even lesser than that of SBN. This indicates that the forecasts made by *semBnet* are the closest to the observed time

Table 5
Comparative study of proposed *semBnet* based multivariate prediction approach.

Prediction variable: *Soil moisture* (S)					
Loc.	Prediction techniques	Prediction year 2015			
		RMSE	MAE	MAPE	RVAR
Location-1	HW	54.128	46.797	11.616	≈ 0
	A-ARIMA	54.129	46.797	11.616	≈ 0
	ST-OK	32.326	28.007	01.620	00.793
	ANN	54.378	46.431	13.275	00.223
	HBAR	22.523	19.048	05.581	01.190
	SBN	22.725	19.861	01.947	01.412
	Proposed ***semBnet***	**21.978**	**18.972**	**01.635**	**01.390**
Location-2	HW	53.473	46.392	10.601	≈ 0
	A-ARIMA	53.473	46.392	10.601	≈ 0
	ST-OK	31.124	26.703	01.825	00.789
	ANN	40.986	36.177	08.264	00.439
	HBAR	23.660	20.200	06.164	01.157
	SBN	22.999	20.034	04.667	01.479
	Proposed ***semBnet***	**20.837**	**17.736**	**03.270**	**01.443**
Location-3	HW	46.906	39.031	03.417	≈ 0
	A-ARIMA	46.906	39.031	03.416	≈ 0
	ST-OK	46.467	40.844	13.049	00.402
	ANN	39.654	30.940	00.735	00.034
	HBAR	38.347	32.146	10.994	01.228
	SBN	21.883	19.779	00.290	01.239
	Proposed ***semBnet***	**18.452**	**15.895**	**00.191**	**01.175**
Location-4	HW	66.396	56.116	04.215	≈ 0
	A-ARIMA	62.399	51.311	01.587	≈ 0
	ST-OK	46.959	41.237	13.266	00.394
	ANN	51.808	40.390	03.705	00.066
	HBAR	30.972	26.014	08.825	01.075
	SBN	23.603	19.363	07.232	00.960
	Proposed ***semBnet***	**20.938**	**16.769**	**05.976**	**00.988**

series, and proves the worth of using spatial semantics in the prediction process.

- Similarly, it is also evident from the Tables 2–5 (second column) that the proposed *semBnet* provides the best prediction, producing the minimum MAE in all the cases.
- The MAPE measures, as presented in third column of the Tables 2–5, also show a significantly lesser value of 0.025% – 10.58% for the proposed *semBnet*, especially for *Study Area-2*. These also demonstrate the effectiveness of considering semantic information in meteorological time series prediction.
- The measure of RVAR, that basically quantifies the ratio of variation in the observed series and the predicted series, becomes unity in an ideal situation. The last column in the Tables 2–5 show that the values of RVAR for proposed *semBnet* based prediction are always towards 1, however those for other techniques (especially ARIMA models and ANN) are far from 1. This indicates that in case of *semBnet*, the predicted time series are more likely to maintain the similar variations as in the original series.

The overall percentage improvement (error reduction percentage) in proposed *semBnet* based approach, compared to standard Bayesian network (SBN) based prediction, has been summarized in Fig. 5. It shows that, for *Study Area-1, semBnet* provides overall ≈ 5% improvement in RMSE and MAE, and ≈ 17% improvement in MAPE. Further, for the *Study Area-2*, the percentage improvement of *semBnet* in RMSE, MAE, and MAPE are 11%, 10% and ≈ 31% respectively, which are significantly better than those in case of *Study Area-1*. The reason is that the *Study Area-1* is comparatively smaller, and most (≈ 73.6%) of its parts belong to agricultural land, thereby contributing uneven semantic knowledge during the *semBnet* learning. On the other hand, the training locations chosen from the Study Area-2 have comparatively more homogeneous distribution of LULC, and thus, aid in better knowledge incorporation.

