International Journal of the Physical Sciences Vol. 6(18), pp. 4412-4418, 9 September, 2011
Available online at http://www.academicjournals.org/IJPS
DOI: 10.5897/IJPS11.631
ISSN 1992 - 1950 ©2011 Academic Journals

Full Length Research Paper

Bayesian networks for defining relationships among climate factors

Guillermo De la torre-Gea, Genaro M. Soto-Zarazúa, Ramón G. Guevara-González and Enrique Rico-García*

Department of Biosystems, School of Engineering, Queretaro State University, C.U. Cerro de las Campanas, Querétaro, México.

Accepted 14 July, 2011

Factors that shape the weather are studied to design models to make predictions. Understanding the relationships among these factors contribute to a better knowledge of atmospheric phenomena. However, conventional statistical techniques do not take into account the dependent relationships among these factors. Bayesian learning models are used in learning processes, which consist in quantification of conditional probability, resulting in the identification of causal relationships between the variables. In this paper the use of Bayesian networks for probabilistic analysis allows us to determine spatial and temporal dependencies among climatic variables not observable with other methods. Considering an incomplete meteorological data set from three years and three sites with distinct climates, the dependent relationships between climatic variables are observable in different proportions for each type of climate. It is possible to determine the influences among variables, temperature, humidity, dew point, pressure, wind speed and precipitation, through the use of Bayesian networks that permit us to understand their interaction in different climates.

Key words: Probability models, climate, forecast, data mining, K2 algorithm.

INTRODUCTION

The climate is a combination of atmospheric phenomena (temperature, pressure, rainfall, wind, radiation and humidity) that characterizes a place for a long time. The factors that form the climate have been studied to develop forecasting models. For this propose, it is necessary to obtain data sets by a systemic and homogeneous method from meteorological stations during periods of minimum 30 years considered representative. Climatology is based on a statistical analysis of meteorological information which is gathered, temporal variations that occur in climate parameters are incorporated into statistical averages (Landsberg, 1955). Climate databases contain the statistical properties of different local climatology that exist in an area of study, whereas a reanalysis of these databases could contain the evolution of the atmosphere simulated by numerical models. Thus, this information can be used to analyze problems related to the atmosphere's dynamic and to the possible impacts it may cause on the climatology of local regions. However statistical techniques in these problems, including linear regression methods for weather forecast, clustering methods and principal component analysis for identification of representative atmospheric patterns, fragment information and assume *ad-hoc* spatial independencies in order to simplify the resulting problem-driven models (Easterling and Peterson, 1995; Cofiño et al., 2002). On the other hand, climate data obtained from weather stations are often incomplete, thus necessitating the use of data mining techniques for analysis (Cano et al., 2004; Hruschka et al., 2007). For these reasons, it is necessary to apply techniques to analyze the climatic variables without affecting its inherent properties.

Numerical models for climate forecasting are abstract representations of the real world. These models discretize areas or bodies in two or three dimensions respectively using approximate functions to describe the behavior of the climatic variables of interest studies.

*Corresponding author. E-mail: ricog@uaq.mx or garciarico@yahoo.com.mx. Tel: (52) (442) 1921200 ext. 6016. Fax: (52) (442) 1921200 ext. 6015.

Nowadays, numeric models are indispensable to climate forecast. According to Lima and Lall (2010), time varying scaling parameters could be estimated and used to assess whether there are statistically significant trends in the climate data using Bayesian networks (BNs). Moreover, Tae-wong et al. (2008) proposed a space-time stochastic model that represents the temporal and spatial dependences of daily rainfall occurrence. A BN precipitation model based on hidden Markov was developed by HongRui et al. (2010) using the maximum likelihood estimation method with incomplete data. In accordance with Kazemnejad et al. (2010), using the Bayesian analysis is recommended, especially when a small data sample set is available.

The objective of this study was to demonstrate that BN's could be used to find a structure that best describes the relationships among climate variables, in meteorological stations in which incomplete and limited, in time, data exist. On the other hand, with this model a precipitation forecast was performed in order to compute the conditional probability of raining.

Weather elements

The constituent elements of weather are temperature, pressure, wind, humidity and rainfall. To define the climate in a particular place, it requires a record for many years (Guttman, 1989). Temperature and precipitation are the most important climate elements, because the other three elements depend of them.