In summary, the proposed *semBnet* based prediction shows improved performance from all aspects, and proves to become superior to the linear statistical and non-linear prediction techniques. Considering the training set from a larger spatial region with sufficient evidence for each LULC category may lead to even better prediction results.

It must also be mentioned here that the proposed *semBnet* is a generic model which can be used for diverse set of applications. Any problem/application, that can be modeled using standard Bayesian network, can also be modeled in terms of *semBnet*, if and only if the application involves at least one variable having different interpretations, and the relevant domain knowledge is available in the form of some semantic hierarchy. In *semBnet*, the semantic hierarchy plays a significant role. Since the *semBnet* assumes that the similar concepts (semantic interpretations) of a variable have similar influence to or from the other variables in a causal dependency graph, the semantic hierarchy should be formed accordingly. Therefore, different application may need different hierarchy, based on appropriate domain knowledge. For example, in our experimental study we have used the domain knowledge that similar land-use/land-cover (LULC) category has similar effect on the meteorological variables, like *land surface temperature, precipitation, soil moisture* etc. Moreover, the semantic similarity measure also may need to be varied from one application to another [3]. Same semantic similarity measure may not be applicable for incorporating every kind of domain knowledge. However, whatever be the similarity measure, the proposed *semBnet* is a generic approach that shows how to incorporate such domain knowledge in a standard Bayesian analysis.

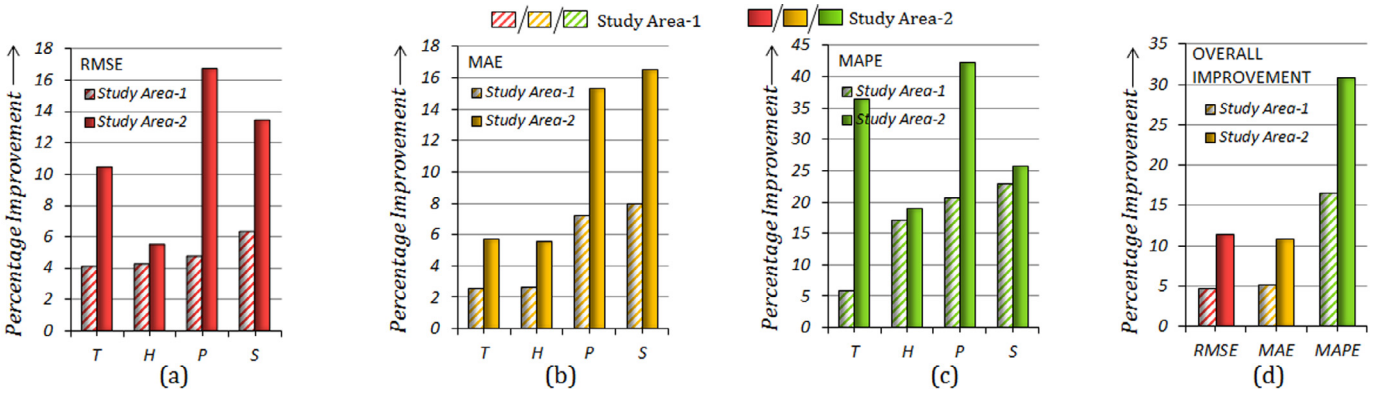


Fig. 5. Percentage improvement (% error reduction) of *semBnet* in comparison with standard Bayesian network (SBN).

5. Conclusions

In this paper, we have proposed a semantic Bayesian network based approach for multivariate prediction of meteorological time series data. The key contribution in this work is to define a variant of Bayesian network, termed as *semBnet*, that can incorporate the *spatial semantics* in modeling the probabilistic relationships among meteorological variables. The semBnet has been analyzed from both theoretical and empirical perspectives in comparison with the linear statistical models (ARIMA, ST-OK), *hierarchical Bayesian autoregressive* model (HBAR), and non-linear machine learning techniques (ANN, SBN). Case studies have been performed to predict *Temperature, Relative humidity, Rainfall,* and *Soil moisture* for two spatial regions in India. The prediction performance in terms of four different statistical measures (RMSE, MAE, MAPE, and RVAR) proves and validates the superiority of proposed *semBnet* based prediction approach. The improved accuracy also establishes the significance of incorporating *domain knowledge* in meteorological time series prediction.