Relative humidity

Is the relationship between the water that is contained in an air mass and the maximum water that can be admitted without condensation, maintaining the same conditions of temperature and atmospheric pressure. This is obtained from Equation 1:

$$RH = P(H_2O) / P^*(H_2O) \times 100\% \quad (1)$$

where $P(H_2O)$ is the partial pressure of water vapor in the mixture of air $P^*(H_2O)$ is the saturation pressure of water vapor at the temperature of the mixture of air and RH is relative humidity of the air mixture being considered.

Dew point

It is defined as the condensing vapor temperature. Dew point depends on relative humidity and air temperature, Therefore varies with the amount of water in the atmosphere, pressure and temperature (Martines and Lira, 2008).

It is calculated from Equation 2:

$$Dp = \sqrt[8]{(RH/100) \times (110 + T)} - 110 \quad (2)$$

where Dp = Dew point, T = Temperature °C and RH = Relative humidity

Atmospheric pressure

Is the air pressure at any point in the atmosphere, which shows variations, associated with the weather changes. The pressure can be defined by reference to the microscopic properties of the gas. For an ideal gas with N molecules, the mass moving with an average random speed contained in a volume V, gas particles that impact can be calculated by the gas pressure using Equation 3:

$$P = (Nmv^2_{rms}) / 3V \text{ (ideal gas)} \quad (3)$$

Equation 3 tells us that the pressure of a gas depends directly on the molecular kinetic energy. The ideal gas law allows us to ensure that the pressure is proportional to absolute temperature. These two statements allow one of the most important statements of the kinetic theory: average molecular energy is proportional to temperature (Wash et al., 1990).

Atmospheric temperature

In accordance to Karl et al. (1986) atmospheric temperature refers to the degree of specific heat of air in a particular place and time as well as the temporal and spatial evolution of this element in different climatic zones.

1. Maximum temperature is the largest air temperature reached at a place in a day, usually between 14:00 and 17:00 h.
2. Minimum temperature, is the lowest temperature reached at a place in one day, which is usually recorded at dawn.
3. Average temperature, are statistical averages derived from maximum and minimum temperatures.

Relations between climatic variables

Temperature and air pressure

These are two elements of climate that vary inversely to each other when temperature of the air is higher and the pressure is lower. Inversely, when the air is cooler the atmospheric pressure increases. Thus, where the air temperature rises, the weather tends to be unstable and may produce rain and even thunderstorms. And where the air temperature decreases, the weather will be more stable and present no clouds on sunny days and dry

environment (Wash et al., 1990).

Humidity and dew Point

Relative humidity is the ratio of vapor in relation to that required to reach the saturation point, expressed in percentage. When the air is saturated, the relative humidity equals 100% and reaches the dew point.

Temperature and relative humidity

Relative humidity is a measure of air moisture content and is an indicator of evaporation, transpiration and convective rainfall probability. Relative humidity depends strongly to the temperature, which changes considerably during the day.

Air temperature and relative humidity kept a very close relationship with real constant tension of water vapor. There is an inversely proportional linear relationship which can be demonstrated both theoretically and empirically. This is explained, that as the temperature increases, the saturation level of water vapor in the atmosphere increases (Martines and Lira, 2008).

Bayesian networks theory

BNs are types of knowledge representation developed in the field of artificial intelligence for approximate reasoning (Pearl, 1988; Mediero, 2007; Gámez et al., 2011; Zaidan et al., 2011). A BN is a direct acyclic graph whose nodes correspond to concepts or variables and whose links correspond to relationships or functions (Correa et al., 2009). Variables are defined in a discrete or qualitative domain, and functional relationships describe causal inferences expressed in terms of conditional probabilities that are shown in Equation 4.

$$P(x_1,\ldots,x_n) = \prod_{i=1}^{n} P(x_i \mid \text{parents } (x_i)) \quad (4)$$

BNs can be used to identify previously undetermined relationships among variables or to describe and quantify these relationships even with an incomplete data set (Hruschka et al., 2007; Reyes, 2010). The solution algorithm of BNs allows the computation of the expected probability distribution of output variables. The result of this calculation is dependent on the probabilities distribution of input variables. Globally, BNs can be perceived as a joint probabilities distribution of a collection of discrete random variables (Garrote et al., 2007).

$$P(c_j \mid x_i) = P(x_i \mid c_j)\, P(c_j) \,/\, \sum_k P(x_i \mid c_k)\, P(c_k) \quad (5)$$

A priori probability $P(c_j)$ is the probability that a sample x_i belongs to class c_j, which is gives no information on their characteristic values, as shown in Equation 5. Machine learning in artificial intelligence is closely related to data mining, classification or clustering methods in statistics, inductive reasoning and pattern recognition. Statistical machine learning methods can apply the framework of Bayesian statistics; however, machine learning can employ a variety of classification techniques to produce models other than BNs (Subramaniam et al., 2010; Naveed et al., 2011). The objective of BN structure learning is to find a configuration that best describes the observed data. The number of possible structures of direct acyclic graph for searching is exponential in the number of variables in the domain, defined in Equation 6:

$$f(n) = \sum_{i=1}^{n} (-1)^{i+1} C_i^n\, 2^{i(n-i)}\, f(n-i) \quad (6)$$

The most representative method of the score-and-search-based approach is the K2 algorithm. The algorithm starts by assigning each variable without parents. It then incrementally adds a parent to the current variable which mostly increases the score of the resulting structure. When any addition of a single parent cannot increase the score, it stops adding parents to the variable. Since an ordering of the variables is known beforehand, the search space under this constraint is much smaller than the entire structure space, and there is no need to check cycles in the learning process. If the ordering of the variables is unknown, we can search over orderings (Guoliang, 2009).

We obtain a data set from three meteorological stations which represent three different types of climate. The data set was composed for the variables temperature, dew point, humidity, pressure, wind speed and precipitation, with three values: maximum, average and minimum. The data set was discretized and used to develop a BN model that describes the relationships among all variables. The model shows differences that allow us to identify the independence and dependence variables as well as quantify the degree of influence between them.

METHODS

Studied area description

The studied area covers the state of Queretaro, Mexico which has 11,689 km^2 of surface area and where three different climatic areas are identified: the south portion that covers part of the physiographic province of Eje Neovolcánico; the central region that comprises areas of the Eje Neovolcánico, Sierra Madre oriental and Mesa central; and the north area that corresponds to a portion of the Sierra Madre oriental. Three climatic stations in Queretaro state were chosen to obtain the climate data which represent the three different climates (Figure 1) located in Jalpan that corresponds to a climate ACw, colon that corresponds to a climate BS1k and Amealco with a climate Cw in accord to Koppen classification (Garcia, 1988), which are shown in Table 1.

Figure 1. Meteorological stations of Comisión Estatal de Aguas of Queretaro (CEA, 2010).

Data collection

For the purposes of this study we obtained daily data from the period 01/01/2007 to 31/12/2010 through the network of climatic stations of comision Estatal de Aguas of Queretaro, through the portal:

http://www.wunderground.com/weatherstation/ListStations.asp?selectedCountry=Mexico.

For these stations only have daily weather records since august 2006.

The data sets were composed for the values maximum, average and minimum of the variables temperature, dew point humidity, pressure, wind speed, gust wind speed and precipitation. We decided to use maximum and minimum values because it expresses more realistic relationships.

BNs analysis

The analysis of BNs was performed by the ELVIRA system version 0.162 in three stages suggested (Mediero, 2007):

Pre-processing

We used the average imputation algorithm to complete the partial data sets. This algorithm replaces missing values / unknown to the average of each variable. This method does not need parameters. Massive data for each variable were discretized using the equal frequency algorithm with two intervals.

Machine learning

According to Wang et al. (2006), for best regulate of Bayesian network structure we used the K2 algorithm learning method with Bayesian estimation and the maximum number of parents equal to 5, with no restrictions.

Post-processing

A dependency analysis was done to obtain the topological structure of the network, which represents the variables and their causal dependencies. After obtaining the parametric learning network, the conditional probabilities were calculated showing the relationships of influences between variables.

RESULTS AND DISCUSSION

Three BN topologies that correspond to each location (Amealco, Colon and Jalpan) were obtained. Each BN was organized according to similar variables in five groups: temperature, dew point, humidity, pressure and wind speed with three values: Maximum, average and minimum, respectively.