In future, the work can be extended to incorporate the climate change pattern information [11] in the proposed framework to further improve the meteorological prediction accuracy.

References

[1] P. Aguilera, A. Fernández, R. Fernández, R. Rumí, A. Salmerón, Bayesian networks in environmental modelling, Environ Modell Software 26 (12) (2011) 1376–1388.
[2] J.E. Bagley, A.R. Desai, K.J. Harding, P.K. Snyder, J.A. Foley, Drought and deforestation: has land cover change influenced recent precipitation extremes in the amazon? J Clim 27 (1) (2014) 345–361.
[3] S. Bhattacharjee, S.K. Ghosh, Measurement of semantic similarity: a concept hierarchy based approach, in: Proceedings of 3rd international conference on advanced computing, networking and informatics, Springer, 2016, pp. 407–416.
[4] Bhuvan, *Indian Geo-Platform of ISRO*, 2015. (http://bhuvan.nrsc.gov.in/bhuvan_links.php#). [Online; Accessed 08-Nov-2015]
[5] C.J. Butz, H. Yao, S. Hua, A join tree probability propagation architecture for semantic modeling, J Intell Inf Syst 33 (2) (2009) 145–178.
[6] R. Cano, C. Sordo, J.M. Gutiérrez, Applications of Bayesian networks in meteorology, in: Advances in Bayesian networks, Springer, 2004, pp. 309–328.
[7] L. Cao, Support vector machines experts for time series forecasting, Neurocomputing 51 (2003) 321–339.
[8] R. Chang, W. Brauer, M. Stetter, Modeling semantics of inconsistent qualitative knowledge for quantitative bayesian network inference, Neural Netw 21 (2) (2008) 182–192.
[9] C. Chatfield, The analysis of time series: an introduction, CRC press, 2013.
[10] N. Cressie, C.K. Wikle, Statistics for spatio-temporal data, John Wiley & Sons, 2015.
[11] M. Das, S. Ghosh, Spatio-temporal pattern analysis for regional climate change using mathematical morphology, ISPRS Ann Photogramm Remote Sens Spatial Inf Sci 2 (4) (2015) 185–192.
[12] M. Das, S.K. Ghosh, A probabilistic approach for weather forecast using spatio-temporal inter-relationships among climate variables, in: Industrial and information systems (ICIIS), 2014 9th international conference on, IEEE, 2014, pp. 1–6.
[13] M. Das, S.K. Ghosh, Short-term prediction of land surface temperature using multifractal detrended fluctuation analysis, in: India conference (INDICON), 2014 annual IEEE, IEEE, 2014, pp. 1–6.
[14] M. Das, S.K. Ghosh, V. Chowdary, A. Saikrishnaveni, R. Sharma, A probabilistic nonlinear model for forecasting daily water level in reservoir, Water Resour Manage 30 (9) (2016) 3107–3122.
[15] B. Ding, K. Yang, J. Qin, L. Wang, Y. Chen, X. He, The dependence of precipitation types on surface elevation and meteorological conditions and its parameterization, J Hydrol 513 (2014) 154–163.
[16] Y.Y. Haimes, Risk modeling, assessment, and management, John Wiley & Sons, 2015.
[17] T. Hengl, G.B. Heuvelink, M.P. Tadić, E.J. Pebesma, Spatio-temporal prediction of daily temperatures using time-series of modis lst images, Theor Appl Climatol 107 (1–2) (2012) 265–277.
[18] C.C. Holt, Forecasting seasonals and trends by exponentially weighted moving averages, Int J Forecast 20 (1) (2004) 5–10.
[19] K.-M. Kim, J.-H. Hong, S.-B. Cho, A semantic Bayesian network approach to retrieving information with intelligent conversational agents, Inf Process Manag 43 (1) (2007) 225–236.