The conditional probability of precipitation was

Table 1. Meteorological data summary (precipitation was averaged annually).

Climates variable	ACw			BS1k			C(w)		
	Maximum	Minimum	Average	Maximum	Minimum	Average	Maximum	Minimum	Average
Temperature (°C)	42	0.5	19.2	35.0	-4.1	16.8	28.1	-3.4	12.4
Dew point (°C)	25.0	-73.3	8.1	19.1	-73.3	8.4	16.0	-73.3	2.7
Humidity (%)	100	6.0	62.6	100.0	7.0	63.8	100.0	10.0	62.1
Wind speed	24.1 km/h from the Oeste	-	1.0 km/h	20.9 km/h from the ONO	-	2.0 km/h	46.7 km/h from the SO	-	-
Gust speed	53.1 km/h from the North	-	-	51.5 km/h from the Oeste	-	-	94.7 km/h from the Oeste	-	-
Wind	-	-	SE	-	-	SE	-	-	SSE
Pressure (hPa)	939.6	909.1	-	1036.5	994.5	-	1036.1	996.5	-
Precipitation (mm)	1909.75			362.7			726.4		

calculated for the three climates, obtaining the following results shown in Figure 3.

For the three climates, the BNs show some similar patterns: maximum temperature influences average temperature and minimum temperature. Maximum pressure impacts minimum pressure and maximum wind speed influence maximum gust speed. These behaviors are present in variables form the same groups. Temperature affects humidity, because with increasing temperature, the relative humidity decreases how it is exposed (Martines and Lira, 2008), their results are expected but this relationship is more clearly with dew point in Equation 2. Minimum temperature impacts maximum dew point, since this variable means the temperature at which it begins to condense the water vapor in the air, producing fog when the temperature is low enough, mainly at mornings, when minimum temperatures are present (Karl et al., 1986).

The influences specific for each climate are the following:

1. For humid temperate climate (Cw) showing in Figure 2a, we observed a positive association between minimum temperature -3.4°C and maximum pressure 1036.1 hPa, how is presented in Table 1, which indicates when the air is cooler, atmospheric pressure increases (Wash et al, 1990). Average temperature of 12.4°C impact to minimum humidity 10% indicating that weather tends to be moderately dry, which is confirmed by the precipitation of 726.4 mm. Therefore precipitation is affected by average dew point (2.7°C), this could be temperature confirming temperate ambient conditions, where condensation of vapor water occurs at low temperatures.
2. For semi-dry temperate climate (Bs1k) shown in Figure 2b, we observed an association between average temperature 16.8°C with maximum pressure 1036.5 hPa and average humidity to 63.8%, suggesting that atmospheric pressure increases not at morning and the weather is more stable and present no clouds (Wash et al., 1990) but more fresh than Cw climate. However maximum dew point (19.1°C) is presents with minimum temperature (-4.1°C) which indicates that condensation water vapor is at morning. Precipitation and maximum wind speed are impacted by the maximum dew point suggesting that this temperature is easy to reach and wind is an important factor in precipitation, taking into account that this climate is drier than the other two.
3. For warm humid climate shown in Figure 2c, there are no influences between dew point and humidity, indicating a difference with the other climates. It is possible that this behavior is typical to warm humid climates, where precipitation is more abundant. Precipitation (1909.7 mm) is negatively influential to maximum humidity (96%), which means that a more humid climate, precipitation is influenced for humidity value. Moreover, maximum dew point (25.0°C) is impacted by minimum temperature (0.5°C) indicating much humidity at the morning (Karl et al., 1986). Pressure is autonomous from temperature, which is a characteristic temperature greater variation equal to 41.5°C as shown in Table 1.

Because the complexity of the relationships obtained for each variable, only the calculation of the conditional probability of precipitation was performed, since this variable is defined for two states (rain present and absent), hence its discretization is simpler and its probability distribution function is binomial. Figure 3 shows that 13°C is the temperature with the most probability for Cw climate, with 16°C of maximum dew point confirmed in Table 1. The same calculus for semi-dry temperature climate was 18°C, with 19.1°C of maximum dew point. For warm humid climate we obtained the most probability of precipitation in 94% of humidity, with 100% of maximum humidity.

Conclusion

Three different climates have been analyzed to discover

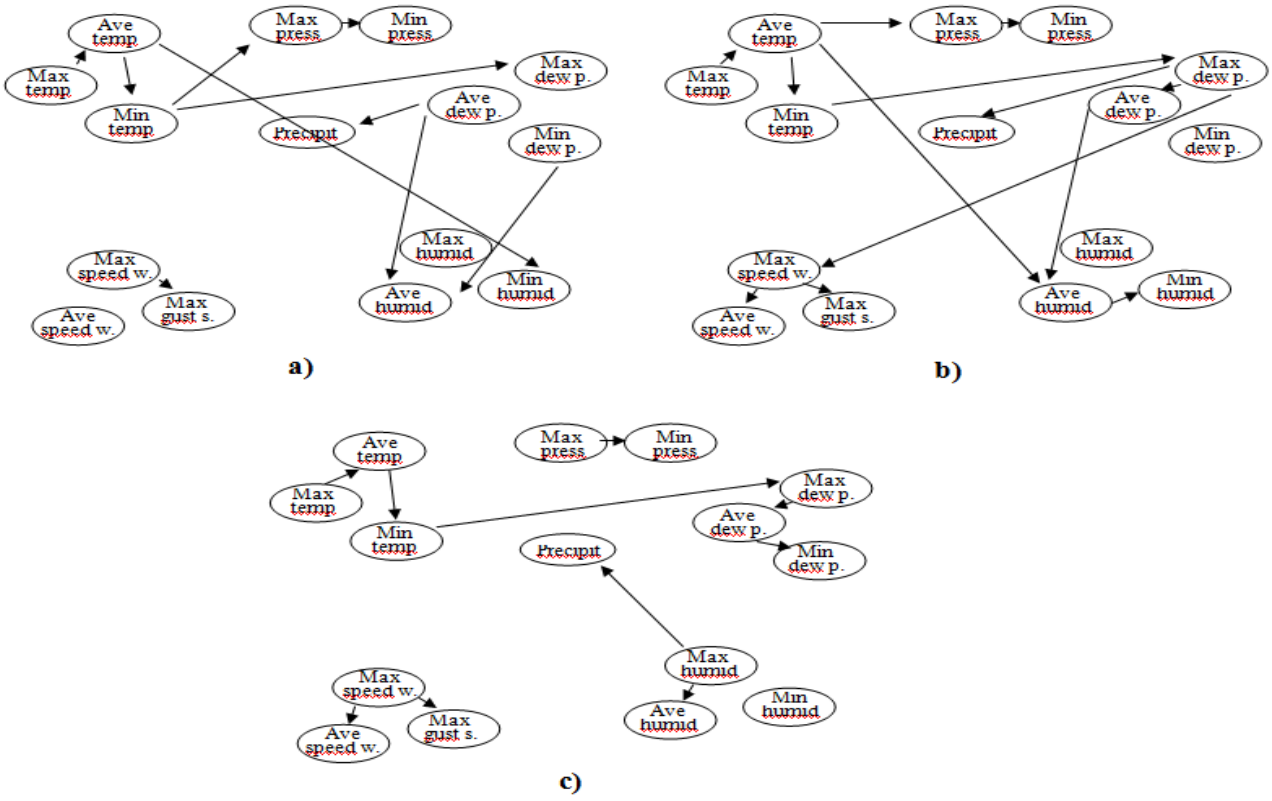

Figure 2. Bayesian network of: (a) Humid temperate climate, (b) Semi-dry temperate climate and (c) Warm humid climate.