[20] J. Lawry, A framework for linguistic modelling, Artif Intell 155 (1) (2004) 1–39.
[21] J. Lawry, Y. Tang, Uncertainty modelling for vague concepts: a prototype theory approach, Artif Intell 173 (18) (2009) 1539–1558.
[22] Y. Li, Z.A. Bandar, D. McLean, An approach for measuring semantic similarity between words using multiple information sources, IEEE Trans Knowl Data Eng 15 (4) (2003) 871–882.
[23] D.P. Loucks, E. Van Beek, J.R. Stedinger, J.P. Dijkman, M.T. Villars, Water resources systems planning and management: an introduction to methods, models and applications, Paris: Unesco, 2005.
[24] J. Luo, A.E. Savakis, A. Singhal, A bayesian network-based framework for semantic image understanding, Pattern Recognit 38 (6) (2005) 919–934.
[25] A.L. Madsen, C.J. Butz, Exploiting semantics in bayesian network inference using lazy propagation, in: Canadian conference on artificial intelligence, Springer, 2015, pp. 3–15.
[26] A.L. Madsen, F.V. Jensen, Lazy propagation: a junction tree inference algorithm based on lazy evaluation, Artif Intell 113 (1) (1999) 203–245.
[27] M. Marszałek, C. Schmid, Semantic hierarchies for visual object recognition, in: Computer vision and pattern recognition, 2007. CVPR'07. IEEE conference on, IEEE, 2007, pp. 1–7.
[28] A. Mellit, A.M. Pavan, M. Benghanem, Least squares support vector machine for short-term prediction of meteorological time series, Theor Appl Climatol 111 (1–2) (2013) 297–307.
[29] Microsoft-Research, FetchClimate, 2015. (http://research.microsoft.com/en-us/um/cambridge/projects/fetchclimate/app/). [Online; Jan-2016].
[30] NRSC/ISRO, Natural resource census—land use land cover database, 2016. (http://bhuvan.nrsc.gov.in/gis/thematic/tools/document/2LULC/lulc1112.pdf). [Online; Accessed 2-Sep-2016].
[31] S.K. Sahu, K.S. Bakar, Hierarchical bayesian autoregressive models for large space–time data with applications to ozone concentration modelling, Appl Stoch Models Bus Ind 28 (5) (2012) 395–415.
[32] G. Simon, A. Lendasse, M. Cottrell, J.-C. Fort, M. Verleysen, Time series forecasting: obtaining long term trends with self-organizing maps, Pattern Recognit Lett 26 (12) (2005) 1795–1808.
[33] R.S. Teegavarapu, Floods in a changing climate: extreme precipitation, Cambridge University Press, 2012.
[34] L. Uusitalo, Advantages and challenges of bayesian networks in environmental modelling, Ecol Modell 203 (3) (2007) 312–318.
[35] F. Van Harmelen, V. Lifschitz, B. Porter, Handbook of knowledge representation, 1, Elsevier, 2008.
[36] S. Venkadesh, G. Hoogenboom, W. Potter, R. McClendon, A genetic algorithm to refine input data selection for air temperature prediction using artificial neural networks, Appl Soft Comput 13 (5) (2013) 2253–2260.
[37] H. Xiao, Q. Weng, The impact of land use and land cover changes on land surface temperature in a karst area of china, J Environ Manage 85 (1) (2007) 245–257.
[38] C. Zhou, H. Chen, Z. Peng, Y. Ni, G. Xie, A semantic bayesian network for web mashup network construction, in: Green computing and communications (GreenCom), 2010 IEEE/ACM Int'l conference on & Int'l conference on cyber, physical and social computing (CPSCom), IEEE, 2010, pp. 645–652.