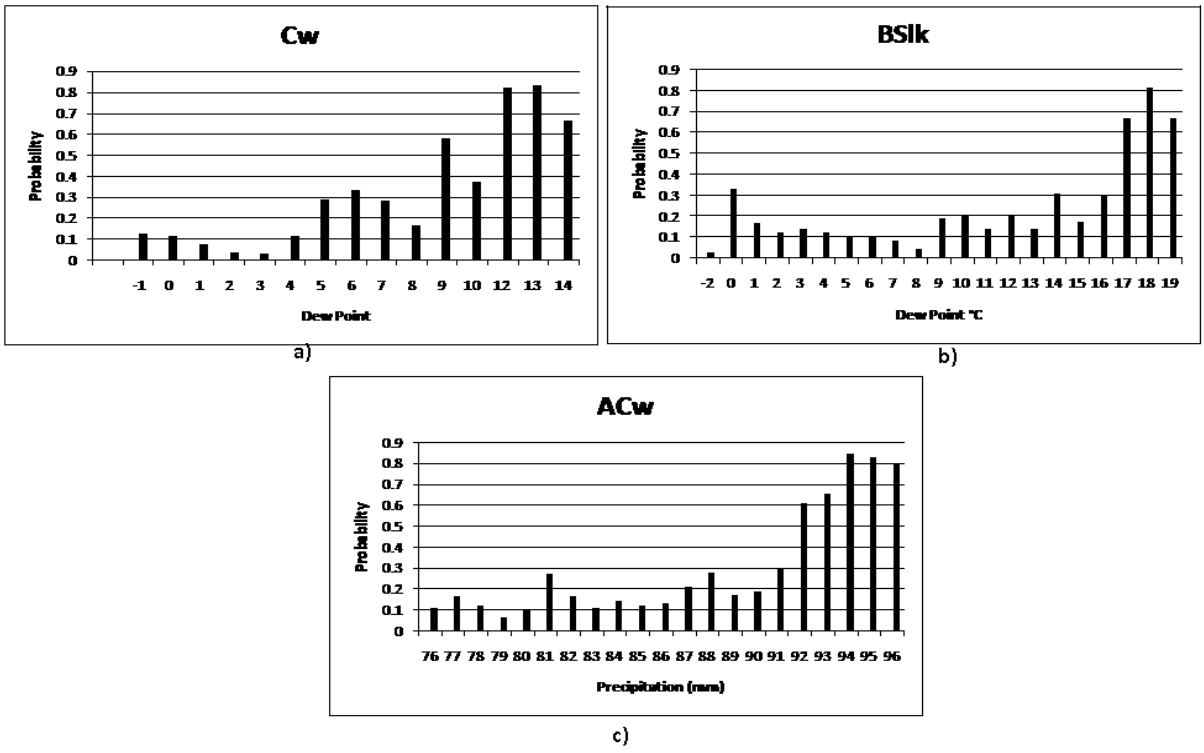

Figure 3. Conditional probability of: (a) Precipitation/dew point from humid temperate climate, (b) Precipitation/dew point from semi-dry temperate climate and (c) Precipitation/humidity from warm humid climate.

the relationships among variables using Bayesian network, with the objective to express causal-effect dependences. The analysis was performed on a set of incomplete data taking into account of 14 variables grouped by similarity weather into five groups.

These observations confirm that BNs can be used to identify previously undetermined relationships among variables or to describe and quantify these relationships even with an incomplete data set. Although to make predictions is advisable to define the domain using fewer variables and reducing the states of each variable. Moreover, considering of maximum and minimum values, we obtain a more precise analysis. It is possible to determine the influences among variables by the use of BNs that allow us to understand their interaction in different climates. The BNs allow the computation of the expected probability distribution function of precipitation, to starting from an incomplete and limit in time set of meteorological data. Currently, the use of new methods to determine relationships between variables in different phenomena, BNs applied to data mining can be a valuable tool for analysis.

Future directions

Not all relationships were explained, dew point and humidity relationship in ACw climate required a deeper analysis and to be compared with other humid climates. This new methodology can be applied to other climates and also incorporate other climatic variables like solar radiation.

ACKNOWLEDGEMENTS

This research was partially supported by the Mexican National Council of Science (CONACYT) and FIFI 2010, Engineering Department, Queretaro State University.

REFERENCES

Cano R, Sordo C, Gutierrez JM (2004). Applications of Bayesian Networks in Meteorology. In Gámez et al. (eds) Advances in Bayesian Networks, Springer, pp. 309-327.

Cofıño AS, Cano R, Sordo C, Gutierrez JM (2002). Bayesian Networks for Probabilistic Weather Prediction: Proceedings of the 15th European Conference on Artificial Intelligence, pp. 695-700.

Correa M, Bielza C, Paimes-Teixeira J, Alique JR (2009). Comparison of Bayesian networks and artificial neural networks for quality detection in a machining process. Expert Syst. Appl., 36(3): 7270-7279.

Easterling DR, Peterson TC (1995). A new method for detecting undocumented discontinuities in climatological time series. Int. J. Climatol., 15(4): 369-377.

Gámez JA, Mateo JL, Puerta JM (2011). Learning Bayesian networks by hill climbing: efficient methods based on progressive restriction of the neighborhood. Data Min. Knowl. Discov., 22: 106-148.

García E (1988). Modifications to Köppen classification system climate to suit the conditions of the Mexican Republic. Offset Larios S.A. Mexico, pp. 201-217.

Garrote L, Molina M, Mediero L (2007). Probabilistic Forecasts Using Bayesian Networks Calibrated with Deterministic Rainfall-Runoff Models. In Vasiliev et al. (eds.), Extreme Hydrological Events: New Concepts for Security, Springer, pp. 173-183.

Guoliang L (2009). Knowledge Discovery with Bayesian Networks. PhD dissertation, National University of Singapore, Singapore.

Guttman NB (1989). Statistical descriptors of climate. Bull. Am. Meteorol. Soc., 70: 602-607.

HongRui W, LeTian Y, XinYi X, QiLei F, Yan J, Qiong L, Qi T (2010). Bayesian networks precipitation model based on hidden Markov analysis and its application. Sci China Tech. Sci., 53(2): 539-547.

Hruschka E, Hruschka E, Ebecken NFF (2007). Bayesian networks for imputation in classification Problems. J. Intell. Inform. Syst., 29: 231-252.

INEGI 2011 www.inegi.org.mx

Karl RT, Williams NC, Young P, Wendland W (1986). A model to estimate the time of observation bias associated with monthly mean maximum, minimum and mean temperatures for the United States. J. Clim. Appl. Meteorol., 25: 145-160.

Kazemnejad A, Zayeri F, Hamzah NA, Gharaaghaji R, Salehi M (2010). A Bayesian analysis of bivariate ordered categorical responses using a latent variable regression model: Application to diabetic retinopathy data. Sci. Res. Essays, 5(11): 1264-1273.

Landsberg HE (1955). Weather 'normals' and normal weather. Wkly. Weather Crop Bull., 42: 7-8.

Lima CHR, Lall U (2010). Spatial scaling in a changing climate: A hierarchical bayesian model for non-stationary multi-site annual maximum and monthly streamflow. J. Hydrol., 383(3): 307-318.

Martines E, Lira L (2008). Dew Point Calculation at Different Pressures. Metrology Symposium. SM2008-M117-1098-1.

Mediero OL (2007). Probabilistic forecast flood flows Through Bayesian Networks Applied to a Distributed Hydrological Model. PhD dissertation, Polytechnic University of Madrid, Madrid, Spain.

Naveed N, Choi MTS, Jaffar A (2011). Malignancy and abnormality detection of mammograms using DWT features and ensembling of classifiers. Int. J. Phy. Sci., 6(8): 2107-2116.

Pearl J (1988). Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann, San Mateo CA, United States, pp. 64-70.

Reyes P (2010). Bayesian networks for setting genetic algorithm parameters used in problems of geometric constraint satisfaction. Intell. Artificial., 45: 5-8.

Subramaniam T, Jalab HA, Taqa AY (2010). Overview of textual anti-spam filtering techniques. Int. J. Phys. Sci., 5(12): 1869-1882.

Tae-wong K, Hosung A, Gunhui CH, Chulsang Y (2008). Stochastic multi-site generation of daily rainfall occurrence in south Florida, Stoch. Environ. Res. Risk Assess., 22: 705-717.

Wang S, Li X, Tang H (2006). Learning Bayesian Networks Structure with Continous Variables. In Li et al. (eds).Lecture Notes in Computer Science, Heidelberg: Springer-Verlang, pp. 448-456.

Wash CH, Heikkinen HS, Liou C, Nuss AW (1990). A Rapid Cyclogenesis Event during GALE IOP 9. Mon. Weather Rev., 118(2): 234-257.

Zaidan AA, Ahmed NN, Abdul HK, Gazi MA, Zaidan BB (2011). Spam influence on business and economy: Theoretical and experimental studies for textual anti-spam filtering using mature document processing and naive Bayesian classifier", Afr. J. Bus. Manag., 5(2): 596-607.